Performance Assessment: Exploratory Data Analysis

Gabriela Howell

Master of Science Data Analytics, Western Governors University

D20 – Exploratory Data Analysis

Professor Gagner

March 19, 2024

A. Organizational Situation or Issue in the Data Dictionary:

A1. Question for Data Set

The Journal of Family Medicine and Primary Care highlights the importance of comprehending the factors influencing patient readmission in modern healthcare research (Samuel et al., 2022). Guided by this, my research poses the question: " Is there a significant association between timely admission (‘Timely_admission’) and patient readmission (‘ReAdmis’)?" Recognizing the contributing factors to readmission is crucial, as they significantly impact patient outcomes. Addressing this challenge can lead to more effective approaches for patient care and management.

A2. Stakeholder Benefits

Analyzing the relationship between ‘Timely_admission’ and ‘Readmis’ can offer valuable insights to a wide range of stakeholders, including healthcare administrators, policymakers, and medical professionals. This examination may reveal patterns or risk factors linked to readmission rates, enabling targeted interventions and improvements in patient care.

A3. Relevant Data Identification

The goal is to investigate the factors influencing hospital readmission by analyzing diverse medical datasets that encompass continuous and discrete numerical data, as well as categorical information. Specifically, we focus on the dependent variable “ReAdmis,” which is a binary categorical variable indicating whether a patient was readmitted within a month of their initial discharge. In line with the chi-square analysis test, we have included only the pertinent variables, along with their corresponding data types:

- Timely_admission (Categorical)
- ReAdmis (Categorical)

This approach ensures a focused examination of the association between timely admission and patient readmission, contributing to a deeper understanding of readmission risk factors.

B. Data Analysis Description:

The approach is to use statistical tests. This includes the Chi-square test for categorical variables to find relationships and anything significant. By identifying these dynamics, we can enhance our insights into hospital readmission patterns.

B1. Data Analysis Code

The code for this analysis is included with this document titled ‘PA D20 V3.ipynb.’ Please refer to it for your review.

