D207- Exploratory Data Analysis

Western Governors University

- Oviya Selvaraj

A1. Research question:

Is there an association between hospital readmission and gender?

A2. Benefit of analysis for the Stakeholders

The analysis suggests no significant association between hospital readmission and gender, meaning both genders are equally likely (or unlikely) to be readmitted. This indicates that female and male patients receive similar care and support during and after hospitalization. Thus, gender alone should not be a focus for readmission prevention strategies and any efforts designed to reduce readmission by the stakeholders do not need to be specifically around gender-based differences.

A3. Data used in analysis

The data columns used in this analysis are 'ReAdmis' (Yes/No) whether the patient was readmitted within a month of release or not and 'Gender' (Male, Female and Nonbinary).

B1. & B2. Code and output can be found in 'D207_PA.ipynb'.

B3. Justification of analysis

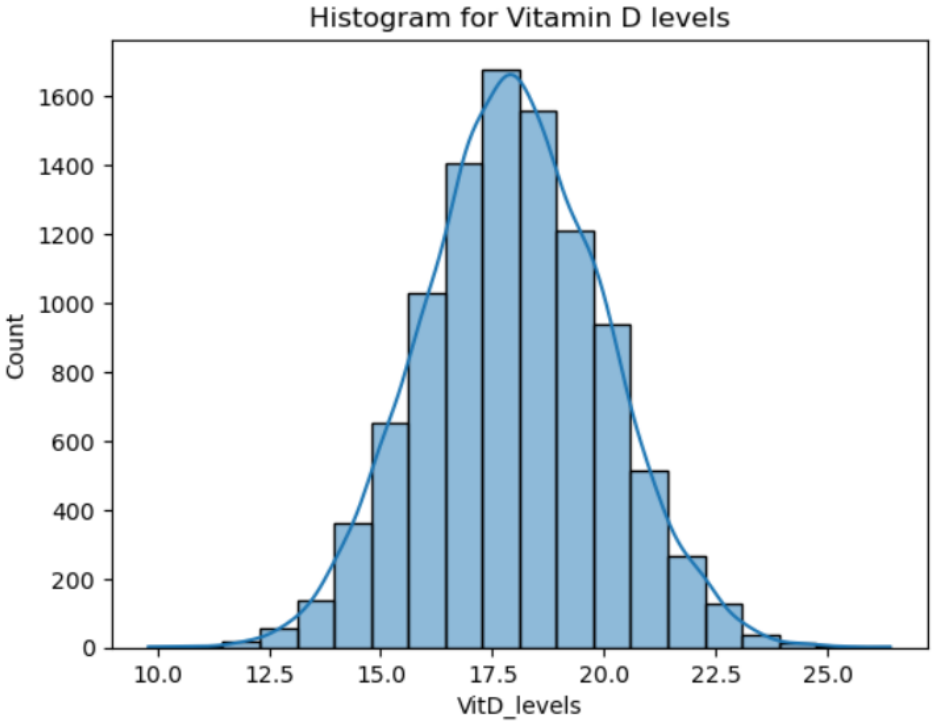
The chi-squared test is used to determine significant association between categorical variables. It is particularly helpful when we want to test whether the distribution of categorical data differs from what you would except under certain hypothesis. This test is well suited because it compares these categories in relation to one another. In this case, distribution of Gender is different for patients who were readmitted versus those who were not.

