EDA Exploratory Data

Matthew E. Heino

D 207 Exploratory Data Analysis

EDA Exploratory Data

Introduction

This document coincides with the assessment for D207 Exploratory Data Analysis. The topics that are covered in this assessment are the following be able to present an organizational question or issue that can be answered using a dataset that has been supplied by the university, be able to perform data analysis using an appropriate test (e.g. chi-square, t-test, or ANOVA), identify and graphically depict the distribution of continuous and categorical values both univariate and bivariate.

Note: The code used for this assessment can be found in the accompanying Jupyter Notebook. The name of the file is "Heino D207 Assessment.ipynb".

Background

The dataset that was chosen for this assessment was the medical dataset. This data set was provided by the university. It is composed of 50 columns or features along with 10,000 rows or observations. The dataset is composed of information about the patient. It has information about address, charges, and other variables. The dataset has both categorical and continuous variables. Examples of categorical variables in the dataset are gender and complication risk.

Each of these categorical variables can be grouped and only have a finite set of values. For instance, the Gender column only takes the values of Male, Female, and nonbinary. An example of a continuous variable would be the Income column. The values that are contained within this column can take on a wide range of values. There are no bounds to these values.

It is assumed that the data is "cleaned." This assumption is based on the requirements document that was used for the assessment. Based on this assumption there will be no formal data cleaning of the supplied CSV file.

Section A

The following sections will discuss a question that is of interest to the medical organization. The question that is proposed in **Section A1** will later be answered in a subsequent section of the paper. There will be an explanation of how the stakeholders could benefit from an analysis of the question that has been proposed. The final section of **A** will discuss the data that will be required to answer the question and will be enumerated in this section. There will be a brief explanation of why this data is needed and what role this data plays in answering the question.

A1. The Research Question: The question that would be of interest to the organization would be "Are patients that are overweight have a higher rate of readmission? " This question can be answered using the appropriate test. The test that can be used to answer this question would be the chi-square test. The actual test will be performed in **Section B** of this document. The relationship between the chosen variables can be expressed in the following terms:

H_0: $readmission_{overweight} = readmission_{population}$

H_1: $readmission_{overweight} \neq readmission_{population}$

A2. How do stakeholders benefit? The stakeholders benefit from the analysis as there will be insight into what caused a patient to be readmitted to the hospital based on the criteria or variables that are discussed below. The organization wants to assess if there is any correlation between the variables stated below and readmission to the hospital or medical

center. If there is a correlation is there something the medical organization can do to lessen the chance of readmission using the values that are found in the analysis?

A3. Identification of the data. This section will discuss the data that is required to answer the question that was raised in **Section A1** of the assessment. The data that is needed is summarized in the chart below.

Variable Name	Description	Type of Variable
ReAdmis	Has the patient been readmitted within the last month? (Yes/No)	Categorical
Overweight	Is the patient overweight? (Yes/No)	Categorical

Table 1. Hypothesis variables.

Section B

This section will be a description of the results of the data analysis along with what test was performed to see if there is any correlation between the variables stated in **Section A3**. In **Section B2** there will be a display of the output of the results from the test and a brief discussion of what the results mean. The final section of **Section B** will be a justification of why the given test was chosen.

B1. Performing the test. The test that was performed was the chi-square. The **α** that was used for this test was 0.05. This will equate to 95% certainty. The code that was used to

perform the test is given in the snippet below (*Scipy.stats.chi2_contingency — SciPY v1.11.3 Manual*, n.d.)[1].

```
# Using crosstab and the counts ReAdmis and Overweight
cross_table = pd.crosstab(medical_df.ReAdmis,medical_df.Overweight)
print("\nThe Contents of the cross table: \n ", cross_table)

# Conduct the chi-squared using chi2_contingency.
alpha = 0.05
result = chi2_contingency(cross_table)

# section B2
print("\nResults: ", result)
print("\nThe p-value  is the following:", result[1])
if result[1] > alpha:
        print("The null hypothesis is accepted!")
else:
        print("The null hypothesis is rejected!")
```

B2. Output of the results of the test. The output of the calculation after executing the code from **Section B1** is shown below.

Output:

```
The Contents of the cross table:
 Overweight    No   Yes
ReAdmis
No        1821  4510
Yes       1085  2584
```

[1] For full code please see the accompanying Jupyter Notebook and the appropriate section – Section B.

