D207 – Data Exploration

WGU M.S. Data Analytics

Lyssa Kline

July 19, 2024

A, Research Question

A1, Question

The hypothesis research question in scope for this assessment is, “Are hospital readmissions related to high blood pressure?” Hospital readmissions is a recurring problem across the country and hospitals are currently working on reducing the amount of patient readmissions. This research question is crucial to the business because it identifies the chances of hospital readmissions in relation to patients with high blood pressure. This could help determine the care that a patient with high blood pressure may receive initially to insure they won’t be readmitted.

In the future this question could be expanded to explore additional areas of opportunities to cut back on hospital readmissions.

A2, Stakeholders

As mentioned in the medical dataset data dictionary, in the medical industry, readmission of patients is such a problem that an external organization penalizes hospitals for excessive readmissions (Centers for Medicare and Medicaid Services or CMS). This organization has deemed to reduce the amount of hospital readmissions due to hospitals being overconfident and underprepared. Since this fine was put in place, hospitals have been working to reduce the number of patients getting readmitted.

The research question in scope and the reason for this study is one, out of many ways, that a hospital can begin to research and analyze areas that greatly impact hospital readmissions. The hope is that we can begin to gain an understanding on factors that influence readmission in patients. In doing so, we gain the opportunity to give greater care to those with high risk of readmission upfront, which will both help hospital readmissions as well as customer satisfaction.

This analysis is important to reduce the number of admissions a single patient occurs in each period. In doing so, hospitals will occur less fees and overall customer care, health, and satisfaction will improve. The various stakeholders who may benefit from such a study include hospital administration, clinical staff, patients and families, regulatory bodies, patient advocacy groups, etc.

A3, Data Output

The medical dataset chosen for this analysis contains 49 columns and 10,000 rows. each containing variables relating to a patient’s medical history and the services required.

The research question is directly related to the readmission of patients and the indication of high blood pressure. Thus, the relevant data for this analysis are as follows.

- ReAdmis – (qualitative) field containing ‘Yes’ or ‘No’ categories. This field indicates whether the patient was readmitted within a month of release or not.
- HighBlood – (qualitative) field containing ‘Yes’ or ‘No’ categories. This field indicates whether the patient has high blood pressure or not.

It should be noted that both these fields inside the medical clean dataset have a data type of object which represents a categorical field. Both fields contain qualitative nominal values that can be divided into distinct categories based on attributes (i.e. yes and no).

To adequately evaluate whether there is a significant association between the two 'yes or no' variables. A Chi-Squared test would be most appropriate when testing categorical data. This type of test will directly address the research question and prove whether there is a relationship between the two chosen variables. In this specific analysis, a Chi-Squared test will help to prove if there is a relationship between customers re-admission and customers with high blood pressure.

B, Describe Analysis

The below sections describe the analysis performed on the hypothesis in scope.

B1, Dataset Analysis Techniques

A chi-squared test is a statistical test used to determine if there is a significant association between two categorical variables. In python you can perform a chi-squared test using the 'scipy.stats' module. This analysis compared the observed frequencies of events to the expected frequencies if the variables were independent.

The full set of code used to create the chi-squared analysis shown below starts with inputting the necessary packages, reading in the medical dataset, creating a contingency table, then using chi-squared analysis to output the results. See the results outputted under B2 below.

[34]:
```
# Importing all packages used in this workflow.
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency
import matplotlib.pyplot as plt
import seaborn as sns
```