C1. Univariate Analysis

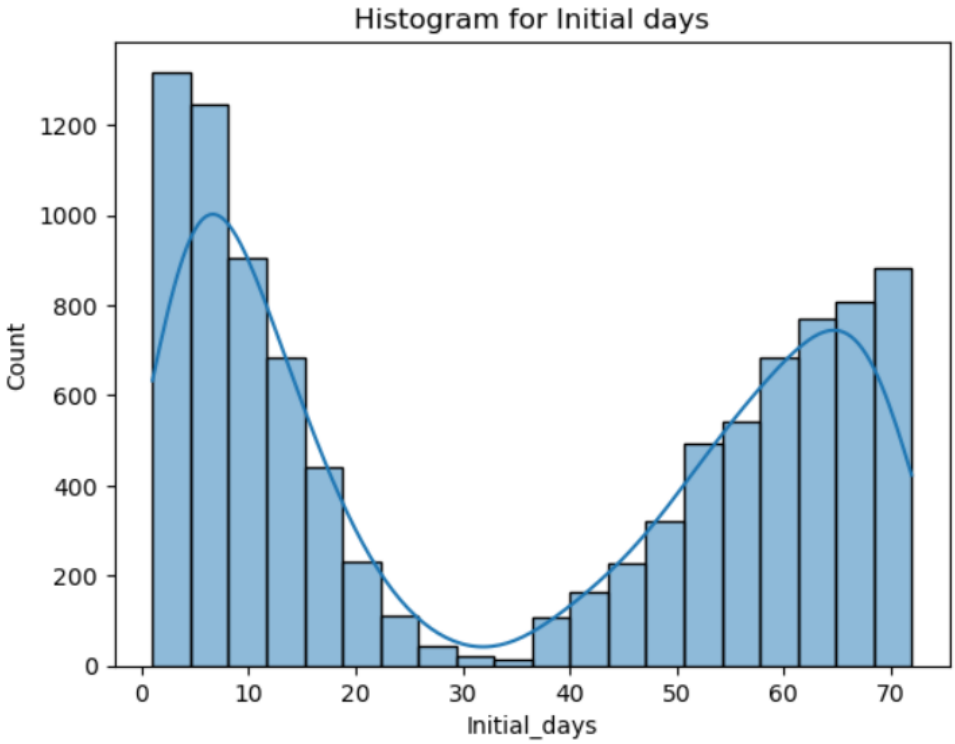
For the univariate analysis the two continuous variables used are ‘Vitamin D levels’ and ‘Initial days’, the number of days the patient stayed in the hospital during the initial visit.

Observation:

- The histogram for Vitamin D levels (‘VitD_levels’) shows a uniform distribution curve.



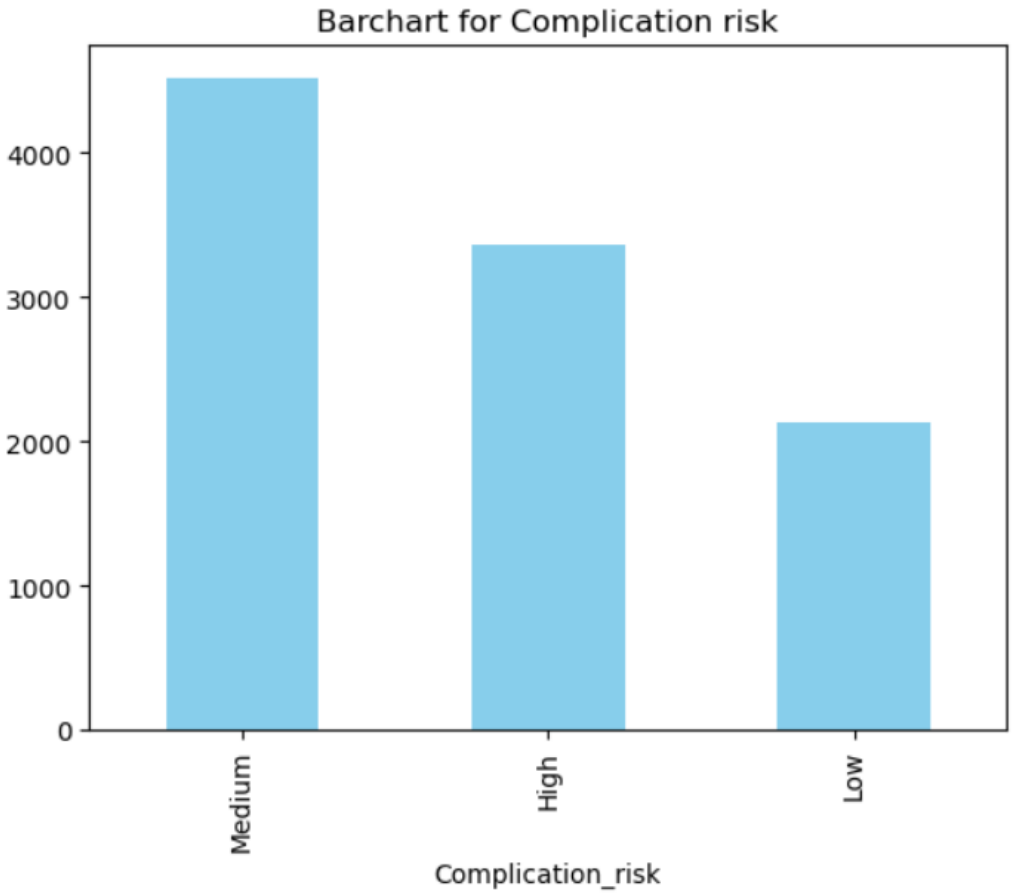
- The histogram for Initial days (‘Initial_days’) shows a bimodal distribution.