Results: (0.6984802059617877, 0.4032948387365497, 1, array([[1839.7886, 4491.2114],
[1066.2114, 2602.7886]]))

The p-value is the following: 0.4032948387365497

The null hypothesis is accepted!

B3 Justification of the test used. The reason for using chi-square is that this type of test is best used for the types of data that were discussed in **Section A** of the paper. The variables that will be used in this test are ReAdmiss and Overweight. The test was composed of two categorical variables. This is the only suitable test that can be performed when we need to examine the relationship between two categorical variables. In this case the two variables are the ReAdmiss and Overweight. We want to see if there is a relationship in which we can see a significant association among the two variables. Other tests would require transforming the categorical values into a numeric form and then performing the test.

There are a few assumptions that can be made with these categorical values. The first is that these variables are mutually exclusive. These variables seem to be in line with this assumption. Readmission is independent and so is overweight. One does not rely on the value of the other (*Chi-Square*, 2018).

C Section

In this section, there will be an identification of two continuous and two categorical variables. The analysis will use univariate statistics and there will be a visualization of these

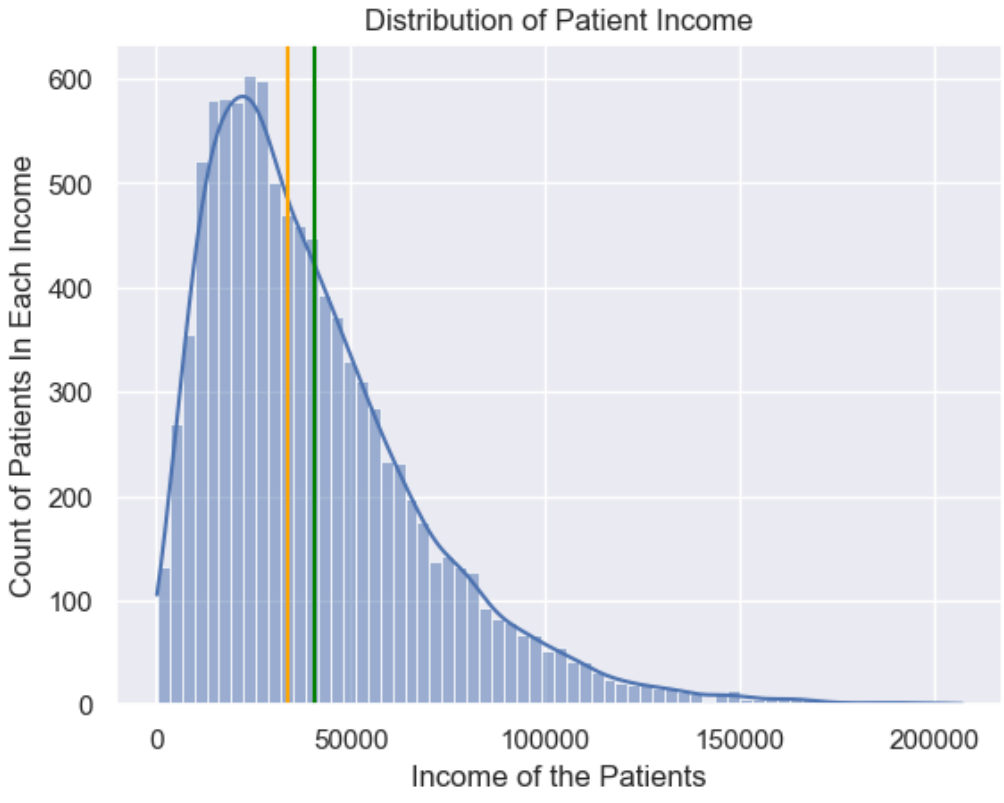
variables[2]. There will be a look into the distribution of both the categorical and continuous variables.

The continuous variables. The two continuous variables that were selected for this section are:

Variable Name	Description	Type of Variable
Income	Annual income of the patient as recorded at the time of admission.	Continuous
TotalCharge	The amount charged to the patient.	Continuous

Table 1. Univariate continuous variables.

The Jupyter Notebook was used to create some summary statistics about the continuous variables for the continuous and categorical variables. Please refer to the Notebook for these values in section C. The graph of the variables is shown below.



2

Figure 1. Univariate patient income graph.

