Master's degree thesis

LOG950 Logistics

Data Mining in Health Management

Saeed Raeisi

Number of pages including this page: 82

Molde, 23.05.2022

Mandatory statement

Each student is responsible for complying with rules and regulations that relate to examinations and to academic work in general. The purpose of the mandatory statement is to make students aware of their responsibility and the consequences of cheating. Failure to complete the statement does not excuse students from their responsibility.

*Please complete the mandatory statement by placing a mark **in each box** for statements 1-6 below.*		
1.	**I/we hereby declare that my/our paper/assignment is my/our own work, and that I/we have not used other sources or received other help than mentioned in the paper/assignment.**	☒
2.	**I/we hereby declare that this paper** 1. Has not been used in any other exam at another department/university/university college 2. Is not referring to the work of others without acknowledgement 3. Is not referring to my/our previous work without acknowledgement 4. Has acknowledged all sources of literature in the text and in the list of references 5. Is not a copy, duplicate or transcript of other work	Mark each box: 1. ☒ 2. ☒ 3. ☒ 4. ☒ 5. ☒
3.	I am/we are aware **that any breach of the above will be considered as cheating, and may result in annulment of the examination and exclusion from all universities and university colleges in Norway for up to one year, according to the Act relating to Norwegian Universities and University Colleges, section 4-7 and 4-8 and Examination regulations** section 14 and 15.	☒
4.	**I am/we are aware that all papers/assignments may be checked for plagiarism by a software assisted plagiarism check**	☒
5.	**I am/we are aware that Molde University College will handle all cases of suspected cheating according to prevailing guidelines.**	☒
6.	**I/we are aware of the University College's rules and regulation for using sources**	☒

Personal protection

Personal Data Act

Research projects that processes personal data according to Personal Data Act, should be notified to Data Protection Services (NSD) for consideration.

Have the research project been considered by NSD? ☐yes ☒no

- If yes:

Reference number:

- If no:

I/we hereby declare that the thesis does not contain personal data according to Personal Data Act.: ☒

Act on Medical and Health Research

If the research project is effected by the regulations decided in Act on Medical and Health Research (the Health Research Act), it must be approved in advance by the Regional Committee for Medical and Health Research Ethic (REK) in your region.

Has the research project been considered by REK? ☐yes ☒no

- If yes:

Reference number:

Publication agreement

ECTS credits: 30

Supervisor: João Carlos Ferreira and Berit Irene Helgheim

Agreement on electronic publication of master thesis

Author(s) have copyright to the thesis, including the exclusive right to publish the document (The Copyright Act §2).

All theses fulfilling the requirements will be registered and published in Brage HiM, with the approval of the author(s).

Theses with a confidentiality agreement will not be published.

I/we hereby give Molde University College the right to, free of charge, make the thesis available for electronic publication: ☒yes ☐no

Is there an agreement of confidentiality? ☐yes ☒no

(A supplementary confidentiality agreement must be filled in)

- If yes:

Can the thesis be online published when the period of confidentiality is expired? ☒yes ☐no

Date: 23.5.2024

PREFACE AND ACKNOWLEDGEMENTS

It would be my pleasure to thank my supervisors, Dr João Carlos Ferreira and Dr Berit Irene Helgheim for the assistance and advice they have given throughout my Master thesis. I appreciate all the understanding and support you provided during the crucial stages of this journey.

1. Introduction Part

ABSTRACT

In this research, we explore the use of past hospital collected data to improve hospital management. A data analytics process was implemented with data cleaning and data visualization to identify major variables' performance taking into account the hospital management process, using less resources with a good service to the population on the waiting time at their appointment at Hospital.

In this study, the impact of physicians' performance on key indicators like the length of stay and waiting time in the Emergency Department (ED) are discussed. Further, the effects of the performance of physicians in waiting time and Length of Stay (LOS) considering the triage colour.

A data analytic investigation of the impact of patients' average waiting for time and length of stay (LOS) was performed to compare the efficiency of different groups of physicians. In this research, based on the CRISP-DM method by using an emergency department's (general surgery unit) real-life data from 2015 to 2017, through the One-Way Analysis of Variance (ANOVA) method, it was found that the waiting time and average LOS belong to experienced physicians (who visit more patients) are longer than the less experienced physicians (who visit less patients). Using the Two-Way ANOVA method, the emerging and very urgent patients (Red and Orange triage colour) who were visited by experienced physicians have longer average LOS than the same level patients who were visited by less experienced physicians. On the other hand, not urgent patients (Blue triage colour) who were visited by high experienced physicians have shorter average LOS than the same patients who were visited by less experienced physicians. This study was performed using two scenarios of physicians' grouping, including Pareto grouping (80-20) and Frequency grouping ("Very High", "High", "Medium", "Low" and "Very Low").

Keywords: Waiting Time, Length of Stay, Emergency Department, Data Mining, CRISP-DM, ANOVA

List of Figures

List of Tables

Table of Contents

Chapter 1

INTRODUCTION

1.1 Background

These days, most hospitals have started to digitize their records, as a result of this data, the big data phenomenon has become more apparent in the healthcare system. With the appearance of big data, through the applying data mining and data analysis, ED (Emergency Department) hospital management process could be improved. Therefore, is important to understand the clinical process and when data is available a data analytics process over this data allows us to understand the process, and inefficiencies, check the lack of resources and so on.

Health analytics (data analytics over health data) has evolved as a major field of study and application, demonstrating the extent of the impact of data and information-based management on problem-solving and decision-making in modern healthcare organizations. Hospitals and other healthcare organizations have been adopting descriptive health analytics for medical data. Moreover, Healthcare organizations throughout the globe are interested in improving quality and performance; it is essential to define healthcare performance and establish quality improvement dimensions and strategies. Numerous studies describe how healthcare performance improvement encompasses the combined and ongoing efforts of all healthcare stakeholders, including healthcare professionals, patients and their families, researchers, payers, planners, and educators, in order to implement the changes that will result in improved patient outcomes, system performance, and professional development.

More recently, healthcare data warehouses became available by aggregating many sorts of data from various systems and sources in order to generate operational healthcare dashboards, strategic scorecards, and data storage. The primary aims of health analytics are to identify performance gaps and provide the most effective solutions. Health analytics should aid healthcare professionals and organizations in monitoring performance on a continuous and consistent basis, as well as in diagnosing poor performance and identifying the underlying causes of issues. Health analytics assist users in designing, developing, implementing, and evaluating several key performance indicators that may improve continuous monitoring, uncover causes for performance variations, and ultimately optimize performance (Fisher, 2013).

Health analytics is a business-driven phrase that spans a broad range of business intelligence applications and big data analysis. This new concept is predicated primarily on the availability and accessibility of data and information pooled through the effective integration and interoperability of a vast array of systems and tools, including hospital information systems, electronic medical records, clinical decision support systems, and other specialized medical systems (Madsen, 2012).

The emergency care system is on the front lines of the healthcare system and in terms of access to care, Emergency Departments (Eds) are an impressive public health success story (Institute of Medicine, 2007). Service quality in the EDs is measured for waiting time and LOS (Xu, Wong, & Chin, 2013), therefore these two variables are the most important factors to measure the physician's efficiency. No systematic and exhaustive review has been provided to cover the physician efficiency considering LOS and waiting time and there is no study to show the interactions between independent variables like physicians' performance and triage colour.

In this master thesis, the improvement of the hospital management process is investigated by applying data mining and investigating the main factors that indicate the physicians' efficiency. For that, the author started by studying big data analytics in the healthcare system, learning about the variables of the data, manipulating the data, and producing descriptive plots in order to find the best way of comparison methods.

The received data contains a big data which is related to the ED (General Surgery Unit) of a Portuguese hospital. Then, through data manipulation it is observed that the main variables that are influenced by physicians' performance are waiting time and length of stay and it was found that the physicians could split into different groups based on the number of visited patients.

1.2 Problem Statement

The potential for improvement in the hospital (ED) management process needs to be explored through accessibility to relevant data which are available today thanks to the advances in big data. The flow of the hospital management process which is the case of this study consists of a sequence of several steps namely admission, triage colour, physician first note, medical discharge, and administrative discharge. Based on past data and visualizing them, it can be identified the variables that need to be investigated to suggest improvements.

1.3 Goal and Research Questions

1.3.1 Research Goal

The main goal of this research is to apply the data mining process to extract knowledge from past data to improve hospital management. From data received from a Portuguese hospital, it is observed that physician's performance has a meaningful effect on the proficiency of the ED management process. From the hospital data visualization performed in chapter 3, the aim is to explore if the length of stay and waiting time-as two important indicators of the efficiency in hospital management-depends on physicians considering their performance (based on the number of visited patients).

1.3.2 Research Questions

The main research question is the investigation of hospital management improvement based on knowledge extraction through a data analytics process. From the data available from a Portuguese hospital, the below operational questions are as followed:

RQ1- How does physicians' performance impact important indicators such as length of stay and waiting time in the ED?

RQ2- How is the performance of physicians influential in Length of Stay (LOS) considering the triage colour?
RQ3- How is the performance of physicians influential in waiting time considering the triage colour?

1.4 Structure of the Thesis

Chapter 1 has provided the context of the thesis subject and has presented the research gap regarding physician efficiency in ED through the data mining method. This chapter has justified why closing the research gap is necessary then the rest of the thesis is organized as below:

Chapter 2: Literature reviews Including, the importance of the emergency department, waiting time and length of stay concept, the importance of the physician's role in ED, big data analytics in healthcare, operation, and logistic perspective.

Chapter 3: Research Methodology, where data mining techniques were applied to the dataset to improve the ED process and understand the main factors that influence it. The work methodology used in this research is based on Cross-Industry Standard Process for Data Mining (CRISP-DM) method.

Chapter 4: In this chapter statistical analytics of the study which covers the fourth and fifth of the CRISP-DM method are examined and analyzing the results and comparing them to the predetermined research goals are done.

Chapter 5: where the results are analyzed and compared with the predefined goals of the research and research questions.

Chapter 6: This chapter provides the conclusion.

Chapter 7: This chapter provides a summary of the research, limitation of the study and suggestions for future research.

Chapter 2

LITERATURE REVIEW

Several studies were conducted on PubMed, ScienceDirect, Scopus, SpringerLink and ResearchGate, using the keywords "Emergency Department Length of Stay", "Emergency Department waiting time" and "Big data analytics in the healthcare system".

2.1 Big Data Analytics in Healthcare System

In this respect, most of the authors developed frameworks or models on how to use big data analytics in healthcare.

There is still a lot of work to be done in developing big data analytics and applications in healthcare, but rapid advances in platforms and tools can make progress more quickly. Big data in healthcare is defined as "electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware" (Raghupathi & Raghupathi, 2014). In addition, they described big data as "extremely large data sets that can be analyzed computationally to find patterns, trends, and associations, visualization, querying, information privacy and predictive analytics on large widespread collection of data."

In 2015, Caetano et. al, based on the CRISP-DM method predicted LOS. they claimed that by having access to better estimations (more likely to occur in the future) of the LOS, hospital managers can make more accurate decisions. This resulted in better planning of the hospital resources, consequently in a better hospital management performance, with an increase in the number of available beds for new patients and a reduction in surgical waiting time (Caetano, Cortez, & Laureano, Using Data Mining for Prediction of Hospital Length of Stay: An Application of the CRISP-DM Methodology, 2015).

Chauhan and Jangade provided an architectural framework for big data analytics in healthcare, studying tools, data mining techniques and data sources (Chauhan & Jangade, 2016).

Kankanhalli et al. found that big data analytics can be effectively used in the healthcare industry defining that it has three main dimensions: volume, velocity, and variety (Kankanhalli, Hahn, & Tan, 2016).

Asri et al. described some of the big data analytics challenges like data sources, data quality and human resources and managing big data analytics systems and highlight some challenges that big data analytics faces in healthcare (Asri, Mousannif, Moatassime, & Noel, Big data in healthcare: Challenges and opportunities, 2015).

The obtained DM predictive and explanatory knowledge results were considered credible by the hospital specialists and are valuable for hospital managers. By having access to better estimates of the LOS that is more likely to occur in the future and which factors affect such estimates, hospital managers can make more informed decisions. Such informed decisions can lead to better planning of the hospital resources, resulting in a better hospital management performance, with an increase in the number of available beds for new admissions and a reduction of surgical waiting lists.

2.2 Importance of Emergency Department

The emergency care system is on the front lines of the healthcare system and in terms of access to care, EDs are an impressive public health success story. Moreover, the emergency care system has become the “safety net of the safety net” providing an excellent setting for urgent and lifesaving care, as well as ambulatory care for many communities (Institute of Medicine, 2007).

In ED as a complex system understanding the path of the patient from the gate to the exit door (Length of Stay) are key to improving patient flow (Krall, Cornelius, & Addison, 2014).

Emergency Department (ED) crowding has been identified as a worldwide public health issue as well as an important safety issue (Pines & Griffey, 2015).

In an emergency department (ED), factors that influence quality and efficiency include the patient length of stay (LOS) and admission, prompt ambulance diversion, quick and accurate triage, nurse and physician assessment, diagnostic and laboratory services, consultations, and treatment (Gul & Celik, 2017).

Bittencourt et al, through 15 systematic reviews found that there are four feasible types of interventions in the ED throughput component, with positive results on patients’ length of stay in the ED. Four proven effective interventions Considering the quantity and quality of systematic reviews are: (1) the use of a physician/nurse to perform and supervise triage and flow of patients; (2) strengthening the care team using nurse practitioners; (3) implementation of new areas for caring of patients with acute non-critical conditions or areas to medicate and observe patients before assessing severity and (4) use of the Lean methodology and full capacity protocols (Bittencourt, Stevanato, Bragança, Gottems, & O’Dwyer, 2020).

2.3 Waiting Time and Length of Stay

Studying time is a valuable way to identify areas in the patient care process that are causing delays (Kyriacou, Ricketts, Dyne, McCollough, & Talan, 1999). Service quality in the Eds is measured by waiting time and LOS (Xu, Wong, & Chin, 2013). The waiting time is one of the most common complaints in ED (Lee, Endacott, Flett, & Bushnell). Therefore, waiting time is the most important factor directly influencing patient satisfaction in ED (Nhdi, Asmari, & Thobaity, 2021).

LOS plays a critical role in healthcare organizations and the complexity of fully identifying all factors that may impact LOS remains (Kudyba & Gregorio, 2011). LOS is defined in terms of the inpatient days which are computed by subtracting the day of admission from the day of discharge (Caetano, Cortez, & Laureano, Using Data Mining for Prediction of Hospital Length of Stay: An Application of the CRISP-DM Methodology, 2015). Song et. al found that patients experience shorter LOS when physicians work in a dedicated queuing system with a fairness constraint as opposed to a pooled queuing system with the same fairness constraint using a hospital’s ED data (Song, Tucker, & Murrell, 2015).

2.4 The Importance of the Physicians' role in ED

The staffing of physicians has been considered a factor that affects ED flow. Emergency physicians (EPs) play a critical role in the evaluation process. The length of time that takes the physician to assess the patient once a room is assigned to the patient directly affects the evaluation interval. (Krall, Cornelius, & Addison, 2014).

Hemmati et. al considered Low skillfulness, experience, and knowledge of the staff as one of the main reasons that affect physician decision time (Hemmati, Mahmoudi, Dabbaghi, Fatehi, & Rezazadeh, 2018).

Nhdi et al. expected that the highest effective index for LOS (with a strong positive significance correlation) is the decision to disposition time in EDs (Nhdi, Asmari, & Thobaity, 2021).

The data in the healthcare system reflects the reactions of physicians and healthcare professionals in emergency situations. Unexpected patterns in the data are driven by the physicians' behaviour, which highlights the importance of considering contextual information during the process of mining analyses in healthcare (Andel, Beerepoot, Lu, Weerd, & Reijers, 2021).

2.5 Logistic Services

Managing patients' length of stay plays a crucial role in healthcare organizations. Using analytic methods, providers can better manage the processes impacting LOS by leveraging data resources describing the network of activities that affects the patients' LOS. Organizational productivity can be enhanced by strategic initiatives aimed at increasing resource allocation to address process throughput and bottlenecks (Kudyba & Gregorio, 2011).

To better manage the processes impacting LOS, providers can leverage data resources describing the network of activities that impact a patient's LOS with analytic methods. Studying waiting time and length of stay (LOS) ED-which is defined as the time from triage to the discharge-helps management plan and manage ED operations and resources (Sariyer, Taşar, & Cepe, 2019) (Nhdi, Asmari, & Thobaity, 2021).

Chapter 3

DATA AND METHODS

The research method in this study is a quantitative approach. Through this study, the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach is used to improve the reliability, repeatability, and manageability of data mining projects, resulting in a more efficient process (Wirth & Hipp, 2000). The CRISP-DM process consists of 6 phases, and we adapted it for analyzing the data. The first and second phases are including business understanding and data understanding, which involve collecting, describing, exploring, and verifying the data. Step three includes data preparation, which begins with selecting, cleaning, exploring, verifying the quality, integrating, and formatting data. Then the visualization to identify operations patterns is performed that raise the operation goals to be answered in chapter four. The fourth phase called modelling which comprises selecting a modelling technique, generating a test design, building a model, and finally assessing the model in this study one-way analysis of variance (ANOVA) and two-way ANOVA is applied. In the fifth phase (Evaluation), the model (s) are evaluated and Interpreted. The fourth and fifth phases are written in chapter 4.

Figure 1 illustrates the Cross-Industry Process for Data Mining (CRISP-DM) cycle.

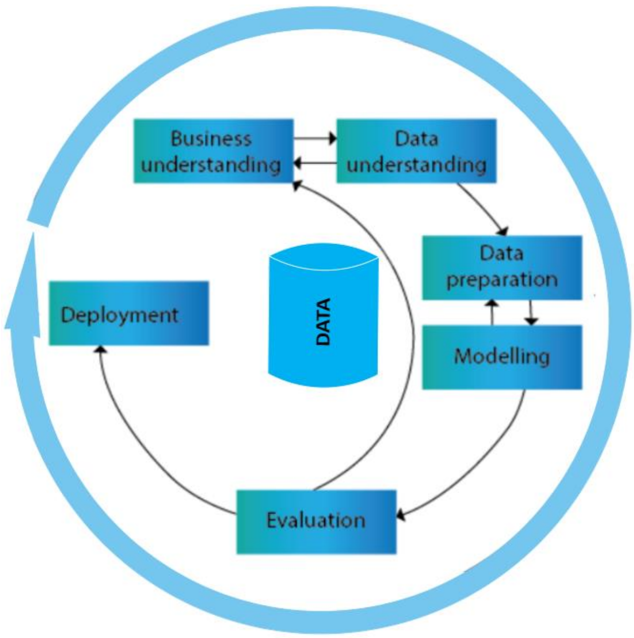


Figure 1 Cross-Industry Process for Data Mining (CRISP-DM) (Wirth & Hipp, 2000)

3.1 Business Understanding

The improvement in the ED process needs to be explored through accessibility to relevant data. The flow of the hospital management process which is the case of this study consists of the sequence of admission, triage colour, physician first note, medical discharge, and administrative discharge.

Moreover, two important indicators of the efficiency in hospital management are the length of stay and waiting time which can be considered in order to evaluate the efficiency of physicians' performance.

3.2 Data Understanding

The data used in this study comes from a Portuguese hospital near Lisbon (Garcia de Orta Hospital), that provides ED data between January 1st, 2015 and December 31st, 2017.

This real data contains 273,712 observations and each record is stored in the hospital's database based on visiting a patient in the general surgery of the Emergency Department. There are 12 variables to describe the interaction of a patient with the hospital ED.

Five of the twelve features (variables) in the original dataset are DateTime (YYYY-MM-dd hh:mm:ss) variables, including; Admission Date, Triage Date, Examination (first note) Date, Medical Discharge Date and Discharge Administrative Date.

The patients, triage nurses, and doctors involved in the event are identified with identifiers (IDs). Patient ID is a feature to identify the patient and track his/her history (A patient may be visited several times). There is also an id to identify the nurse responsible for the patient triage. There are two key ids in the dataset for the doctors who are responsible for the first note and medical discharge.

The dataset also includes an attribute called Triage Colour coding, which represents the triage colour assigned to the patient during the triage process. This variable involves five possible levels, each colour indicates what level of treatment is required.:

- RED: emerging patients
- ORANGE: very urgent patients
- YELLOW: urgent patients
- GREEN: less urgent patients
- BLUE: not urgent patients

There is also a variable regarding the discharge named Destiny which shows the patients' fate after LOS in ED.

Figure 2 illustrates the original variables of the dataset.

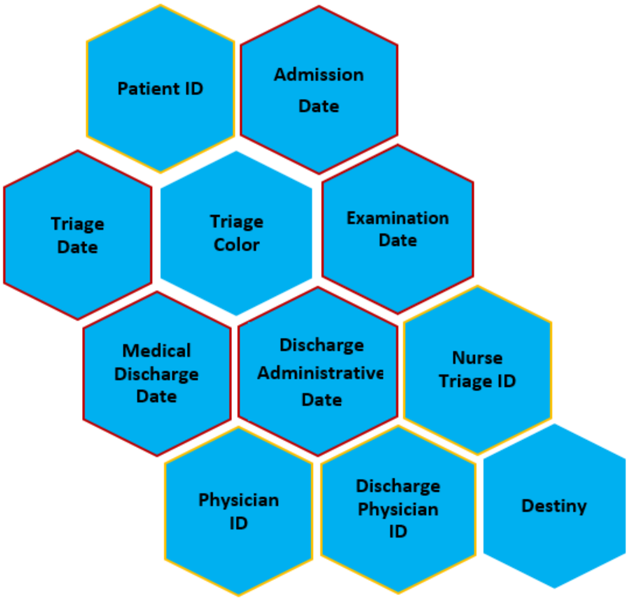


Figure 2 Original dataset variables

As can be seen from Figure 3, the ED process flow includes five main steps. It starts with the patient's admission where the patient is admitted to the ED. Next is the triage, in which the patient is classed based on the Manchester Triage Protocol (MTP). This is followed by the patient waiting to see a doctor, which is the third ED process. After the physician notes or gets treatment, the patient will be discharged by a doctor based on the patient's health state and finally, in the last step which is called administrative discharge the patient effectively leaves the hospital.