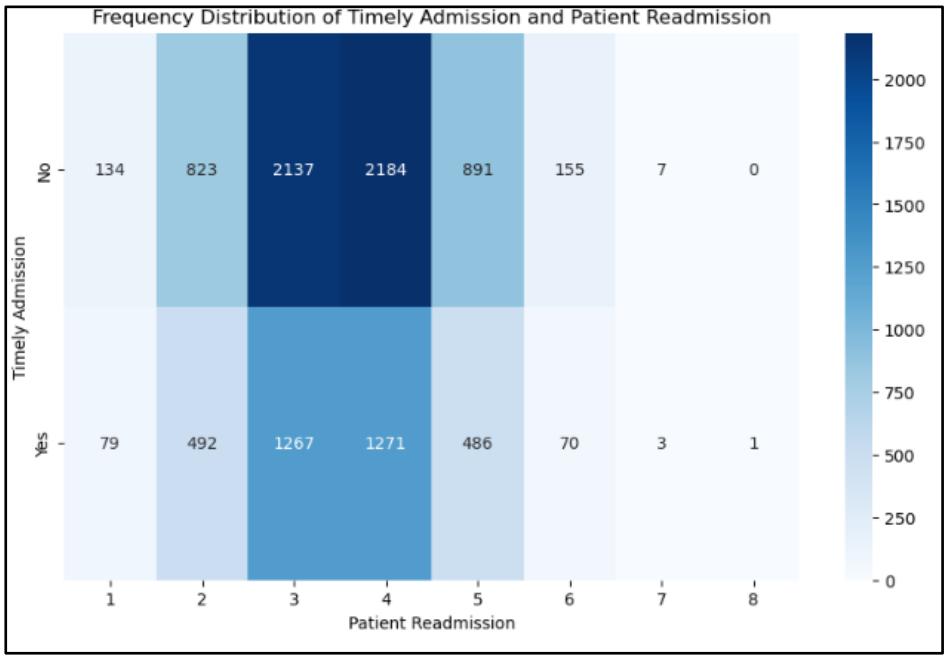
B2. Analysis Output and Results

I conducted a Chi-square analysis to investigate the relationship between 'Timely_admission' and 'ReAdmis', utilizing a contingency table to showcase the frequency distribution of 'Timely_admission' levels within each readmission group. As well as a visualize in the heatmap located below The table revealed that, for patients with “No” readmission, the highest frequency occurred at 'Timely_admission' rating 3, followed by 4 and 2. Similarly, in patients with “Yes” readmission, the most frequent 'Timely_admission' ratings were 3, 4, and 2. This analysis serves as a foundational exploration of the association between these variables.

Contingency Table:

Timely_admission	1	2	3	4	5	6	7	8
ReAdmis								
No	134	823	2137	2184	891	155	7	0
Yes	79	492	1267	1271	486	70	3	1

Further insights were gained by examining the percentage breakdown, indicating the proportion of each 'Timely_admission' rating within each readmission group. Approximately 33.8% of patients in the “No” readmission group rated 'Timely_admission' as 3, while around 34.5% of patients with “Yes” readmission gave a rating of 4. Visualizing these findings through a heatmap provided a quick identification of relationships, with darker cells signifying higher frequencies. Overall, this comprehensive analysis enhances my understanding of factors influencing readmission, offering valuable insights for informed decision-making in the healthcare domain.



The Chi-square test results further corroborated these findings, showing a chi-square statistic of 6.83 with a p-value of 0.45 and 7 degrees of freedom. The expected frequencies for both 'No readmission' and 'Readmission' categories were calculated, indicating no significant association between 'ReAdmis' and 'Timely_admission', as the p-value exceeds 0.05.

Chi-square test results:

- Chi-square statistic: 6.83
- P-value: 0.45
- Degrees of freedom: 7
- Expected frequencies:
- No readmission: [134.85, 832.53, 2155.07, 2187.36, 871.78, 142.45, 6.33, 0.63]
- Readmission: [78.15, 482.47, 1248.93, 1267.64, 505.22, 82.55, 3.67, 0.37]

```
Contingency Table:
Timely_admission    1    2     3     4    5    6  7  8
ReAdmis
No                134  823  2137  2184  891  155  7  0
Yes                79  492  1267  1271  486   70  3  1

Chi-square test results:
Chi-square statistic: 6.8269565562495975
P-value: 0.44711691481022053
Degrees of freedom: 7
Expected frequencies:
[[1.3485030e+02 8.3252650e+02 2.1550724e+03 2.1873605e+03 8.7177870e+02
  1.4244750e+02 6.3310000e+00 6.3310000e-01]
 [7.8149700e+01 4.8247350e+02 1.2489276e+03 1.2676395e+03 5.0522130e+02
  8.2552500e+01 3.6690000e+00 3.6690000e-01]]
```

B3. Analysis Technique Justification

The choice of employing a chi-square test in this analysis stems from the nature of the variables involved. The chi-square test of independence is particularly appropriate for exploring

the association between two categorical variables, making it an appropriate choice for evaluating the relationship between patients' readmission status ('ReAdmis') and their ratings for timely admission ('Timely_admission'). By using this non-parametric test, we can evaluate whether there is a statistically significant connection between these variables without making assumptions about their distribution. It allows me to assess whether there is a statistically significant connection between these variables. By choosing the chi-square test, I ensure the analytical approach aligns with the research question, enhancing the validity of our findings.

