Mercedes Joseph

D207

Professor Sewell

Exploratory Data Analysis Task 1

June 18, 2023

1. Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:

A1. **Research question**

What conditions and other factors contribute to the readmission of patients?

A2.

Knowing how medical conditions and other factors contribute to patient readmission benefits healthcare providers, policymakers, and patients. Enables personalized care planning, efficient resource allocation, targeted interventions, improved healthcare planning and policy decisions, and patient empowerment. These actions of interventions and planning have often led to, “successfully reduced readmission rates for patients discharged to home” ( Kriplani, S., 2014). By addressing the underlying factors associated with readmission, stakeholders can work collaboratively towards achieving better patient outcomes and a more efficient healthcare system.

A3.

I have selected a set of variables related to medical conditions and other factors to investigate their influence on patient readmission. Below is a list of these variables, excluding the dependent variable "ReAdmis," which I have identified below as the outcome variable for the independent factors. By examining the relationships between these variables and the dependent variable, my goal is to identify the factors that have a significant impact on patient readmission.

Independent Variables:

HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, Initial days, Total charge, and additional charges.

A screenshot of a computer code

Description automatically generated with low confidence

Data types and examples for the original factors.

A picture containing text, font, screenshot

Description automatically generated

I conducted a correlation analysis to examine the relationship between the dependent variable, "ReAdmis," and the selected factors. Among these factors, the variable "Initial\_days" exhibited the strongest correlation with a coefficient of 0.8, indicating a significant positive relationship. Additionally, I observed a moderate correlation between the variable "TotalCharge" and "ReAdmis." These findings suggest that both "Initial\_days" and "TotalCharge" may have a notable influence on patient readmission. Further investigation into these factors can provide valuable insights into their impact on readmission rates.

A screenshot of a computer screen

Description automatically generated with low confidence

B1.

Based on the results obtained from the correlation matrix, I formulated a null hypothesis stating that the variable "Initial\_days" does not have any influence on patient readmission. To test this hypothesis, I performed a t-test on the two variables using the following code:A screen shot of a computer code

Description automatically generated with low confidence

B2.

The t-test gave the following results:

A picture containing text, screenshot, font

Description automatically generated

The results from the t-test show that I have a significant amount of evidence to reject my null hypothesis. The p-value of (< 2.2e-16) indicates that the likelihood of observing such a large difference in means between the two groups (readmis\_0 and readmis\_1) if there were truly no influence of initial days on ReAdmis is highly unlikely. Additionally, the 95% confidence interval (-41.91645 to -40.84716) provides further support for the significant difference, as it does not include zero. I can conclude from these results that the factor of initial days has a substantial amount of influence on the variable ReAdmis, suggesting that the duration of initial days plays a crucial role in determining the likelihood of readmission.

A screenshot of a computer

Description automatically generated

B3.

I chose to use a t-test because it is the best way to compare the average initial days a (numeric value) between two groups: patients who were readmitted and those who were not. I wanted to see if the length of initial days had any impact on whether a patient was readmitted or not. I also chose the to use t-test analysis method due to its ability to create, “statistical models that take data characteristics into account will allow for better interpretation of data outcome” (Yu et al., 2022).

The t-test results showed a very strong and highly significant difference between the two groups. This means that the average initial days for readmitted patients was significantly different from the average initial days for non-readmitted patients.

The t-test confirms that the length of initial days does have a significant influence on whether a patient is likely to be readmitted. Patients with longer initial days tend to have a higher chance of being readmitted compared to those with shorter initial days.

C1.

A picture containing text, font, screenshot, line

Description automatically generated

This is the code that I used to summarize the stats of additional charges and used the code to create a visualization.

Min. 1st Qu. Median Mean 3rd Qu. Max.

3126 7986 11574 12935 15626 30566

The code that allowed for me to view that stats of Additional charges.

A picture containing text, diagram, screenshot, plot

Description automatically generated

The histogram displays the frequency distribution of the column "Additional Charges," which is a continuous variable. Based on the results, the histogram shows a prominent peak around the 13000 mark. This indicates that a significant number of instances have an additional charge value near that point. The histogram provides valuable insights into the distribution and concentration of additional charges within the given dataset.

A picture containing text, font, screenshot, white

Description automatically generated

This code allowed me to summarize the stats of the column of continuous variable Initial days and created the histogram for visualization of the data.

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.002 7.896 35.836 34.455 61.161 71.981

The code allowed for me to look at the stats that are listed above.