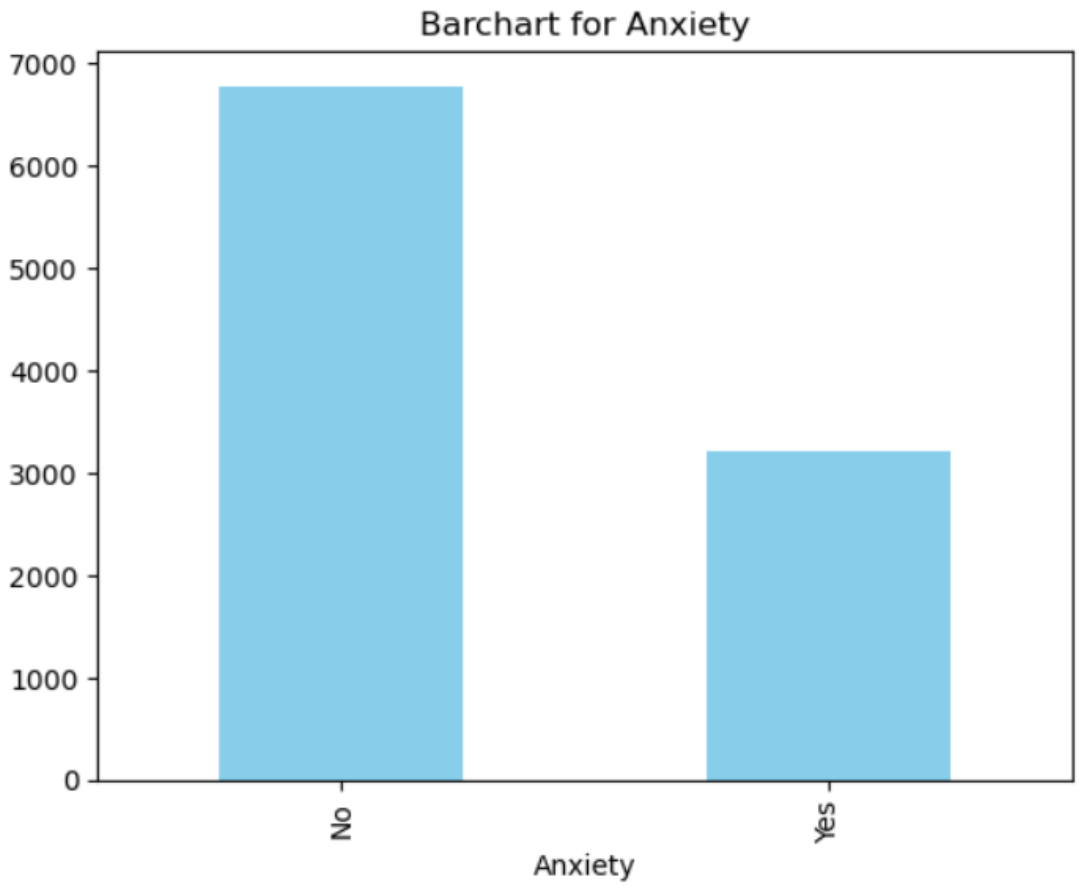
The two categorical variables used are ‘Complication risk’ which is the level of complication risk for the patient as assessed by a primary patient assessment (high, medium and low) and Anxiety (Yes/No).