C. Univariate Statistics for Continuous and Categorical Variables:

To gain insights into the distribution of both continuous and categorical variables through univariate statistics. For the categorical variables "Timely_admission" and "Timely_treatment," the following descriptive statistics were obtained:

Timely_admission:

- Count: 10,000
- Mean: 3.5188
- Standard Deviation: 1.031966
- Minimum: 1
- 25th Percentile (Q1): 3
- Median (50th Percentile): 4
- 75th Percentile (Q3): 4
- Maximum: 8

Timely_treatment:

- Count: 10,000
- Mean: 3.5067
- Standard Deviation: 1.034825
- Minimum: 1
- 25th Percentile (Q1): 3
- Median (50th Percentile): 3
- 75th Percentile (Q3): 4
- Maximum: 7

For the continuous variables “Initial_days” and “TotalCharge,” the descriptive statistics are as follows:

Initial_days:

- Count: 10,000
- Mean: 34.455299
- Standard Deviation: 26.309341
- Minimum: 1.001981
- 25th Percentile (Q1): 7.896215
- Median (50th Percentile): 35.836244
- 75th Percentile (Q3): 61.161020
- Maximum: 71.981490

TotalCharge:

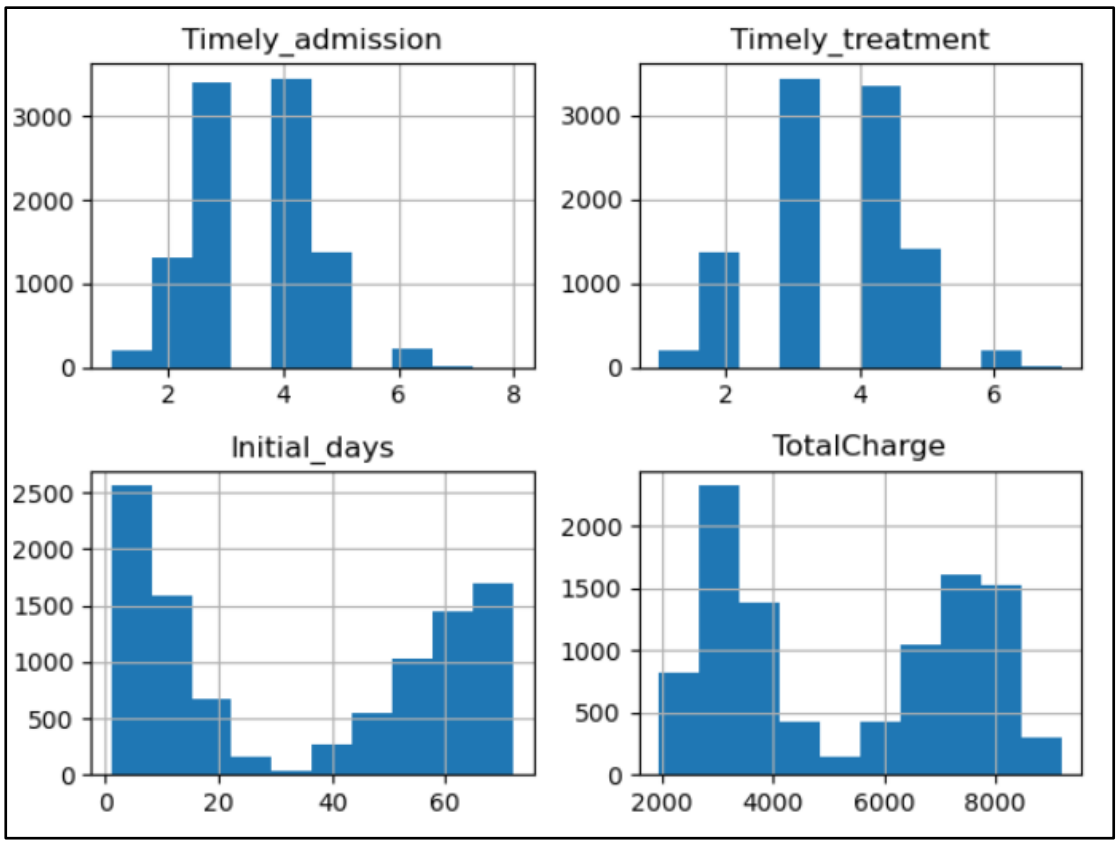
- Count: 10,000

- Mean: 5312.172769
- Standard Deviation: 2180.393838
- Minimum: 1938.312067
- 25th Percentile (Q1): 3179.374015
- Median (50th Percentile): 5213.952000
- 75th Percentile (Q3): 7459.699750
- Maximum: 9180.728000