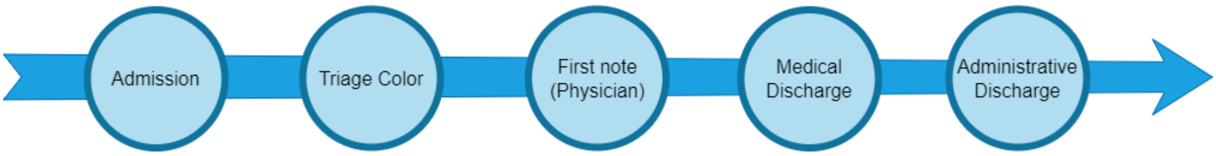


Figure 3 ED process flow

3.3 Data Preparation

This is the third step of the CRISP-DM method. The main goal of this process is to prepare the data for the rest of the research.

Research results and analysis can be compromised if the data is faulty, so the data must be accurate and veracious, respecting some of the characteristics of big data (Asri, Mousannif, Moatassime, & Noël, Big data in healthcare: Challenges and opportunities, 2015).

Data preprocessing is not restricted to classical data mining tasks such as classification or regression. In novel data mining fields, data preprocessing is increasingly becoming a tool for improving models (García, Ramírez-Gallego, Luengo, Benítez, & Herrera , 2016).

The author developed a Rmarkdown in R programming language for analyzing, manipulating and statistical computing, using several packages to analyze, manipulate and visualise the data such as; dplyr, tidyverse, MASS, ggplot2 and so on.

Several data cleansing steps were taken to ensure that the data was valid and won't compromise the results of the study. The author considered inconsistent, the records that have missing values and the minus periods, but outliers did not remove due to the nature of the data.

3.3.1 Cleaning Data

The following steps have been done to clean the data:

- All the rows where column DateTime had empty values were removed.
- The patients who are not assigned triage colour are omitted from the dataset
- As in this research the department of general surgery has been studied the other departments were removed from the dataset

Outliers are a crucial part of data analysis. Removing or retaining an observation can be a challenging decision, particularly if outliers are influential. In some cases, outliers simply occur when observations are different from the norm. In such cases, it may not be wise for these observations to be removed. Therefore, it is important to analyze the causes behind an observation that has been identified as an outlier (Hyndman & Athanasopoulos, 2018). As outliers of the data in this study are an integral part of the data and are unavoidable, they should not be deleted.

After the cleaning steps, the number of observations is reduced to 278,584. Then some informative plots are created to describe better the dataset.

The below bar graph shows the number of patients during the day based on triage colour coding. As shown in the graph, the (5)Blue (not urgent) patients went to the emergency department less than the other patients and the distribution of (1)Red and (2)Orange during the night are more than the other levels.

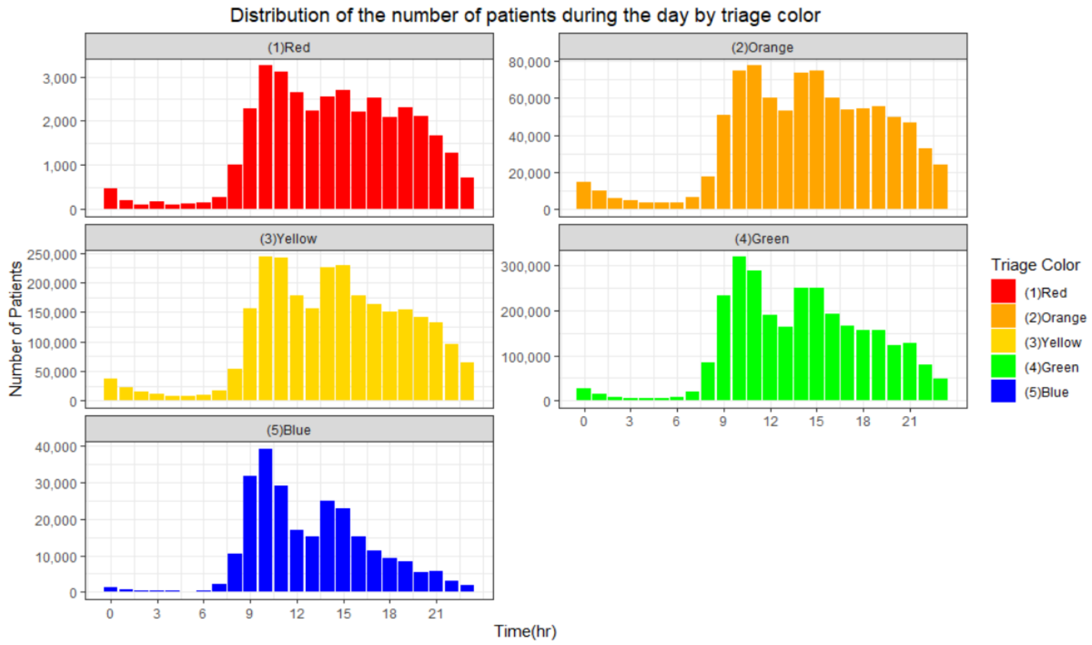


Figure 4 Distribution of the number of patients during the day based on triage colour

3.3.2 Construct Data

In this part, operations such as creating derived attributes, creating new records, or converting values for existing attributes are performed.

Based on the data, two important indicators of the efficiency in the emergency department which are the length of stay and waiting time that can be considered in order to evaluate the efficiency of physicians' performance should be specified.

LOS and waiting time are determined by considering the segments where the physicians have a direct effect on them.

As can be seen from Figure 5, According to the physicians’ performance length of stay is split into three segments, LOS1, LOS2 and LOS3. Based on the goals of this research, by deducting “Medical Discharge Date” from “Triage Colour Date” and by deducting “First Note Date” from “Triage Colour Date” the “LOS2” (Because only LOS2 is affected by a physician's performance) and “Waiting Time” are obtained.

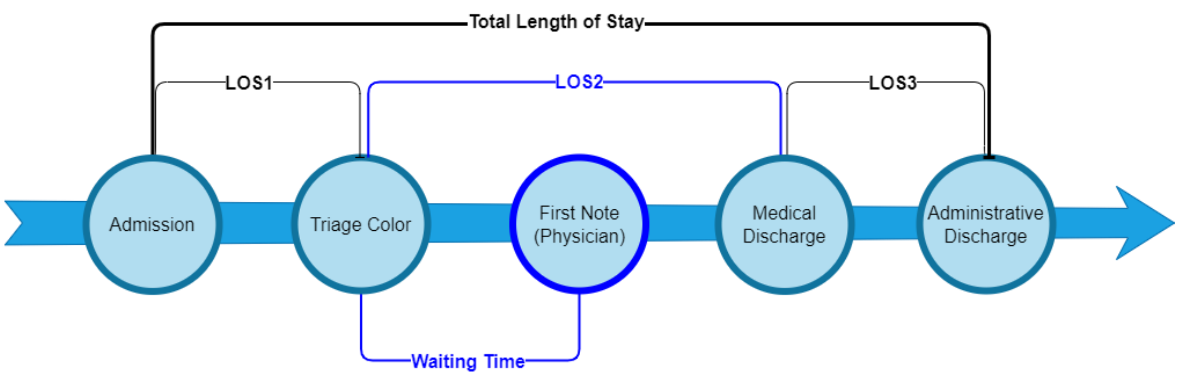


Figure 5 LOS and Waiting Time in the ED process flow

In this step, the final dataset with the two new important variables including "LOS2" and "Waiting Time" is prepared.

To extract more information from the dataset and make some descriptive plots, some variables should have been added to the original dataset, like "season", "month", "weekdays", "parts of day", "hours", "number of doctors" and "number of patients".

Figure 6 illustrates the seasonal average waiting time considering admission periods (morning: from 8:00 to 17:59, evening: from 16:00 to 23:59, night: from 00:00 to 07:59). Looking into Figure 6 the highest value of patients' waiting time (between triage and receiving the doctor's first note) belongs to the fall season in the morning period.

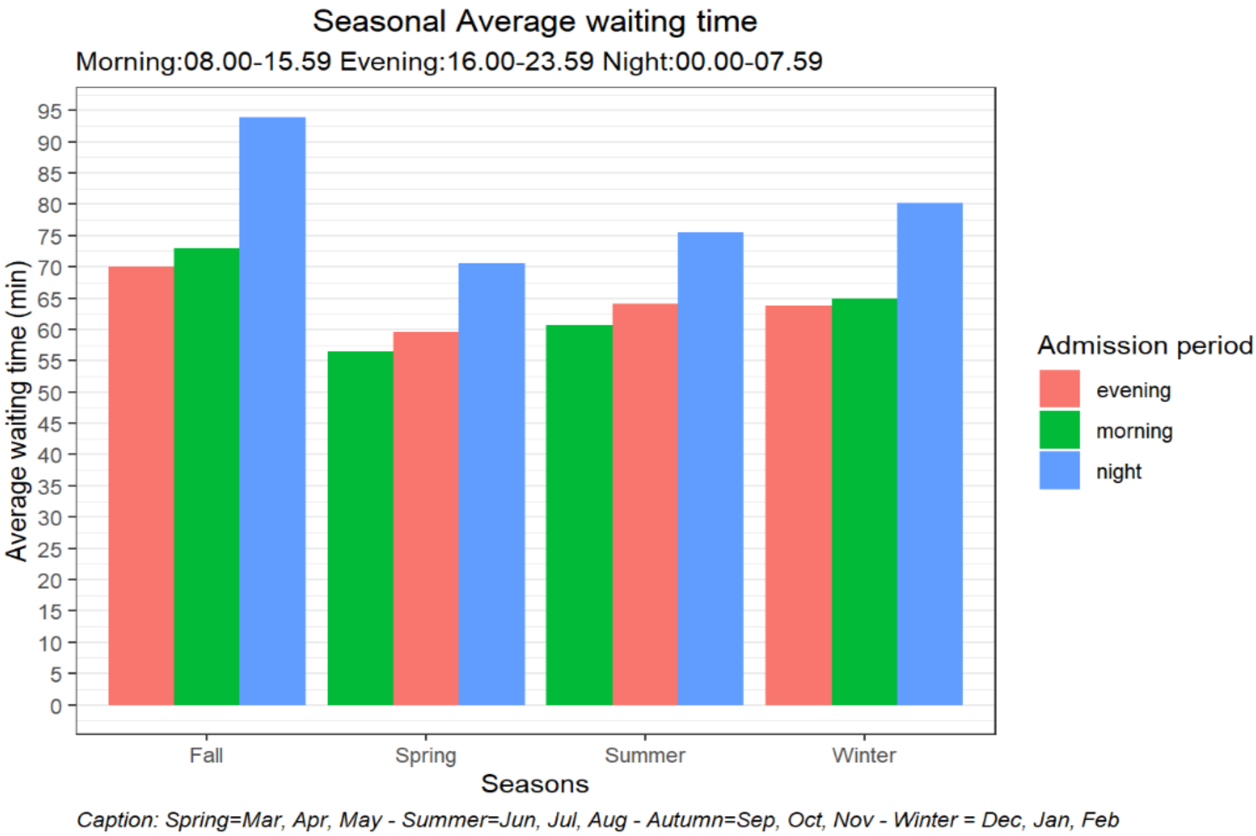


Figure 6 Average waiting time (Triage-first note) during the Seasons by daily period

Figure 7 provides information on the monthly patients' waiting time beside the graph demonstrates it separately based on the triage colour. The graph reveals a trend of fluctuation in the waiting time during the year; it hits the lowest value in May and the highest value at the end of the year (December). Moreover, the right side of the graph shows that the lowest waiting time belongs to emerging patients (Red triage colour); it refers that the first note issued by physicians has the minimum duration.

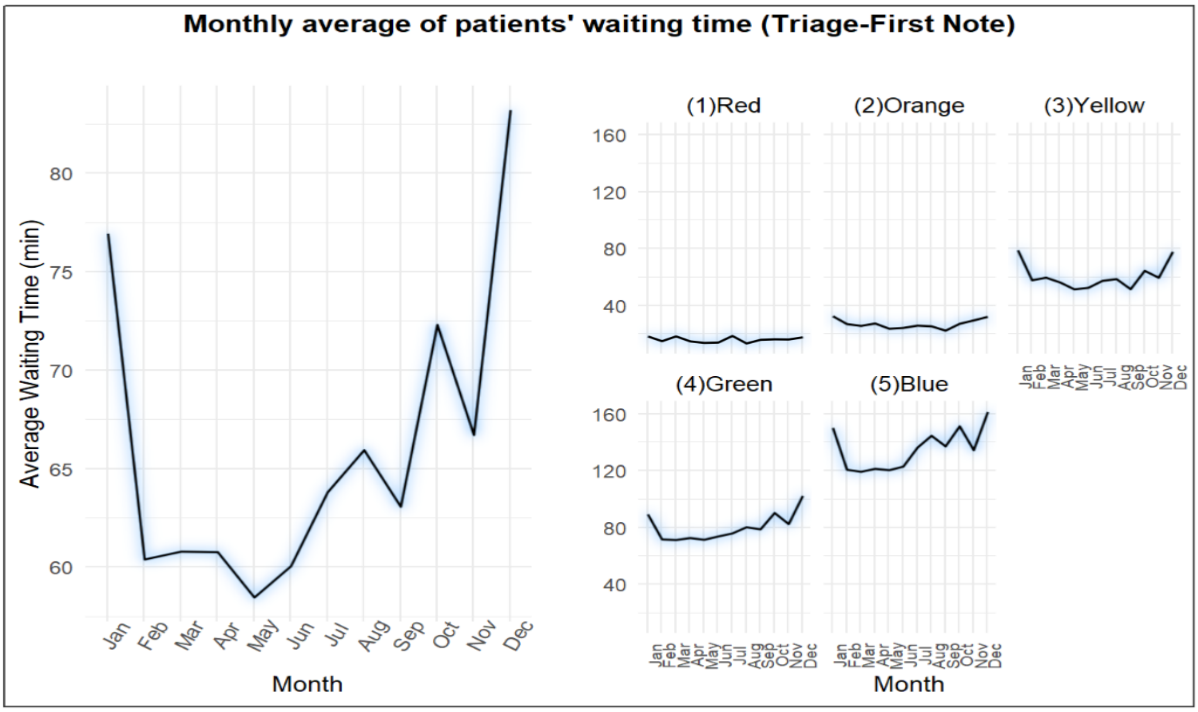


Figure 7 Monthly average of patients' waiting time (Triage-First Note)

The status of weekly waiting time (from triage colour to doctor first note) is shown in Figure 8. As it turns out, the waiting time for red patients (emerging patients) is shorter and for blue patients (not urgent patients) it is longer. The shortest time for the doctor's note is recorded over the weekend. In the early part of the week waiting time has the highest value.

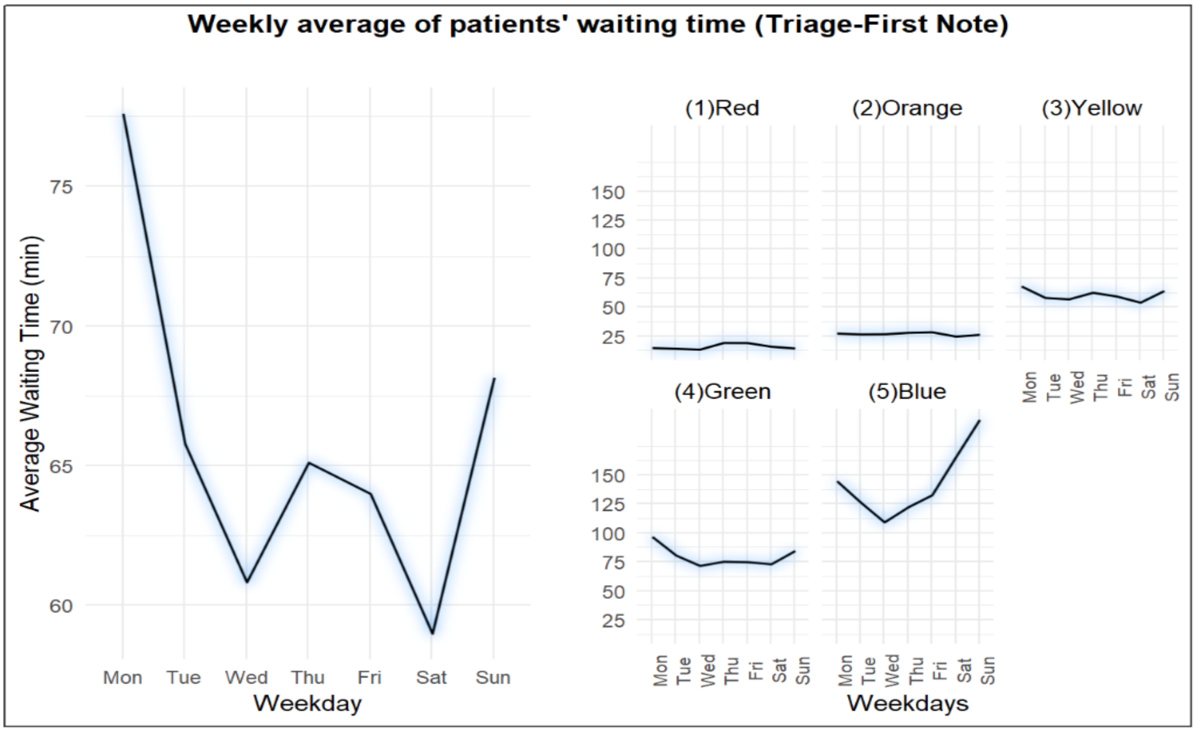


Figure 8 Weekly average of patients' waiting time (Triage-First Note)

The graph in Figure 9 represents the evolution of the average waiting time during the year. The average LOS2 stood at 318 in January. However, over the following four months, there is a sharp fall to 272 minutes. Same as the monthly waiting time, LOS2 hits the lowest point in May. The average LOS2 increases gradually from May to August, then it rises significantly from September to December to 323 minutes.

The Red triage patients have a fluctuation in the average LOS2 at ED. Although the lowest recorded value of average LOS2 belongs to emerging patients (Red), during some parts of the year average LOS2 of less urgent (Green) and not urgent (Blue) patients are less than the Red one.

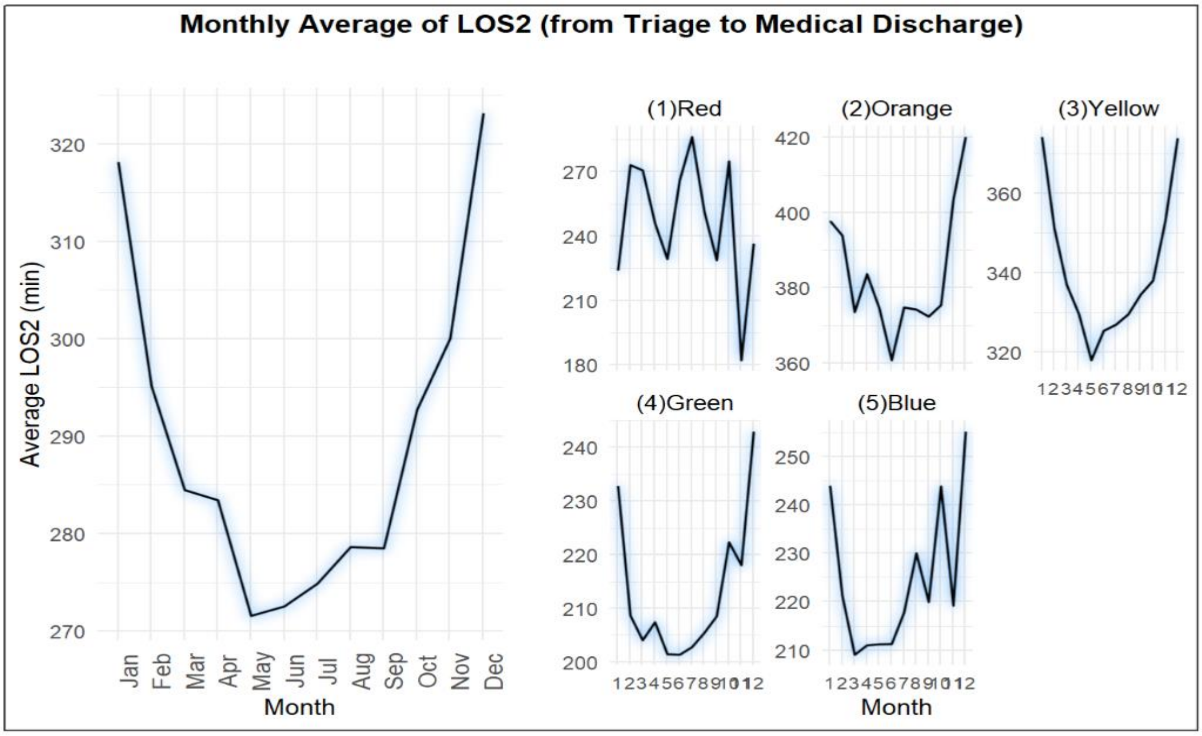


Figure 9 Monthly Average of LOS (from Triage to Medical Discharge)

Making the same analysis but during the week, Sunday is the day with the longest average waiting time, 298 minutes (around 5 hours). The average waiting time decreases throughout the week, hitting a low of 277 minutes on Wednesday and Saturday has the second-lowest value. The line graph illustrates the second-highest value on Monday.