[36]:
```
# Using pandas to read the loaded csv file, indicating the first column is an index which pandas does not need.
df = pd.read_csv('./medical_clean-Copy.csv', index_col=0)
# Checking the info on the loaded dataset, indicating columns, data types, and size.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 10000 entries, 1 to 10000
Data columns (total 49 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Customer_id         10000 non-null  object
 1   Interaction         10000 non-null  object
 2   UID                 10000 non-null  object
 3   City                10000 non-null  object
 4   State               10000 non-null  object
 5   County              10000 non-null  object
 6   Zip                 10000 non-null  int64
 7   Lat                 10000 non-null  float64
 8   Lng                 10000 non-null  float64
 9   Population          10000 non-null  int64
 10  Area                10000 non-null  object
 11  TimeZone            10000 non-null  object
 12  Job                 10000 non-null  object
 13  Children            10000 non-null  int64
 14  Age                 10000 non-null  int64
 15  Income              10000 non-null  float64
 16  Marital             10000 non-null  object
 17  Gender              10000 non-null  object
 18  ReAdmis             10000 non-null  object
 19  VitD_levels         10000 non-null  float64
 20  Doc_visits          10000 non-null  int64
 21  Full_meals_eaten    10000 non-null  int64
 22  vitD_supp           10000 non-null  int64
 23  Soft_drink          10000 non-null  object
 24  Initial_admin       10000 non-null  object
 25  HighBlood           10000 non-null  object
 26  Stroke              10000 non-null  object
 27  Complication_risk   10000 non-null  object
 28  Overweight          10000 non-null  object
 29  Arthritis           10000 non-null  object
 30  Diabetes            10000 non-null  object
 31  Hyperlipidemia      10000 non-null  object
 32  BackPain            10000 non-null  object
 33  Anxiety             10000 non-null  object
 34  Allergic_rhinitis   10000 non-null  object
 35  Reflux_esophagitis  10000 non-null  object
 36  Asthma              10000 non-null  object
 37  Services            10000 non-null  object
 38  Initial_days        10000 non-null  float64
 39  TotalCharge         10000 non-null  float64
 40  Additional_charges  10000 non-null  float64
 41  Item1               10000 non-null  int64
 42  Item2               10000 non-null  int64
 43  Item3               10000 non-null  int64
 44  Item4               10000 non-null  int64
 45  Item5               10000 non-null  int64
 46  Item6               10000 non-null  int64
 47  Item7               10000 non-null  int64
 48  Item8               10000 non-null  int64
dtypes: float64(7), int64(15), object(27)
memory usage: 3.8+ MB
```

```
[38]: # input the two columns into a contingency table field to count the number of yes or no values that each field contains.
contingency_table = pd.crosstab(df['ReAdmis'], df['HighBlood'])
print(contingency_table)
```

```
HighBlood    No   Yes
ReAdmis
No         3747  2584
Yes        2163  1506
```

```
[40]: # Put contingency table values into a observed field, each row represents one of the variables.
# The following counts represent each yes or no count in the variables under observation.
observed = np.array(contingency_table)
print(observed)
```

```
[[3747 2584]
 [2163 1506]]
```

```
[42]: # Conduct Chi-Squared test using chi2_contingency founction from scipy.stats module
# This will calculate the chi-square statistic, compare it to the chi-square distribution, and compute the associated p-valu.e
chi2, p, dof, expected = chi2_contingency(observed)
```

B2, Dataset Analysis Output

The output of the analysis is shown below. The chi-squared test performed returned a chi-squared statistic of .04 and a p-value of 0.83. This is the probability that the observed distribution is due to chance. A low P-value typically <= 0.05 would indicate that there is a significant association between the variables. Due to the p-value being less than the significant threshold level we will reject this null hypothesis, indicating that the variables are independent from one another.

Although the P-value is the key observed distribution, the chi-squared analysis will also allow us to review the degrees of freedom and the expected frequencies. Degrees of freedom show the number of independent values or quantities which can be assigned to a statistical distribution. This is calculated in a chi-squared test based on the dimensions of the contingency table. Our degree of freedom is 1, indicating that there is a simple relationship between the variables. With degree of freedom being 1, the critical value is lower than for higher degrees of freedom, making it easier to detect a significant association if one exists.

```
[49]: # Interprete results using an if else statement.
alpha = 0.05
if p < alpha:
    print("There is a significant associate between the variables.")
else:
    print("There is no significant association between the variables.")
```

```
There is no significant association between the variables.
```

```
[51]: # Print out all the values between calculated for additional analysis.
print(chi2, p, dof, expected)
```

```
0.04239657973011679 0.8368656684578771 1 [[3741.621 2589.379]
 [2168.379 1500.621]]
```

```
[57]: # Print out all the values and expected frequencies between calculated for additional analysis.
# Examine expected frequencies
print(f"Chi-squared statistic:\n", chi2)
print(f"P-value:\n", p)
print(f"Degrees of freedom:\n", dof)
print("Expected frequencies:\n", expected)
```

```
Chi-squared statistic:
 0.04239657973011679
P-value:
 0.8368656684578771
Degrees of freedom:
 1
Expected frequencies:
 [[3741.621 2589.379]
 [2168.379 1500.621]]
```

