***Western Governors University***

***MSDA***

***D207 – Exploratory Data Analysis***

*Natallia Zimnitskaya*

*Student ID:  
012247127*

**A1: QUESTION FOR ANALYSIS**

Is the readmission rate for patients with diabetes greater than that of the general population?

**A2: BENEFIT FROM ANALYSIS**

Exploring whether patients with diabetes have a higher readmission rate than the general population can provide valuable insights for an organization and its stakeholders. This information can help healthcare providers identify specific needs and risk factors associated with diabetic patients, allowing for the development of targeted interventions and personalized care plans. By addressing these needs proactively, hospitals can reduce readmission rates, improve patient outcomes, and enhance overall care quality. Additionally, stakeholders such as insurers and policymakers can use this data to allocate resources more effectively, design better health policies, and support initiatives that promote better disease management and preventive care, ultimately leading to a more efficient and cost-effective healthcare system.

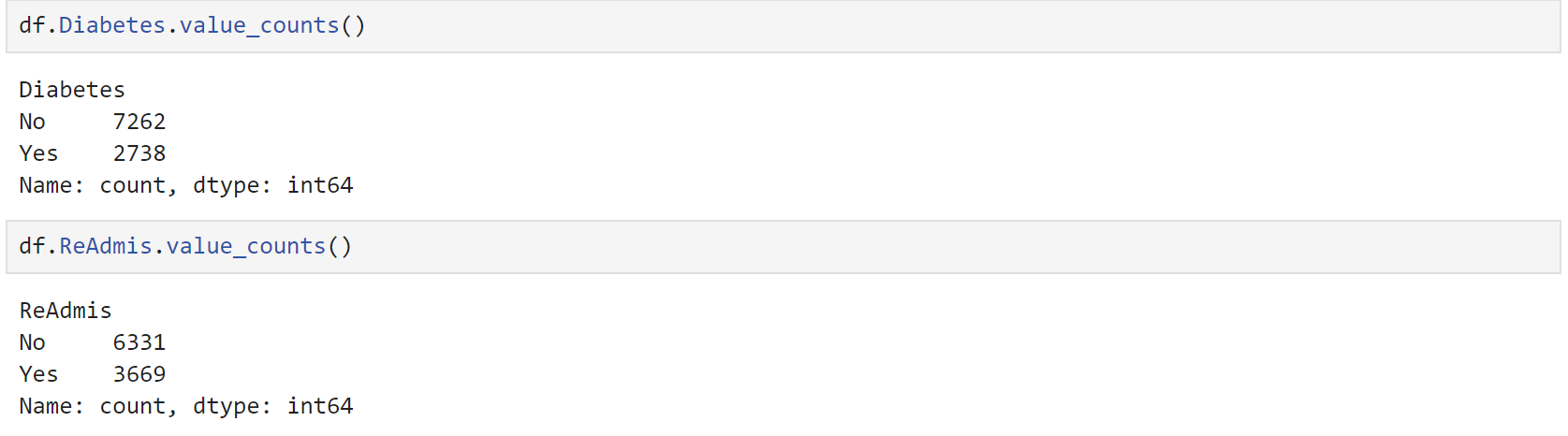
**A3: DATA IDENTIFICATON**

To answer the question above the medical\_clean.csv dataset is loaded to the Jupyter Notebook.

The dataset contains 49 columns and 10000 entries.

The variables listed below are relevant to this research question:

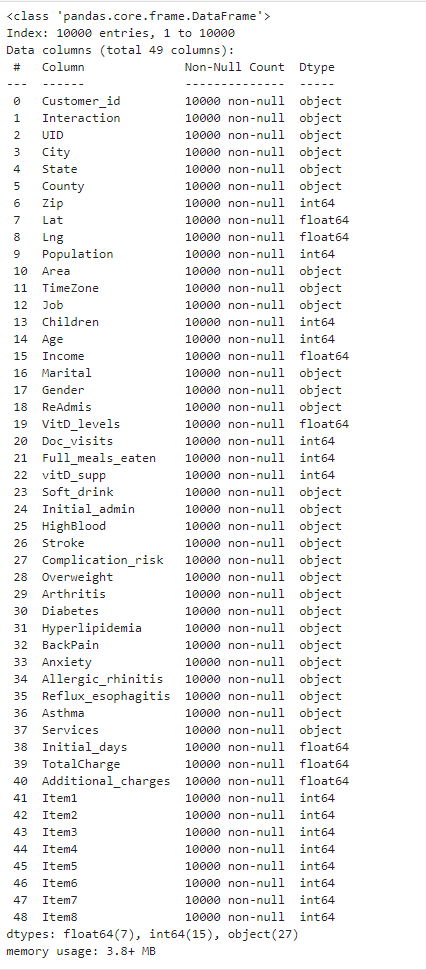
1. Diabetes (binary, qualitative). It shows whether diabetes is present in the patient or not.
2. ReAdmis (binary, qualitative). It shows if the patient was readmitted within a month of being discharged.