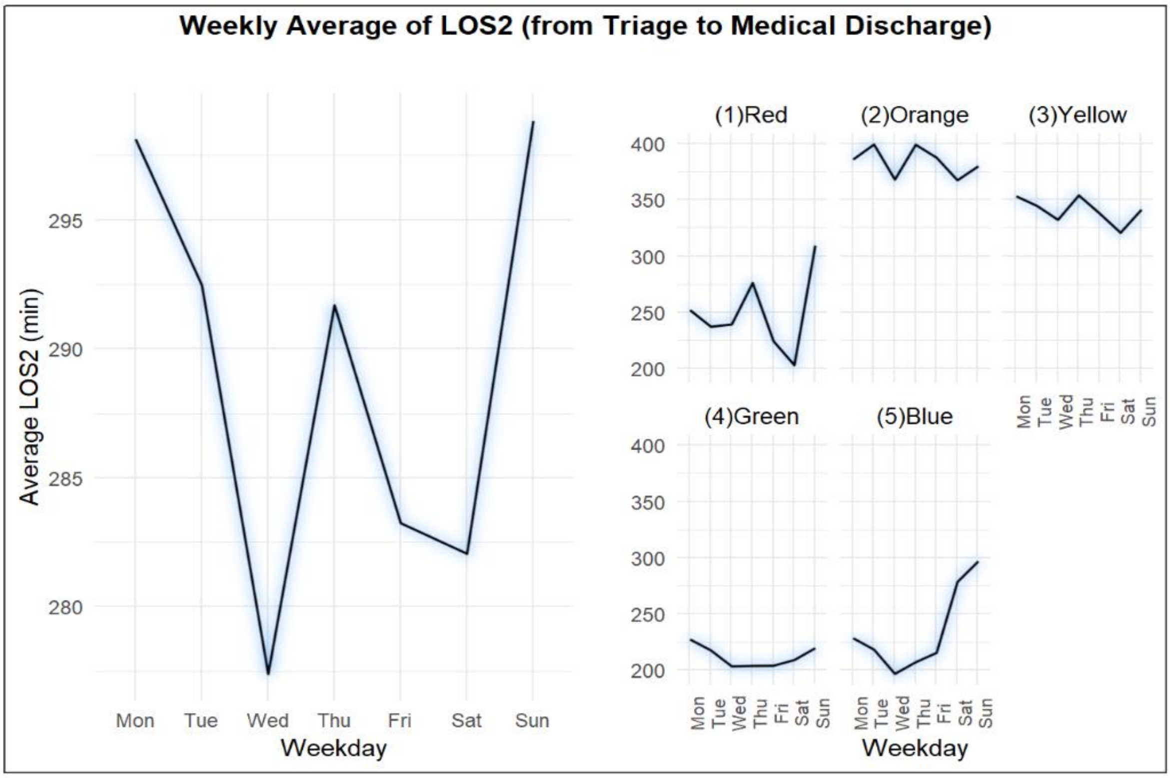


Figure 10 Weekly Average of LOS (from Triage to Medical Discharge)

3.3.3 Format Data

After data manipulation and data visualization from Figure 4 to Figure10, it is found that considering our goals it is possible to split physicians into different groups based on the number of visited patients, such that the physicians who visit more patients are highly experienced and those who visit less patients have less experience.

The data is still based on Patient ID, since the goal of the research is to evaluate the effect of physician groups on waiting time/LOS2 it needs to be manipulated according to the Doctor ID.

In formatting transformation, the content of the data does not change, while syntactic modifications are made to the data.

In other to compare different groups of physicians, the author considers 2 grouping scenarios including Grouping based on frequencies and Pareto (80/20).

- ***The Frequencies Grouping***

In Table 1 and Table 2, In this scenario, physicians (observations) split into 5 groups based on the cumulative sum of Patients. According to this method, the physicians were placed in one of the "Very High", "High", "Medium", "Low" and "Very Low" groups. Hence, the physicians who visited more patients fell into the "Very High" group provided they are part of the first 20 per cent of the cumulative sum of Patients.

- ***The Pareto (80/20) Grouping***

Also, in Table 1 and Table 2, This scenario is similar to the previous grouping, but in this scenario, physicians are divided into two main groups the first group are 131 doctors named "%80" who visited 80 per cent of the patients, and the rest are 671 doctors called "%20" who visited 80 per cent of the patients.

Descriptive analysis of waiting time in two different scenarios is reported below in Table 1. It reveals the number of physicians in a column called "count", it also shows the mean, standard deviation (SD), median, minimum, and maximum value according to the different groups in each scenario.

Table 1 Waiting Time descriptive analysis in 2 scenarios

Waiting Time (Triage to Doctor's First Note)

GroupName	count	mean	SD	median	min	max
Pareto Group						
%80	131	65.12281	27.61280	58.76568	21.368773	153.38463
%20	675	52.27813	35.09815	47.07393	1.033333	195.95417
Frequency Group						
Very High	7	72.54060	19.62908	74.92488	31.701508	88.57063
High	18	70.61331	25.05513	70.70084	22.754964	112.22747
Medium	35	70.53001	31.84072	53.67956	21.368773	152.71363
Low	71	60.33400	26.20997	56.25849	23.352705	153.38463
Very Low	675	52.27813	35.09815	47.07393	1.033333	195.95417

As can be seen in Table (2) descriptive analysis of the patient's LOS2 including count (number of physicians), mean, SD, median, minimum, and maximum values described and summarized based on 2 different groping scenarios that fulfil every condition of the data.

Table 2 Length of Stay (LOS2) descriptive analysis in 2 scenarios

Length Of Stay (LOS2)

GroupName	count	mean	SD	median	min	max
Pareto Group						
%80	131	269.0511	111.89118	285.3529	21.368773	153.38463
%20	675	254.1960	135.89086	239.2833	1.033333	195.95417
Frequency Group						
Very High	7	318.1871	100.99907	354.8269	31.701508	88.57063
High	18	305.3763	80.34895	310.1657	22.754964	112.22747
Medium	35	282.0673	100.28010	293.8712	21.368773	152.71363
Low	71	248.5810	121.79356	242.7016	23.352705	153.38463
Very Low	675	254.1960	135.89086	239.2833	1.033333	195.95417

Eventually, the LOS2, waiting times and physicians' groups in order to obtain major operative goals in the next steps are prepared.

Finally, the new dataset including 806 observations, and 6 variables (features) for modelling are indicated in Figure 11.

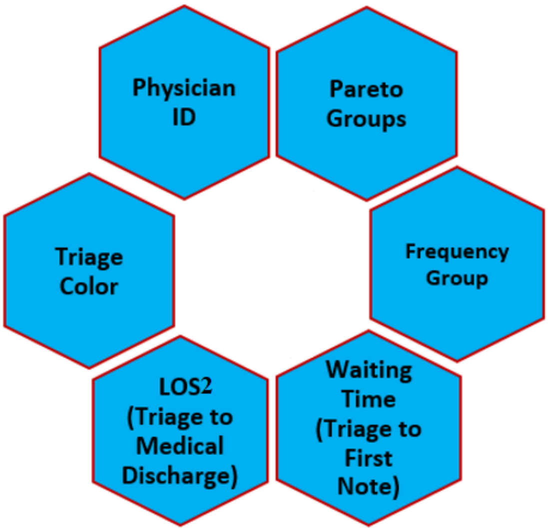


Figure 11 Variables for modelling

Chapter 4

STATISTICAL ANALYSIS

In this chapter, the fourth and fifth phases of the CRISP-DM method, including Modeling and Evaluation are examined.

4.1 Modeling

In this research, the one-way analysis of variance (ANOVA), also known as one-factor ANOVA is applied to compare different groups of physicians based on their patient's waiting time/LOS. Furthermore, the effect of physician groups and triage colour on waiting time/LOS is evaluated simultaneously through a Two-way ANOVA test.

4.1.1 One-way ANOVA Test

Firstly, in this comparing mean modelling, the data is visualized based on the groups by the response variables(Waiting time and LOS).

Visualizing the relationship between variables

Figure 12 demonstrates the distribution of physicians’ average waiting time (from triage colour to their first note) in the Pareto grouping scenario. From the graph, it is clear that the average waiting time for the physicians’ group named “%80” (who visit 80 per cent of the patients) is higher than the “%20” group.

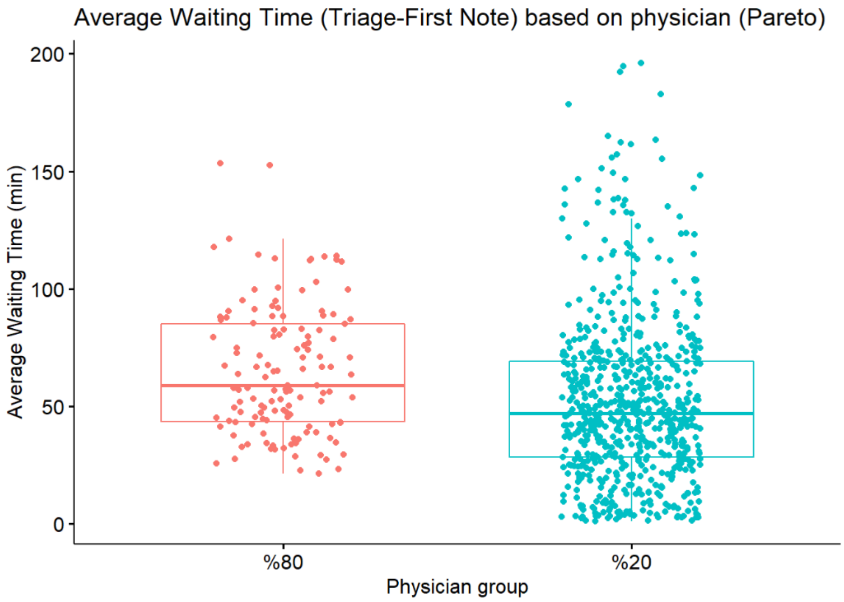


Figure 12 Distribution of Average Waiting Time (Pareto scenario)

Figure 13 illustrates the distribution of physicians' average waiting time (from triage to first note) based on the Frequency grouping scenario. "Very Low" group has the lowest average waiting time, which means the physicians (675 out of 806 physicians) who visiting less patients than the other groups have the lowest average time. However, the "Very High" group has the highest average waiting time meaning that the physicians (7 out of 806 physicians) who visit the largest number of patients than the other groups have the highest average time.

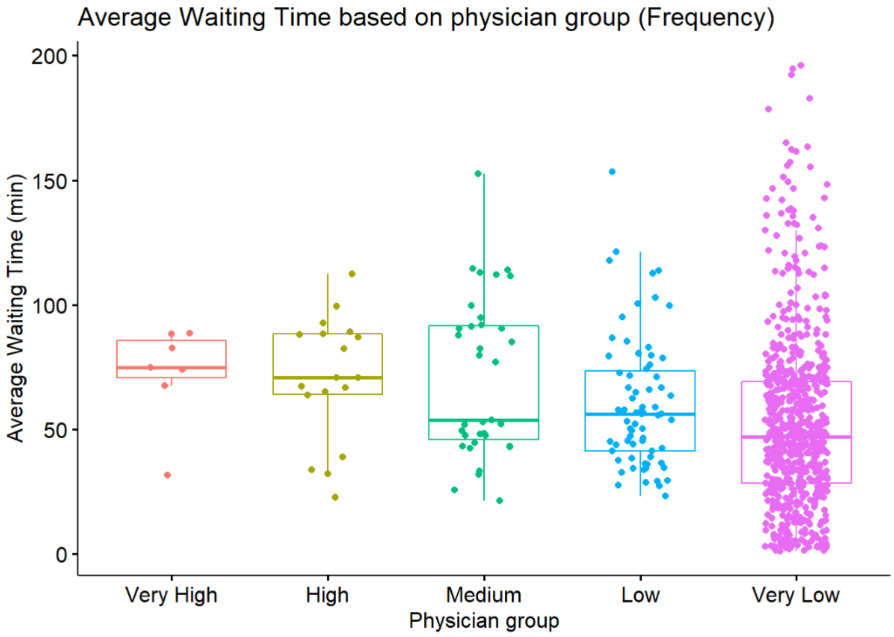


Figure 13 Distribution of Average Waiting Time (Frequency scenario)

The distribution of average LOS physicians' average LOS (from triage colour to medical discharge) in the Pareto grouping scenario is as below (Figure 14).

As can be seen from the graph, the LOS for the physicians' group named "%80" (who visit 80 per cent of the patients) is higher than the "%20" group. It means the physicians (131 out of 806 physicians) who visit more patients than the other group have more LOS.

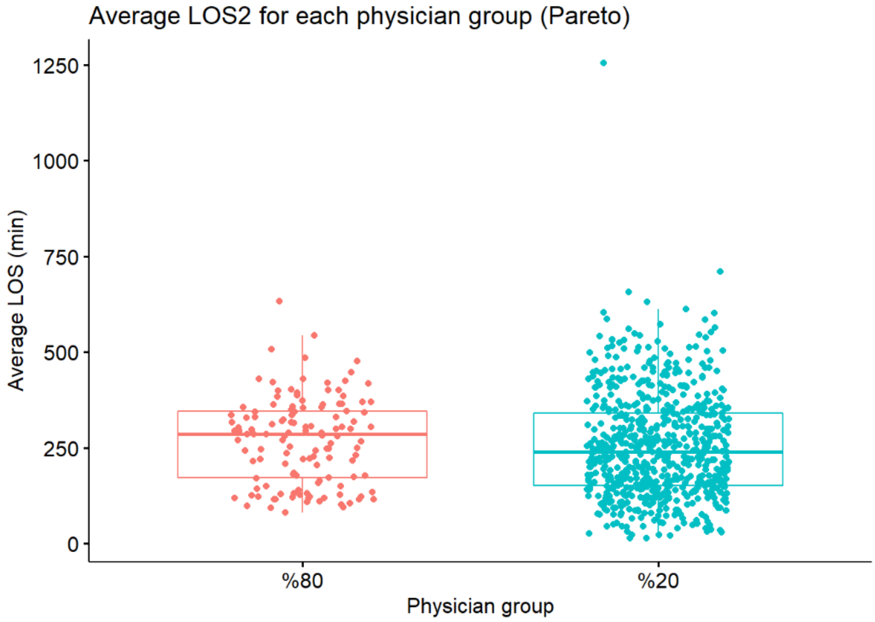


Figure 14 Distribution of Average LOS (Pareto scenario)

Figure 15 shows the distribution of physicians' average LOS (from triage colour to medical discharge) based on the Frequency grouping scenario. "Very Low" and "Low" groups have the lowest average LOS meaning that the physicians (746 out of 806 physicians) who visiting less patients than the other groups have the lowest LOS. However, the "Very High" group has the highest average waiting time meaning that the physicians (7 out of 806 physicians) who visit the largest number of patients than the other groups have the highest LOS.

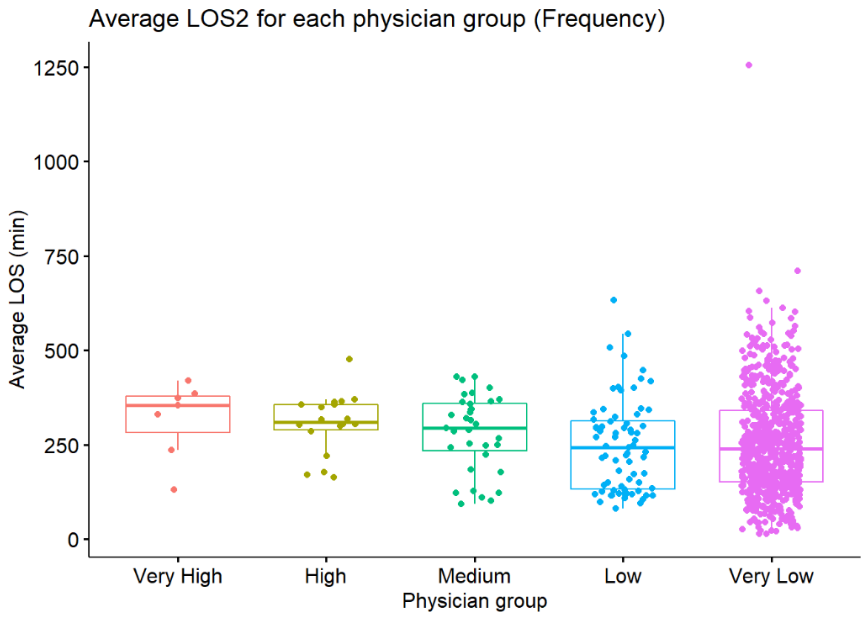


Figure 15 Distribution of Average LOS (Frequency scenario)

In a more detailed review, it can be investigated mean plot with 95% confidence interval (CI) in Figure 16 and Figure 17. A 95% confidence interval means that 95% of the time, the "true" population mean will be within that interval and 5% of the time, the population mean will be outside of that interval. This means you can be 95% confident that the true population mean is within your confidence interval.

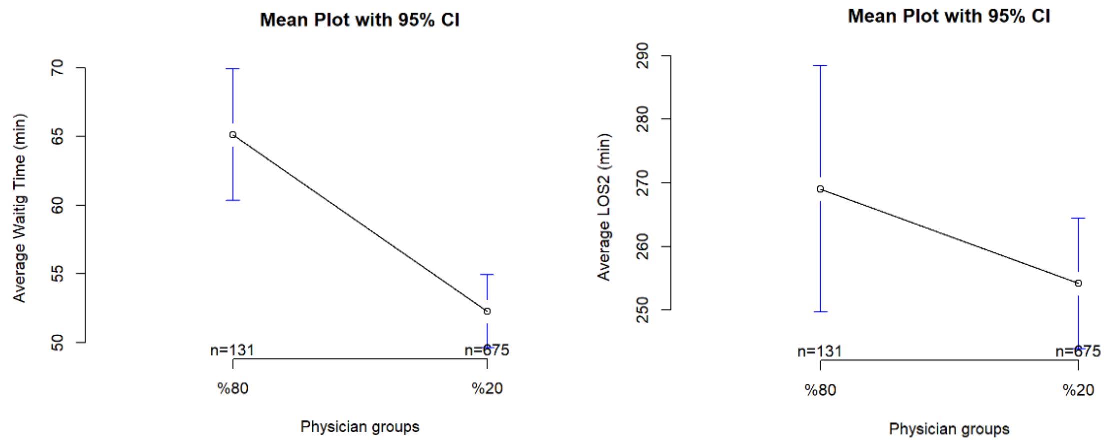


Figure 16 Mean plots with 95% confidence interval (Pareto Grouping)

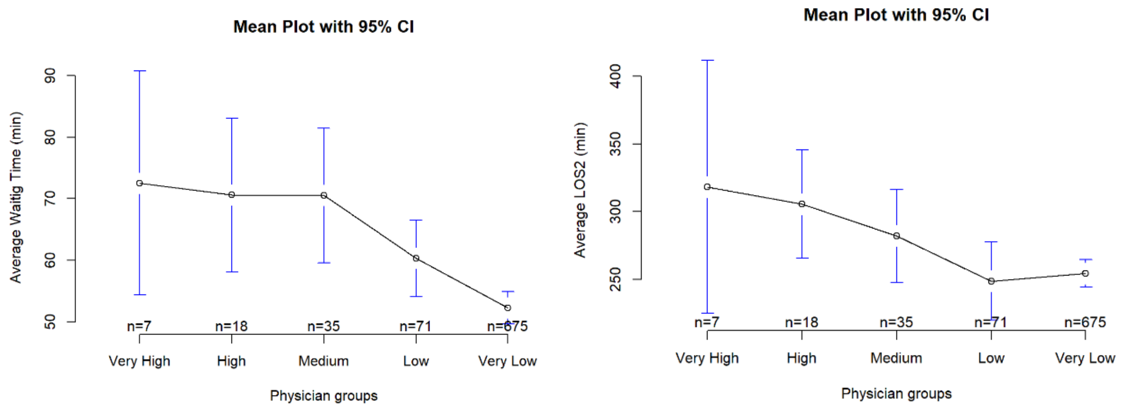


Figure 17 Mean plots with 95% confidence interval (Frequency Grouping)

Assumptions of One-way ANOVA test

As an Assumptions of the ANOVA test, the data of each factor level are normally distributed. Therefore, by applying the Jarque-Bera which is a goodness-of-fit test, we can determine whether sample data fit a normal distribution in terms of skewness and kurtosis or not.

Table 3 Jarque-Bera test for the waiting time

Test statistic	P value	Alternative hypothesis
224.8	0 * * *	greater

Table 4 Jarque-Bera test for the LOS2

Test statistic	P value	Alternative hypothesis
521.2	0 * * *	greater

According to the Jarque-Bera results (Tables 3 and 4), the p-value for both Waiting Time and Length of Stay is less than 0.05, two significant results at the standard level of 0.05, implying rejection of the null hypotheses of the normal distributions.

To tackle this issue a suite of transformation-estimating functions that can be applied to normalize data named “bestNormalize” package in R is used. The best transformation method which is Ordered Quantile (ORQ) normalization (which is introduced as OrderNorm Transformation in the package) with the estimated normality statistic of 0.0048 is picked automatically via this method.

As Figure 18 illustrates the “OrderNorm Transformation” is the suites one visually.