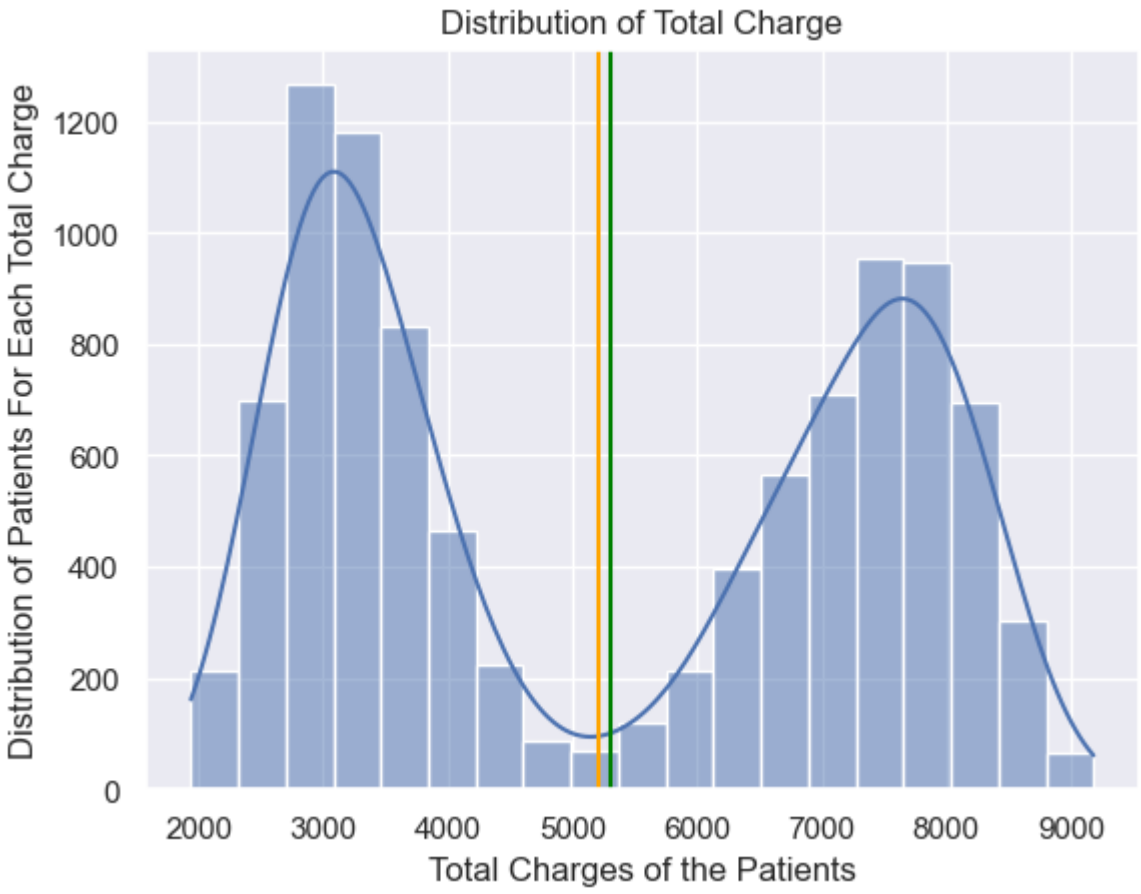


Figure 2. Univariate total charge graph.

The green vertical line represents the mean and the orange line is the median for these variables.

The categorical variables. The two categorical variables that were selected for this section are:

Variable Name	Description	Type of Variable
Gender	Gender of the patient.	Categorical

Complication_risk	Level of complication risk.	Categorical

Table 2. Univariate categorical variables.

The graph of the categorical variables is shown below.

