7COM1079-0901-2024 - Team Research and Development Project

Final report title: How does the average length of hospital stay differ between patients admitted for an acute care and childbirth care?

Group ID: A283

Dataset number: DS239 (length\_of\_hospital\_stay.csv)

Prepared by: Ankit Joshi 22032491

Piyush Kumar Mallick 23079409

Ahmad Mujtaba Khan 2311130

Haider Abid 23081929

Jeswanth Kumar Botcha

University of Hertfordshire

Hatfield, 2024

Table of Contents

1. Introduction …………………………………………………………………………...3
   1. Problem statement and research motivation ………………………………….3
   2. The data set …………………………………………………………………...3
   3. Research question……………………………………………………………..3
   4. Null hypothesis and alternative hypothesis (H0/H1) …………………………3
2. Background research………………………………………………………………….4
   1. Research papers (at least 3 relevant to your topic / DS)………………………4
   2. Why RQ is of interest (research gap and future directions according to the literature)………………………………………………………………………4

1. Visualisation…………………………………………………………………………..5
   1. Appropriate plot for the RQ *output of an R script (NOT a screenshot)……….5*
   2. Additional information relating to understanding the data (optional)………..5

1. Analysis………………………………………………………………………………5
   1. Statistical test used to test the hypotheses and output………………………..5
   2. The null hypothesis is rejected /not rejected based on the p-value…………. 6

1. Evaluation – group’s experience at 7COM1079……………………………………...6
   1. What went well……………………………………………………………….6
   2. Points for improvement………………………………………………………6
   3. Group’s time management…………………………………………………....6
   4. Project’s overall judgement…………………………………………………..6
   5. Comment on GitHub log output……………………………………………...7

1. Conclusions…………………………………………………………………………..7
   1. Results explained. …………………………………………………………...7
   2. Interpretation of the results ………………………………………………….7
   3. Reasons and/or implications for future work, limitations of your stud……...7

1. Reference list………………………………………………………………………...8

Harvard (author, date) format.

1. Appendices ………………………………………………………………………….9
2. R code used for analysis and visualisation………………………………………9
3. GitHub log output………………………………………………………………..9

1. Introduction

* 1. Problem statement and research motivation **(100 words)**

We want to learn more about the reasons behind the variation of hospital stays of which following could be some of the reasons.

* **Rise in healthcare costs**: Healthcare prices have grown dramatically during the last several decades. Hospital stays account for a large portion of these expenses, thus it is vital to study changes in typical stay durations and the reasons driving them.
* **Advancement in healthcare**: Advances in medical technology, outpatient treatment, and clinical standards may have resulted in shorter hospital stays. (Shah, 2024)
* **Efficiency of hospitals**: Shorter hospital stays are sometimes viewed as a sign of healthcare efficiency, but they might jeopardize quality of treatment if not handled effectively.
  1. The data set **(75 words)**

The data set provided to us contains the length of hospital stays for acute disease and childbirth for various nations across the globe for past few decades.The dataset contains columns of interest which are “subject” which will be our independent variable of nominal type. The other column will be “value" which is our dependent variable of ordinal type.

The data set is intriguing to work on, as it would be great to know the factors that might be responsible for the variation of hospital stays over the years which could be technological advancements, rise in healthcare cost or other factors.

* 1. Research question **(50 words).**

Our research question is “How does the average length of hospital stay differ between patients admitted for an acute care and childbirth care?”

We will be doing a comparison of means analysis to answer our RQ, if there is a difference of hospital length stays between acute disease and childbirth.

* 1. Null hypothesis and alternative hypothesis (H0/H1) **(100 words)**

Below is our hypothesis:

* Null hypothesis (H0): There is no difference in the mean of the hospital stay between acute care and childbirth care.
* Alt hypothesis (H1): There is a difference in the mean of the hospital stay between acute care and childbirth care.