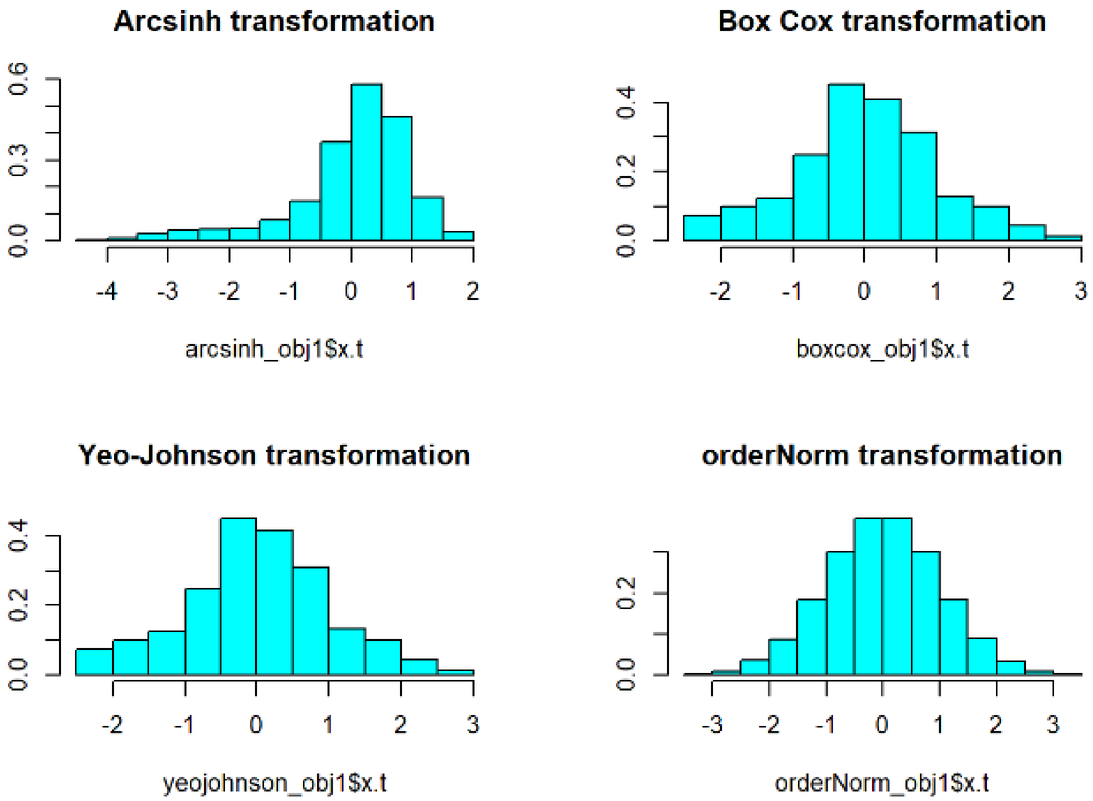


Figure 18 Best Normalizing via the bestNormalize for one way ANOVA

After BestNormalize Transformation the results show that the data is normally distributed (Tables 5 and 6).

Table 5 Jarque test for Waiting Time

Test statistic	P value	Alternative hypothesis
0.03666	0.9818	greater

Table 6 Jarque test for LOS

Test statistic	P value	Alternative hypothesis
0.03664	0.9818	greater

ANOVA test hypotheses

The hypotheses of one-way ANOVA are as below.

- Null hypothesis: The means waiting time/LOS for the different physician groups are the same
- Alternative hypothesis: At least one sample mean is not equal to the others.

Compute the One-way ANOVA test

In Table 7 and Table 8, it is determined if there is any significant difference between the average waiting time/LOS in the 5 experimental conditions (or 2 experimental conditions in the Pareto scenario).

By standard statistical software (R), two ANOVA tables are obtained as below.

Table 7 Analysis of Variance (Waiting Time ~ Pareto)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pareto	1	22.77	22.77	23.41	1.57e-06
Residuals	804	781.9	0.9726	NA	NA

Table 8 Analysis of Variance (Waiting Time ~ Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
visited_patient_class	4	25.55	6.388	6.567	3.347e-05
Residuals	801	779.1	0.9727	NA	NA

Through the visualization of data for one-way ANOVA, in the Pareto scenario indicates that the average waiting time for the physician in the “%80” group (high experienced) is higher than the “%20” group (high experienced). Similarly, in the Frequency scenario, the average waiting time decreases as moving away from the "Very High" physician group to the "Very Low" physician group. These arguments are supported by the ANOVA test where the p-value is less than the significance level of 0.05, it can be concluded that there are significant differences between the average waiting time of the different groups of physicians.

As described in Table 9 and Table 10, in the case of the LOS2, the average LOS2 for the physician in the "%80" group (high experienced) is higher than the "%20" group (high experienced) and the same result is obtained for the Frequency scenario. These results are not statistically significant at an alpha level of 0.05. Therefore, results show that there are no significant differences between the average length of stay of physician groups in the Pareto and Frequency groups.

Table 9 Analysis of Variance (LOS2 ~ Pareto)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pareto	1	2.38	2.38	2.384	0.1229
Residuals	804	802.3	0.9979	NA	NA

Table 10 Analysis of Variance (LOS2 ~ Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
visited_patient_class	4	7.448	1.862	1.871	0.1136
Residuals	801	797.3	0.9953	NA	NA

Tukey multiple pairwise-comparisons

In the one-way ANOVA test, a significant p-value indicates that some of the group means are different, but it does not clear which pairs of groups are different. It's possible to perform multiple pairwise-comparison through Tukey HSD (Tukey Honest Significant Differences), to determine if the mean difference between specific pairs of groups is statistically significant.

Since the Waiting Time ANOVA test is significant, we can compute Tukey HSD for performing multiple pairwise-comparison between the means of groups.

It can be seen from the outputs (Table 11), that only the difference between "Very Low" and "Medium" is significant with an adjusted p-value of 0.0068.

Table 11 multiple pairwise-comparisons (Frequency grouping)

	diff	lwr	upr	p adj
High-Very High	-0.1141	-1.315	1.087	0.999
Medium-Very High	-0.1592	-1.276	0.9573	0.9951
Low-Very High	-0.4117	-1.48	0.6566	0.8301
Very Low-Very High	-0.7368	-1.761	0.2876	0.2835
Medium-High	-0.04507	-0.8272	0.737	0.9999
Low-High	-0.2976	-1.009	0.414	0.7834
Very Low-High	-0.6228	-1.267	0.02122	0.06361
Low-Medium	-0.2525	-0.8094	0.3044	0.7281
Very Low-Medium	-0.5777	-1.045	-0.1102	0.006824
Very Low-Low	-0.3252	-0.6616	0.01123	0.06381

- **diff**: the difference between means of the two groups
- **upr**: Upper bound of the confidence interval at 95% in mean between group
- **lwr**: Lower bound of the confidence interval at 95% in mean between group
- **p adj**: Adjusted p-value when there are multiple groups

4.1.2 The Two-way ANOVA Test

The author used Two-way ANOVA test to evaluate simultaneously the effect of two variables named physicians group and triage colour (or weekdays) on a response variable which can be waiting time or LOS.

Check the data

For two different scenarios, there are 5×5 and 5×2 design cells with the factors being triage colour and physician group and the different number of physicians in each cell.

Table 12 The two-way descriptive data analysis

The number of Physicians

	Very High	High	Medium	Low	Very Low
(1)Red	8	17	25	40	150
(2)Orange	5	11	25	53	483
(3)Yellow	5	10	22	66	602
(4)Green	7	17	30	49	601
(5)Blue	10	14	20	39	340

The number of Physicians (Pareto)

	%80	%20
(1)Red	90	150
(2)Orange	94	483
(3)Yellow	103	602
(4)Green	103	601
(5)Blue	83	340

Visualizing interaction effects

Basically, an interaction effect occurs when the effect of one factor depends on the effect of other factors, and this is seen by the lines in the plot are not parallel and nonparallel lines indicate interaction. In other words, the interaction between two factors occurs when the differences between the mean values of one factor are not consistent across levels of the other factor (Alin & Kurt, 2006).

The Two-way interaction plot visualizes a line plot that shows group differences. In this two-way plot, it can be seen possible interactions, which plot the mean of the response (waiting time) for two-way combinations of factors called physician group (Pareto) and triage colour.

Figure 20 shows that the effect of triage colour is not meaningfully different for the "%80" physician group (who visit 80 per cent of the patients at each triage colour level) and the "%20" physician group. It illustrates a slight difference between the Green triage colour and the Blue one where the physicians who see more patients ("%80" group) have more average waiting time for Green colour patients than the "%20" group.

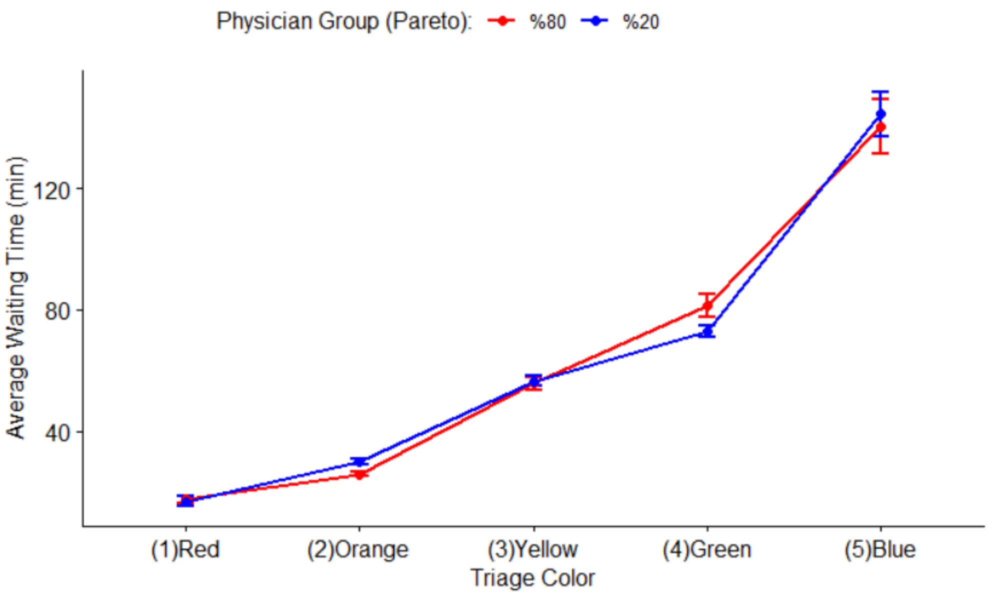


Figure 19 Interaction plot for average waiting time (Pareto grouping)

Figure 21 shows the effect of triage colour and physician groups simultaneously on a response variable which is waiting time. Green and Blue triage colours affect the physician groups differently. For Blue patients, the "Very High" and the "High" groups of physicians have the lowest average waiting time. For Green patients, the "High" group has the highest amount of waiting time. The average waiting time for the Red and Orange patients in different groups of physicians is almost the same.

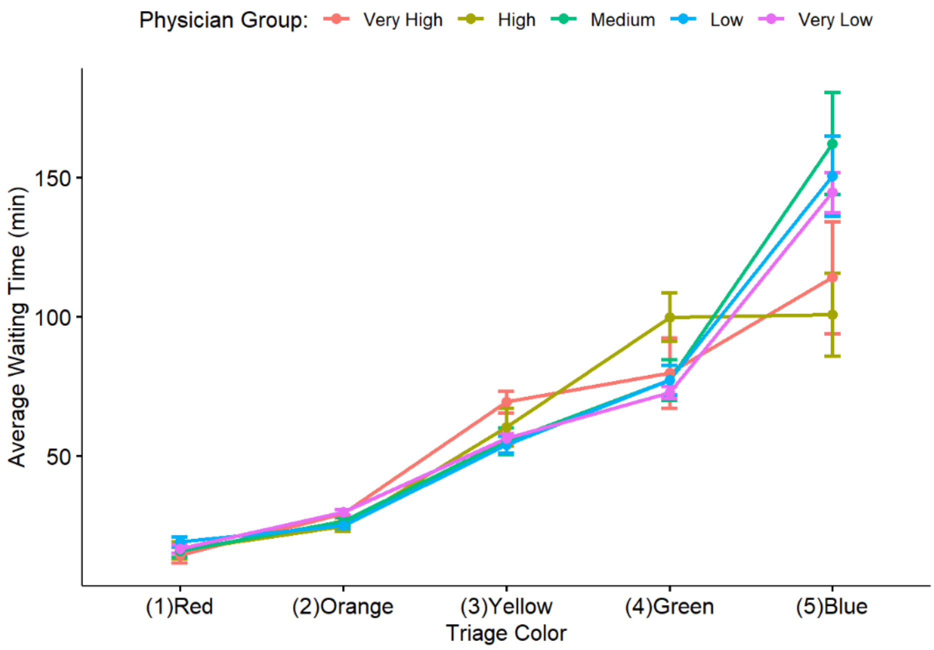


Figure 20 Interaction plot for average LOS

As illustrated by the graph below (Figure 22), in LOS, the effect of triage colour interacts with the physician group. That is, triage colour affects the "%80" physician group differently than the "%20" physician group.

Briefly, we see the red line (the "%80" group) for the emerging, very urgent and urgent patients are higher than the blue line (the "%80" group) whereas the blue line is higher than the red one for the less urgent and not urgent patients. It means that the patients with high urgency levels who visited by the "%80" physician group stayed at the ED more than the patients with high urgency level who are visited by the "%20" physician group.

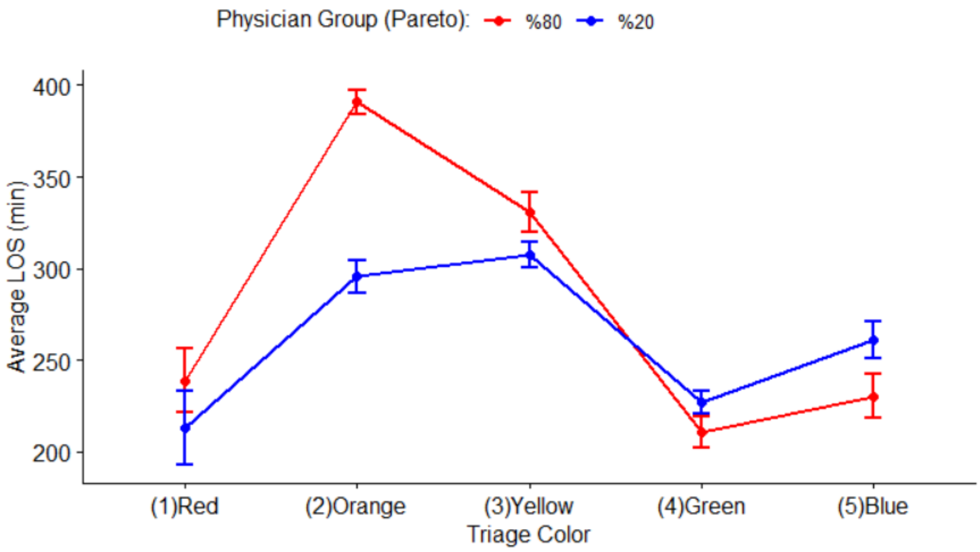


Figure 21 Interaction plot for average LOS (Pareto grouping)

Figure 23 shows how triage colour and physician group affect the average length of stay (LOS). As nonparallel lines indicate interaction, the effect of "Triage Colour" is different for the 5 levels of the physician group. There is a change when going from one triage colour to the next one, and the type of change depends on the physicians' group. For physicians' "High" and "Very High" groups, Red patients experienced longer length of stay than Blue patients.

Furthermore, the emerging patients (Red triage colour) who are visited by high experienced physicians ("Very High" group) have longer LOS than the same patients who are visited by less experienced physicians ("Very Low" group).

On the other hand, not urgent patients (Blue triage colour) who are visited by high experienced physicians ("Very High" group) have shorter LOS than the same patients who are visited by less experienced physicians ("Very Low" group).

For the "Medium" physicians' group, Blue patients experienced less length of stay than did Green.

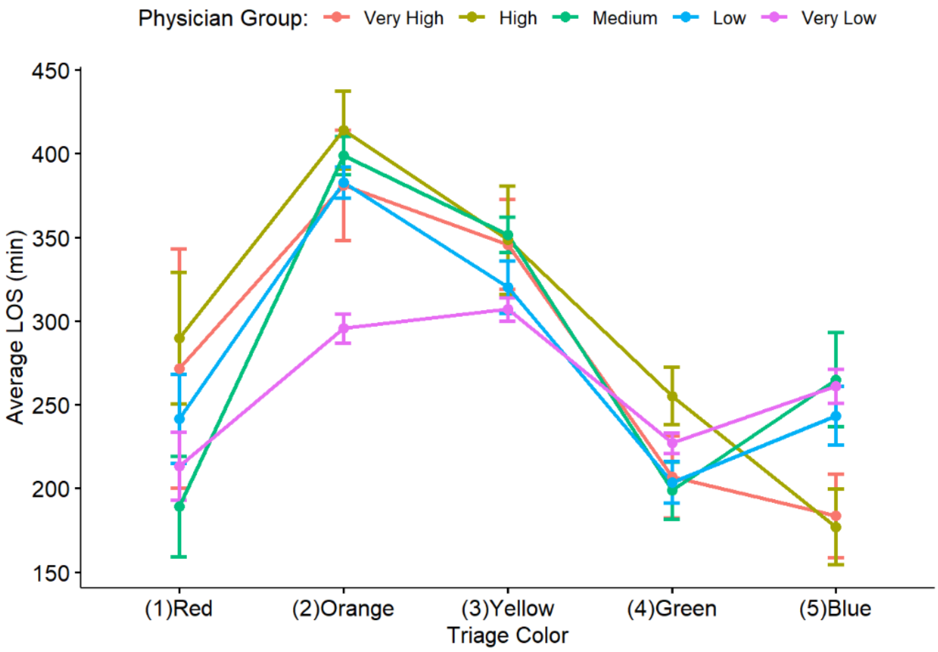


Figure 22 Interaction plot for average LOS2

Assumptions of Two-way ANOVA test

According to the results below (Table 13 and Table 14) A routine check of normality by the test of Jarque-Bera in the P-value 0 for both waiting time and LOS, significant results at the standard level of 0.05, implying rejection of the null hypotheses of the normal distributions.

Table 13 Jarque-Bera test for the waiting time

Waiting Time

Test statistic	P value	Alternative hypothesis
68843	0 * * *	greater

Table 14 Jarque-Bera test for the LOS2

LOS2

Test statistic	P value	Alternative hypothesis
68843	0 * * *	greater

In order to normalize the data same as the One-way ANOVA, the “bestNormalize” package in R is used. Again, the best transformation method which is Ordered Quantile (ORQ) normalization (which is introduced as OrderNorm Transformation in the package) with the estimated normality statistic of 1.101 is picked automatically via this method.

Figure 24 shows the “OrderNorm Transformation” is the suit one visually.

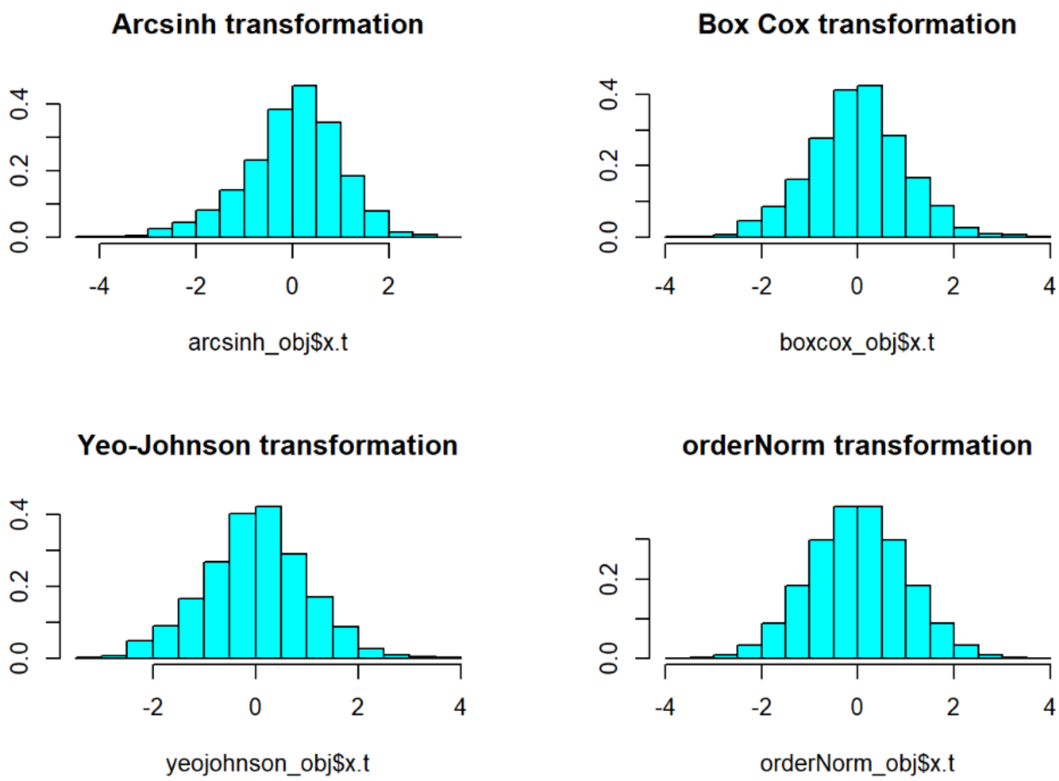


Figure 23 Best Normalizing via the bestNormalize for Two-way ANOVA

Test hypotheses

The hypotheses of Two-way ANOVA are as below.

1. Physician Groups have similar means
2. Triage Colour (or weekdays) has similar means
3. There is no interaction between Physician Groups and Triage Colour

The alternative hypothesis of cases 1 and 2 is “the means are not equal” and the alternative hypothesis for case 3 is “there is an interaction between Physician Groups and Triage Colour”.

Compute The Two-way ANOVA test

Using two-way ANOVA, there are two possible means models: the additive model and the interaction model.

Additive model

From the ANOVA Table 15, we can conclude that both “Triage Colour” and “Physician group” are statistically significant. “Triage Colour” is the most significant factor variable. These results would lead us to believe that changing patients’ urgency level or the group of doctors, will impact significantly the mean waiting time.