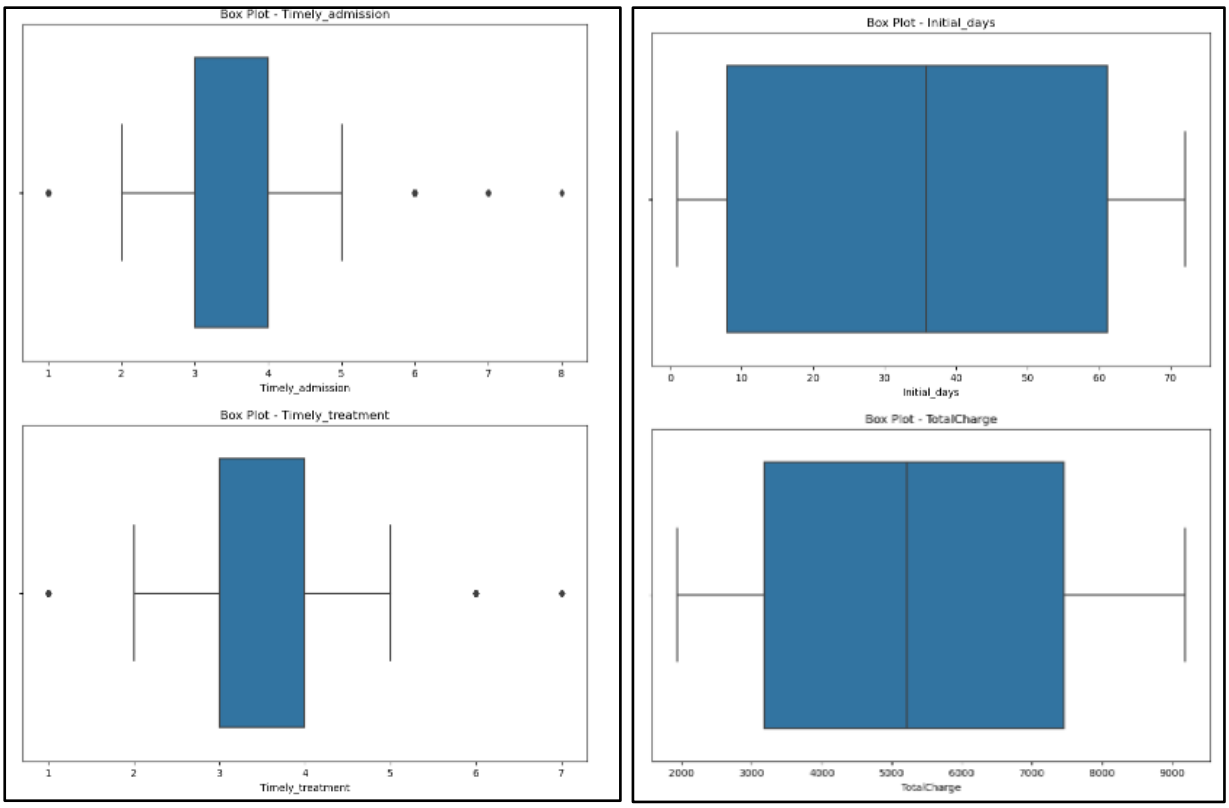
These statistics offer a comprehensive overview of each variable's distribution, including measures such as mean, median, count, minimum, and maximum values. Next, visual representations such as histograms will be used to provide a clearer view of the frequency distribution within each variable, aiding in understanding the spread and concentration of data points.

C1. Visualization of Findings

I conducted histograms for all four variables: "Timely_admission", "Timely_treatment", "Initial_days" and "TotalCharge" to explore their distributions. Particularly, the distributions of "Timely_admission" and "Timely_treatment" exhibit a normal pattern, displaying gaps between the values 3-4 and 5-6. In contrast, "Initial_days" and "TotalCharge" display a bimodal distribution.