A picture containing text, diagram, screenshot, plot

Description automatically generated

The histogram generated by the code provides a visual representation of the frequency distribution of the variable "Initial days." The histogram reveals distinct peaks around the 10-day mark and another peak at the 70-day mark. Between these peaks, the frequency gradually decreases until reaching a trough around the 35-day mark. This distribution pattern allows for a clear understanding of the distribution of initial days among the patients. The code below allowed me to create a bar plot to visualize the variables properties.

barplot(table(MD$Marital), main = "Distribution of Marital")

The graph below illustrates the distribution of patients' marital status. The graph reveals that the majority of the marital status categories have a count just below or at the count of 2000. This suggests that there is a concentration of patients within these specific marital status groups.A picture containing text, screenshot, rectangle, design

Description automatically generated

The code below allowed me to create a bar plot to visualize the variables properties.

barplot(table(MD$Area), main = "Distribution of Area")

The graph represents the distribution of the categorical variable "Area" properties, including "Suburban," "Urban," and "Rural." It is evident from the bar graph that the number of patients is evenly distributed across all three categories. Each category has approximately 3500 patients, indicating a balanced distribution among the different areas.A picture containing text, screenshot, rectangle

Description automatically generated

D1.

The continuous variables that are represented in the scatterplot are Initial\_days and Income:

A picture containing text, screenshot, diagram

Description automatically generated

I chose the categorical variables Stroke and HighBlood and created a stacked bar graph:

A screenshot of a graph

Description automatically generated with low confidence

E1.

Hypothesis: There is a significant correlation between the variable "Initial\_days" and the variable "ReAdmis."

Null – hypothesis: The factor Initial\_days does not have any influence on the readmission of a patient.

I conducted a t-test to examine the relationship between the variable "Initial\_days" and the variable "ReAdmis" and evaluate my hypothesis. My null hypothesis stated that the factor "Initial\_days" does not have any influence on the readmission of a patient. However, the results from the Welch Two Sample t-test provided strong evidence against the null hypothesis.

The t-test compared the means of "Initial\_days" for two groups: "readmis\_0" (no readmission) and "readmis\_1" (readmission). The test statistic was calculated as t = -151.72 with a degrees of freedom (df) value of 9967.4. The p-value was found to be less than 2.2e-16, indicating a small probability of observing such a large difference in means under the assumption of no true difference.

Based on the hypothesis, which states that the true difference in means is not equal to 0, the t-test results provide strong evidence that there is a significant difference in the average "Initial\_days" between patients who were readmitted and those who were not readmitted. The 95 percent confidence interval (-41.91645, -40.84716) further supports this hypothesis. The estimated mean of "Initial\_days" for patients who were not readmitted was 19.2491, while the estimated mean for those who were readmitted was 60.6309.

In conclusion, the t-test results suggest that the variable "Initial\_days" plays a significant role in the readmission of patients. The findings reject the null hypothesis and indicate a substantial difference in "Initial\_days" between patients who were readmitted and those who were not.

E2.

One limitation of the exploratory data analysis is the lack of guidance from an expert. Without an expert to consult with, it becomes challenging to determine which factors should be evaluated and to interpret the results accurately. Having an expert's input would be valuable in understanding the significance of each variable and answering any questions that may arise during the analysis process.

Another limitation is the potential for misleading information in the dataset. Without additional information or expert guidance, it is difficult to assess the accuracy and reliability of the data. This uncertainty raises concerns about the validity of the analysis and the conclusions drawn from it.

Despite these limitations, I conducted a correlation matrix to identify the factors that most strongly influence the variable "ReAdmis." According to my results, the two most influential factors are "Initial\_days" and "Total Charge." These variables demonstrate a significant impact on patient readmission. However, further discussions with an expert would be beneficial to provide a deeper understanding of why these variables have such a pronounced effect on readmission.

In closing, while the exploratory data analysis has provided insights into the potential contributing factors, the limitations of not having expert guidance and uncertainty about data accuracy underscore the need for further investigation and consultation with experts to ensure the accuracy of the analysis.

E3.

Based on the findings regarding patient readmission, I would strongly recommend that the hospital's operations and data teams collaborate to conduct a more in-depth analysis of the data. It is crucial to investigate the factors that commonly contribute to patients requiring longer initial visits or hospital stays. By identifying these factors, the hospital can potentially implement strategies to address them and reduce the need for readmission.

Additionally, it is important to explore the average Total Charge associated with patients who are readmitted. Understanding the financial implications of readmission can provide valuable insights for the hospital's financial planning. By examining the patterns and trends in Total Charge, the hospital can identify opportunities to optimize costs and enhance patient outcomes.

By bringing together the expertise of the operations team and the insights gained from the data analysis, the hospital can develop targeted interventions and improvements to reduce readmission rates and enhance the overall quality of care. This effort will enable the hospital to make data-driven decisions and implement evidence-based strategies that can positively impact patient outcomes and resource utilization.

References

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists : 50+ essential concepts using r and python*. O'Reilly Media, Incorporated.

Kripalani, S., Theobald, C. N., Anctil, B., & Vasilevskis, E. E. (2014). Reducing hospital readmission rates: current strategies and future directions. *Annual review of medicine*, *65*, 471-485.

Masi, A. T. (1965). Potential uses and limitations of hospital data in epidemiologic research. *American Journal of Public Health and the Nations Health*, *55*(5), 658-667.

Yu, Z., Guindani, M., Grieco, S. F., Chen, L., Holmes, T. C., & Xu, X. (2022). Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. *Neuron*, *110*(1), 21-35.