Observation:

- Barchart for Complication Risk. We can observe an ordinal distribution.



- Barchart for Anxiety. We can observe Bernoulli distribution due to the binary nature.



The visualization can be found in 'D207_PA.ipynb'.

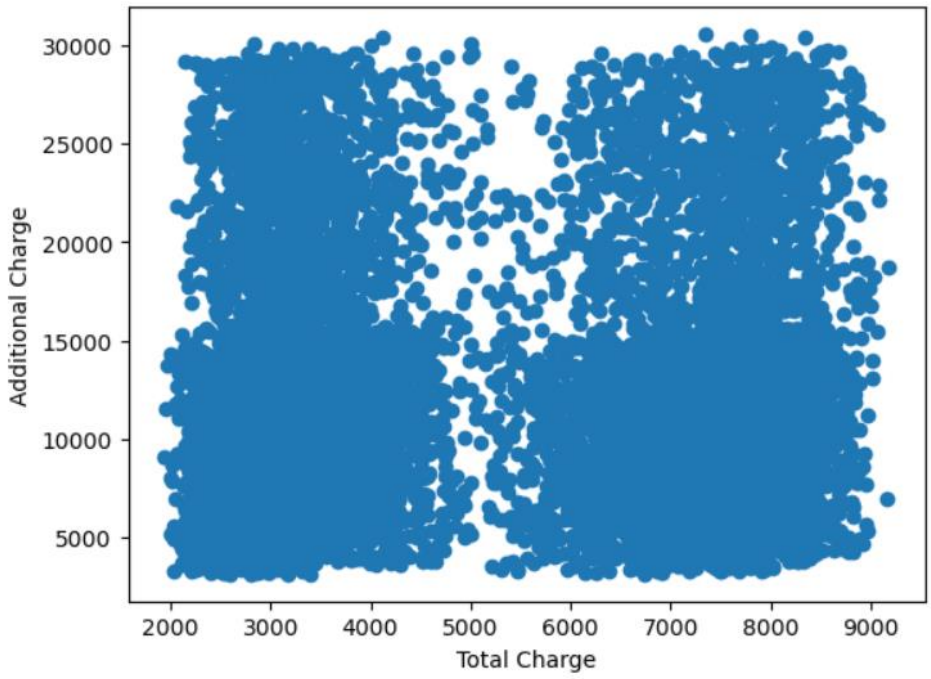
D. Bivariate Statistics

For the bivariate analysis, the two continuous variables used are:

- **Total Charge**: the average charge per patient based on the total charge divided by the number of days in the hospital.
- **Additional Charge**: The average amount charged to the patient for miscellaneous procedures, treatments, medicines etc.

Bivariate Statistics for Continuous Variables:

Using the Pearson Correlation Coefficient, we can measure the strength and direction of the linear relationship between the two continuous variables. Upon calculation, we get the value of 0.292, which indicates a weak positive correlation between the two variables.