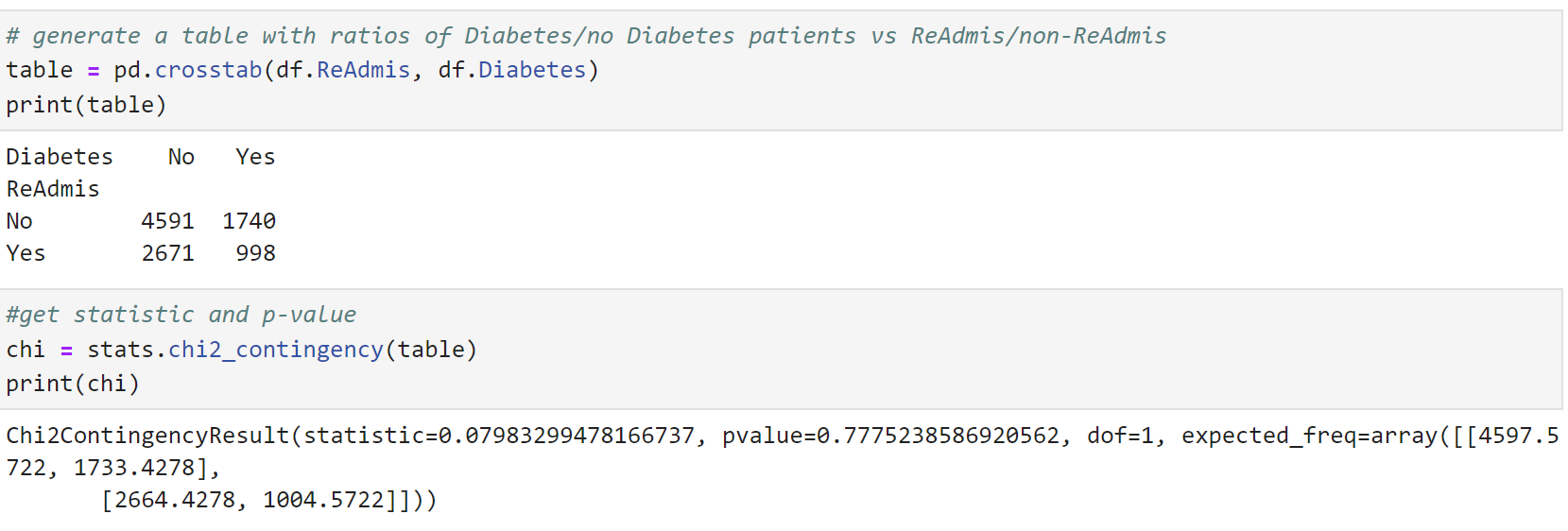


The chi-square test of Independence determines whether there is an association between categorical variables (i.e., whether the variables are independent or related) (*LibGuides: SAS Tutorials: Chi-Square Test of Independence*, n.d.). In the context of a chi-squared test between ReAdmis (readmission status) and Diabetes, the dependent variable is ReAdmis and the independent variable is Diabetes. This is because we are investigating whether the presence of diabetes (independent variable) has an effect on the likelihood of readmission (dependent variable).

**B1 & B2: CODE & OUTPUT**







Based on the chi-squared test results, with a chi-squared statistic of 0.0798 and a p-value of 0.7775, we conclude that there is no significant association between readmission status and diabetes. The high p-value indicates that any observed differences in readmission rates between diabetic and non-diabetic patients are likely due to random chance rather than a true underlying relationship. Therefore, diabetes does not appear to be a significant factor influencing hospital readmissions in this dataset. This finding suggests that other factors may be more critical in determining readmission rates and should be explored further to improve patient outcomes and healthcare strategies.