Table 15 Two-way ANOVA test for Waiting Time by Triage Colour and Physician group (Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	936.2	234.1	363.4	1.57e-249
visited_patient_class	4	11.22	2.805	4.356	0.001636
Residuals	2640	1700	0.644	NA	NA

As Table 16 depicts, both "Triage Colour" and "Physician group" (Pareto) are statistically significant it means changing the triage colour or Pareto group (20-80), will impact significantly the mean Waiting Time.

Table 16 Two-way ANOVA test for Waiting Time by Triage Colour and Physician group (Pareto)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	936.2	234.1	363.8	8.151e-250
Pareto	1	11.22	11.22	17.44	3.068e-05
Residuals	2643	1700	0.6433	NA	NA

As can be seen from Table 17 and Table 18, "Triage Colour" and "Physician group" (Frequency/Pareto) are statistically significant it is defined changing the triage colour or doctor group, will impact significantly the mean length of stay.

Table 17 Two-way ANOVA test for LOS by Triage Colour and Physician group (Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	190.8	47.71	51.75	5.163e-42
visited_patient_class	4	22.91	5.726	6.211	5.681e-05
Residuals	2640	2434	0.9219	NA	NA

Table 18 Two-way ANOVA test for LOS by Triage Colour and Physician group (Pareto)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	190.8	47.71	51.78	4.845e-42
Pareto	1	21.67	21.67	23.51	1.312e-06
Residuals	2643	2435	0.9214	NA	NA

Interaction effects

The primary aim of performing two-way ANOVA is to find out whether there is an interaction between the two independent features on the dependent feature. In this test, one of these two independent features (Variables) acts as a central feature and the other as a moderator feature.

The resulting ANOVA table of the two-way ANOVA interaction model for waiting time is shown in Table 19 and Table 20. When the result does not show a statistically significant interaction effect, this indicates that the effect of an independent variable is the same for each level of the other independent variable. According to the tables, the first impression is that the main effects including “Triage Colour” and “Physician group” are statistically significant but the interaction (“Triage Colour” * “Physician group”) is not significant at the alpha level of 0.05. It means the joint effect of “Triage Colour” and “Physician group” is not statistically higher than the sum of both effects individually on waiting time.

Table 19 Interaction effects of Two-way ANOVA test for Waiting Time (Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	936.2	234.1	363.4	3.22e-249
visited_patient_class	4	11.22	2.805	4.355	0.001637
Triage_Color:visited_patient_class	16	10.13	0.633	0.9827	0.4728
Residuals	2624	1690	0.6441	NA	NA

Table 20 Interaction effects of Two-way ANOVA test for Waiting Time (Pareto scenario)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	936.2	234.1	364.2	6.204e-250
Pareto	1	11.22	11.22	17.45	3.042e-05
Triage_Color:Pareto	4	4.166	1.041	1.62	0.1663
Residuals	2639	1696	0.6427	NA	NA

If there is no statistically significant interaction effect, it means that the effect of “physician groups” on average LOS is the same for a different level of “Triage Colour”. According to Table 21, the two main effects and the interaction are all statistically significant. The interaction significance test evaluates whether it is valid to conclude that the lines in the population are not parallel. The test result below describes, the significant effect as the “Triage Colour” * “Physicians groups” interaction. Therefore, there is a substantial interaction effect between “Triage Colour” and “Physicians groups” on LOS.

Table 21 Interaction effects of Two-way ANOVA test - for LOS2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	190.8	47.71	52.42	1.54e-42
visited_patient_class	4	22.91	5.726	6.292	4.902e-05
Triage_Color:visited_patient_class	16	45.72	2.857	3.139	2.388e-05
Residuals	2624	2388	0.9101	NA	NA

Same as the previous result, the two main effects and the interaction are all statistically significant on LOS for the Pareto grouping scenario, the results demonstrated in Table 22.

Table 22 Interaction effects of Two-way ANOVA test - for LOS2 (Pareto)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	190.8	47.71	52.42	1.517e-42
Pareto	1	21.67	21.67	23.8	1.131e-06
Triage_Color:Pareto	4	33.16	8.29	9.108	2.636e-07
Residuals	2639	2402	0.9102	NA	NA

4.2 Evaluation

In other to evaluate ANOVA models, the residuals must be checked using the normality test. Here Jarque-Bera normality test is considered. The null hypothesis in this test is that the data follow normal distribution, in other words, there is no trend in residuals. Naturally, if the p-value exceeds the significance level (0.05). the normality assumption is accepted, and we argue that the model is appropriate, and the results are solid. The visualization result of the normality check for each test is provided in the appendix.

4.2.1 One-way ANOVA

Check the normality assumption

Table 23 represents the normality test result on one-way ANOVA models. The null hypothesis states that the residuals are normally distributed.

Table 23 Normality assumption (Jarque-Bera test)

Waiting Time ~ Pareto

Test statistic	P value	Alternative hypothesis
1.092	0.5794	greater

Waiting Time ~ Frequency

Test statistic	P value	Alternative hypothesis
1.165	0.5585	greater

LOS2 ~ Pareto

Test statistic	P value	Alternative hypothesis
0.07363	0.9639	greater

LOS2 ~ Frequency

Test statistic	P value	Alternative hypothesis
0.2188	0.8964	greater

In order to verify that the residuals are normally distributed, a normal probability plot of residuals is used. In the QQ plots below (Figure 25), the quantiles of the residuals are plotted against the quantiles of the normal distribution. Since all the points in all plots are distributed approximately along the reference line, the normality requirement is met. This conclusion is supported by the Jarque-Bera test in which the p-values are not significant ($p > 0.05$), therefore we can assume normality.

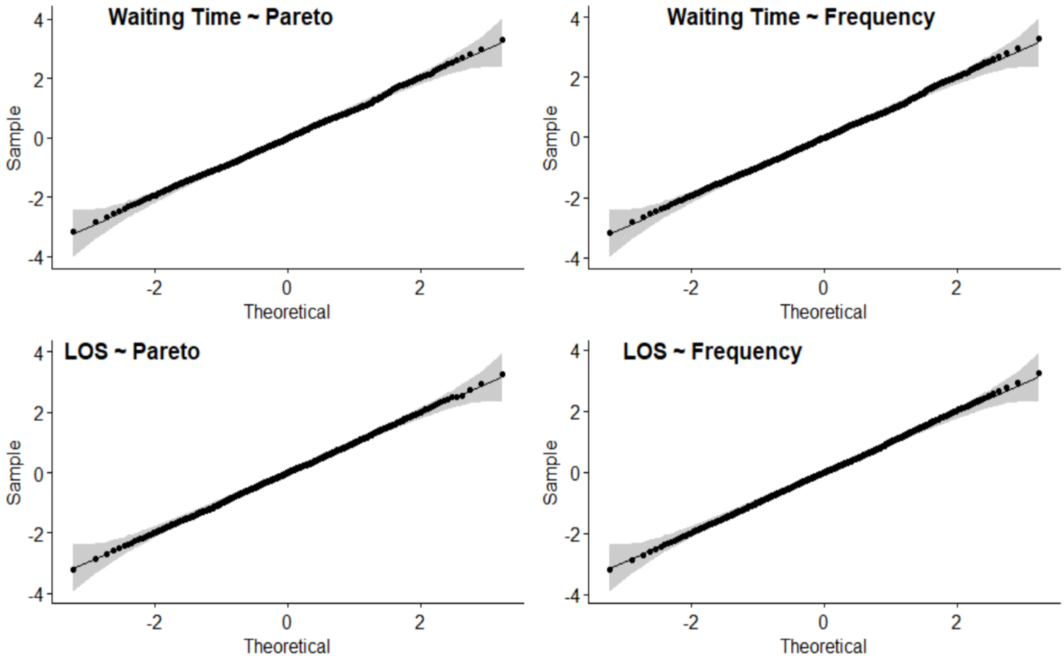


Figure 24 Q-Q plot for residuals of the One-way ANOVA models

From the auto-correlation charts (Appendix) it can be seen that residuals are also approximately normal.

4.2.2 Two-way ANOVA

With a statistical argument similar to that of one-way ANOVA, none of the p-values in Table 24 is significant ($p > 0.05$), therefore we can assume normality.

Table 24 Normality assumption (Jarque-Bera test)

ue-Bera Normality Test: `aov_wt_fr$residuals`

Test statistic	P value	Alternative hypothesis
402.5	0 * * *	greater

Jarque-Bera Normality Test: `aov_wt_p$residuals`

Test statistic	P value	Alternative hypothesis
402.5	0 * * *	greater

Jarque-Bera Normality Test: `aov_los_fr$residuals`

Test statistic	P value	Alternative hypothesis
25.9	2.38e-06 * * *	greater

Jarque-Bera Normality Test: `aov_los_p$residuals`

Test statistic	P value	Alternative hypothesis
25.68	2.656e-06 * * *	greater

Figure 26 illustrates Q-Q plot for residuals of the two-way ANOVA mode.

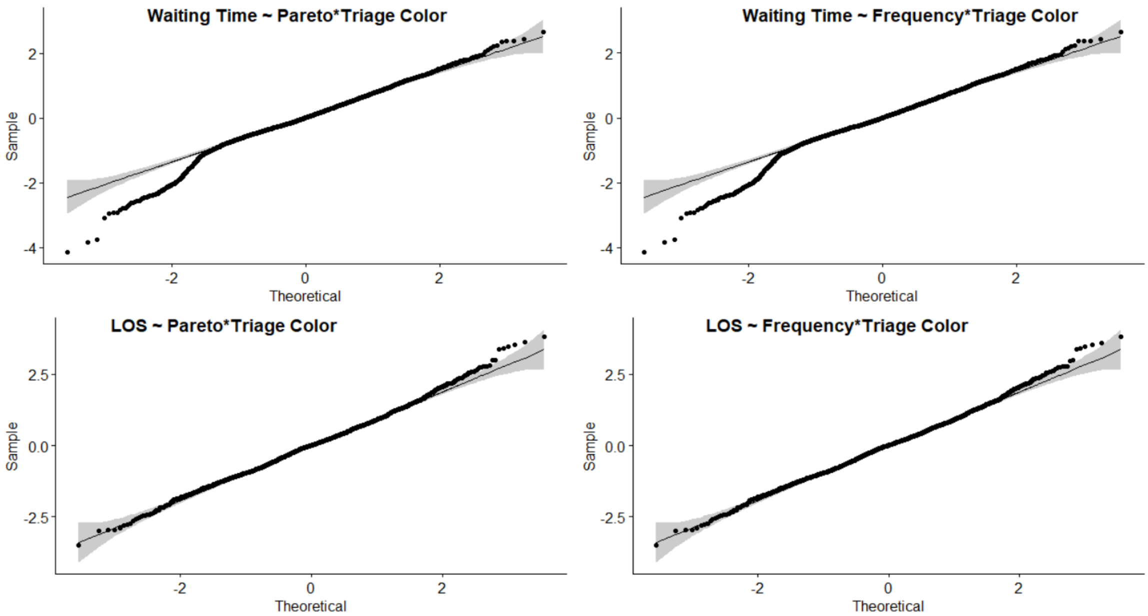


Figure 25 Q-Q plot for residuals of the Two-way ANOVA models

Chapter 5

DISCUSSION

During this research, the author aimed to study the effect of physicians' performance (measured by Group variable in the data-analyzed for the Pareto group and Frequency group separately) on their patients' waiting times and length of stay in the ED (general surgery unit) of a Portuguese hospital using real data.

The present research is a multidisciplinary work that joins a management approach with a computer science (data science) to create knowledge about the ED process. Some initial arguments are provided using data manipulation and data visualization.

5.1 Answer to the Research Questions

By conducting the data manipulation, data visualization and data modeling through CRISP-DM method, the research questions are answered as follows.

RQ1-Different groups of physicians (based on their experiences) have a significant effect on the average waiting time. In other words, it can be concluded that there are significant differences between the average waiting time of the different group of physicians. But different groups of physicians do not have a significant effect on the average LOS2. These arguments obtained from Figure 16 and Figure 17 (Mean plots with 95% confidence interval) and they supported by the one-way ANOVA test (From Table 7 to Table 10).

RQ2-The results show that changes in triage levels and physicians' groups significantly affect the average waiting time as the response variables. This argument is derived from Figure 20 and Figure 21 and it supported by the two-way ANOVA test (Table 19 and Table 20).

RQ3-Moreover, the results demonstrate that changes in the triage level and physician group have a significant effect on the average LOS2. Figures 22 and 23 support the argument, as well as two-way ANOVA results (Table 21 and Table 22), confirm that.

5.2 Analysis of The Findings

One-way ANOVA modelling for the Pareto scenario indicates that the average waiting time for the physician in the "%80" group (high experienced) is higher than that of the "%20" group (less experienced). The results of the ANOVA test show that physicians' groups have a significant effect on the average waiting time in the Pareto scenario. Similarly, in the Frequency scenario, the group variable (physicians' performance) has a significant effect on the average waiting time at an alpha level of 0.05.

Considering LOS2 as the response variable, the physician in the "%80" group (high experienced) is higher than that of the "%20" group (less experienced). The results of the ANOVA test show that, the physicians' group has not a significant effect on the average LOS2 in the Pareto scenario. Similarly, in the Frequency scenario, the group variable (physicians' performance) has not a significant effect on the average LOS2.

The results from two-way ANOVA show that the triage colour and the physician groups are statistically significant on both average waiting time and average LOS2. It shows that changes in triage colour or physicians' performance will significantly impact the average waiting time and average LOS2 as the response variables.

In addition, regarding LOS2 interactions, we have a statistically significant interaction effect between "physician groups" and "Triage Colour" in their relationship with average LOS2. It determines that the effect of different "physician groups" in the presence of the different levels of "Triage Colour" is significant on the average LOS2. But we do not have a statistically significant interaction effect in the case of the model with average waiting time as the response variable. It indicates that the effect of different "physician groups" considering "Triage Colour" on the average waiting time is the same.

According to the visualization and with the support of ANOVA results, the emerging patients (Red triage colour) who are visited by high experienced physicians (the "Very High" group) have longer LOS2 than the same patients' triage level who are visited by less experienced physicians ("Very Low" group). And not urgent patients (Blue triage colour) who are visited by high experienced physicians ("Very High" group) have shorter LOS than the same patients who are visited by less experienced physicians ("Very Low" group).

"Triage Colour" and "Physician group" (in both Frequency and Pareto scenarios) are statistically significant is indicate changing the triage colour or doctor group, will impact significantly the mean length of stay. And regarding waiting time interactions, Green and Blue triage colours affect the physician group differently. For Blue patients, "Very High" and "High" groups of physicians have the lowest average waiting time. For Green patients, the "High" group has the highest amount of waiting time. The average waiting time for the Red and Orange patients in different groups of physicians is almost the same.

The results of this study are in line with the findings of Hemmati et. al as they claimed that considered Low skillfulness, experience, and knowledge of the staff as one of the main reasons that affect physician decision time (Hemmati, Mahmoudi, Dabbaghi, Fatehi, & Rezazadeh, 2018). Also, here the achievements support the argument of Krall et al. which represented waiting time as an effective factor in the patients' evaluation interval (Krall, Cornelius, & Addison, 2014).

Chapter 6

CONCLUSIONS

Data mining through extracting the key pattern and information from a large amount of data has the potential to improve the hospital (emergency department) management process and make informed decisions for healthcare providers.

In this study, the possibility of the impact of different groups of physicians on the ED process from the perspective of their proficiency comparison is investigated. For that, the corresponding metrics indicate the efficiency of this process including the patients' length of stay and waiting time.

From the results, it is indicated that there are significant differences between physician groups in terms of average waiting times (in both Pareto and Frequency grouping scenarios). As such, the average waiting time of physicians who visit more patients (high experienced physicians) is more than the average waiting time of physicians who visit less patients (less experienced physicians). Although there are no significant differences between the average LOS2 of different physician groups.

It also indicates that the impact that physicians' groups (based on their experiences) have on the patients' LOS2 depends on the triage colour. while the impact that physicians' groups (based on the experiences) have on the patients' waiting time does not depend on the triage colour.

Chapter 7

RESEARCH SUMMARY

In this study, the possibility of the impact of physicians on the ED process from the perspective of their proficiency comparison is investigated. For that, the corresponding metrics which indicate the efficiency of this process should be found in the analyzing data.

The main goal of this research is to improve the ED process using the CRISP-DM method over real-world data that got from an emergency department (General Surgery Unit) of a Portuguese hospital.

Within the data visualization, it is discovered that by splitting the physicians into different groups, the working quality of different groups of physicians can be compared considering the effective variables (waiting time and length of stay).

From the data, it is possible to group physicians into different classes from high experienced (high hours working) to less experienced (less hours working) according to the number of visited patients at the general surgery of emergency department. Two grouping scenarios were applied, Pareto and Frequency.

In the Pareto grouping scenario, the physicians are divided into two groups, “80” and “20”. Group named “80” consist of 131 physicians who visited 80 per cent of the patients and Group “20” includes 675 physicians who visited 20 per cent of the patients during the three years.

In the Frequency grouping scenario, physicians split into 5 groups based on the cumulative sum of Patients. According to this method, the physicians were placed in one of the "Very High", "High", "Medium", "Low" and "Very Low" groups. Hence, the physicians who visited more patients fell into the "Very High" group provided they are part of the first 20 per cent of the cumulative sum of Patients.

During this research, the author applied data mining techniques with several input variables and data manipulation to find the most accurate results. The CRISP-DM process consists of 6 phases, and in this research, the first 5 phases of the method are applied to analyse the data.

In the first and second phases, business understanding and data understanding including collecting, describing, exploring, and verifying the data were performed. Then the Data preparation, which is the most time-consuming phase and includes selecting, cleaning, exploring, verifying the quality, integrating, and formatting data was done. In this phase, some informative plots were created to describe better the dataset. The manipulated data shows how a hospital works and can be used to understand staff efficiency. The fourth phase called modelling which comprises selecting a modelling technique, generating a test design, building a model, and finally evaluating the model was performed. In this research, the one-way analysis of variance (ANOVA), was applied to compare different groups of physicians based on their patient's waiting time/LOS. Furthermore, the effect of physician groups and triage colour on waiting time/LOS is evaluated simultaneously through the Two-way ANOVA test. Finally, in the fifth phase, the models were evaluated and interpreted.

Through the one-way ANOVA method, it was found that there are significant differences between the average waiting time of physician groups (in both Pareto and Frequency grouping scenarios). As such, the average waiting time of physicians who visit more patients is more than the average waiting time of physicians who visit less patients. Moreover, the average LOS belongs to the physicians who visit more patients are more than the average LOS belongs to the physicians who visit less patients (less experienced physicians).

As a result of two-way ANOVA, we discovered that the “Triage Colour” and the “Physician group” are statistically significant which means changing the triage colour or Physician group, will significantly impact the average Waiting Time (or average LOS).

In addition, regarding LOS interactions, we have a statistically significant interaction effect on average LOS at an alpha level of 0.05. This determines that the effect of different “physician groups” considering “Triage Colour” on the average LOS are different. But we do not have a statistically significant interaction effect on average waiting time at an alpha level of 0.05. This indicates that the effect of different “physician groups” considering “Triage Colour” on the average waiting time is the same. In other words, the emerging patients (Red triage colour) who were visited by high experienced physicians (“Very High” group) have longer LOS than the same patients' triage level who were visited by less experienced physicians (“Very Low” group).

To sum up, average waiting time and LOS belong to different groups of physicians considering triage colour are compared to each other. Then the main effect of the physicians’ group and triage colour and their interaction are highlighted. The findings assist to plan and manage emergency department operations and resources.

7.1 LIMITATIONS OF THE STUDY

Due to privacy issues, it is not easy to get hospital data, but we should invest in an automatic process to extract this data from the hospital in an anonymized process. This allows expanding easy this approach to other hospitals.

7.2 SUGGESTIONS FOR FURTHER RESEARCH

This approach can be replicated in other hospitals and also in other departments. From this, it is possible to improve the management process and check other variables such as weekdays. Also, it is important to merge this hospital data with external data, like weather data, and local events to increase the variables to be analysed.

References

Alin, A., & Kurt, S. (2006). Testing non-additivity (interaction) in two-way ANOVA tables with no replication. *Statistical Methods in Medical Research*, 63-85. doi:10.1191/0962280206sm426oa

Andel, V. v., Beerepoot, I., Lu, X., Weerd, I. v., & Reijers, H. A. (2021). The data in healthcare system reflects the reactions of physicians and healthcare professionals in the emergency situations. *European Conference on Information Systems (ECIS)* .

Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T. (2015). Big data in healthcare: Challenges and opportunities. *2015 International Conference on Cloud Technologies and Applications (CloudTech).* Marrakech. doi:10.1109/CloudTech.2015.7337020

Au-Yeung, S. W., Harder, U., McCoy, E. J., & Knottenbelt, W. J. (2009). Predicting patient arrivals to an accident and emergency department. *Emergency Medicine Journal*, 241–244. doi:10.1136/emj.2007.051656

Awangga, R. M., Pane, S. F., Tunnisa, K., & Suwardi, I. S. (2018). K Means Clustering and Meanshift Analysis for Grouping the Data of Coal Term in Puslitbang tekMIRA. *Telkomnika*, 1351-1357.

Bittencourt, R. J., Stevanato, A. d., Bragança, C. T., Gottems, L. B., & O'Dwyer, G. (2020). Interventions in overcrowding of emergency departments: an overview of systematic reviews. *Revista de Saúde Pública*. doi:https://doi.org/10.11606/s1518-8787.2020054002342

Caetano, N., Cortez, P., & Laureano, R. M. (2015). Using Data Mining for Prediction of Hospital Length of Stay: An Application of the CRISP-DM Methodology. *International Conference on Enterprise Information Systems*, 149-166.