1. Background research
   1. Research papers (at least 3 relevant to your topic / DS) **(200 words)**

Upon searching for the dataset being used for research papers, I didn’t find the exact dataset but did find some research papers along this topic of length of hospital stays, mentioning few of them below:

* **Statistical analysis on length of stay in hospital**

The rising financial problems of healthcare institutions make studies of resource distribution more and more important and valuable. Among these studies, identification of length of stay of hospital patients (LOS) has attracted many scientists recently since it contributes to better knowledge of hospital costs and helps these institutions control the costs. This paper is devoted to study the length of stay of inpatients in hospital. Although predicting the length of stay is difficult, it is actually useful and beneficial if some key factors that have influence on patient length of stay could be determined. (Nguyen and Dang, 2021)

* **Evaluation of factors that influenced the length of hospital stay using data mining techniques**

In hospitals, there are major performance indicators, including LOS, bed occupancy rate, bed turnover, bed turnover interval, and mortality rates that should be determined and monitored regularly [[15](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-02027-w#ref-CR15)]. This study aimed to evaluate each inpatient’s LOS and its associated factors in six different wards of the hospital using statistical analysis along with data mining by the R-studio program. The average LOS in the study population was about two days. Results also demonstrated that about 60% of patients were discharged after one night of hospitalization. (skandari, M., Alizadeh Bahmani, A.H., Mardani-Fard, 2022)

* **Predicting hospital length of stay using machine learning on a large open health dataset**

The study showcases the practical utility of machine learning models in predicting LoS during patient admittance. The emphasis on interpretability ensures that the models can be easily comprehended and replicated by other researchers. Healthcare stakeholders, including providers, administrators, and patients, stand to beneﬁt signiﬁcantly. The ﬁndings oﬀer valuable insights for cost estimation and capacity planning, contributing to the overall

enhancement of healthcare management and delivery. (Jain, Singh, Rao and Garg, 2024)

* 1. Why RQ is of interest (research gap and future directions according to the literature) **(100 word**s)

Understanding trends in the average length of hospital stays for acute diseases and births is relevant to healthcare planning, resource allocation, and patient outcomes. Existing study frequently focuses on overall hospital efficiency or specific circumstances, but it lacks a full historical examination of these critical areas. Exploring this gap reveals information on how advances in medical treatment, regulatory changes, and socioeconomic variables have affected hospital stays. This research also informs future healthcare initiatives, such as increasing patient flow and adapting treatment to changing requirements, while also considering the possible effect of technological and healthcare delivery advancements.

1. Visualisation
   1. Appropriate plot for the RQ(**50 words)**

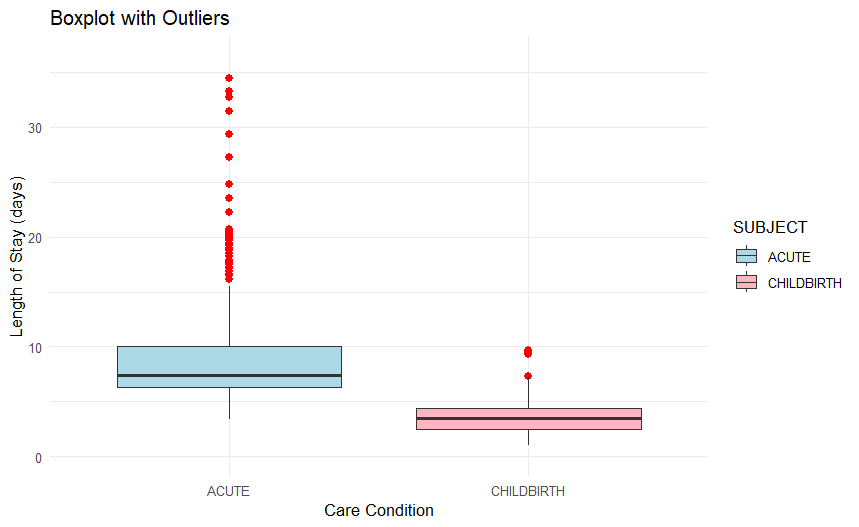


Figure Boxplot with outliers

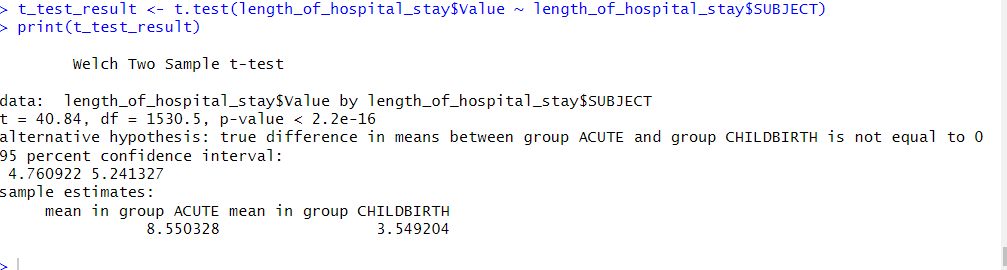
We chose to plot a boxplot with outlier for our visualisation because they are very useful for immediately determining the location, dispersion, and symmetry or skewness of a collection of data, as well as comparing these qualities across two or more data sets.