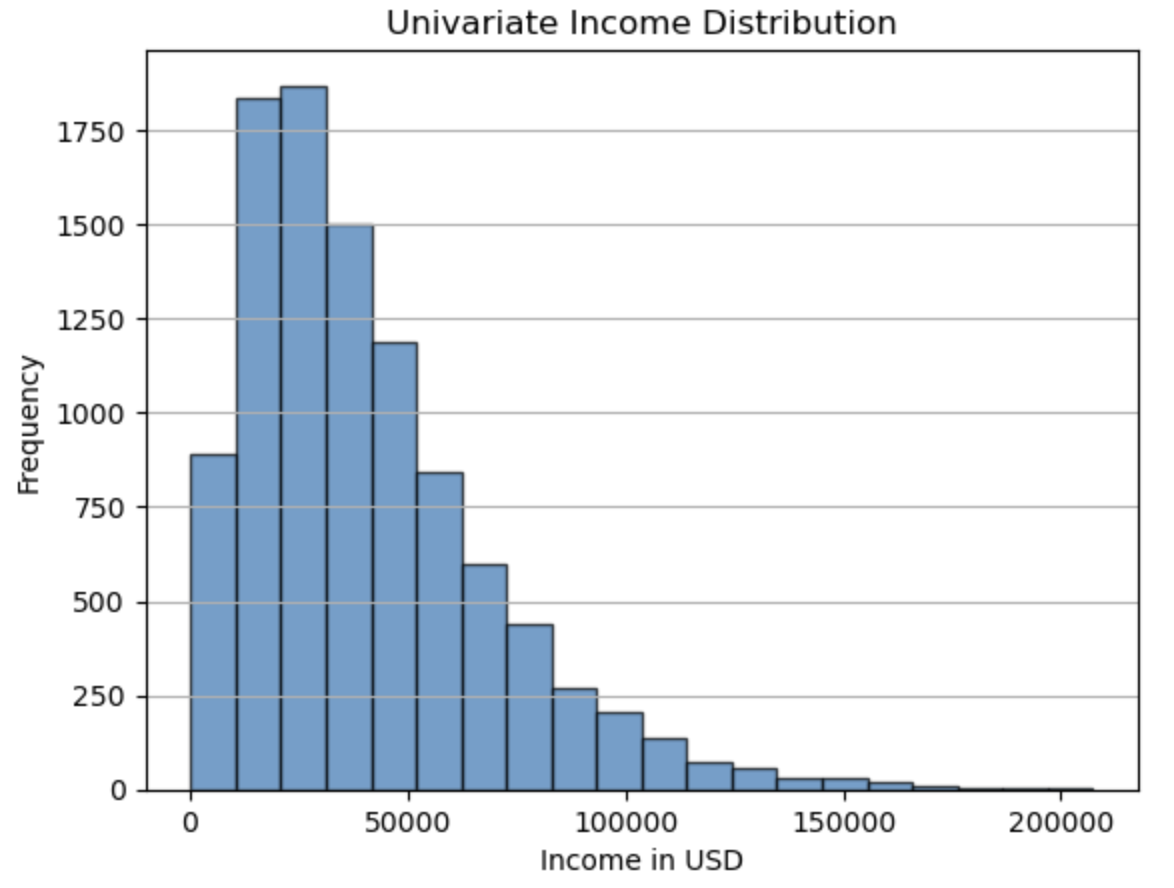
**B3: JUSTIFICATION**

Chi-square is performed using the **scipy** package and requires no normality test like other methods (*D207 Webinar EP. 1T.pptx*, n.d.). It enables the analyst to assess the relationship between two categorical variables efficiently. This test is particularly useful when dealing with large datasets, as it can handle a significant amount of data without the need for complex assumptions about the underlying distribution. By using the chi-square test, analysts can determine whether there is a statistically significant association between variables, such as readmission status and diabetes, providing valuable insights for healthcare decision-making and policy development. This method is straightforward to implement and interpret, making it a practical choice for exploratory data analysis and hypothesis testing in various fields.

**C & C1: UNIVARIATE STATISTICS AND VISUAL OF FINDINGS**

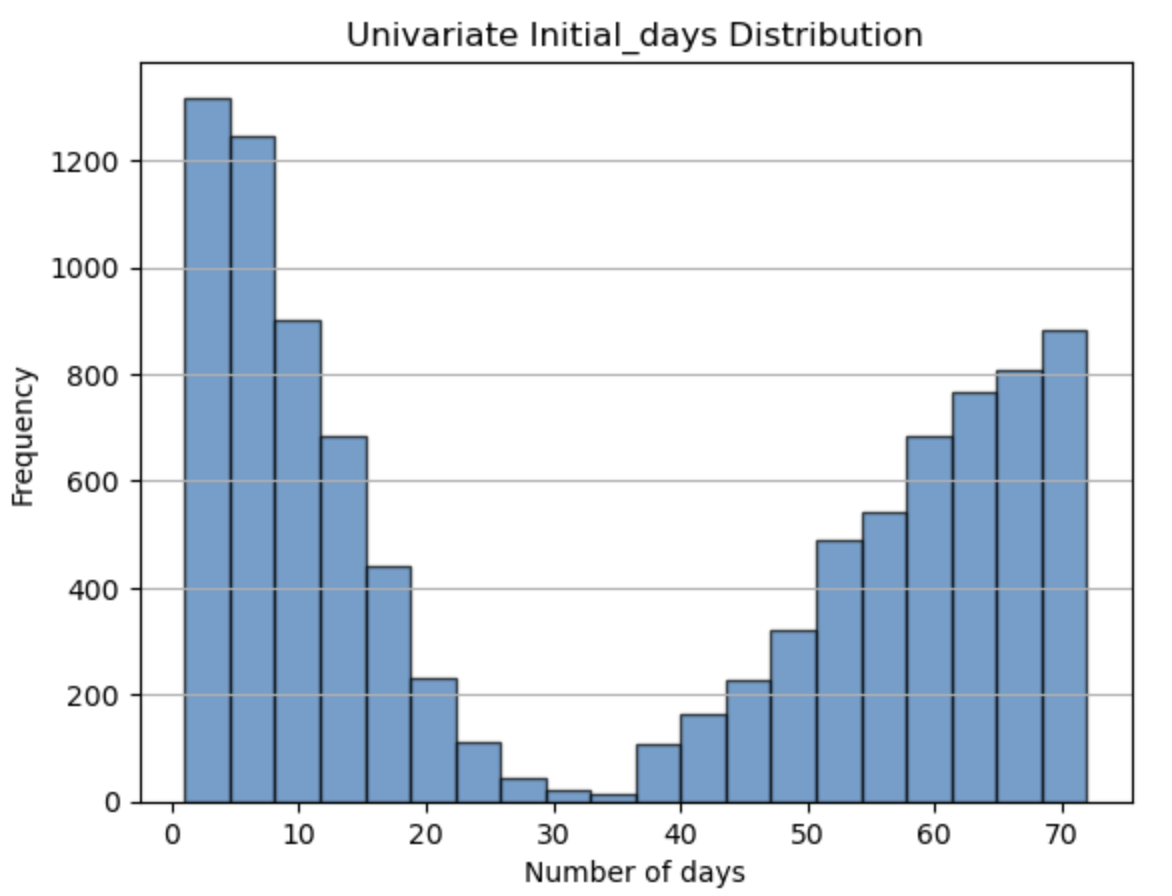
For a univariate exploration we selected two continuous variables “Income” and “Initial\_days” (the number of days the patient stayed in the hospital during the initial visit), and two categorical variables “Services” (primary service the patient received while hospitalized) and “Initial\_admin” (the means by which the patient was admitted into the hospital initially).