B3, Dataset Analysis Explanation

The Chi-squared test was chosen among other statistical analysis options due to its simplicity and advantages when assessing categorical variables. A chi-squared test enables analysts to perform several important tasks on categorical variables. This includes determining whether two categorical variables are independent of each other and/or determining if an observed frequency distribution matches an expected distribution. One of the advantages of the Chi-squared test is that is does not require the assumption of normality in the data. Additionally, it is non-parametric, it does not assume a specific distribution for the population.

As you can see in the code attached to B1 above, the Chi-squared test uses the scipy package in python. This test was chosen to be the appropriate choice to use against two categorical variables to test for independence. The goal of the test was to determine if the variables in scope were independent of each other or not.

C, Univariate Exploration

When the chi-squared test does not show a significant association, it means there is no statistically significant relationship between the categorical variables. However, understanding the distribution of your continuous and categorical variables using statistics is essential for further analysis and insights.

Univariate analysis can be done on continuous and categorical variables. In this case, we will be performing it on both types of variables for further analysis. We will use the matplotlib.pyplot and seaborn packages in python. Univariate analysis for categorical variables involves using a frequency distribution or visualizations to summarize and visualize the frequency of each category. Frequency distribution counts the occurrences of each category. Visualizations use bar plots and pie charts to visualize the distribution.

In this analysis, visualizations have been deemed the best way to understand the distribution of each of the variables. An analysis has been done in python, see results and output of the code underneath C1 below.

C1, Visually Represented Findings

For the performed Univariate Exploration on Categorical Variables, the two categorical variables in scope are ReAdmis (hospital readmission indicator) and HighBlood (High blood pressure indicator).

The bar chart on the left is being used to display the frequency or proportion of patients with readmissions. When looking at this bar chart you can begin to understand the distribution of the variable by visually seeing which bar is higher. Each bar in the bar chart represents a category and the height of the bars correspond of the count or proportion of observations in that category. The x-axis represents the number of readmissions, and the y-axis represents the number of patients. This indicated that out of the total number of patients, there is a high proportion of patients who did not get readmitted vs did.

The pie chart on the right is being used to display the proportion of each category as a slice of the whole pie, each section is proportional to the category's overall frequency. When looking at this pie chart you can start to get a picture of how many patients were admitted that have high blood pressure. In this case, the chart is showing the percent of patients with high blood pressure within the overall count of patients. Each of the slices in this pie chart represent a category of the variable, with 40.9% of patients having high blood pressure and 59.1% not having high blood pressure, it can be observed that there is a higher proportion of patients without high blood pressure then with. However, it can also be observed that the % of patients with high blood pressure is very high in relation to the total number of patients.

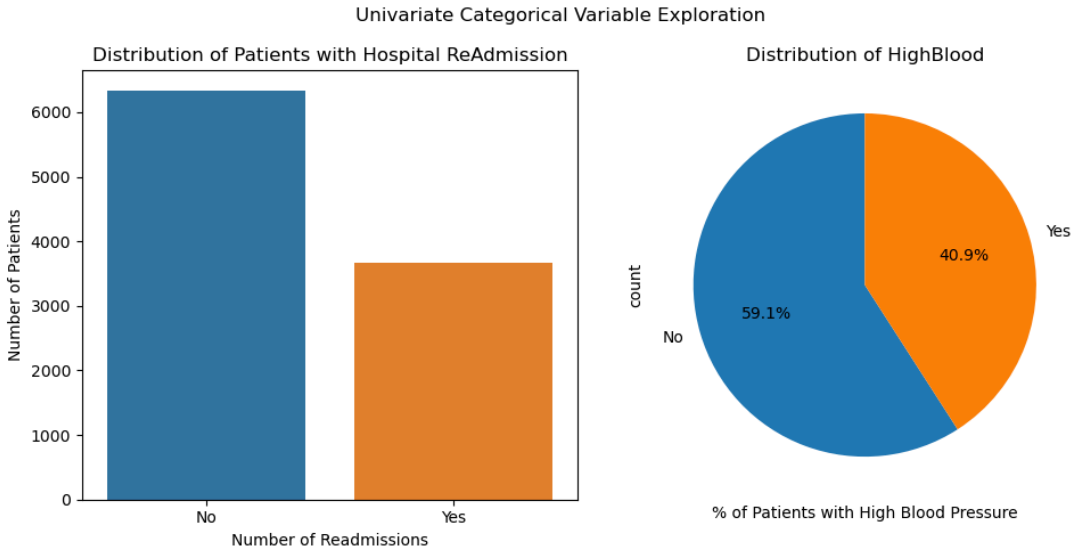
Both charts displayed below help identify patterns and trends in the data, even though neither of the categories in scope high a higher proportion of 'yes', the overall counts still could seem alarming and indicate potential areas for further investigation or intervention.