Caetano, N., Cortez, P., & Laureano, R. M. (2015). Using Data Mining for Prediction of Hospital Length of Stay: An Application of the CRISP-DM Methodology. *International Conference on Enterprise Information Systems*, 149–166.

Chauhan, R., & Jangade, R. (2016). A robust model for big healthcare data analytics. *6th International Conference - Cloud System and Big Data Engineering (Confluence)*.

Fisher, N. I. (2013). *Analytics for leaders: A performance measurement system for business success.* Cambridge University Press.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera , F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*.

Gul, M., & Celik, E. (2017). An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *HEALTH SYSTEMS*, 263-284.

Hemmati, F., Mahmoudi, G., Dabbaghi, F., Fatehi, F., & Rezazadeh, E. (2018). The factors affecting the waiting time of outpatients in the emergency unit of selected teaching hospitals of Tehran. *Electron J Gen Med*. doi:10.29333/ejgm/93135

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice (2nd ed).* Melbourne: Monash University, Australia.

Institute of Medicine. (2007). *Hospital-Based Emergency Care: At the Breaking Point.* Washington, DC: National Academies Press.

Kankanhalli, A., Hahn, J., & Tan, S. (2016). Big data and analytics in healthcare: Introduction to the special section. *Inf. Syst. Front*, 233–235. doi:10.1007/s10796-016-9641-2

Krall, S. P., Cornelius, A. P., & Addison, J. B. (2014). Hospital Factors Impact Variation in Emergency Department Length of Stay more than Physician Factors. *Western J Emerg Med*, 158–164.

Kudyba, S., & Gregorio, T. (2011). Identifying factors that impact patient length of stay metrics for healthcare providers with advanced analytics. *Health Informatics Journal*, 235–245.

Kyriacou, D. N., Ricketts, V., Dyne, P. L., McCollough, M. D., & Talan, D. A. (1999). A 5-Year Time Study Analysis of Emergency Department Patient Care Efficiency. *Annals of Emergency Medicine*, 326-335.

Lee, G., Endacott, R., Flett, K., & Bushnell, R. (n.d.). Characteristics of patients who did not wait for treatment in the emergency department: A follow up survey. 56-62.

Madsen, L. B. (2012). *Healthcare business intelligence: a guide to empowering successful data reporting and analytics.* John Wiley & Sons.

Nhdi, N. A., Asmari, H. A., & Thobaity, A. A. (2021). Investigating Indicators of Waiting Time and Length of Stay in Emergency Departments. *Open Access Emergency Medicine* , 311-318. doi:https://doi.org/10.2147/OAEM.S316366

Pines, J. M., & Griffey, R. T. (2015). What we have learned from a decade of ED crowding research. *Academic Emergency Medicine*, 985-987.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst.* doi:10.1186/2047-2501-2-3

Remesan, R., & Mathew, . (2014). *Hydrological Data Driven Modelling: A Case Study Approach.* Springer.

Sariyer, G., Taşar, C. Ö., & Cepe, G. E. (2019). Use of data mining techniques to classify length of stay of emergency department patients. *Bio-Algorithms and Med-Systems*, 35-37. doi:15:10.1515/bams-2018-0044

Song, H., Tucker, A. L., & Murrell, K. L. (2015). The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay. *Institute for Operations Research and the Management Sciences (INFORMS)*, 3032–3053.

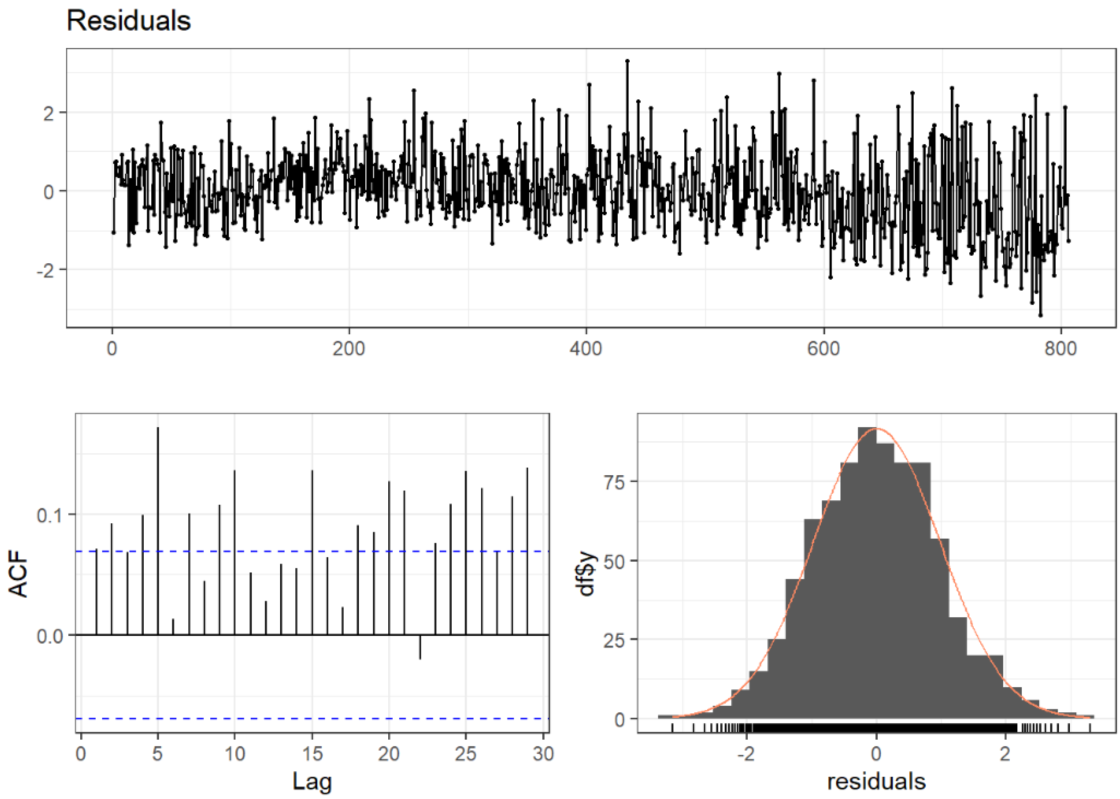
Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, *1*, pp. 29-40.

Xu, M., Wong, T., & Chin, K. (2013). Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network. *Decision Support Systems*, 1488-1498.

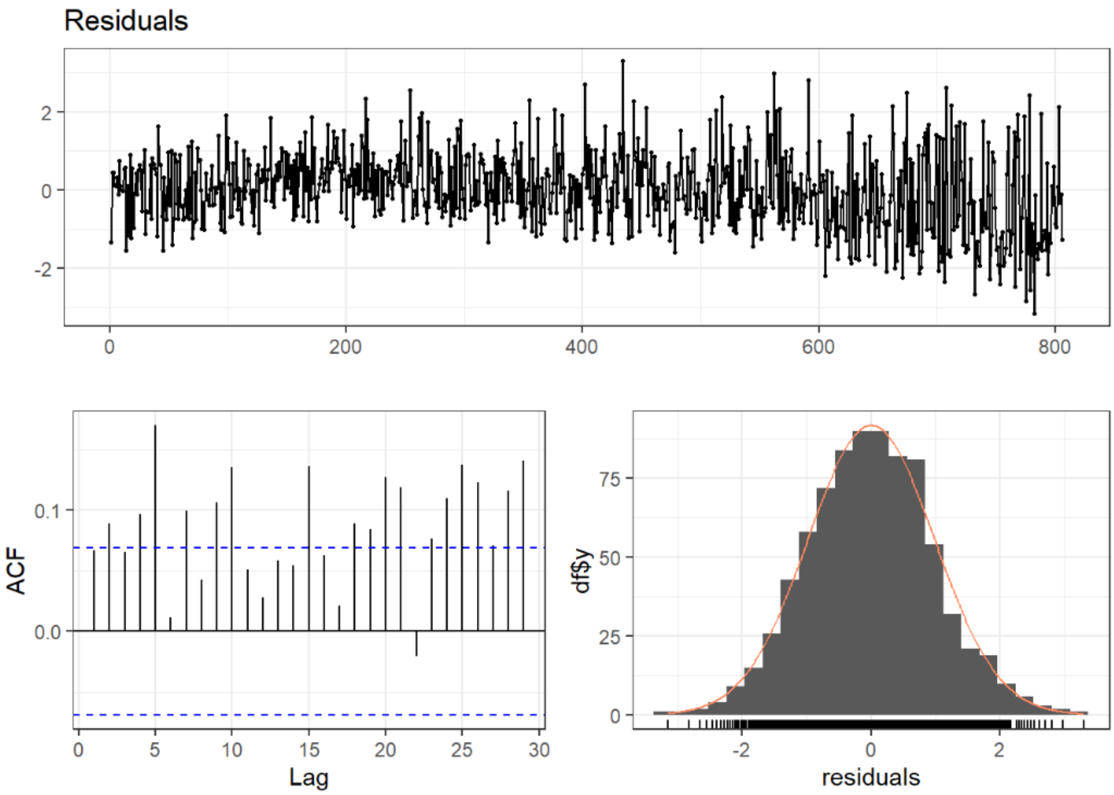
Appendix A

Residuals (One-way ANOVA)

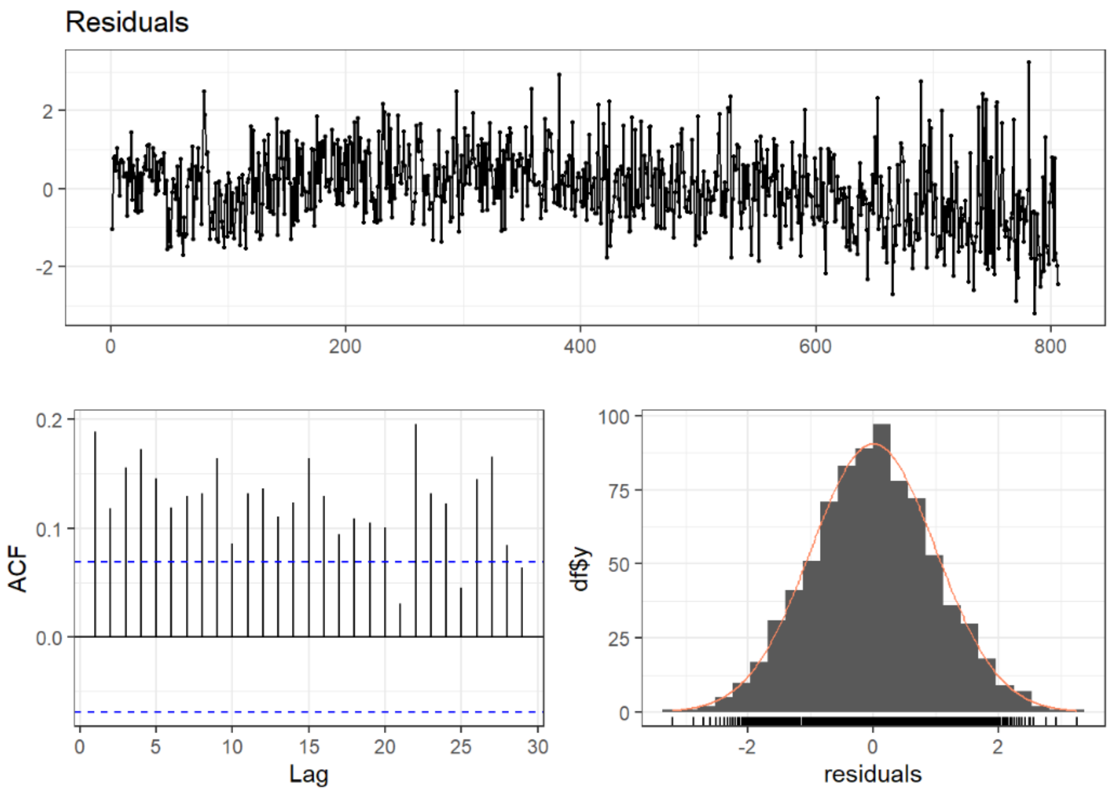
Waiting Time ~ Pareto



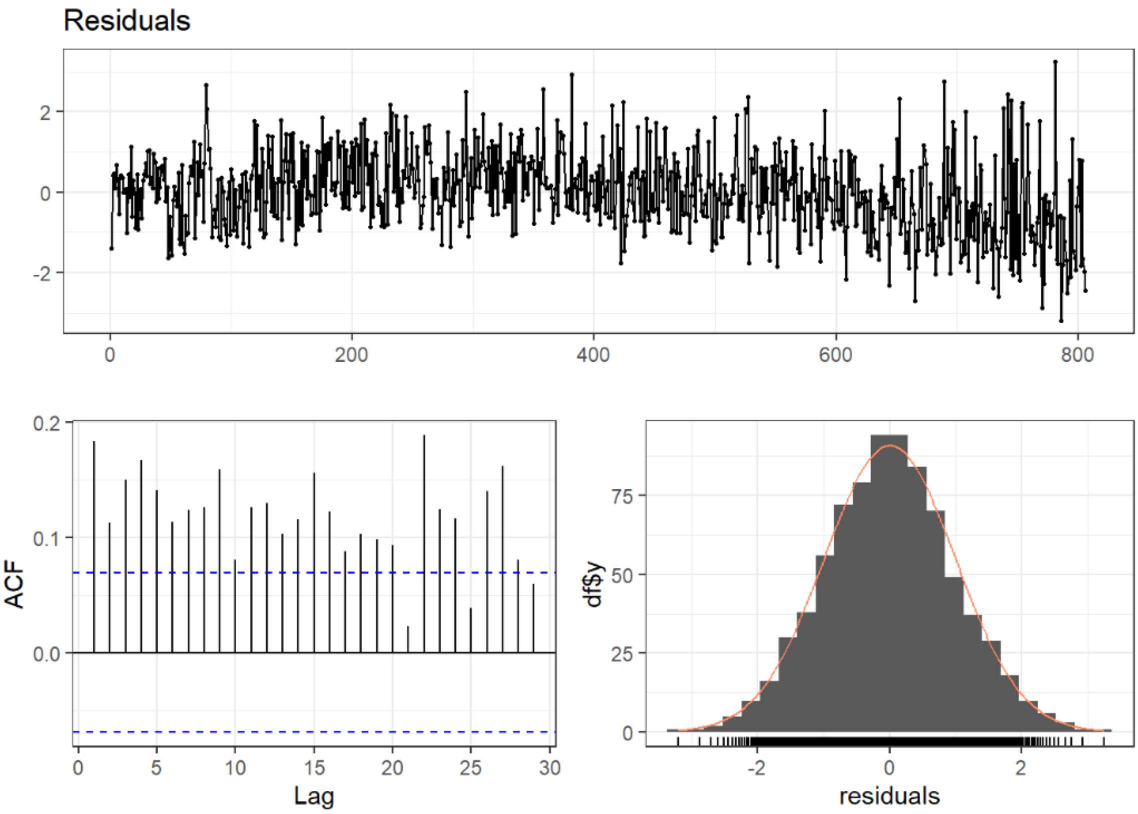
Waiting Time ~ Frequency



Length of Stay ~ Pareto



Length of Stay ~ Frequency

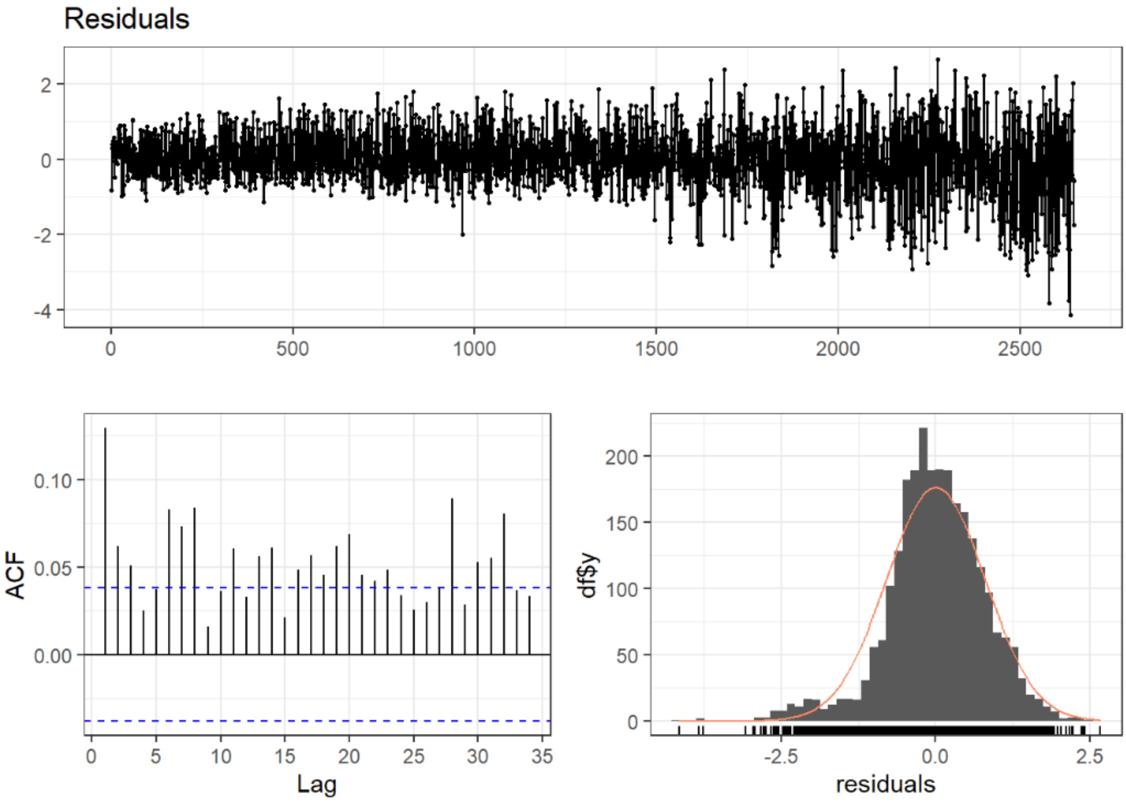


Residuals
2
0
-2
0
200
400
600
800
0.2
0.1
0.0
ACF
0
5
10
15
20
25
30
Lag
75
50
25
0
df$y
-2
0
2
residuals

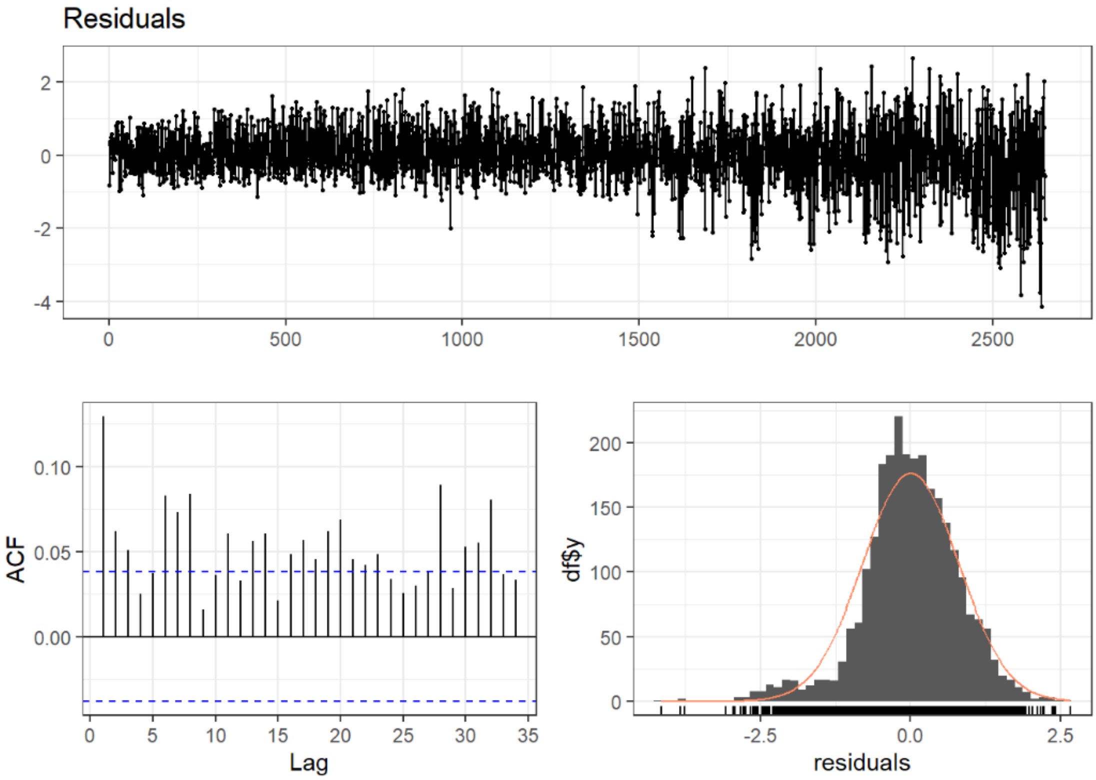
Appendix B

Residuals (Two-way ANOVA)