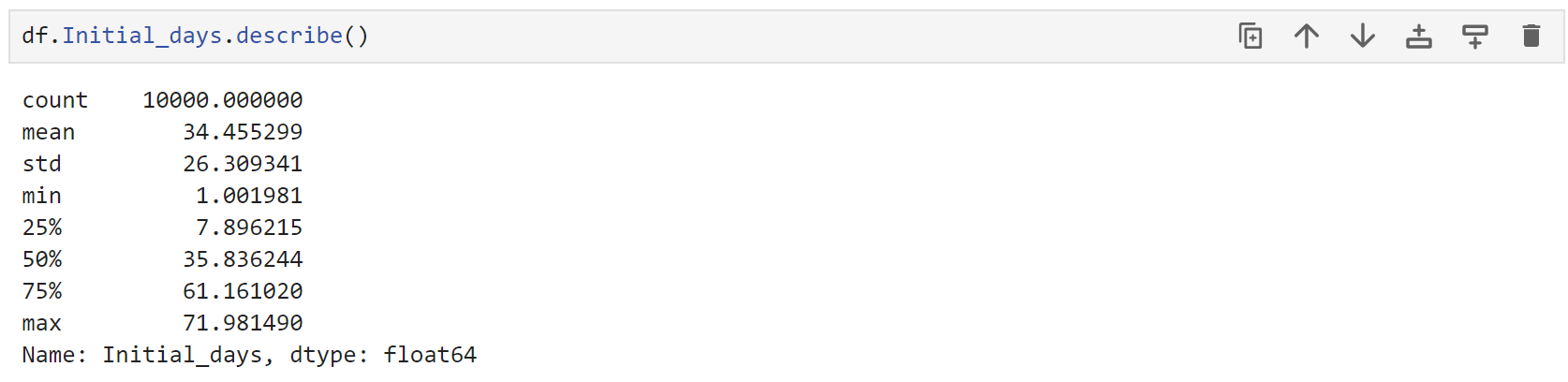


The distribution of the Income variable is skewed. The mean income ($40490.50) is higher than the median ($33768.42), indicating a right skew. The wide range between the minimum income ($154.08) and the maximum ($207249.10) suggests the presence of high-income outliers that pull the mean to the right.

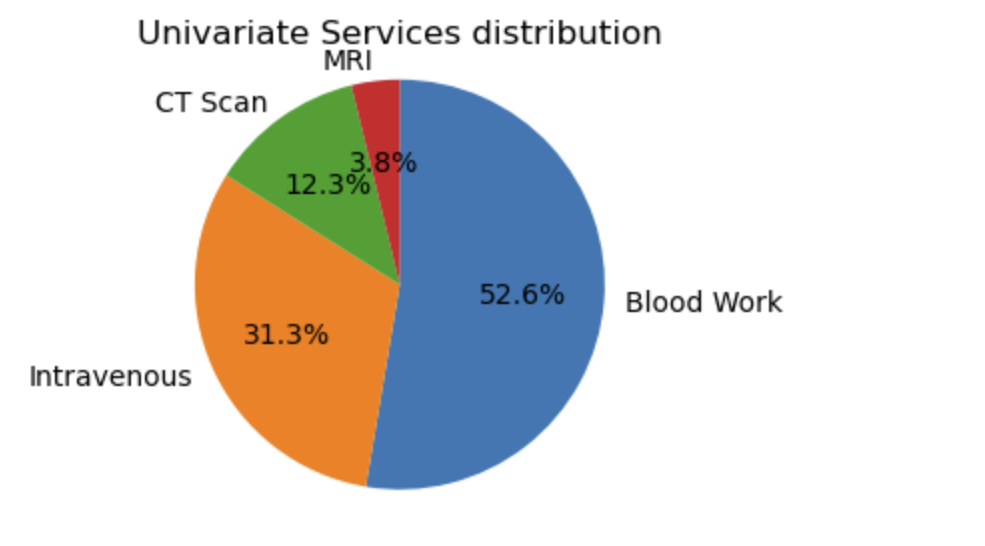




The distribution of Initial\_days appears bimodal, there are two distinct peaks. The bimodal nature suggests that there are two prevalent groups within the dataset: one with shorter initial periods and another with longer initial periods. This could reflect different patient categories or treatment protocols, highlighting the need for further investigation to understand the underlying factors contributing to these two distinct groups.

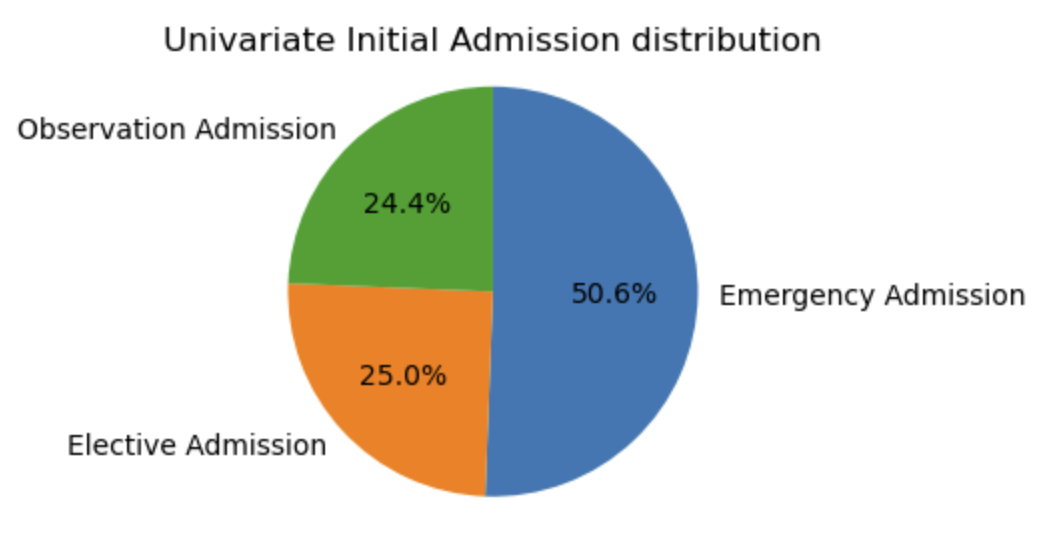






The univariate distribution of the Services variable reveals that blood work is the most common service (52.6%), indicating its routine necessity for diagnostics. Intravenous services (31.3%) reflect a significant need for IV treatments. CT scans (12.3%) and MRI scans (3.8%) are the least frequent services.