[275]:
```
# Univariate analysis for categorical variables. Requires the use of matplotlib.pyplot and seaborn packages.
# Use figure to create a new figure with a specified size to help the proportion and clarity of the plots.
fig, axes = plt.subplots(1, 2, figsize=(10, 5))
plt.suptitle("Univariate Categorical Variable Exploration")

# First plot: this divides the figure into a 1x2 grid and selects first subplot. Uses ReAdmis categorical variable in bar plot.
# Bar Chart Analysis on ReAdmis.
sns.countplot(ax=axes[0], x='ReAdmis', data=df)
axes[0].set_title('Distribution of Patients with Hospital ReAdmission')
axes[0].set_xlabel("Number of Readmissions")
axes[0].set_ylabel("Number of Patients");

# Second plot: this divides the figure into a 1x2 grid and selects the second subplot. This uses HighBlood in a pie chart.
# Pie Chart Analysis on HighBlood.
df["HighBlood"].value_counts().plot.pie(ax=axes[1], autopct='%1.1f%%', startangle=90)
axes[1].set_title('Distribution of HighBlood')
axes[1].set_xlabel("% of Patients with High Blood Pressure");

# Set layout for charts.
plt.tight_layout()
plt.show()
```



For the performed Univariate Exploration on Continuous Variables, the two continuous variables in scope are TotalCharge (total amount charged to a patient) and Age (patients age).

The histogram on the left is being used to display the distribution of the total amount charged to patients. When looking at this histogram you can begin to understand the frequency of

data points within specified intervals. The continuous variable is divided into intervals (or bins) and the height of each bar represents the number of observations in the bin. The x-axis represents the total amount charged to the patient, and the y-axis represents the frequency of that charge. This indicated that out of the total number of patients, there most patients fall into the $3000-$4000 range with a significant number of patients also falling into the $7000-$8000 range.

The box plot on the right is providing a summary of the distribution of the age of customers. When looking at this box plot you can start to get a picture what the age group ranges are of patients. In this case, the box plot is showing the quartile ranges of patient's age. Each quartile in the plot represents 25% of the data. It can be observed that the median age or central tendency is at the age of 52 with the middle 50% of patients aged between 36 and 70. The spread of ages is slightly wider below the median with 25% of patients falling between ages 18-52. On the other end of the median 25% of patients fall between ages 52-90. There are outliers that sit below the age of 20 and the at age of 90.

