1. Real-World Organizational Situation or issue in the Data Dictionary
   1. Do patients with higher complication risk have a higher rate of readmission?
   2. Stakeholders can use the finding s to enhance care protocols for patients identified as having a higher complication risk. By addressing potential risk factors early and implementing targeted interventions, providers can potentially reduce the likelihood of complications and subsequent readmissions, thereby improving overall care quality.
   3. The pertinent variables under examination for this research inquiry encompass ReAdmis and Complication\_risk.
2. Describe the Data Analysis
   1. See code attached (Saemi Ramirez (011926418) D207 PA Part B v3.ipynb)
   2. See the screenshot below

A screenshot of a computer

Description automatically generated

* 1. I used chi2\_contingency function because the dependent variable is categorical and this function calculates both the chi-square statistic and the corresponding p-value to test the hypothesis of independence concerning the observed frequencies within the contingency table (Scipy).

1. Univariate Statistics (Saemi Ramirez (011926418) D207 PA Part C v3.ipynb)

I selected VitD\_levels and Initial\_days as continuous variables and HighBlood and Overweight for categorical variables.  
VitD\_levels has normal distribution, mean of approximately 17.96, standard deviation of 2, minimum of 9.81, 25% of 16.63, 50% of 17.95, 75% of 19.35, and maximum of 26.39.  
  
A graph with numbers and lines

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated

Initial\_days has bimodal distribution, non-symmetric of two modes. The mean of approximately 34.46, standard deviation of 26.31, minimum of 1.00, 25% of 7.89, 50% of 35.84, 75% of 61.16, and maximum of 71.98.  
A graph with blue and white bars

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generatedA screenshot of a graph

Description automatically generated

The variable HighBlood has 2 unique values which are “Yes” and “No”. Top is “No” with the count of 5910 and “Yes” is counted for 4090.

A screenshot of a computer code

Description automatically generated A screenshot of a graph

Description automatically generated

The variable Overweight also has 2 unique values like HighBlood which are “Yes” and “No”. The most answered was “Yes” with 7097 and “No” was counted for 2906.

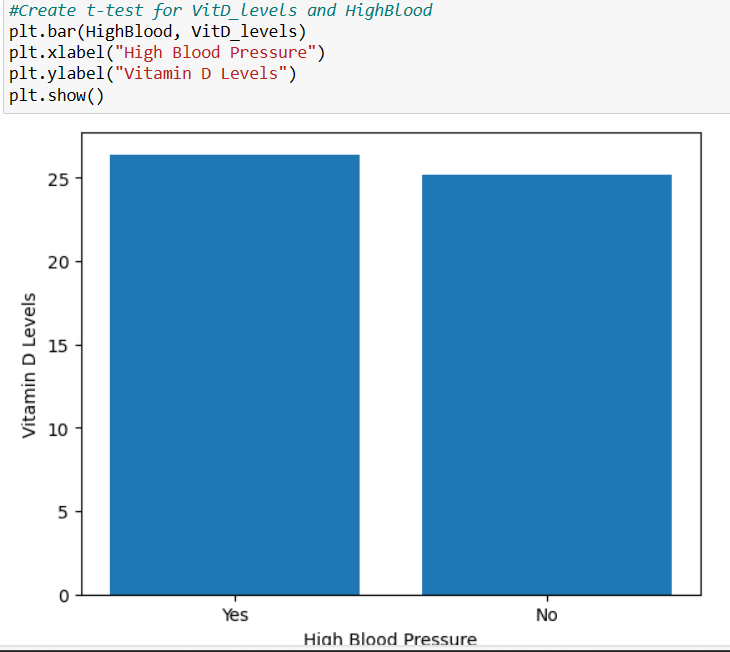
1. Bivariate Statistics (Saemi Ramirez (011926418) D207 PA Part D v3.ipynb)

I chose the same variables as section C which are VitD\_levels and Initial\_days as continuous variables and HighBlood and Overweight for categorical.

* 1. The data presented in the graph below illustrates the relationship between VitD\_levels and Initial\_days. Pearson’s correlation coefficient (r) was employed to determine the statistical significance of the numeric variables, yielding a p-value of 0.715757394908708. Given that this value exceeds the conventional threshold of 0.05, the null hypothesis cannot be deemed statistically significant and thus is not rejected.   
     A blue dots on a white background

     Description automatically generated  
     A screenshot of a computer

     Description automatically generated
  2. The analysis encompassed both numeric and categorical variables, including VitD\_levels and HighBlood, VitD\_levels and Overweight, Initial\_days and HighBlood, and Initial\_days and Overweight, all of which yielded consistent findings. The calculated p-values for each comparison were uniformly 0, indicative of complete independence between the numeric and categorical variables. Consequently, the null hypothesis, suggesting no association between these variables, can be confidently rejected.

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* 1. A chi-square test was conducted to analyze the relationship between the categorical variables HighBlood and Overweight. The resulting p-value, calculated at 0.009304497772567753, falls below the predetermined significance level of 0.05. Thus, the observed association between these variables is deemed statistically significant, warranting rejection of the null hypothesis.

A graph with different colored squares

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Summary
   1. As previously detailed in each respective section, the p-value comparing ReAdmis and Complication\_risk in Section B is calculated at 0.923567890607327. This value exceeds the established threshold of 0.05, indicating a lack of statistical significance and thus precluding rejection of the null hypothesis.  
      In Section C, univariate statistics were performed on numeric values, revealing measures such as mean, standard deviation, minimum, maximum, quartiles, and percentiles. Conversely, for categorical variables, unique values, modes, and their respective frequencies were provided.  
      Moving to Section D, bivariate statistics were conducted, notably examining the relationship between VitD\_levels and Initial\_days. The resulting p-value, approximately 0.7, signifies a lack of statistical significance, and consequently, the null hypothesis cannot be rejected. Furthermore, the analysis revealed that the p-values associated with the comparison between numeric and categorical variables were uniformly 0, suggesting complete independence between these variable types, warranting rejection of the null hypothesis. Lastly, the p-value observed for the comparison between HighBlood and Overweight was approximately 0.01, indicating statistically significance and thus supporting rejection of the null hypothesis in this case as well.
   2. The better result could be retrieved if there were more samples. There were total of 10,000, but when it was broken down into six different categories in ReAdmis/no ReAdmis and Complication\_risk (high, low, medium). These may not be representative of the larger population, and any conclusions drawn from the analysis may not be applicable beyond the specific sample studied.   
      Also, the high p-value from Section B (approximately 0.92) might have a higher risk of committing a Type 2 error, which occurs when a true effect exists but is not detected by analysis.
   3. Increase the sample size to improve the reliability and generalizability of the result. A larger sample size can reduce variability, increase the statistical power, and enhance the precision of estimates.
2. Panopto Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1318f207-22ea-4abb-85bb-b16d01827d6e>
3. Reference for 3rd party code
   1. Keith, M (2020, February 15). *Python Practice: Housing Prices: Bivariate Statistics*. YouTube. Retrieved May 11, 2024, from https://www.youtube.com/watch?v=HryVz06U5ps&t=417s
   2. *Scipy.stats.chi2\_contingency* – SciPy Documentation. (n.d.). Retrieved May 10, 2024, from <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html>
4. Reference for in-text citation
   1. *Scipy.stats.chi2\_contingency* – SciPy Documentation. (n.d.). Retrieved May 10, 2024, from <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html>