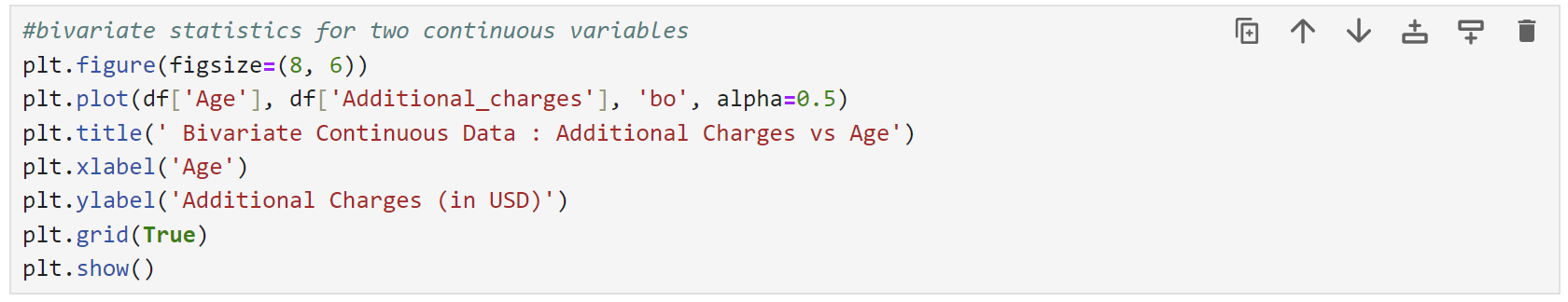


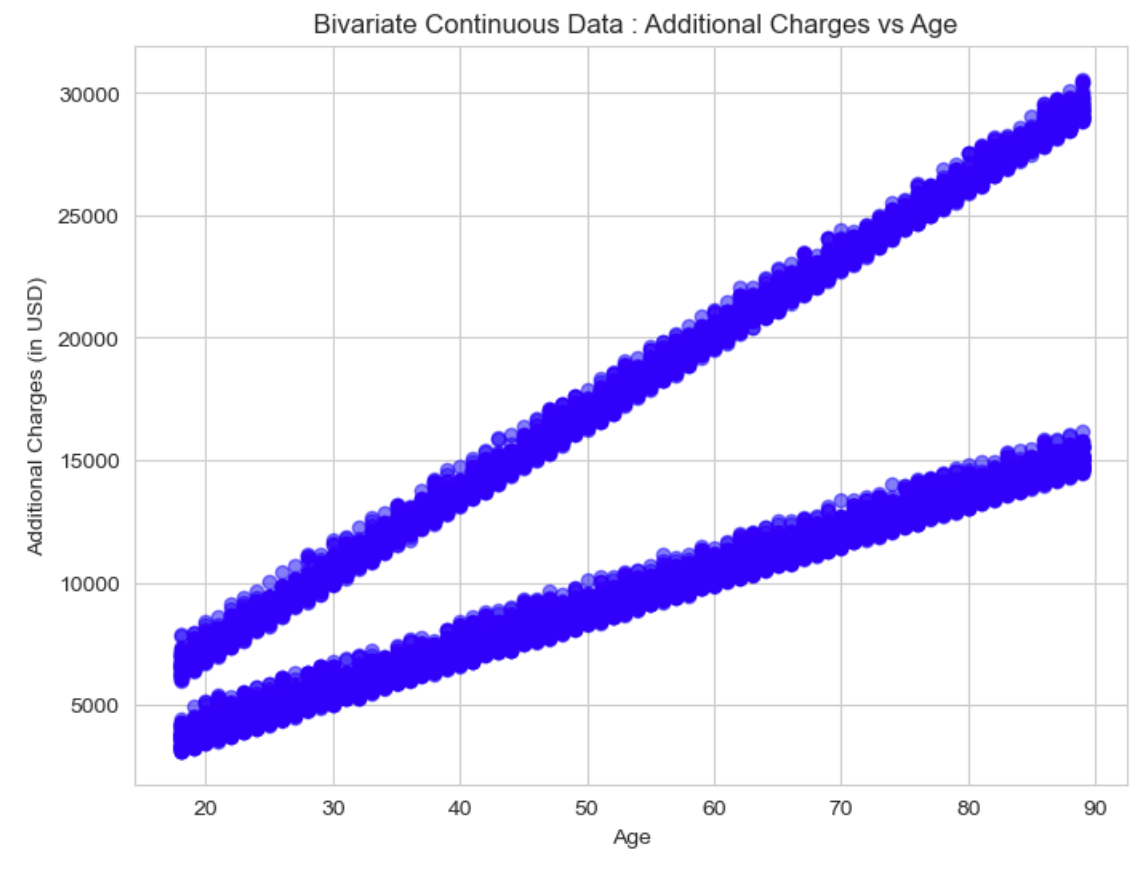
The univariate distribution of the Initial\_admin variable shows that the majority of admissions are emergency admissions (50.6%), indicating that a significant number of patients require immediate and urgent care. Elective admissions (25.0%) reflect planned and non-urgent procedures. Observation admissions (24.4%) suggest a considerable number of patients are admitted for monitoring and evaluation without immediate intervention. This distribution highlights the predominance of emergency situations in the patient population.