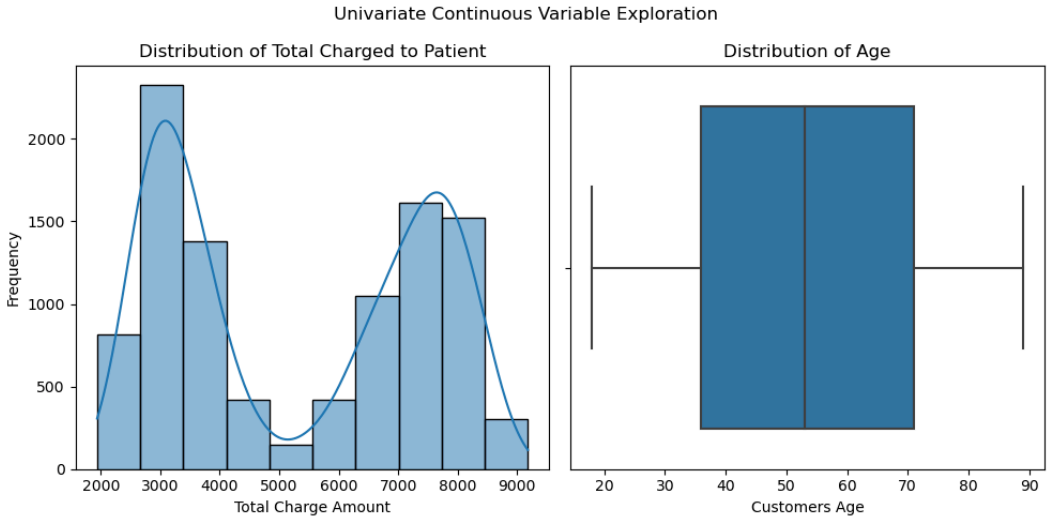
Both charts displayed below provide valuable insights into the distribution of continuous variables; by using these tools we were able to effectively explore and understand characteristics of these variables.

[291]:
```
# Univariate analysis for continuous variables. Requires the use of matplotlib.pyplot and seaborn packages.
fig, axes = plt.subplots(1, 2, figsize=(10, 5))
plt.suptitle("Univariate Continuous Variable Exploration")

# First plot: this divides the figure into a 1x2 grid and selects first subplot. Uses Total Charge continuous variable in histogram plot.
# Histogram Analysis on Total Charge.
sns.histplot(df['TotalCharge'], bins=10, kde=True, ax=axes[0])
axes[0].set_title('Distribution of Total Charged to Patient')
axes[0].set_xlabel("Total Charge Amount")
axes[0].set_ylabel("Frequency")

# Second plot: this divides the figure into a 1x2 grid and selects the second subplot. This uses Age continuous variable in a bot plot.
# Box Plot Analysis on Age.
sns.boxplot(x='Age', data=df, ax=axes[1])
axes[1].set_title('Distribution of Age')
axes[1].set_xlabel("Customers Age")

# Set layout for charts.
plt.tight_layout()
plt.show()
```



D, Bivariate Exploration

After univariate analysis is performed, an additional analysis can be performed to continue to address the variables at hand and the correlations between them. Bivariate analysis can be done on continuous and categorical variables. We will be performing bivariate analysis on all the same variables chosen for univariate, for further analysis.

Bivariate analysis is a statistical method that involves the simultaneous analysis of two variables to determine the empirical relationship between them. This type of analysis may help understand how variables interact or influence each other. This type of analysis uses scatter plots, correlation coefficients, or linear regression for continuous variables. For categorical variables, this analysis will use a contingency table or a stacked bar chart. Additionally, t-tests or chi-squared tests can be performed to calculate bivariate statistics.

Two categorical variables and two continuous variables were chosen for analysis to answer the hypothesis in scope. The two categorical variables in scope are ReAdmis (hospital readmission indicator) and HighBlood (High blood pressure indicator); the two continuous variables in scope are TotalCharge (total amount charged to a patient) and Age (patients age). A T-test was deemed adequate to identify bivariate statistics between the variables in scope. Additionally, a bivariate exploration was performed using these variables and is visually represented under D1 below. The output of the t-tests is shown below.

A t-test is a statistical analysis method used to determine whether there is a significant difference between the means of two groups or variables. The statistical value measures the size of the difference relative to the variation in the sample data. A high t-statistic suggests a significant difference between the two groups, whereas on the other hand a low t-statistic suggests that the sample means are very similar. The p-value represents the probability of observing such an extreme t-statistic under the null hypothesis, which usually states that there is no difference between groups. The output of the t-tests is shown below.

The t-test performed on ReAdmis and TotalCharge returned a statistical value of 157.17 and a p-value of 0.0. The t-statistic in this case is very high, indicating a large difference between the sample means in the data. The p-value of 0.0 indicates that the probability of observing such a large t-statistic by chance is low. These statistics help us identify that overall, the observed difference between the two variables in scope is highly significant. Thus, we can reject this null hypothesis, indicating the variables are independent from one another.