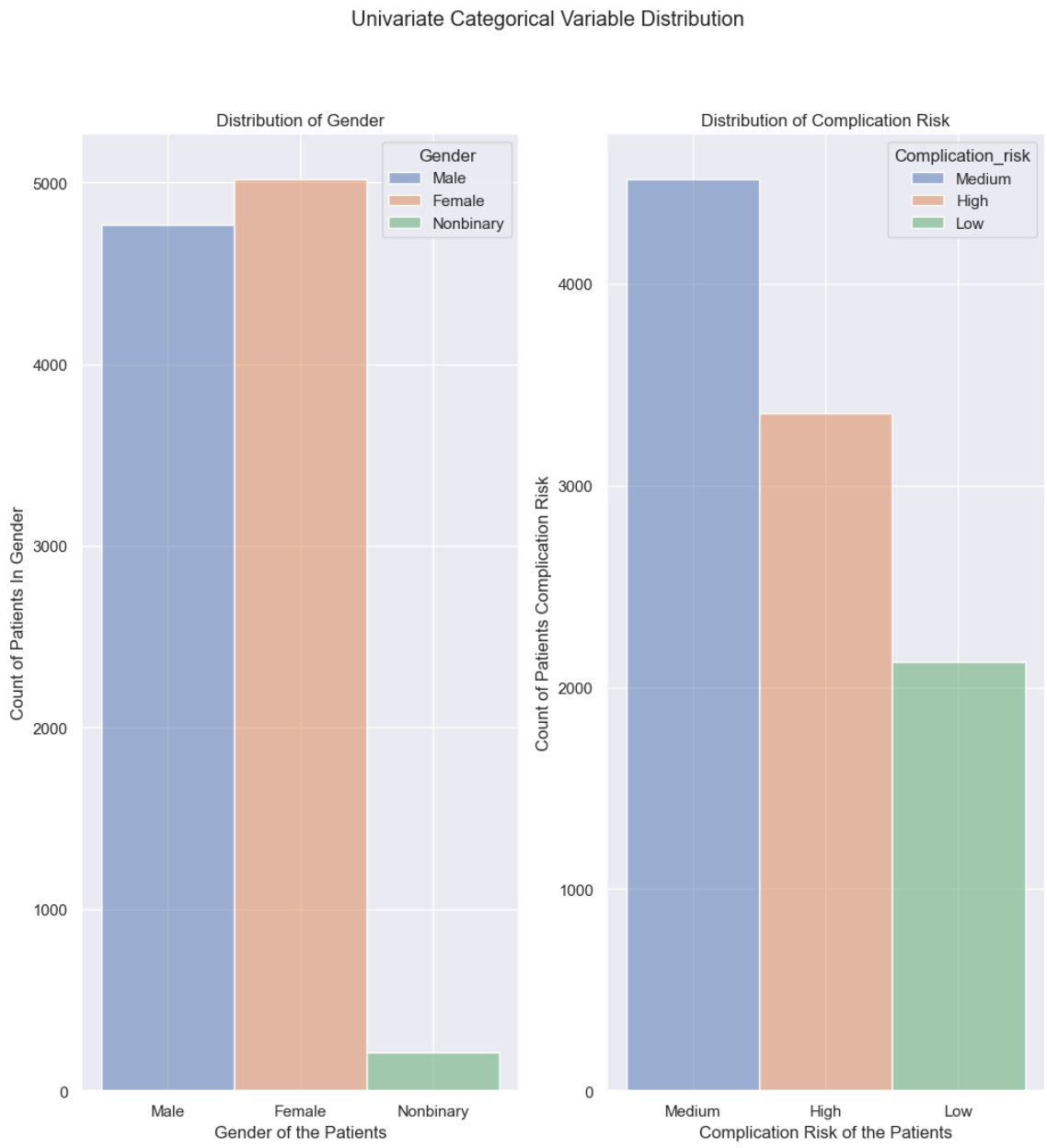


Figure 3. Univariate categorical variables graph.

Section D

In this section, there will be a look at the distribution of two continuous and two categorical variables. These variables will be analyzed using bivariate statistics.

The continuous variables. The two continuous variables that were selected for this section are:

Variable Name	Description	Type of Variable
Age	Age of the patient.	Continuous
Initial_days	The number of days the patient stayed in the hospital	Continuous

Table 3. Bivariate continuous variables.

The graph of the continuous variables is shown below.