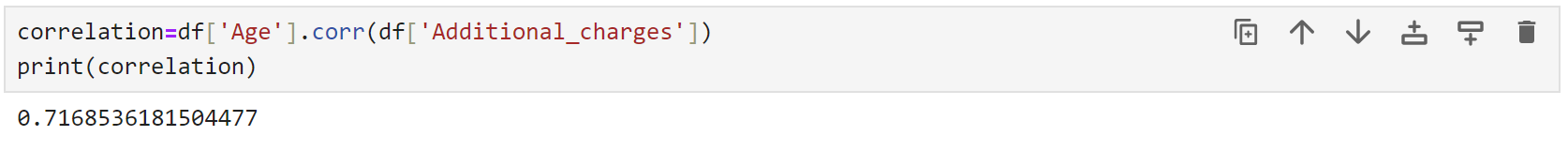
**D & D1: BIVARIATE STATISTICS AND VISUAL OF FINDINGS**

For a bivariate exploration we selected two continuous variables Age and Additional\_charges (the average amount charged to the patient for miscellaneous procedures, treatments, medicines, anesthesiology, etc.), and two categorical variables: Area and Gender.

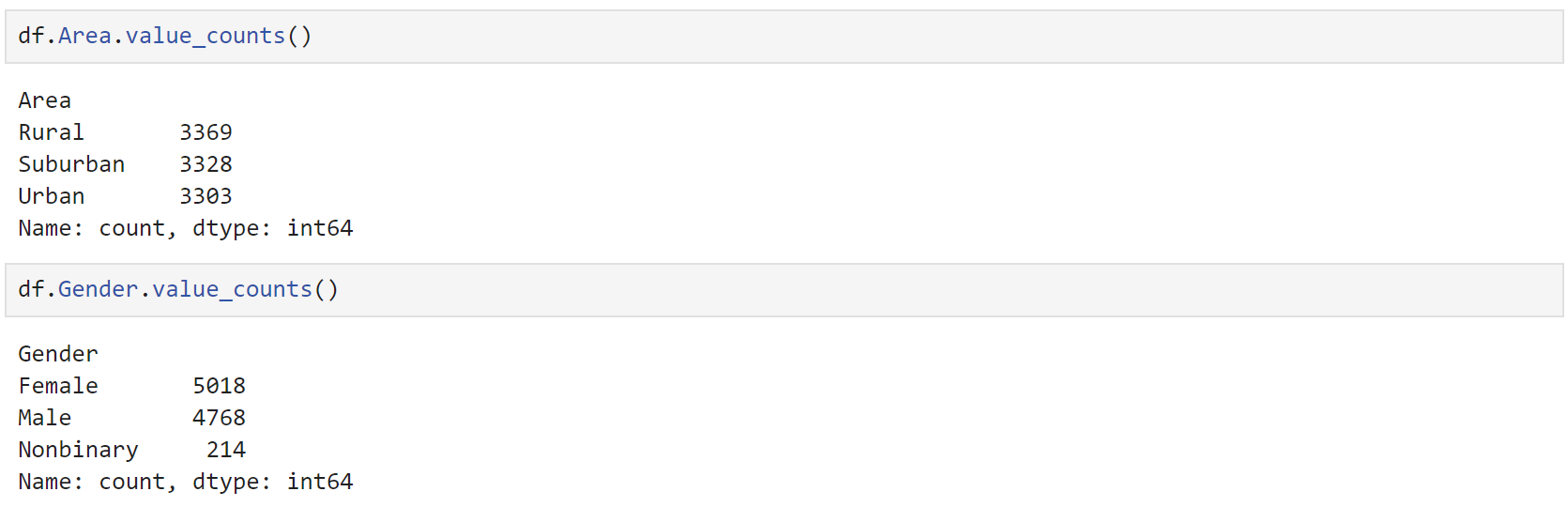




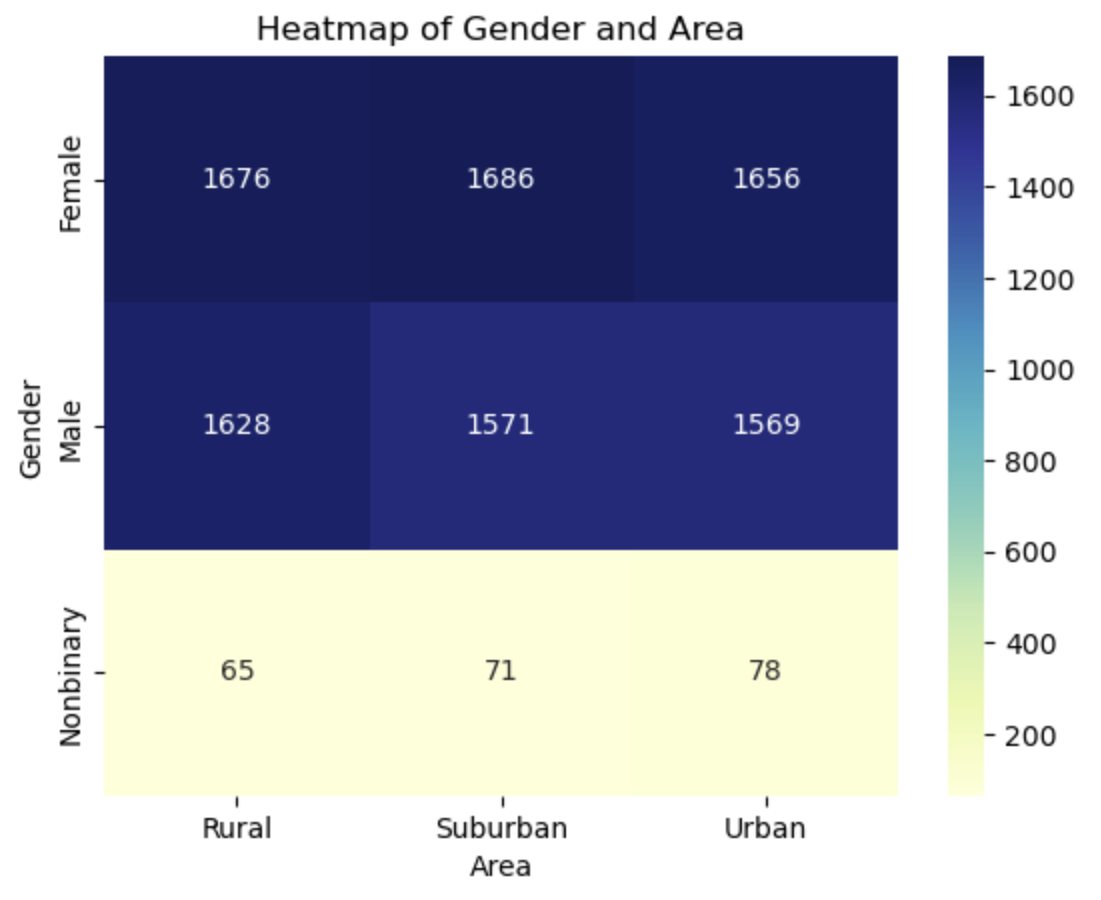


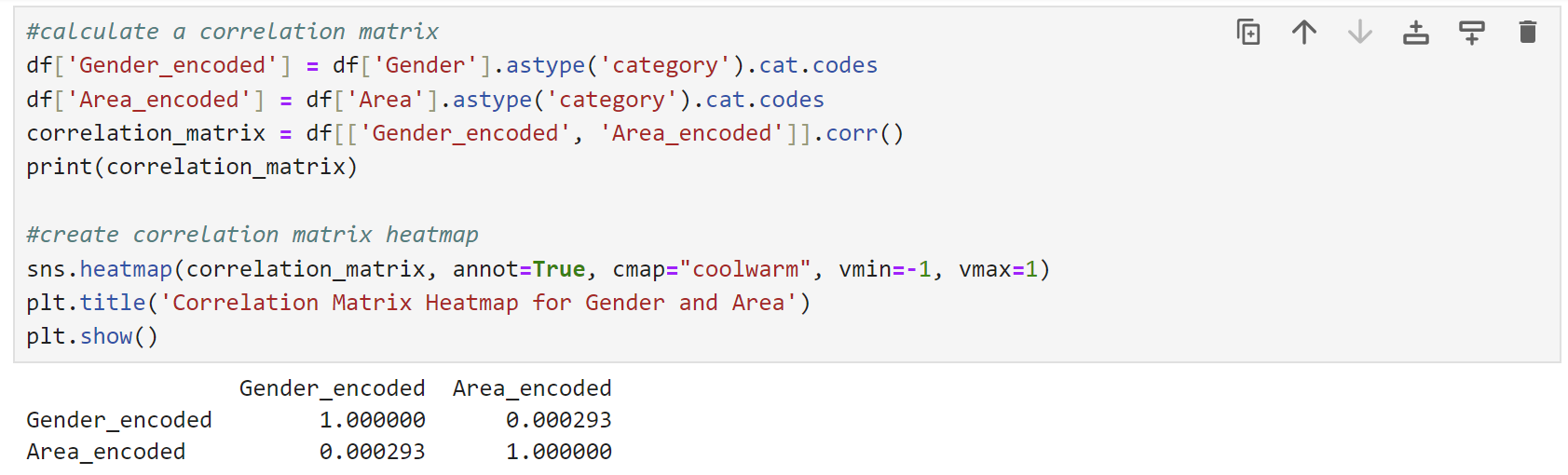


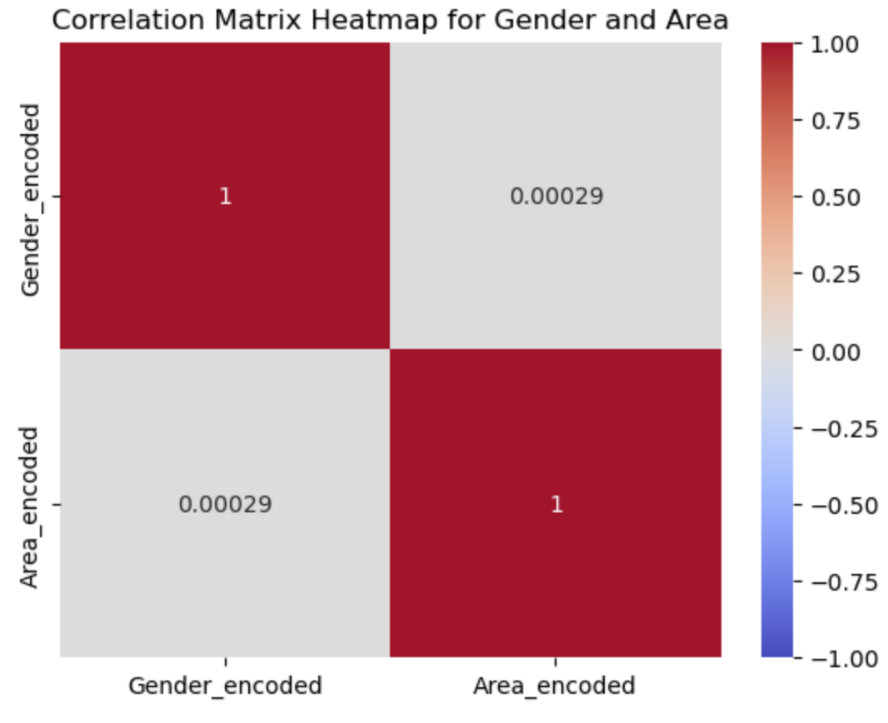
A bivariate distribution describes the relationship between two variables Additional\_charges and Age. The positive correlation is 0.717 and the scatterplot shows a trend where higher additional charges are associated with older ages. For each particular age, we observe two peaks of additional charges. This might be caused by two different types of charges, and their amount increases with the age of the patient. This means that as age increases, additional charges tend to increase as well. The positive correlation suggests a strong linear relationship between these two variables.











The heatmap reveals that there is fairly even gender distribution across Rural, Suburban, and Urban areas. The number of nonbinary people marginally higher in the Urban area (78) compare to the Rural (65). The correlation matrix suggests that there is virtually no relationship between gender and the type of the housing area. The values on the diagonal (1.000000) indicate a perfect positive correlation of each variable with itself, which is expected. The off-diagonal value (0.000293) represents the correlation between Gender and Area, and it is very close to 0.