[193]:
```
# T-test performed for bivariate statistics - Bivariate exploration of Re-Admissions (categorical) & Total Charge Amounts (Continuous)
# Define test groups
group1 = df[df['ReAdmis'] == 'Yes']['TotalCharge']
group2 = df[df['ReAdmis'] == 'No']['TotalCharge']

# Perform t-test
statistic, p_value = ttest_ind(group1, group2)

# Print the results
print(f"t-test: Statistic={statistic}, p-value={p_value}")
```

```
t-test: Statistic=157.16875359158405, p-value=0.0
```

The t-test performed on HighBlood and Age returned a statistical value of .71 and a p-value of 0.47. The t-statistic in this case is close to 0, indicating that the difference between he means of the two variables in scope is small. The p-value of 0.47 means that there is 47.5% chance of observing such a difference if the null hypothesis is true. These statistics help us identify that

overall, the observed difference between the two variables in scope is not statistically significant. Since the p-value is higher than the common significant level (0.05) we can fail to reject the hypothesis indicating that there is not enough evidence to come to a conclusion on the statistically significant difference between the variables.

[195]:
```
# T-test performed for bivariate statistics - Bivariate exploration of High Blood Pressure (categorical) & Age (Continuous)
# Define test groups
group1 = df[df['HighBlood'] == 'Yes']['Age']
group2 = df[df['HighBlood'] == 'No']['Age']

# Perform t-test
statistic, p_value = ttest_ind(group1, group2)

# Print the results
print(f"t-test: Statistic={statistic}, p-value={p_value}")
```

```
t-test: Statistic=0.7146297644481312, p-value=0.47485453255808563
```

In this analysis, visualizations have been deemed the best way to understand the distribution of each of the variables. An analysis has been done in python, see results and output of the code underneath D1 below.

D1, Visually Represented Findings

For the performed Bivariate Exploration on continuous and categorical variables, the two categorical variables in scope are ReAdmis (hospital readmission indicator) and HighBlood (High blood pressure indicator); and the two continuous variables in scope are TotalCharge (total amount charged to a patient) and Age (patients age).

The violin plot on the left combines aspects of a box plot by providing a visualization of the distribution of the continuous variable for each category of a categorical variable. The x-axis represents the categorical re-admission categories of ‘yes’ or ‘no’ which the y-axis represents the continuous variable of total charged to the patient. The length of each category allows us to see the density of the continuous variable while the width of the violin shape is indicating the frequency of the data points at those values. Comparing the shapes of the violins for different categories can reveal differences in the distribution of the continuous variable across categories. Due to the overall shape of each category in the plot, it can be observed that there are a smaller number of patients who have been re-admitted vs not. It can also be observed that the patients who were re-admitted occurred much higher overall charges then those who were not re-admitted.