Subsequently, I created box plots for these variables. The box plot for "Timely_admission" and "Timely_treatment" indicates the presence of outliers, which is expected given their survey nature. It is intentional to retain these outliers as they may convey meaningful insights. Conversely, "Initial_days" and "TotalCharge" do not exhibit any outliers. Examining the interquartile range, upper and lower whiskers, median, and mean provides a comprehensive view of the central tendency and spread of these variables.



D. Bivariate Statistics for Continuous and Categorical Variables:

Upon analyzing the bivariate descriptive statistics, a striking correlation coefficient of 0.9876 is evident between "TotalCharge" and "Initial_days," indicating a strong positive correlation. This suggests that as the number of initial days increases, so does the total charge incurred, implying higher costs associated with longer hospital stays.

As for categorical variables, the focus is on assessing the relationship between timely admission and patient readmission using the variable "Timely_admission," rated from 1 to 8. Descriptive statistics reveal an average rating of approximately 3.52, indicating a moderate level of timeliness observed across most admissions. Notably, among the 10,000 cases examined, the majority (6,331 instances) experienced no readmission. This underscores the importance of

efficient admission processes in influencing readmission rates, highlighting the significance of prompt and effective admission procedures in healthcare settings.

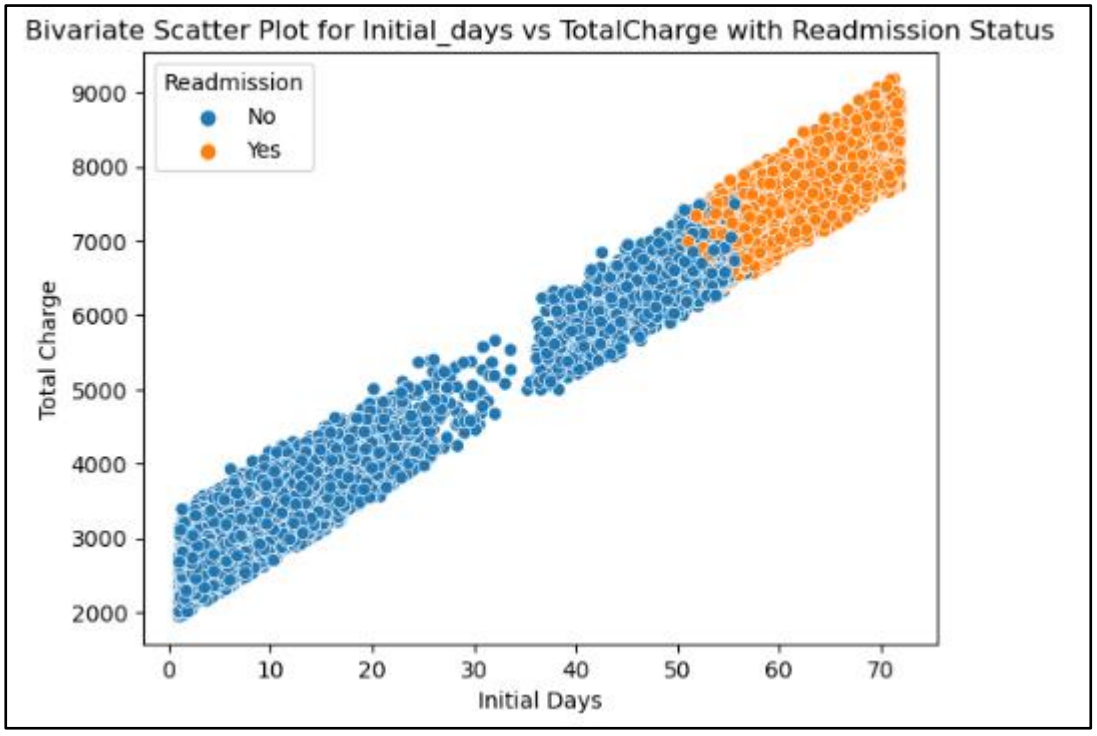
In terms of the financial implications of prolonged hospitalization, a substantial positive correlation (approximately 0.9876) is found between the duration of initial hospital stays ("Initial_days") and the total charges incurred ("TotalCharge"). This correlation emphasizes the significant increase in costs as hospital stays lengthen, underscoring the financial burden associated with prolonged hospitalization. Effective management strategies are crucial for controlling expenses related to extended stays, emphasizing the need for prudent resource allocation and efficient utilization of healthcare resources.