**E1: RESULTS OF ANALYSIS**

The initial inquiry sought to determine whether the readmission rate for patients with diabetes was greater than that of the general population. There is a statistically significant difference, according to the alternative hypothesis, and there is no statistically significant difference, according to the null hypothesis. This relationship was examined using the chi-square test, with an alpha of 0.05. With a chi-squared statistic of 0.0798 and a p-value of 0.7775, the chi-squared test results lead to the conclusion that there is no meaningful correlation between diabetes and readmission status.

The relationship between two continuous variables, Additional\_charges and Age, is described by a bivariate distribution. The scatterplot displays a trend where higher additional charges are linked to older ages, and the positive correlation is 0.717. This implies that extra fees typically rise in tandem with an increase in age. There appears to be a strong linear relationship between these two variables based on the positive correlation.

The correlation matrix heatmap for Gender and Area illustrates the relationship between these two categorical variables. The off-diagonal value of 0.000293 implies that Gender and Area are not significantly correlated.

**E2: LIMITATIONS OF ANALYSIS**

The chi-square test for the relationship between Diabetes and ReAdmis has several limitations. In this dataset, the chi-square statistic is very low (0.08) with a high p-value (0.78), indicating no significant association between diabetes and readmission status. This result could be influenced by other confounding variables not accounted for in the test, and the large sample size (10,000 entries) might mask subtle associations.

**E3: RECOMMENDATIONS**

While the chi-square test provides a basic understanding, it may not capture the complexity of the relationship between these variables and should be complemented with other statistical methods for a more comprehensive understanding.