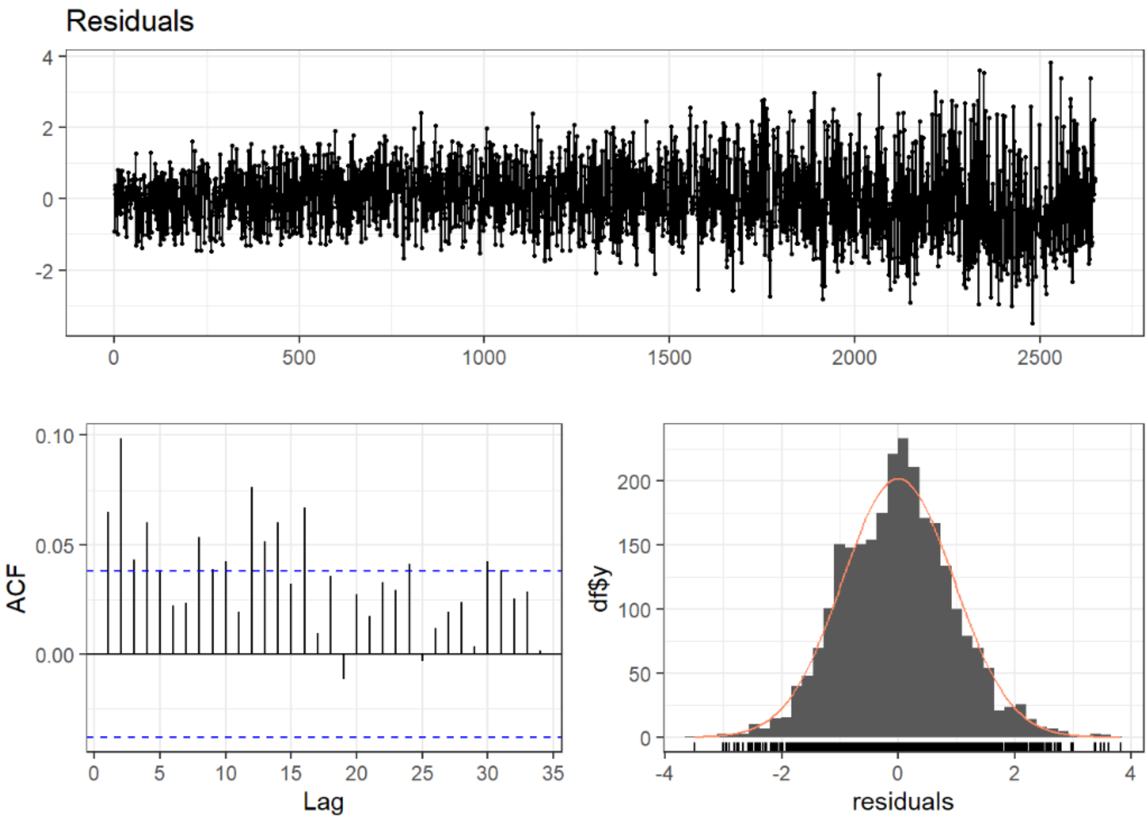
Waiting Time ~ Pareto



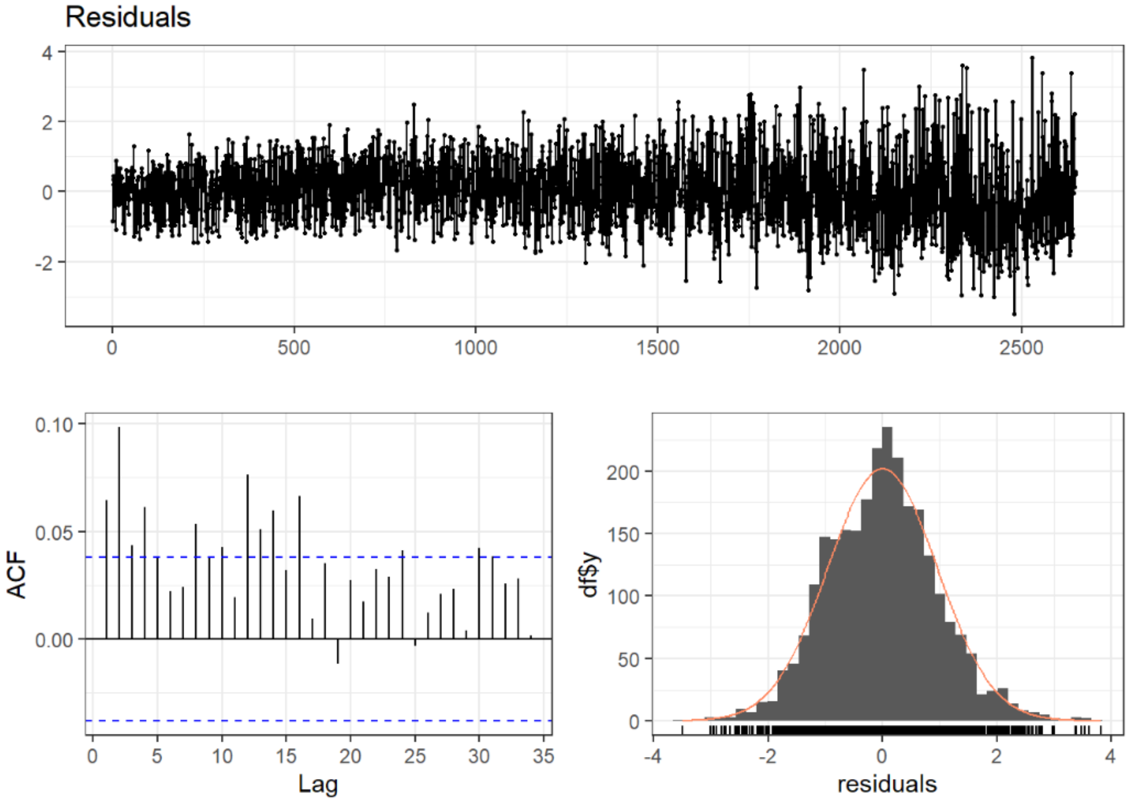
Waiting Time ~ Frequency



Length of Stay ~ Pareto



Length of Stay ~ Frequency


Residuals
4
2
0
-2
0
500
1000
1500
2000
2500
0.10
0.05
0.00
ACF
0
5
10
15
20
25
30
35
Lag
200
150
100
50
0
df$y
-4
-2
0
2
4
residuals

2. The Research Paper

Raeisi Saeed[1], Berit Helgheim[1 *] and João C Ferreira[1,2,3]

[1] Molde College University, Molde, Norway

[2]Inov Inesc Inovação—Instituto de Novas Tecnologias, 1000-029 Lisbon, Portugal

[3] ISTAR, ISCTE-Instituto Universitário de Lisboa, 1649-026 Lisboa, Portugal

***** Correspondence: saeed.raeisi@himolde.no; Tel.: +47 413 77 858

Physician Efficiency Considering Patients' Waiting Time and Length of Stay Using Data Mining (CRISP-DM Method)

Abstract

Background: The digital age, with ICT, IoT, and big data, has opened new opportunities for improving the delivery of healthcare services, in which the data mining approach can help improve the hospital management process by providing a big picture identifying process efficiencies.

Objective: This is a multidisciplinary study where a data mining approach extracting the key pattern and information from a large amount of data has the potential to improve the admission process in the emergency department (ED) and make informed decisions for healthcare providers. From this data, we can provide a big picture of the admission process. Since we get data from the emergency department, we check the process, patients waiting time, length of stay (LOS) and physicians' efficiency.

Methods: A data analytic investigation of the impact of patients' average waiting time and length of stay (LOS) was performed to compare the efficiency of different groups of physicians. This research is performed based on the CRISP-DM method by using an emergency department's (general surgery unit) real-life data from 2015 to 2017, through the One-Way Analysis of Variance (ANOVA) and Two-way ANOVA methods.

Results: From the developed approach using hospital data, it indicates that the waiting time and average LOS belonging to experienced physicians (who visit more patients) are longer than those less experienced physicians (who visit less patients). Using the Two-Way ANOVA method shows that the impact that physicians' groups (based on their experiences) have on the patients' LOS2 depends on the triage colour. while the impact that physicians' groups (based on the experiences) have on the patients' waiting time does not depend on the triage colour.

Conclusions: We show that data visualization tools for healthcare analysts to help them make better decisions, by the big picture identified. Big data offers a lot of potential for improving healthcare administration and taking the sector to the next level, identifying working problems, and reducing costs for academicians, practitioners, and researchers working in the field of healthcare management; above all the findings assist to plan and manage emergency department operations and resources.

Keywords: Healthcare; Data Mining; Waiting Time; Length of Stay; Physicians efficiency

Introduction

Healthcare organizations throughout the globe are interested in improving quality and performance; it is essential to define healthcare performance and establish quality improvement dimensions and strategies. Numerous studies describe how healthcare performance improvement encompasses the combined and ongoing efforts of all healthcare stakeholders, including healthcare professionals, patients and their families, researchers, payers, planners, and educators, in order to implement the changes that will result in improved patient outcomes, system performance, and professional development.

In recent decades, health analytics (data analytics over health data) has evolved as a major field of study and application, demonstrating the extent of the impact of data and information-based management on problem-solving and decision-making in modern healthcare organizations. Hospitals and other healthcare organizations have been adopting descriptive health analytics for medical data.

More recently, healthcare data warehouses became available by aggregating many sorts of data from various systems and sources in order to generate operational healthcare dashboards, strategic scorecards, and data storage. The primary aims of health analytics are to identify performance gaps and provide the most effective solutions. Health analytics should aid healthcare professionals and organizations in monitoring performance on a continuous and consistent basis, as well as in diagnosing poor performance and identifying the underlying causes of issues. Health analytics assist users in designing, developing, implementing, and evaluating several key performance indicators that may improve continuous monitoring, uncover causes for performance variations, and ultimately optimize performance (Fisher, 2013).

Health analytics is a business-driven phrase that spans a broad range of business intelligence applications and big data analysis. This new concept is predicated primarily on the availability and accessibility of data and information pooled through the effective integration and interoperability of a vast array of systems and tools, including hospital information systems, electronic medical records, clinical decision support systems, and other specialized medical systems (Madsen, 2012).

Service quality in EDs is measured by waiting time and LOS (Xu, Wong, & Chin, 2013), therefore these two variables are the most important factors to measure the physician's efficiency. No systematic and exhaustive review has been provided to cover the physician efficiency considering LOS and waiting time. Studying time is a valuable way to identify areas in the patient care process that are causing delays (Kyriacou, Ricketts, Dyne, McCollough, & Talan, 1999). Service quality in the Eds is measured by waiting time and LOS (Xu, Wong, & Chin, 2013). The waiting time is one of the most common complaints in ED (Lee, Endacott, Flett, & Bushnell). Therefore, waiting time is the most important factor directly influencing patient satisfaction in ED (Nhdi, Asmari, & Thobaity, 2021).

LOS plays a critical role in healthcare organizations and the complexity of fully identifying all factors that may impact LOS remains (Kudyba & Gregorio, 2011). LOS is defined in terms of the inpatient days which are computed by subtracting the day of admission from the day of discharge (Caetano, Cortez, & Laureano, 2015). Song et. al found that patients experience shorter LOS when physicians work in a dedicated queuing system with a fairness constraint as opposed to a pooled queuing system with the same fairness constraint using a hospital's ED data (Song, Tucker, & Murrell, 2015).

The staffing of physicians has been considered a factor that affects ED flow. Emergency physicians (EPs) play a critical role in the evaluation process. The length of time that takes the physician to assess the patient once a room is assigned to the patient directly affects the evaluation interval. (Krall, Cornelius, & Addison, 2014).

Hemmati et. al considered Low skillfulness, experience, and knowledge of the staff as one of the main reasons that affect physician decision time (Hemmati, Mahmoudi, Dabbaghi, Fatehi, & Rezazadeh, 2018).

Nhdi et al. expected that the highest effective index for LOS (with a strong positive significance correlation) is the decision to disposition time in EDs (Nhdi, Asmari, & Thobaity, 2021).

The data in the healthcare system reflects the reactions of physicians and healthcare professionals in emergency situations. Unexpected patterns in the data are driven by the physicians' behaviour, which highlights the importance of considering contextual information during the process of mining analyses in healthcare (Andel, Beerepoot, Lu, Weerd, & Reijers, 2021).

The main goal of this research is to apply the data mining process to extract knowledge from past data to improve hospital management. From data received from a Portuguese hospital, it is observed that physician's performance has a meaningful effect on the proficiency of the ED admission process. The aim is to explore if the length of stay and waiting time-as two important indicators of the efficiency in hospital management-depends on physicians considering their performance (based on the number of visited patients).

Data and models

Through this study, the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach is used to improve the reliability, repeatability, and manageability of data mining projects, resulting in a more efficient process (Wirth & Hipp, 2000). The CRISP-DM process consists of 6 phases, and we adapted it for analyzing the data. The first and second phases are including business understanding and data understanding, which involve collecting, describing, exploring, and verifying the data. Step three includes data preparation, which begins with selecting, cleaning, exploring, verifying the quality, integrating, and formatting data. The fourth phase called modelling which comprises selecting modelling techniques, generating a test design, building a model, and finally assessing the model. In the fifth phase (Evaluation), the model (s) are evaluated and Interpreted

Data Understanding

The data used in this study comes from a Portuguese hospital near Lisbon (Garcia de Orta Hospital), that provides ED data between January 1st, 2015 and December 31st, 2017.

This real data contains 273,712 observations and each record is stored in the hospital's database based on visiting a patient in the general surgery of the ED. There are 12 variables to describe the interaction of a patient with the hospital ED.

Five of the twelve features (variables) in the original dataset are DateTime (YYYY-MM-dd hh:mm:ss) variables, including; Admission Date, Triage Date, Examination (first note) Date, Medical Discharge Date and Discharge Administrative Date.

The patients, triage nurses, and doctors involved in the event are identified with identifiers (IDs). Patient ID is a feature to identify the patient and track his/her history (A patient may be visited several times). There is also an id to identify the nurse responsible for the patient triage. There are two key ids in the dataset for the doctors who are responsible for the first note and medical discharge.

The dataset also includes an attribute called Triage Color coding, which represents the triage colour assigned to the patient during the triage process. This variable involves five possible levels, each colour indicates what level of treatment is required.:

RED: emerging patients

ORANGE: very urgent patients

YELLOW: urgent patients

GREEN: less urgent patients

BLUE: not urgent patients

There is also a variable regarding the discharge named Destiny which shows the patients' fate after LOS in ED. Figure 1 illustrates the original variables of the dataset.

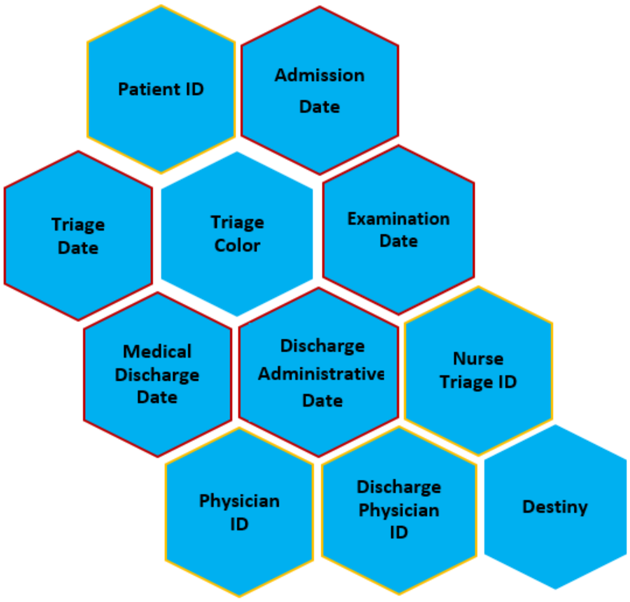


Figure 1 Original dataset variables

Data Preparation

This is the third step of the CRISP-DM method. The main goal of this process is to prepare the data for the rest of the research.

Research results and analysis can be compromised if the data is faulty, so the data must be accurate and veracious, respecting some of the characteristics of big data (Asri, Mousannif, Moatassime, & Noël, Big data in healthcare: Challenges and opportunities, 2015).

Data preprocessing is not restricted to classical data mining tasks such as classification or regression. In novel data mining fields, data preprocessing is increasingly becoming a tool for improving models (García, Ramírez-Gallego, Luengo, Benítez, & Herrera, 2016).

Several data cleansing steps were taken to ensure that the data was valid and won't compromise the results of the study. The author considered inconsistent, the records that have missing values and the minus periods, but outliers did not remove due to the nature of the data.

Cleaning Data - The following steps have been done to clean the data. All the rows where column DateTime had empty values were removed. The patients who are not assigned triage colour are omitted from the dataset as in this research the department of general surgery has been studied the other departments were removed from the dataset

Outliers are a crucial part of data analysis. Removing or retaining an observation can be a challenging decision, particularly if outliers are influential. In some cases, outliers simply occur when observations are different from the norm. In such cases, it may not be wise for these observations to be removed. Therefore, it is important to analyze the causes behind an observation that has been identified as an outlier (Hyndman & Athanasopoulos, 2018). As outliers of the data in this study are an integral part of the data and are unavoidable, they should not be deleted. After the cleaning steps, the number of observations reduced to 278,584. Then some informative plots are created to describe better the dataset.

Feature Engender - In this part, operations such as creating derived attributes, creating new records, or converting values for existing attributes are performed.

The ED admission process flow includes five main steps. The process begins with the admission of patients to the ED. Next is the triage, in which the patient is classed based on the Manchester Triage Protocol (MTP). Afterwards, the patient waits to visit a doctor, which is the third ED process. After the physician note or gets treatment, the patients are discharged by doctors based on the patient's health state and finally, in the last step, the patient effectively leaves the hospital in the administrative discharge step.

As can be seen from Figure 2, According to the physicians' performance length of stay is split into three segments, LOS1, LOS2 and LOS3. Based on the goals of this research, by deducting "Medical Discharge Date" from "Triage Colour Date" and by deducting "First Note Date" from "Triage Colour Date" the "LOS2" (Because only LOS2 is affected by the physician's performance) and "Waiting Time" are obtained.

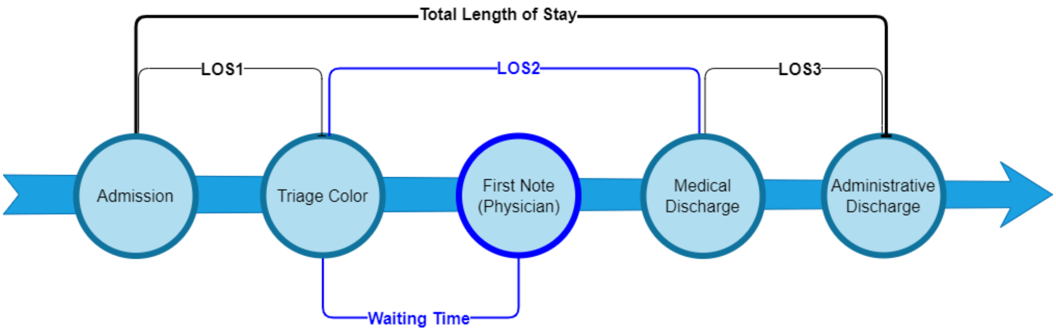


Figure 2 LOS and Waiting Time in admission process flow

In this step the final dataset with the two new important variables includes “LOS2” and “Waiting Time” is prepared.

Format Data- After data manipulation, it is found that considering our goals it is possible to split physicians into different groups based on the number of visited patients, such that the physicians who visit more patients are highly experienced and those who visit less patients have less experience. The data is still based on Patient ID, since the goal of the research is to evaluate the effect of physician groups on waiting time/LOS2 it needs to be manipulated according to the Doctor ID. In formatting transformation, the content of the data does not change, while syntactic modifications are made to the data.

In Table 1, physicians (observations) are split into 5 groups based on the cumulative sum of Patients. According to this method, the physicians were placed in one of the "Very High", "High", "Medium", "Low" and "Very Low" groups. Hence, the physicians who visited more patients fell into the "Very High" group provided they are part of the first 20 per cent of the cumulative sum of Patients.

Table 25 Waiting Time descriptive analysis

Waiting Time (Triage to Doctor's First Note)

GroupName	count	mean	SD	median	min	max
Frequency Group						
Very High	7	72.54060	19.62908	74.92488	31.701508	88.57063
High	18	70.61331	25.05513	70.70084	22.754964	112.22747
Medium	35	70.53001	31.84072	53.67956	21.368773	152.71363
Low	71	60.33400	26.20997	56.25849	23.352705	153.38463
Very Low	675	52.27813	35.09815	47.07393	1.033333	195.95417

As can be seen in Table (2) descriptive analysis of the patient’s LOS2 including count (number of physicians), mean, SD, median, minimum, and maximum values described and summarized to fulfil every condition of the data.

Table 26 Length of Stay (LOS2) descriptive analysis

Length Of Stay (LOS2)

GroupName	count	mean	SD	median	min	max
Frequency Group						
Very High	7	318.1871	100.99907	354.8269	31.701508	88.57063
High	18	305.3763	80.34895	310.1657	22.754964	112.22747
Medium	35	282.0673	100.28010	293.8712	21.368773	152.71363
Low	71	248.5810	121.79356	242.7016	23.352705	153.38463
Very Low	675	254.1960	135.89086	239.2833	1.033333	195.95417

Eventually, the LOS2, waiting times and physicians' groups in order to obtain major operative goals in the next steps are prepared. Finally, the new dataset including 806 observations, and 6 variables (features) for modelling are indicated in Figure 3.

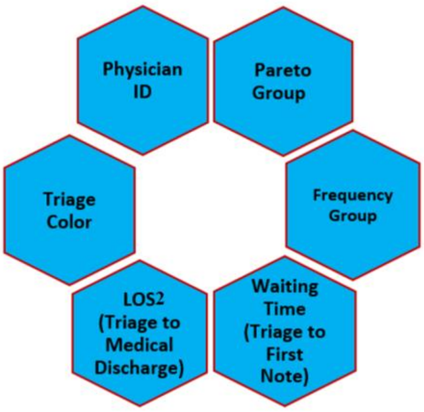


Figure 3 Variables for modelling

Modeling

In this research, the one-way analysis of variance (ANOVA), also known as one-factor ANOVA is applied to compare different groups of physicians based on their patient's waiting time/LOS. Furthermore, the effect of physician groups and triage colour on waiting time/LOS is evaluated simultaneously through the two-way ANOVA test.