Furthermore, a pivot table was generated to review the cross-tabulation of categorical bivariate statistics. This approach, as outlined in a guide on bivariate analysis in Python (Analytics Vidhya, 2022), facilitates the rearrangement of data to streamline analysis and gain valuable insights into the interconnectedness of timely admissions and readmissions. The bivariate correlation analysis confirms a robust relationship between "Initial_days" and "TotalCharge," reinforcing earlier observations and providing valuable insights for optimizing hospital stays and managing readmission rates. These findings contribute to fostering improved healthcare outcomes by highlighting the importance of timely admissions and prudent resource management in healthcare facilities.

D1. Visualization of Findings

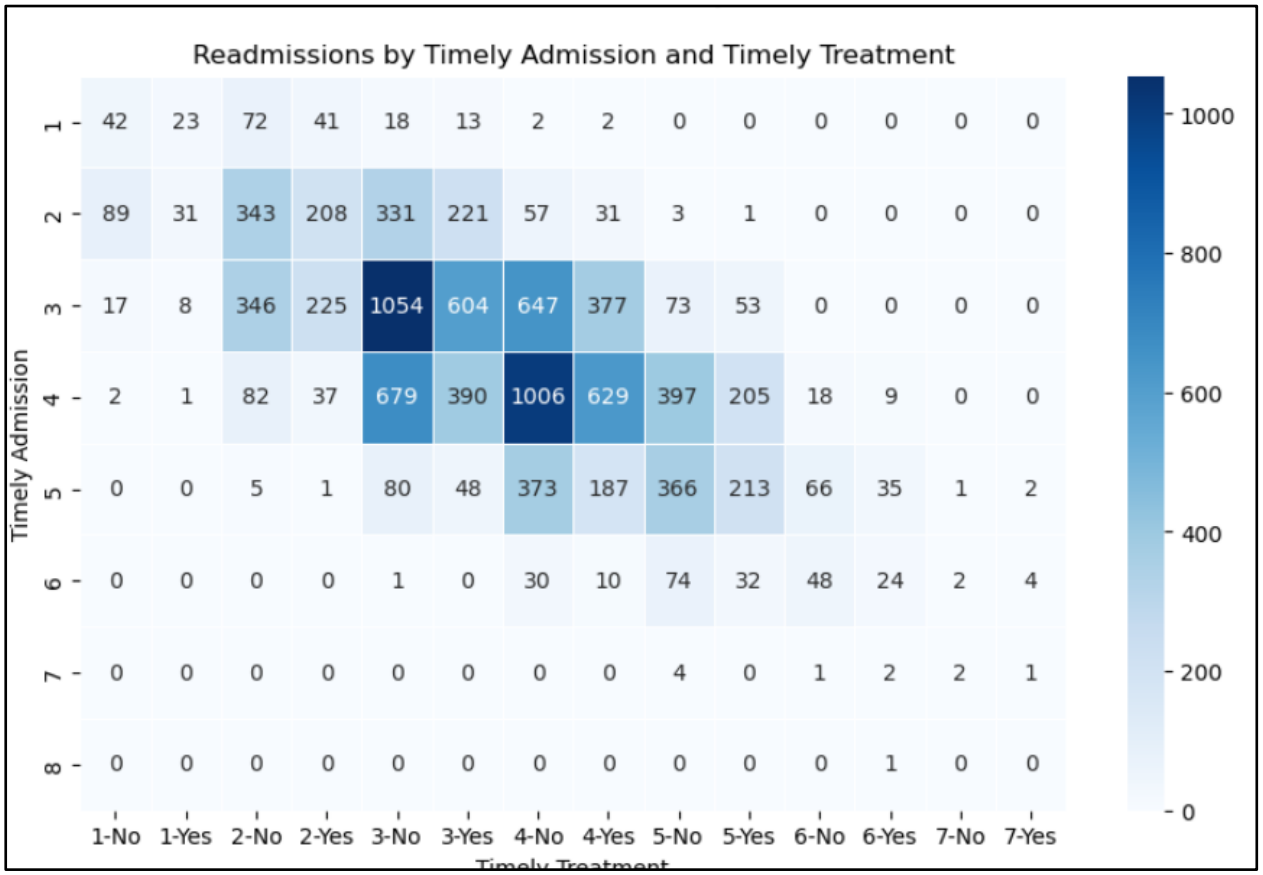
In exploring bivariate relationships through scatter plots and heat map plots using Python, I focused on two continuous variables, "TotalCharge" and "Initial_days," as well as two categorical variables, "Timely_treatment" and "Timely_admission."