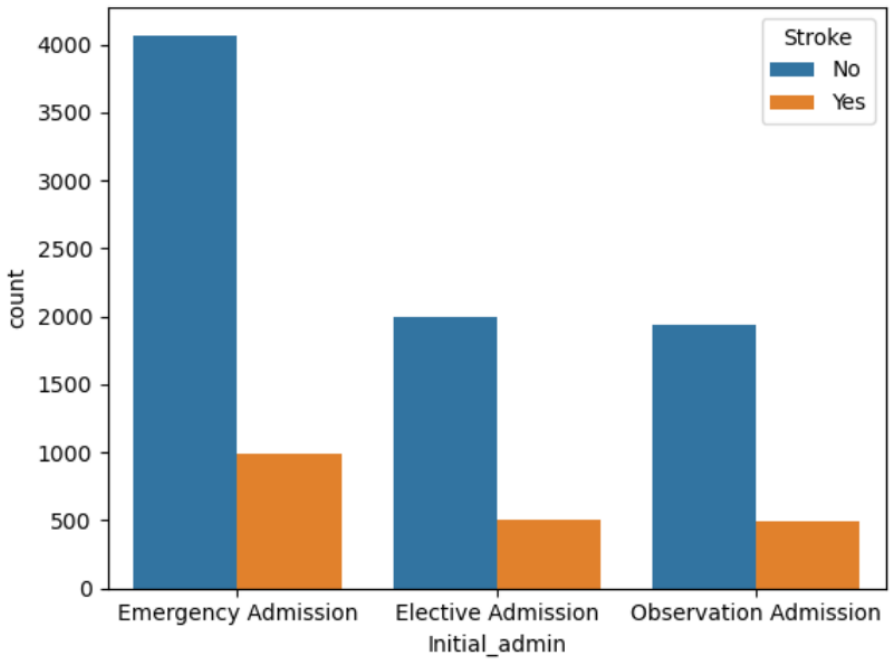


The two categorical variables are

- **Initial Admission** (emergency admission, elective admission, observation)
- **Stroke** (Yes/No)

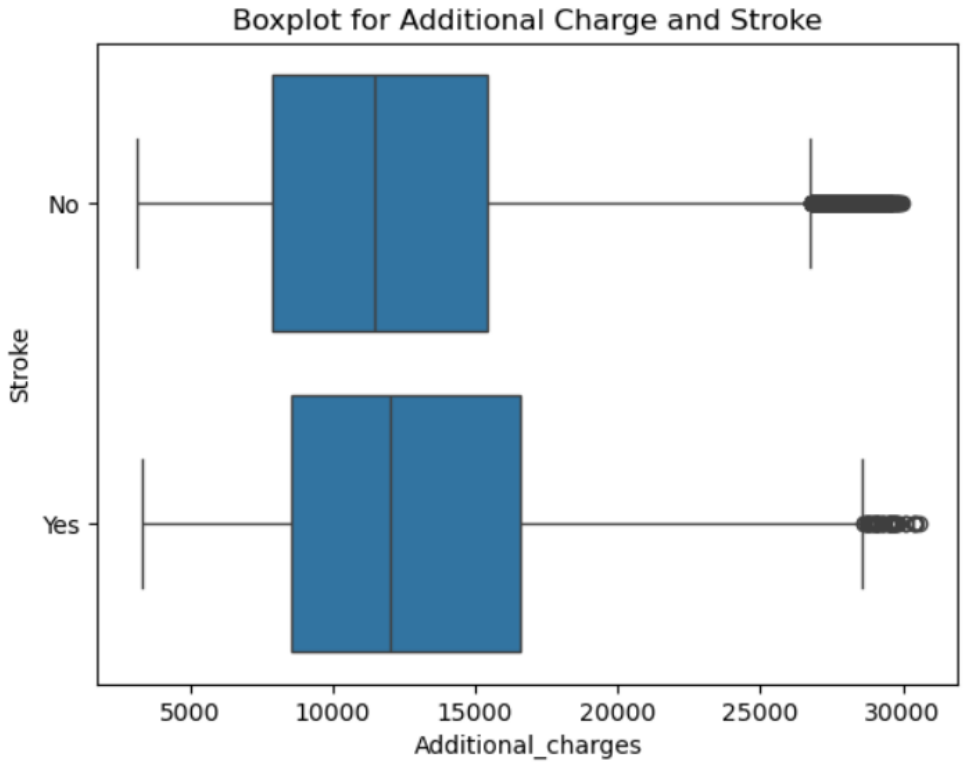
Bivariate Statistics for Categorical Variables:

Using Cramer's V we can measure the strength of two categorical variables. Commonly used after the Chi-Square test to quantify the strength of association. The value we get is 0.0097, indicating that the fields are weakly associated.