* 1. Useful information for the data understanding (**50 words)**

The plot illustrates that the acute care condition has a longer median duration of stay and greater variability than delivery, with more severe outliers. Childbirth requires a shorter and more regular duration of stay. Outliers are common in acute care, suggesting variations in patient conditions or treatment durations.

1. Analysis
   1. Statistical test used to test the hypotheses and output (**75 words)**

Our independent variable was of type nominal and dependent variable was of type ordinal, our dependent variable was normally distributed and our independent variable had exactly two values. Taking all the above conditions into consideration we chose to go with t.test to test our hypothesis.



* 1. The null hypothesis is rejected /not rejected based on the p-value (**100 words)**

The null hypothesis is rejected based on the stated p-value in the t-test (p<2.2×10−16). The null hypothesis states that there is no difference in the average length of hospital stay between the acute and childbirth groups. The tiny p-value (𝛼 = 0.05, α = 0.05) suggests a significant difference in mean hospital stay between the two groups. Therefore, the alternative hypothesis is supported.

1. Evaluation – group’s experience at 7COM1079
   1. What went well **(75 words)**

The group displayed excellent teamwork and communication skills, ensuring that responsibilities were assigned equally and deadlines were reached. The application of theoretical knowledge to actual settings was a standout feature, demonstrating comprehension of fundamental ideas. Members assisted each other in overcoming problems, demonstrating strong peer support. The incorporation of varied ideas enhanced the project, and frequent meetings kept the team on track. The final output fulfilled quality standards by demonstrating substantial research and well-executed execution techniques.

* 1. Points for improvement **(75 words)**

The group might benefit from more defined work assignments to reduce overlap and delays. Time management in the early stages should have been improved, since the team occasionally misjudged the difficulty of assignments. Clearer recording of choices made during meetings would help increase openness. Furthermore, making better use of project management technologies might aid in progress tracking and accountability. Finally, certain team members may benefit from improving their technical abilities so that they can contribute more effectively to specific elements of the project.

* 1. Group’s time management (**50 words)**

The group managed time adequately overall, completing key deliverables on schedule. However, the early stages lacked clear timelines, causing minor delays. Improvements in task prioritization and proactive planning helped mitigate these issues. Towards the end, time allocation was optimized, enabling the team to refine the final deliverable successfully.

* 1. Project’s overall judgement (**50 words)**

The project was a success, reflecting a solid understanding of the subject matter and effective teamwork. The deliverables met expectations, with the group demonstrating critical thinking and problem-solving skills. While some challenges arose, they were addressed efficiently, resulting in a high-quality output that aligns with the course’s objectives.

* 1. Comment on the GitHub log output **(50 words)**

The GitHub log output gave significant information on the group's workflow and collaboration. Frequent commitments indicated continuous progress and work sharing among team members. Clear commit messages increased clarity in modifications, facilitating traceability. However, several commits were huge and less informative, and might have been divided into smaller, more focused improvements.

1. **Commit Message:** [Boxplot with outlier] Boxplots containing outliers give a useful visual overview of data, including probable influence points, which aids in the first investigation and comparison of averages. However, they should be used in combination with other statistical techniques to provide a more robust and trustworthy comparison.
2. **Commit Message:** [T test code snippet] The t-test is a statistical hypothesis test used to determine if there's a significant difference between the means of two groups.
3. **Commit Message:** [R code] This contains all the R code scripts we used while working on the RQ including the plots and tests run.

1. Conclusions
   1. Results explained (**75 words)**

The t-test results indicate a significant difference in hospital stay length between acute care (mean = 8.55 days) and delivery (mean = 3.55 days), with a p-value < 2.2e-16. The 95% confidence range for the mean difference is [4.76, 5.24], indicating that lengthier hospital stays for acute treatment have statistical significance. This is consistent with the boxplot, which highlights acute care's larger mean and variability compared to delivery.