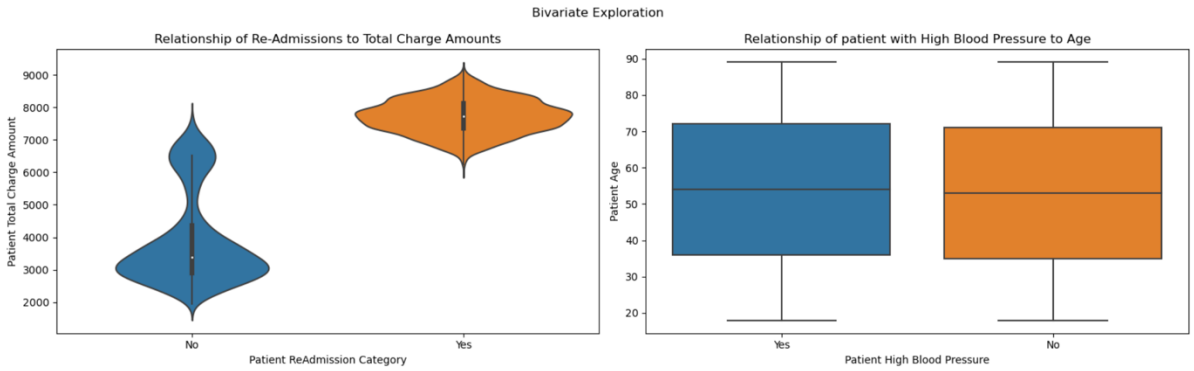
The box plot on the right is providing a summary of the distribution of the continuous variable for each category of a categorical variable. In this case, the x-axis is showing the patients with high blood pressure while the y-axis is showing the age of the patient. When looking at the box plot you can observe that there the interquartile range is higher in patients with high blood pressure then without. The median age for patients with high blood pressure appears to be around the age of 55 with the ICR being between 36-72. Whereas the median age for patients without high blood pressure appears to be around the age of 52 with the ICR being between 35-70. The age of most patients admitted in the hospital with high blood pressure appears to be slightly higher than the age of patients without high blood pressure.

```
[137]: # Bivariate analysis for one Categorical Variable and one Continuous Variable.
plt.figure(figsize = [16,5])
plt.suptitle("Bivariate Exploration")

# LEFT plot: Bivariate exploration of Re-Admissions (categorical) & Total Charge Amounts (continuous)
plt.subplot(1, 2, 1)
plt.title('Relationship of Re-Admissions to Total Charge Amounts')
##plot_order = ["Low", "Medium", "High"]
sns.violinplot(data= df, x= "ReAdmis", y= "TotalCharge")
plt.xlabel('Patient ReAdmission Category')
plt.ylabel("Patient Total Charge Amount")

# RIGHT plot: Bivariate exploration of High Blood Pressure (categorical) & Age (continuous)
plt.subplot(1, 2, 2)
plt.title("Relationship of patient with High Blood Pressure to Age")
sns.boxplot(data=df, x= "HighBlood", y = "Age")
plt.xlabel("Patient High Blood Pressure")
plt.ylabel("Patient Age")

# Set layout for charts.
plt.tight_layout()
plt.show()
```



E, Implications Summary

The below sections summarize the implications of the various analyses performed.

E1, Hypothesis Test

The hypothesis was to determine if hospital readmissions are related to high blood pressure. With the null being that there is no correlation in readmission from patients with high blood pressure against the alternative that there is a correlation. The chi-squared test yielded a chi-squared statistic of .04 with a p-value of .83 and a degree of freedom of 1, indicating that there is not a significant association between hospital readmissions and high blood pressure.

Additional analysis was performed to further enhance this study and explore this relationship further. Univariate statistics showed that 35% of patients were readmitted and 40.9% had high blood pressure. Bivariate analysis through a violin plot indicated that patients with higher total charges were more likely to be readmitted. A box plot revealed that patients with high blood pressure tended to be older on average. These findings suggest that both high blood pressure and age are important factors related to hospital readmissions.

E2, Limitations

There are always limitations when performing analysis on a dataset. Some limitations that could set back this analysis would be sample size or data quality, the sample size of 10,000 patients could potentially need to be expanded to adequately assess the various groups in scope. Another limitation is that a chi-squared test does not account for potential confounding variables that might influence the relationship between the variables. Time-periods and fluctuations in hospital readmissions would not be reviewed into this analysis which could also be a key factor when getting an overall picture of hospital readmissions. Additionally, chi-squared test can indicate

an associate between two variables however cannot establish causality, in other words you cannot assume causation if the two variables were to be correlated. And lastly, although some of the outputs of the univariate and bivariate analysis may have been expected, additional analysis must be performed to come to any conclusion.

E3, Recommended Course of Action

Given that this study indicated that there is no correlation between hospital readmissions and high blood pressure, there is no formal recommendation to make in regard to reducing readmission rates. However, a recommended course of action would be to conduct additional analyses the variables to determine any variables that might influence hospital readmissions, this study could continue to be expanded on to find the variables that do in fact influence this from occurring. Expanding the study to include different populations or timeframes could also be a beneficial enhancement to the study to validate the findings or ensure the results are adequate.

F, Panopto Video Recording Executions

The video recording for this assignment includes a vocalized demonstration of all the code, the code being executed, and the results of the code being represented inside this report.
The video recording for this project can be found inside the Panopto drop box titled "*Exploratory Data Analysis – OEM2 \ D207* "

Panopto video link: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4165fa55-bc9b-4c5d-8d4e-b1ae016bcdff

G, Web Sources

No sources, or segments of third-party code, were used to acquire data or to support the report.

H, Acknowledge Sources

I acknowledge that no sources, or segments of third-party sources, were stated or copied from the web into this report.