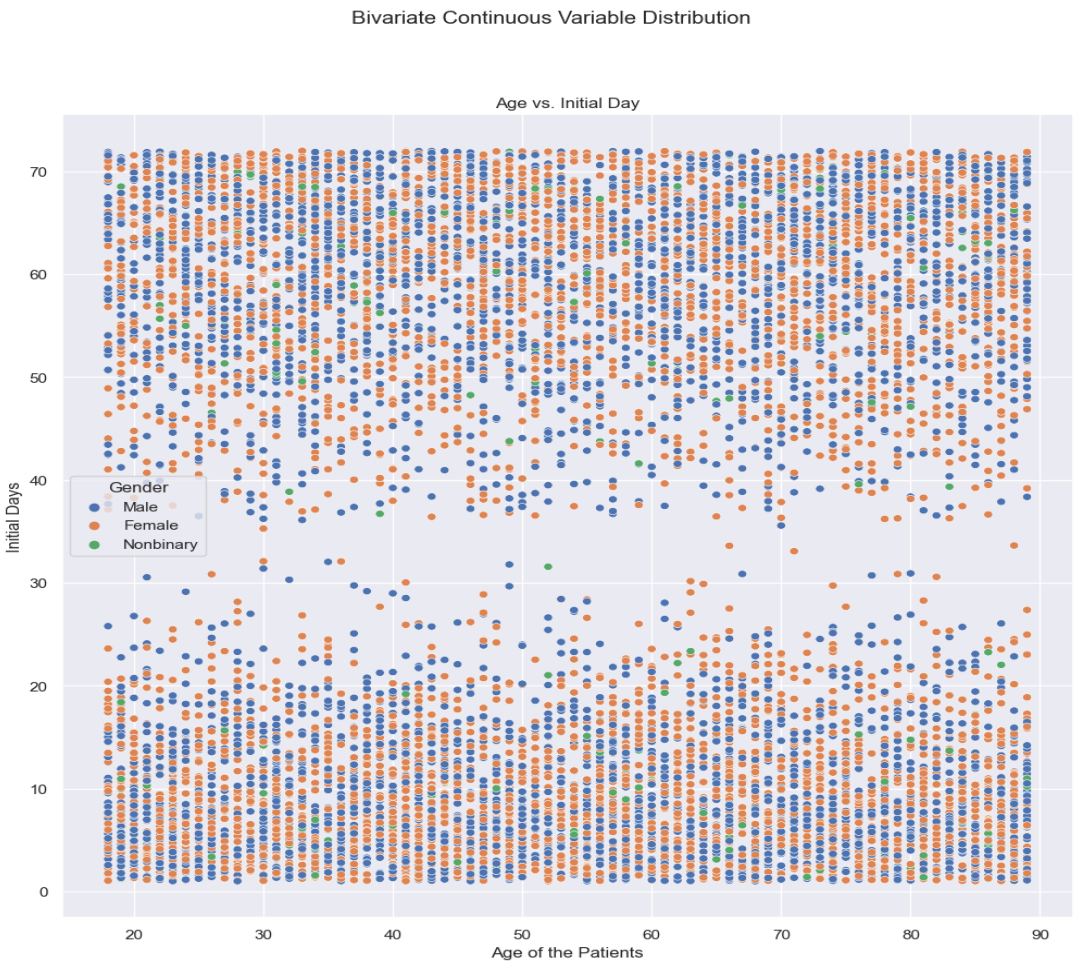


Figure 4. Bivariate continuous variables graph.

The categorical variables. The two categorical variables that were selected for this section are:

Variable Name	Description	Type of Variable
HighBlood	Does the patient have high blood pressure? (Yes/No)	Categorical
Stroke	Did the patient have a stroke? (Yes/No)	Categorical

Table 3. Bivariate categorical variables.

The graph of the categorical variables is shown below.