**F: VIDEO**

Please see attached link to a Panopto video

**G:SOURSES FOR THIRD-PARTY CODE**

*chi2\_contingency — SciPy v1.14.0 Manual*. (n.d.). https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2\_contingency.html

**H: SOURCES**

*D207 - Exploratory Data analysis* (By datacamp). (n.d.). [Video]. datacamp.com. Retrieved August 4, 2024, from https://app.datacamp.com/learn/custom-tracks/custom-d207-exploratory-data-analysis

*D207 Webinar EP. 1T.pptx*. (n.d.-b). <https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/ER_ESbgVK1VMpvpHIVCf0WYBn2BysB8AVdJ25RA3ys1nPg?e=QvmbT5>

GeeksforGeeks. (2022, November 7). *What is Univariate, Bivariate  Multivariate Analysis in Data Visualisation?* GeeksforGeeks. https://www.geeksforgeeks.org/what-is-univariate-bivariate-multivariate-analysis-in-data-visualisation/#

Global Health with Greg Martin. (2019, June 10). *Statistics made easy ! ! !  Learn about the t-test, the chi square test, the p value and more* [Video]. YouTube. <https://www.youtube.com/watch?v=I10q6fjPxJ0>

*LibGuides: SPSS Tutorials: Chi-Square Test of Independence*. (n.d.-b). https://libguides.library.kent.edu/SPSS/ChiSquare#:~:text=The%20Chi%2DSquare%20Test%20of%20Independence%20determines%20whether%20there%20is,It%20is%20a%20nonparametric%20test.

Polat, E. I. (2022, February 26). Hypothesis Testing with Python: Step by step hands-on tutorial with practical examples. *Medium*. <https://towardsdatascience.com/hypothesis-testing-with-python-step-by-step-hands-on-tutorial-with-practical-examples-e805975ea96e>

RegenerativeToday. (2021, July 8). *Correlation Matrix and Heatmap in Python* [Video]. YouTube. https://www.youtube.com/watch?v=2v39dj-PwoA