The scatter plot for “TotalCharge” and “Initial_days” revealed a noticeable linear trend. Particularly, as the number of days spent in the hospital decreased, there was a corresponding absence of readmissions. Around 55 days into the initial hospital stay, a shift from no readmission to yes readmissions were observed. This finding hints at a potential correlation between the length of the initial hospital stay and the likelihood of subsequent readmissions. These insights warrant further investigation.



On the contrary, the heatmap for “Timely_treatment” and “Timely_admission” reveals a predominant presence of “No” responses, particularly noticeable across most categories. However, there was a significant increase in “Yes” responses within the 3 to 4 rating range. This increase indicates an enhancement in timely treatment and admission practices among a subgroup of patients, signaling targeted initiatives to improve timeliness within this rating range. Nonetheless, it is important to note that there is still a significant number of “No” responses within the 3-4 rating range.

The new findings, presented in a heat map format, depict the distribution of readmissions categorized by timely treatment and admission ratings. This visualization provides valuable insights into the relationship between these categorical variables and their impact on readmission rates.



E. Implications of Data Analysis:

E1. Discussion on Hypothesis Test Results

After conducting a thorough analysis of the data to investigate the hypothesis that there is an association between timely admission (“Timely_admission”) and patient readmission (“ReAdmis”), several significant findings have surfaced.

Firstly, the Chi-square resulted in a statistic of approximately 6.83 with 7 degrees of freedom, resulting in a p-value of approximately 0.45. When considering a significance level of 0.05, the obtained p-value exceeds the limit. This implies that there is no statistically significant association between timely admission and patient readmission. Consequently, I do not have sufficient evidence to reject the null hypothesis, suggesting that timely admission does not significantly impact the likelihood of patient readmission.

Upon visual examination, there appears to be a relationship looking at the correlation coefficient of approximately 0.9876 between the duration of initial hospital stays (“Initial_days”) and the total charges incurred (“TotalCharge”). This suggests that as the length of hospitalization increases, so does the total cost, shedding light on the financial implications of prolonged stays.

Moreover, the analysis of admission timeliness indicates that the mean value for “Timely_admission” is approximately 3.52, indicating that the majority of admissions fall within a moderate timeliness range. Additionally, among the 10,000 cases examined, a massive portion (6,331) did not experience readmission (“No” category).

Further exploration into the relationship between timely admission and readmission uncovers a correlation between the length of initial hospitalization and subsequent charges incurred. Specifically, individuals with longer initial hospital stays tend to incur higher expenses.

In summary, after thoroughly analyzing the data, the Chi-square test showed a p-value that exceeds the significance threshold. This means there is not a statistically significant link between timely admission and patient readmission. Essentially, timely admission does not seem to significantly affect the likelihood of patients being readmitted.

However, during the analysis, I did find a connection between the length of initial hospital stays and the subsequent charges incurred. This highlights the financial impact of extended hospitalization. To improve healthcare management, further research is needed to explore other factors contributing to patient readmission.

E2. Discussion on Limitations

In the exploration of factors influencing hospital readmission, my analysis yielded valuable insights. Nevertheless, it is crucial to acknowledge the inherent limitations in my approach, such as the potential for inaccurate decisions or misunderstood information due to dataset constraints (Webber, 2019).