Visualizing Data - Figure 4 demonstrates the distribution of physicians' average waiting time (from triage to first note) based on the physicians' group. The "Very Low" group has the lowest average waiting time, which means the physicians (675 out of 806 physicians) who visit less patients than the other groups have the lowest average time. However, the "Very High" group has the highest average waiting time meaning that the physicians (7 out of 806 physicians) who visit the largest number of patients than the other groups have the highest average time.

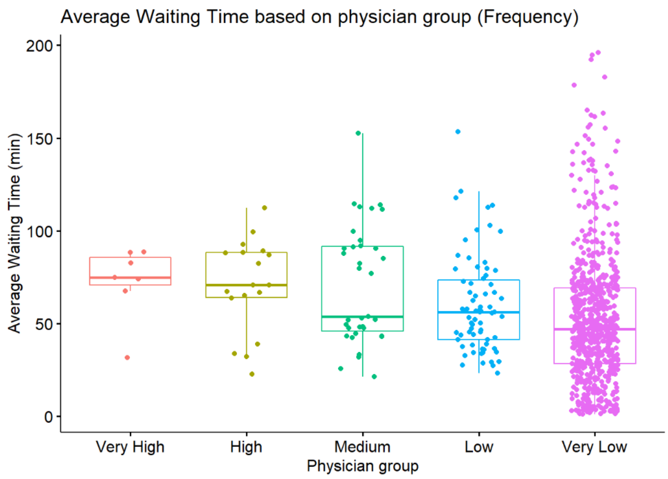


Figure 4 Distribution of Average Waiting Time

Figure 5 shows the distribution of physicians' average LOS (from triage colour to medical discharge).

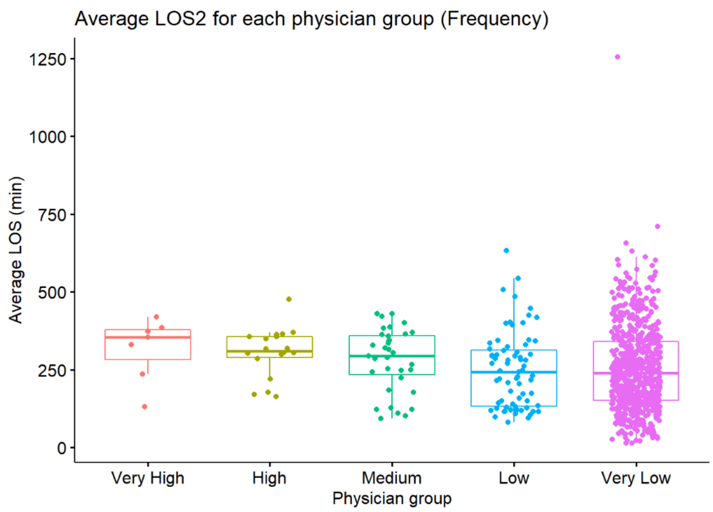


Figure 5 Distribution of Average LOS

In a more detailed review, it can be investigated mean plot with 95% confidence interval (CI) in Figure 6. A 95% confidence interval means that 95% of the time, the "true" population mean will be within that interval and 5% of the time, the population mean will be outside of that interval. This means you can be 95% confident that the true population mean is within your confidence interval.

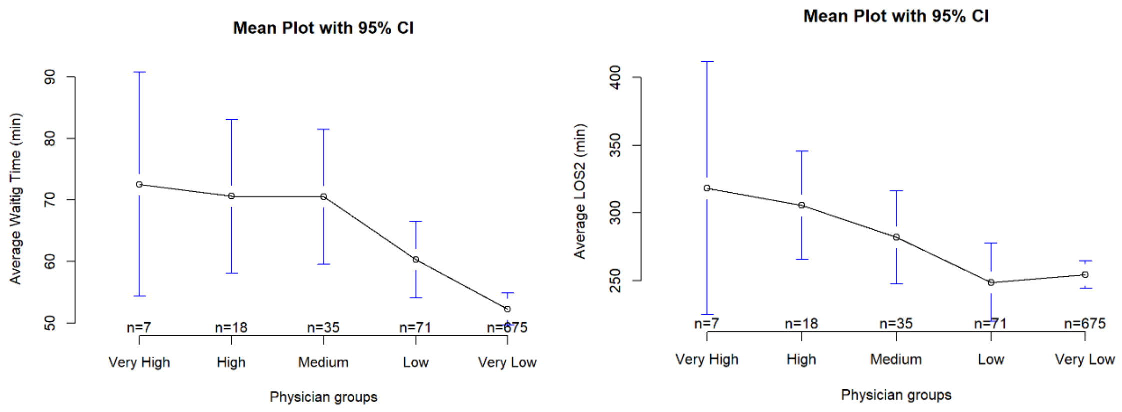


Figure 6 Mean plots with 95% confidence interval

Assumptions of One-way ANOVA test

As an Assumptions of the ANOVA test, the data of each factor level are normally distributed. Therefore, by applying the Jarque-Bera which is a goodness-of-fit test, we can determine whether sample data fit a normal distribution in terms of skewness and kurtosis or not.

According to the Jarque-Bera results, the p-value for both Waiting Time and Length of Stay is less than 0.05, two significant results at the standard level of 0.05, implying rejection of the null hypotheses of the normal distributions.

To tackle this issue a suite of transformation-estimating functions that can be applied to normalize data named “bestNormalize” package in R is used. The best transformation method which is Ordered Quantile (ORQ) normalization (which is introduced as OrderNorm Transformation in the package) with the estimated normality statistic of 0.0048 is picked automatically via this method.

ANOVA test hypotheses

The hypotheses of one-way ANOVA are as below.

- Null hypothesis: The means waiting time/LOS for the different physician groups are the same
- Alternative hypothesis: At least one sample mean is not equal to the others.

Compute the One-way ANOVA test

In Table 3, it is determined if there is any significant difference between the average waiting time/LOS in the 5 experimental. By standard statistical software (R), two ANOVA tables are obtained as below.

Table 3 Analysis of Variance (Waiting Time ~ Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
visited_patient_class	4	25.55	6.388	6.567	3.347e-05
Residuals	801	779.1	0.9727	NA	NA

Through the visualization of data for one-way ANOVA, the average waiting time decreases as moving away from the "Very High" physician group to the "Very Low" physician group. These arguments are supported by the ANOVA test where the p-value is less than the significance level of 0.05, it can be concluded that there are significant differences between the average waiting time of the different groups of physicians.

As described in Table 4, in the case of the LOS2, the average LOS2 for the physician in the "Very High" group (high experienced) is higher than the "Very Low" group (less experienced). These results are not statistically significant at an alpha level of 0.05. Therefore, results show that there are no significant differences between the average length of stay of physician groups.

Table 4 Analysis of Variance (LOS2 ~ Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
visited_patient_class	4	7.448	1.862	1.871	0.1136
Residuals	801	797.3	0.9953	NA	NA

The Two-way ANOVA Test

The author used **the two-way ANOVA test** to evaluate simultaneously the effect of two variables named physicians group and triage colour (or weekdays) on a response variable which can be waiting time or LOS.

As Table 5 shows, there are 5×5 design cells with the factors being triage colour and physician group and the different number of physicians in each cell.

Table 5 The Two-way ANOVA descriptive data analysis

The number of Physicians

	Very High	High	Medium	Low	Very Low
(1)Red	8	17	25	40	150
(2)Orange	5	11	25	53	483
(3)Yellow	5	10	22	66	602
(4)Green	7	17	30	49	601
(5)Blue	10	14	20	39	340

Visualizing interaction effects

Basically, an interaction effect occurs when the effect of one factor depends on the effect of other factors, and this is seen by the lines in the plot are not parallel and nonparallel lines indicating interaction. In other words, the interaction between two factors occurs when the differences between the mean values of one factor are not consistent across levels of the other factor (Alin & Kurt, 2006).

The Two-way interaction plot visualizes a line plot that shows group differences. In this two-way plot, it can be seen possible interactions, which plot the mean of the response (waiting time) for two-way combinations of factors called physician group and triage colour.

Figure 7 shows the effect of triage colour and physician groups simultaneously on a response variable which is waiting time. Green and Blue triage colours affect the physician groups differently. For Blue patients, the “Very High” and the “High” groups of physicians have the lowest average waiting time. For Green patients, the “High” group has the highest amount of waiting time. The average waiting time for the Red and Orange patients in different groups of physicians is almost the same.

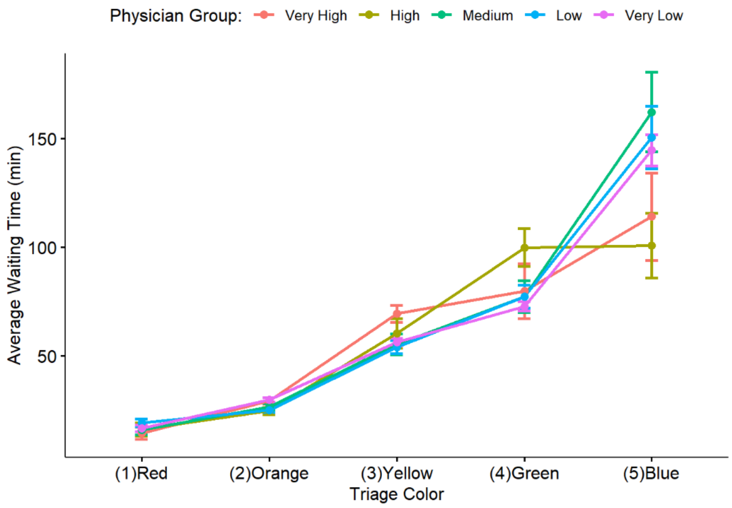


Figure 7 Interaction plot for average LOS

Figure 8 shows how to triage colour and physician group affect the average length of stay (LOS). As nonparallel lines indicate interaction, the effect of “Triage Colour” is different for the 5 levels of the physician group. There is a change when going from one triage colour to the next one, and the type of change depends on the physicians’ group. For physicians’ “High” and “Very High” groups, Red patients experienced longer length of stay than Blue patients. Furthermore, the emerging patients (Red triage colour) who are visited by high experienced physicians (“Very High” group) have longer LOS than the same patients who are visited by less experienced physicians (“Very Low” group).

On the other hand, not urgent patients (Blue triage colour) who are visited by high experienced physicians ("Very High" group) have shorter LOS than the same patients who are visited by less experienced physicians ("Very Low" group). For the "Medium" physicians' group, Blue patients experienced less length of stay than did Green.

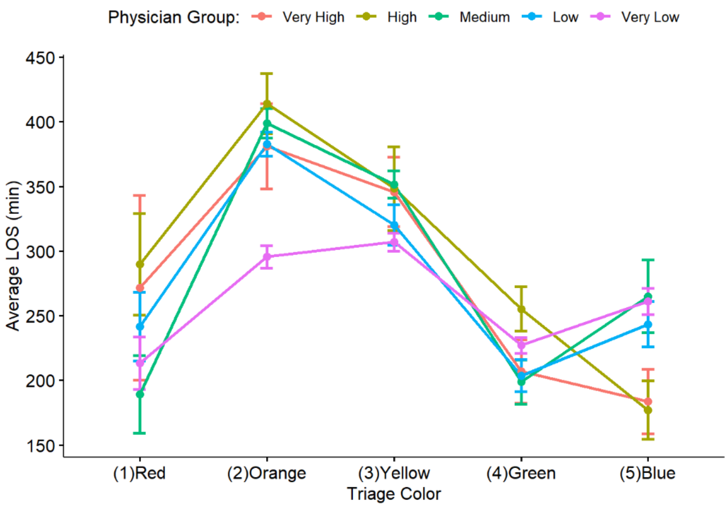


Figure 8 Interaction plot for average LOS2

Assumptions of Two-way ANOVA test

According to the results, a routine check of normality by the test of Jarque-Bera in the P-value 0 for both waiting time and LOS, significant results at the standard level of 0.05, implying rejection of the null hypotheses of the normal distributions.

In order to normalize the data same as the One-way ANOVA, the "bestNormalize" package in R is used. Again, the best transformation method which is Ordered Quantile (ORQ) normalization (which is introduced as OrderNorm Transformation in the package) with the estimated normality statistic of 1.101 is picked automatically via this method.

Test hypotheses

The hypotheses of Two-way ANOVA are as below.

4. Physician Groups have similar means
5. Triage Colour (or weekdays) has similar means
6. There is no interaction between Physician Groups and Triage Colour

The alternative hypothesis of cases 1 and 2 is "the means are not equal" and the alternative hypothesis for case 3 is "there is an interaction between Physician Groups and Triage Colour".

Compute The Two-way ANOVA test

Using two-way ANOVA, there are two possible means models: the additive model and the interaction model.

Additive model

From the ANOVA Table 6, we can conclude that both "Triage Colour" and "Physician group" are statistically significant. "Triage Colour" is the most significant factor variable. These results would lead us to believe that changing patients' urgency level or the group of doctors, will impact significantly the mean waiting time.

Table 6 Two-way ANOVA test for Waiting Time by Triage Colour and Physician group (Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	936.2	234.1	363.4	1.57e-249
visited_patient_class	4	11.22	2.805	4.356	0.001636
Residuals	2640	1700	0.644	NA	NA

As can be seen from Table 7, "Triage Colour" and "Physician group" are statistically significant it is defined changing the triage colour or doctor group, will impact significantly the mean length of stay.

Table 7 Two-way ANOVA test for LOS by Triage Colour and Physician group

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	190.8	47.71	51.75	5.163e-42
visited_patient_class	4	22.91	5.726	6.211	5.681e-05
Residuals	2640	2434	0.9219	NA	NA

Interaction effects

The primary aim of performing two-way ANOVA is to find out whether there is an interaction between the two independent features on the dependent feature. In this test, one of these two independent features (Variables) acts as a central feature and the other as a moderator feature.

The resulting ANOVA table of the two-way ANOVA interaction model for waiting time is shown in Table 19. When the result does not show a statistically significant interaction effect, this indicates that the effect of an independent variable is the same for each level of the other independent variable. According to the tables, the first impression is that the main effects including "Triage Colour" and "Physician group" are statistically significant but the interaction ("Triage Colour" * "Physician group") is not significant at the alpha level of 0.05. It means the joint effect of "Triage Colour" and "Physician group" is not statistically higher than the sum of both effects individually on waiting time.

Table 8 Interaction effects of Two-way ANOVA test for Waiting Time (Frequency)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	936.2	234.1	363.4	3.22e-249
visited_patient_class	4	11.22	2.805	4.355	0.001637
Triage_Color:visited_patient_class	16	10.13	0.633	0.9827	0.4728
Residuals	2624	1690	0.6441	NA	NA

If there is no statistically significant interaction effect, it means that the effect of "physician groups" on average LOS is the same for a different level of "Triage Colour". According to Table 9, the two main effects and the interaction are all statistically significant. The interaction significance test evaluates whether it is valid to conclude that the lines in the population are not parallel. The test result below describes the significant effect as the "Triage Colour" * "Physicians groups" interaction. Therefore, there is a substantial interaction effect between "Triage Colour" and "Physicians groups" on LOS.

Table 9 Interaction effects of Two-way ANOVA test - for LOS2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Triage_Color	4	190.8	47.71	52.42	1.54e-42
visited_patient_class	4	22.91	5.726	6.292	4.902e-05
Triage_Color:visited_patient_class	16	45.72	2.857	3.139	2.388e-05
Residuals	2624	2388	0.9101	NA	NA

Evaluation

In other to evaluate ANOVA models, the residuals must be checked using the normality test. Here Jarque-Bera normality test is considered. The null hypothesis in this test is that the data follow normal distribution, in other words, there is no trend in residuals. Naturally, if the p-value exceeds the significance level (0.05). the normality assumption is accepted, and we argue that the model is appropriate, and the results are solid.

Discussions

During this research, the author aimed to study the effect of physicians' performance (measured by Group variable in the data-analyzed for Frequency group) on their patients' waiting times and length of stay in the ED (general surgery unit) of a Portuguese hospital using real data.

The present research is a multidisciplinary work that joins a management approach with a computer science (data science) to create knowledge about the ED admission process. Some initial arguments are provided using data manipulation and data visualization.

Analysis of The Findings

One-way ANOVA modelling indicates that the average waiting time for the physician in the "Very High" group (high experienced) is higher than that of the "Very Low" group (less experienced). The results of the ANOVA test show that physicians' groups have a significant effect on the average waiting time at an alpha level of 0.05.

Considering LOS2 as the response variable, the physician in the "Very High" group (high experienced) is higher than that of the "Low" and "Very Low" groups (less experienced). The results of the ANOVA test show that the physicians' group has not a significant effect on the average LOS2.

The results from two-way ANOVA show that the triage colour and the physician groups are statistically significant on both average waiting time and average LOS2. It shows that changes in triage colour or physicians' performance will significantly impact the average waiting time and average LOS2 as the response variables.

In addition, regarding LOS2 interactions, we have a statistically significant interaction effect between "physician groups" and "Triage Colour" in their relationship with average LOS2. It determines that the effect of different "physician groups" in the presence of the different levels of "Triage Colour" is significant on the average LOS2. But we do not have a statistically significant interaction effect in the case of the model with average waiting time as the response variable. It indicates that the effect of different "physician groups" considering "Triage Colour" on the average waiting time is the same.

According to the visualization and with the support of ANOVA results, the emerging patients (Red triage colour) who are visited by high experienced physicians (the "Very High" group) have longer LOS2 than the same patients' triage level who are visited by less experienced physicians ("Very Low" group). And not urgent patients (Blue triage colour) who are visited by high experienced physicians ("Very High" group) have shorter LOS than the same patients who are visited by less experienced physicians ("Very Low" group).

"Triage Colour" and "Physician group" are statistically significant indicate changing the triage colour or doctor group, will significantly impact the mean length of stay. And regarding waiting time interactions, Green and Blue triage colours affect the physician group differently. For Blue patients, "Very High" and "High" groups of physicians have the lowest average waiting time.

For Green patients, the “High” group has the highest amount of waiting time. The average waiting time for the Red and Orange patients in different groups of physicians is almost the same.

The results of this study are in line with the findings of Hemmati et. al as they claimed that considered Low skillfulness, experience, and knowledge of the staff as one of the main reasons that affect physician decision time (Hemmati, Mahmoudi, Dabbaghi, Fatehi, & Rezazadeh, 2018). Also, here the achievements support the argument of Krall et al. which represented waiting time as an effective factor in the patients’ evaluation interval (Krall, Cornelius, & Addison, 2014).

Limitations

Due to privacy issues, it is not easy to get hospital data, but we should invest in an automatic process to extract this data from the hospital in an anonymized process. This allows expanding easy this approach to other hospitals.

Conclusions

Data mining through extracting the key pattern and information from a large amount of data has the potential to improve the emergency department (ED) admission process and make informed decisions for healthcare providers.

In this study, the possibility of the impact of different groups of physicians on the ED admission process from the perspective of their proficiency comparison is investigated. For that, the corresponding metrics indicate the efficiency of this process including the patients’ length of stay and waiting time.

Different groups of physicians (based on their experiences) have a significant effect on the average waiting time. In other words, it can be concluded that there are significant differences between the average waiting time of the different groups of physicians. But different groups of physicians do not have a significant effect on the average LOS2. As such, the average waiting time of physicians who visit more patients (high experienced physicians) is more than the average waiting time of physicians who visit less patients (less experienced physicians).

The results show that changes in triage levels (colours) and physicians’ groups significantly affect the average waiting time as the response variables. Moreover, the results demonstrate that changes in the triage level and physician group have a significant effect on the average LOS2.

References

Alin, A., & Kurt, S. (2006). Testing non-additivity (interaction) in two-way ANOVA tables with no replication. *Statistical Methods in Medical Research*, 63-85. doi:10.1191/0962280206sm426oa

Andel, V. v., Beerepoot, I., Lu, X., Weerd, I. v., & Reijers, H. A. (2021). The data in healthcare system reflects the reactions of physicians and healthcare professionals in the emergency situations. *European Conference on Information Systems (ECIS)* .

Caetano, N., Cortez, P., & Laureano, R. M. (2015). Using Data Mining for Prediction of Hospital Length of Stay: An Application of the CRISP-DM Methodology. *International Conference on Enterprise Information Systems*, 149-166.

Hemmati, F., Mahmoudi, G., Dabbaghi, F., Fatehi, F., & Rezazadeh, E. (2018). The factors affecting the waiting time of outpatients in the emergency unit of selected teaching hospitals of Tehran. *Electron J Gen Med*. doi:10.29333/ejgm/93135

Krall, S. P., Cornelius, A. P., & Addison, J. B. (2014). Hospital Factors Impact Variation in Emergency Department Length of Stay more than Physician Factors. *Western J Emerg Med*, 158–164.

Kudyba, S., & Gregorio, T. (2011). Identifying factors that impact patient length of stay metrics for healthcare providers with advanced analytics. *Health Informatics Journal*, 235–245.

Kyriacou, D. N., Ricketts, V., Dyne, P. L., McCollough, M. D., & Talan, D. A. (1999). A 5-Year Time Study Analysis of Emergency Department Patient Care Efficiency. *Annals of Emergency Medicine*, 326-335.

Lee, G., Endacott, R., Flett, K., & Bushnell, R. (n.d.). Characteristics of patients who did not wait for treatment in the emergency department: A follow up survey. 56-62.

Nhdi, N. A., Asmari, H. A., & Thobaity, A. A. (2021). Investigating Indicators of Waiting Time and Length of Stay in Emergency Departments. *Open Access Emergency Medicine* , 311-318. doi:https://doi.org/10.2147/OAEM.S316366

Song, H., Tucker, A. L., & Murrell, K. L. (2015). The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay. *Institute for Operations Research and the Management Sciences (INFORMS)*, 3032–3053.

Xu, M., Wong, T., & Chin, K. (2013). Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network. *Decision Support Systems*, 1488-1498.