* 1. Interpretation of the results (**75 words)**

The study found that acute care patients have much longer hospital stays (mean = 8.55 days) than delivery patients (mean = 3.55 days). This emphasizes the complexity of acute care, which necessitates additional resources and longer treatment times. For the population, this translates into greater healthcare bills and resource demands. In a broader sense, it stresses the need of efficient acute care resource allocation, preventative initiatives to minimize acute care cases, and prospective legislative changes to better balance healthcare delivery and expenditures.

* 1. Reasons and/or implications for future work, limitations of your study (**50 words)**

The study's weaknesses include a limited sample size and potential data collecting biases, which may impair generalizability. To improve results, future research might increase the sample size, investigate new factors, and optimize methodology. Addressing these shortcomings will result in more complete insights and increase the study's usefulness.

1. Reference list

Dr. Julia G., Dr. Alveen S. (2024) Team Research and Development Project

Lecture handouts. University of Hertfordshire. 26 December, 2024.

Eskandari, M., Alizadeh Bahmani, A. H., Mardani-Fard, H. A., Karimzadeh, I., Omidifar, N. & Peymani, P., (2024) Evaluation of factors that influenced the length of hospital stay using data mining techniques. Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-02027-w [Accessed on 28 Dec, 2024]

Jain, R., Singh, M., Rao, R. & Garg, R., (2024). Predicting hospital length of stay using machine learning on a large open health dataset. *BMC Health Services Research*, 24. doi:10.1186/s12913-024-11238-y. Available at: https://www.researchgate.net/publication/382653655\_Predicting\_hospital\_length\_of\_stay\_using\_machine\_learning\_on\_a\_large\_open\_health\_dataset [Accessed on 29 Dec, 2024]

Nguyen, D., Ho Dang, P., Nguyen, T. & Nguyen, V., (2021) Statistical analysis on length of stay in hospital. *Science & Technology Development Journal - Engineering and Technology*, 3, first. doi:10.32508/stdjet.v3iSI3.651. Available at: https://www.researchgate.net/publication/348688920\_Statistical\_analysis\_on\_length\_of\_stay\_in\_hospital/citation/download [Accessed on 29 Dec, 2024]

1. Appendices
2. R code used for analysis and visualisation. Analysis.R code with the appropriate statistics to test the hypotheses.

#Box plot with outliers

ggplot(length\_of\_hospital\_stay, aes(x = SUBJECT, y = Value, fill = SUBJECT)) +

geom\_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2) +

theme\_minimal() +

labs(title = "Boxplot with Outliers ", x = "Care Condition", y = "length of stay(days)") +

scale\_fill\_manual(values = c("lightblue", "lightpink"))

#extended version of box plot

ggplot(length\_of\_hospital\_stay, aes(x = SUBJECT, y = Value, fill = SUBJECT)) +

geom\_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2) +

theme\_minimal() +

labs(title = "Boxplot with Outliers", x = "Care Condition", y = "Length of Stay (days)") +

scale\_fill\_manual(values = c("lightblue", "lightpink")) +

expand\_limits(y = c(0, max(length\_of\_hospital\_stay$Value) + 2) # Extend y-axis upwards

#Histogram with overlay curve

hist\_data <- hist(length\_of\_hospital\_stay$Value,

plot = FALSE) # Keep as frequencies (default behavior)

# Extend the y-axis by adding space above the tallest bar

y\_max <- max(hist\_data$counts) +40 # Adjust "+ 5" to add more space

# Plot the histogram with frequency counts

hist(length\_of\_hospital\_stay$Value,

main = "Histogram",

xlab = "Length of Stay",

col = "lightblue",

border = "black",

ylim = c(0, y\_max))

# Calculate mean and standard deviation

mean\_val <- mean(length\_of\_hospital\_stay$Value)

sd\_val <- sd(length\_of\_hospital\_stay$Value)

# Overlay the normal curve scaled to frequency

curve(dnorm(x, mean = mean\_val, sd = sd\_val) \* length(length\_of\_hospital\_stay$Value) \* diff(hist\_data$breaks)[1],

col = "red",

lwd = 2,

add = TRUE)

# t test

> t\_test\_result <- t.test(length\_of\_hospital\_stay$Value ~ length\_of\_hospital\_stay$SUBJECT)

> print(t\_test\_result)

1. GitHub log output.