One significant limitation revolves around the dataset's potential incompleteness in capturing all relevant factors influencing readmission. This could result in incomplete or overlooked critical variables that play a role in readmission dynamics. To improve this, future research attempts should consider expanding the dataset to incorporate additional key features. This expansion could involve socioeconomic factors, comorbidities, and social support, providing a more holistic understanding of the contributing elements.

Another limitation lies in the retrospective nature of the data. Analyzing past events, while insightful for exploring associations, falls short of establishing causation. To boost the validity of my findings, future studies could adopt a prospective approach, following patients over time. This shift would not only help establish causal relationships but also contribute more healthy evidence to the understanding of readmission dynamics.

The reliance on survey-based responses introduces another layer of limitation due to the inherent subjectivity of such data. Patients' perceptions and self-reporting can introduce biases

into the analysis. To enhance the accuracy of future analyses, a recommended approach would be to integrate objective clinical data with patient-reported outcomes. This combination would provide a more comprehensive and balanced view of the factors influencing readmission.

An additional assumption in my analysis revolves around the independence of observations. However, in healthcare settings, patients within the same hospital or community may share characteristics, challenging this assumption. Future work should explore methodologies that account for clustering or shared characteristics among patients, ensuring a more accurate representation of the complex interactions within healthcare environments.

In summary, while my analysis sheds light on numerous factors influencing hospital readmission, addressing these acknowledged limitations is crucial for developing more comprehensive and reliable insights into this critical healthcare phenomenon.

E3. Recommendations Based on Results

Based on the analysis results, including the findings from the Chi-square test, it is apparent that there is no statistically significant association between timely admission and patient readmission rates. This suggests that timely admission may not have a substantial impact on the likelihood of patient readmission. However, the analysis did reveal a correlation between the duration of initial hospital stays and subsequent charges incurred, indicating a financial implication of extended hospitalization.

To further understand these relationships and their implications, I recommend performing a regression analysis. This analysis would allow for quantifying the influence of timely admission on readmission rates while controlling for other related variables such as patient demographics. Also, since the chi-square test did not show any significant results, it’s a good

idea to try other statistical methods. Regression analysis, logistic regression, or correlation analysis could provide different perspectives on the relationship between timely admission and patient readmission.

Furthermore, considering the observed correlation between prolonged hospital stays and increased total charges, it is crucial to examine deeper into cost management strategies. A cost-benefit analysis should be performed to assess the potential savings from implementing interventions aimed at reducing the length of hospital stays and optimizing resource utilization. This analysis will provide insights into the financial implications of interventions focused on reducing readmission rates and help healthcare facilities make informed decisions regarding resource allocation.

In summary, while the analysis did not find a significant correlation between timely admission and patient readmission rates, it highlighted important insights into the financial implications of prolonged hospital stays. Conducting further regression analysis and exploring cost management strategies will contribute to enhancing patient care and financial efficiency in healthcare facilities.

F. Panopto Video Recording Submission:

Here is the Link:https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1bdf9c38-30a8-40bc-8247-b139006124ec

G. Third-Party Code References

Analytics Vidhya. (2022, February). A quick guide to bivariate analysis in Python. Retrieved from https://www.analyticsvidhya.com/blog/2022/02/a-quick-guide-to-bivariate-analysis-in-python/

Seaborn.heatmap#. seaborn.heatmap - seaborn 0.13.2 documentation. (n.d.). https://seaborn.pydata.org/generated/seaborn.heatmap.html

H. Source Acknowledgments

Limitations in data analytics: Considerations related to ethics ... (n.d.). https://ihe.uga.edu/sites/default/files/inline-files/Webber_2019003_paper_0.pdf

Samuel, S. V., Viggeswarpu, S., Wilson, B. P., & Ganesan, M. P. (2022, September). Readmission rates and predictors of avoidable readmissions in older adults in a tertiary care centre. Journal of family medicine and primary care. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9730993/