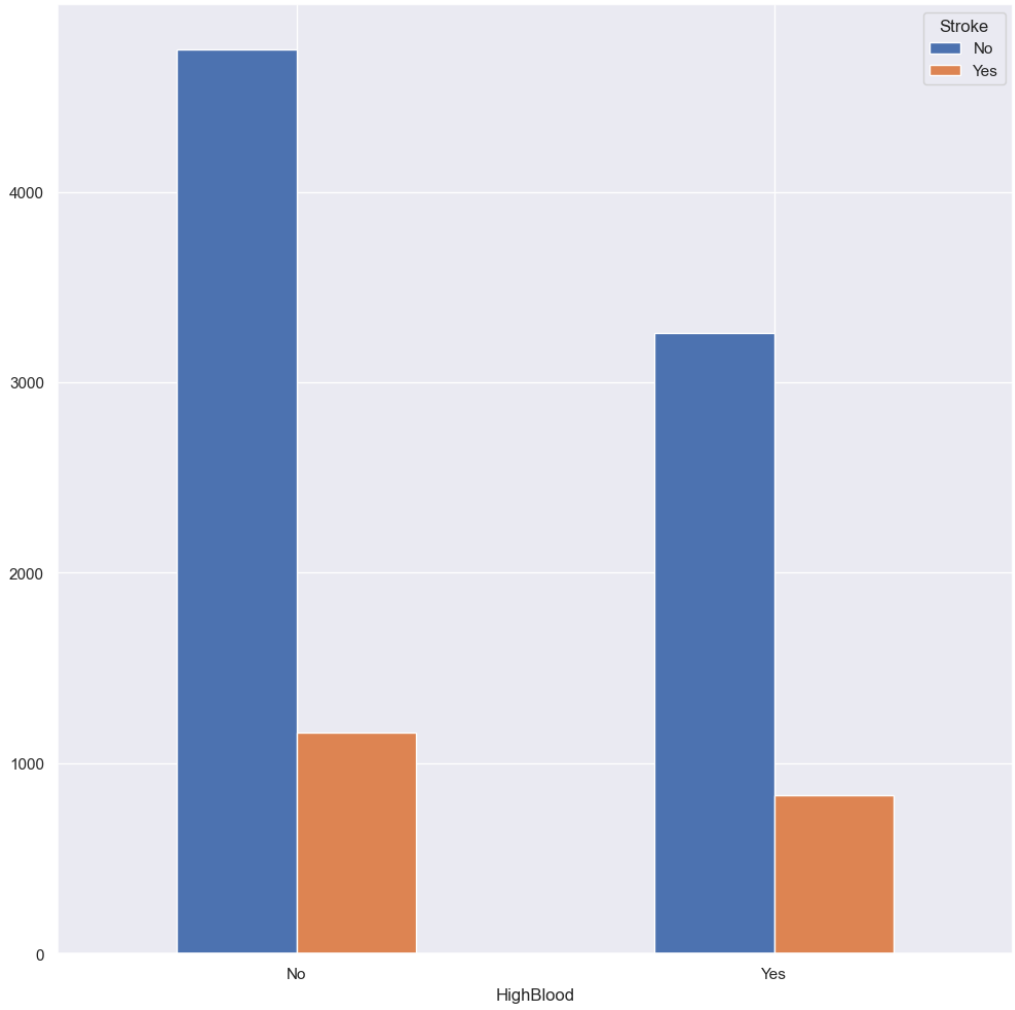


Figure 5. Bivariate categorical variables graph.

Section E

This section will summarize the implications of the data analysis. It will discuss the results of the hypothesis test. It will discuss the limitations of the data analysis. If applicable it will recommend a course of action based on the results of the tests that were performed earlier in the document.

E1. Results of hypothesis test. The results of the yielded the following results. Given the following null and alternative hypotheses:

H_0: $readmission_{overweight} = readmission_{population}$

H_1: $readmission_{overweight} \neq readmission_{population}$

The p-value for the test was ~0.40. The **α** for the test was set at 0.05. Since the p-value is higher than the **α** we must accept the null hypothesis. If the value was less than the **α** then we would be forced to reject the null hypothesis. This finding means that the distribution of the patients who were readmitted being overweight is not statistically significant as compared to the distribution of readmitted patients who are not considered overweight. The finding is that readmission of patients being overweight is equal to those who are considered normal or average weight.

E2. Limitations of the data analysis. The limitations of the analysis are the following the dataset is composed of a limited group of observations. Using a contingency table we can see that the distribution of patients was the following:

Cross Table

		Overweight	
		Yes	No
Readmis	Yes	2584	1085
	No	4510	1821

Table 4. Contingency Table

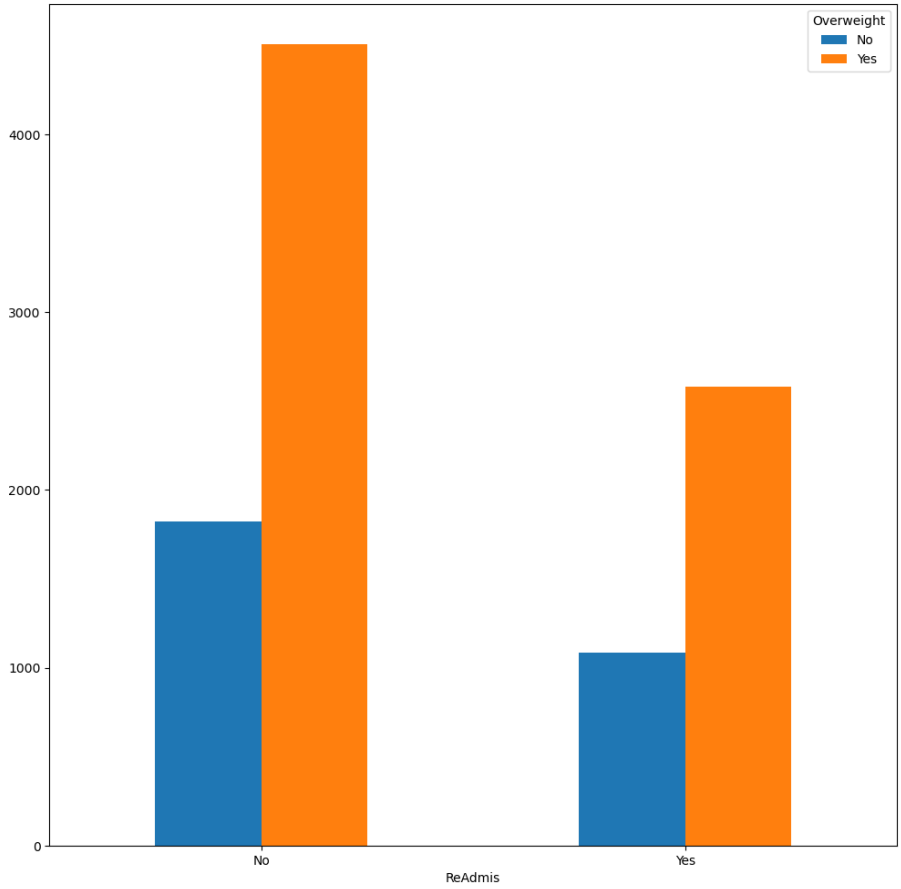


Figure 6. Bivariate categorical variables graph.

The table and graph above show the distribution of the Readmis and the Overweight column. This table indicates that being overweight does not necessarily mean that there is an increase in the likelihood that the patient will be admitted. The caveat here is that this is a relatively small sample set to draw inferences on.

The sample size only consisted of 10,000 patients. It would be appropriate to look at a sample population with more observations. It might be useful to look at other combinations of these categorical values. It might be insightful to look at distributions of outcomes for the

Readmis and Overweight relative to Gender or other categorical to help give this result more credence. Looking at only these categories of patients will likely not give the whole picture of why patients tend to be readmitted.

To interpret these findings another way is that people who are overweight may be more likely to be kept in the hospital longer and have their problems solved because of the additional time in the hospital. The variable that would be needed is available. The additional variable would be the data contained in the Initial_days column to see if there is any statistical difference between the the regular population length of time in the hospital and those patients who are considered overweight.

E3. Recommended course of action. Based on the results of the data analysis the recommended course of action is that nothing can be suggested to correct or remediate the question that was proposed in Section A of this document. A course of action based on the results of the test would be to look further into how the other categories may be used to see if there is a relationship and readmission.

As proposed in the previous section, it might be beneficial to look at how long the stay in the hospital might affect the likelihood of the patient being readmitted. The data for this is already available and no further data gathering is required. This analysis is outside the scope of this paper but is left as a possible future avenue of study.

F. Panopto Video

The link to the Panopto video is the following:

References

G. Code References.

This document made use of the materials that were found in the course content. These materials were provided by WGU and the DataCamp videos. Any code that was directly used or heavily used is acknowledged in this section.

Hashmi, F., & Hashmi, F. (n.d.). *How to visualize the relationship between two categorical variables in Python – Thinking Neuron*. Thinking Neuron – Data Science Application to Real World Problems! Retrieved November 8, 2023, from https://thinkingneuron.com/how-to-visualize-the-relationship-between-two-categorical- variables-in-python/

Holtz, Y. (n.d.). *Basic histogram with Seaborn | The Python Graph Gallery*. The Python Graph Gallery. Retrieved November 8, 2023, from https://python-graph-gallery.com/20-basic-histogram-seaborn/

Scipy.stats.chi2_contingency — SciPY v1.11.3 manual. (n.d.). https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.htm

H. In-text References.

Chi-Square. (2018, April 16). Python for Data Science. https://pythonfordatascienceorg.wordpress.com/chi-square-python/