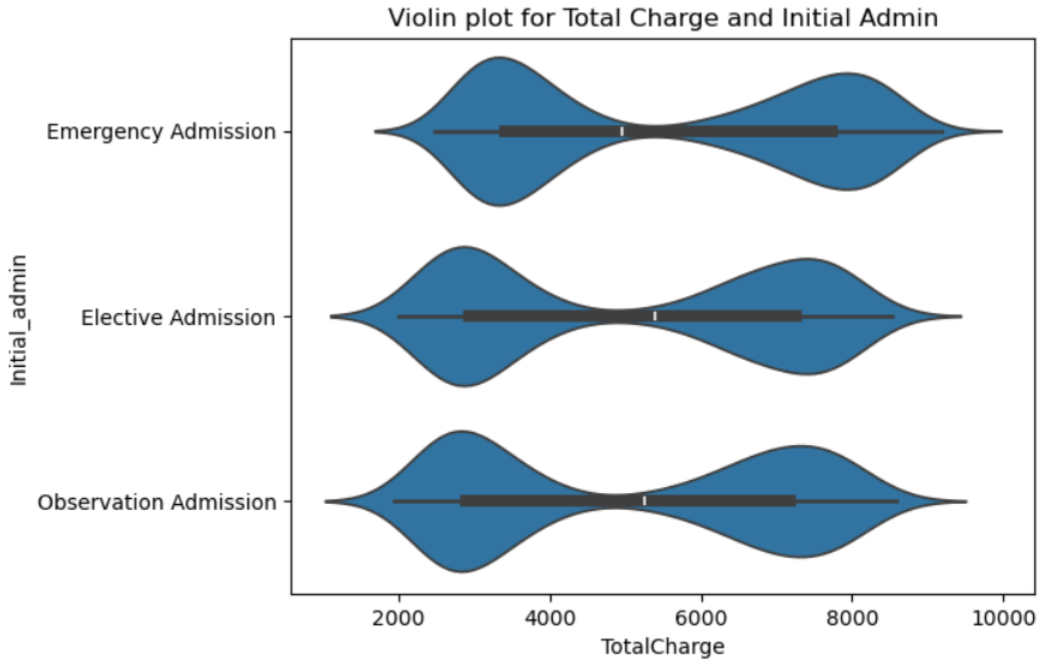


D1. Visualizations:

- Boxplot for Additional charge and Stroke.



- Violin plot for Total charge and Initial admission.



Please refer to the code 'D207_PA.ipynb' for the Bivariate visualizations and statistics.

E1. Result of the analysis

The Chi-squared statistics compare the observed frequency with the expected frequency to tell us how far the observed counts deviate from the expected counts. This helps determine whether the differences are statistically significant, and this method is well-suited for categorical data. The null hypothesis states that there is no association between hospital readmission and gender. On performing the test with an alpha value of 5%, we failed to reject the null hypothesis, which concluded that there is no significant association between hospital readmission and gender.

E2. Limitations of the data analysis

The Chi-squared test is difficult to interpret when there are a large number of categories in the dependent or independent variables. It can show if two variables are related, not if one variable cause another.

E3. Course of action based on analysis

Our analysis indicates gender is not a key driver for hospital readmission. Thus, resources should be reallocated to other factors (e.g., age, diseases, length of hospital stay) that better explain readmission. Programs to prevent readmission could focus on high-risk groups identified by clinical or socioeconomic factors rather than gender. By focusing on the factors most strongly associated with readmissions, we can more effectively allocate resources and improve patient outcomes while reducing hospital costs.

F. Panopto link:
https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bbfefee6-6ec4-4654-befe-b1fe0049ab69#

H. References

In-text citation

It can only establish whether two variables are related.

Reference entry

Adam Hayes (2024, May). Chi-Square (x2) Statistic: What it is, Examples, How and When to use the test. https://www.investopedia.com/terms/c/chi-square-statistic.asp

In-text citation

The Chi-Square test is a statistical procedure for determining the difference between observed and expected data.

Reference entry

Avijeet Biswal (2024, September). What is a Chi-Square Test?

https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test

Analytics Vidya (2024, June). *12 Univariate Data Visualizations with illustrations in python*. https://www.analyticsvidhya.com/blog/2020/07/univariate-analysis-visualization-with-illustrations-in-python/

Kaggle (n.d.). Bivariate plotting with pands. https://www.kaggle.com/code/residentmario/bivariate-plotting-with-pandas

Geeksforgeeks (n.d.). Python – Pearson Correlation Test Between Two Variables. https://www.geeksforgeeks.org/python-pearson-correlation-test-between-two-variables/

Geeksforgeeks (n.d.). How to Calculate Cramer's V in Python? https://www.geeksforgeeks.org/how-to-calculate-cramers-v-in-python/