NOVEMBER 3, 2023

D207 EXPLORATORY DATA ANALYSIS

PERFORMANCE ASSESSMENT

ROBERT PATTON
WESTERN GOVERNORS' UNIVERSITY
Rpatt33@wgu.edu

Table of Contents

Part A: Real World Organizational Situation

1. **Question Relevant to Data Set**

   For this performance assessment, I chose to continue working with the medical data set from D206. My current career as a healthcare professional allows me to easily relate to the data and therefore produce a logical question for this assessment as such.

   Is there a statistical significance between patients who have had a stroke and also have high blood pressure, or are each of those variables independent of one another?

2. **Benefits to Stakeholders**

   Answering this question can help healthcare providers determine how to reduce the risk of strokes, with various interventions for high blood pressure, thus preventing their patients from having a stroke.

3. **Data Relevant to Question**

   This data set contains a total of fifty variables that provide various information for 10000 patient entries. The medical information provided covers patient demographics, medical conditions, admission information, regional information, population information and more. The data we are most focused on for this assessment will be two variables. Since we are trying to determine an outcome, if patients are more at risk for strokes, this will be our dependent variable. The stroke variable is a categorical value with yes or no inputs. Because we are looking at whether high blood pressure is a contributing factor to strokes, this will be our independent value. The high blood value is also a categorical data type with values of yes or no as well.

Part B: Data Analysis

1. **Technique**

   Because my research question involves two categorical variables, I will be using a chi square test to determine a correlation between the variable. The provided code for running a chi square test is as follows:

```
#Import python packages
        import pandas as pd
        import numpy as np
        from scipy.stats import chi2_contingency
#Read in the cleaned CSV file
```

```
        mdf=pd.read_csv('/Users/robertpatton/Desktop/D207/medical_clean.csv')
#Observe unique values for stroke
        mdf.Stroke.unique()
#Observe unique vales for High Blood
        mdf.HighBlood.unique()
#Get value counts for High Blood
        mdf.HighBlood.value_counts()
#Get value counts for Stroke
        mdf.Stroke.value_counts()
#Create data frame for categorical variables
        observed=np.array([[5910,4090],[8007,1993]])
#Run chi square test
        chi2, p, dof, expected=chi2_contingency(observed)
        print("p value is " + str(p))
        print("The degrees of freedom is: " + str(dof))
        print("The Chisquare statistic is: " + str(chi2))
        alpha= 0.05
        if p <= alpha:
        print("Dependent (Reject H0)")
        else:
        print("Independent (H0 holds True)")
```

2. **Code Output**

```
#run chi square test
chi2, p, dof, expected=chi2_contingency(observed)
print("p value is " + str(p))
print("The degrees of freedom is: " + str(dof))
print("The Chisquare statistic is: " + str(chi2))
alpha= 0.05
if p <= alpha:
    print("Dependent (Reject H0)")
else:
    print("Independent (H0 holds True)")
```

```
p value is 1.0462842473664897e-227
The degrees of freedom is: 1
The Chisquare statistic is: 1037.8846970102725
Dependent (Reject H0)
```

3. **Justification for Analysis**

For this performance assessment, we are required to choose from three different techniques for exploratory data analysis. A t-test, ANOVA test, or a chi-square test are the options for this project. Because my question involves two categorical variables, and determining a correlation between them, a chi-square test is the most appropriate of the three options.

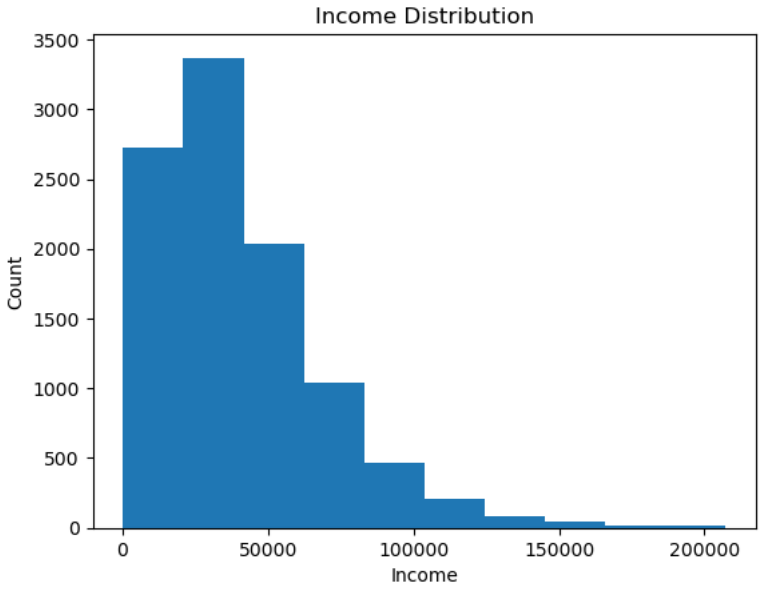
Part C: Univariate Statistics

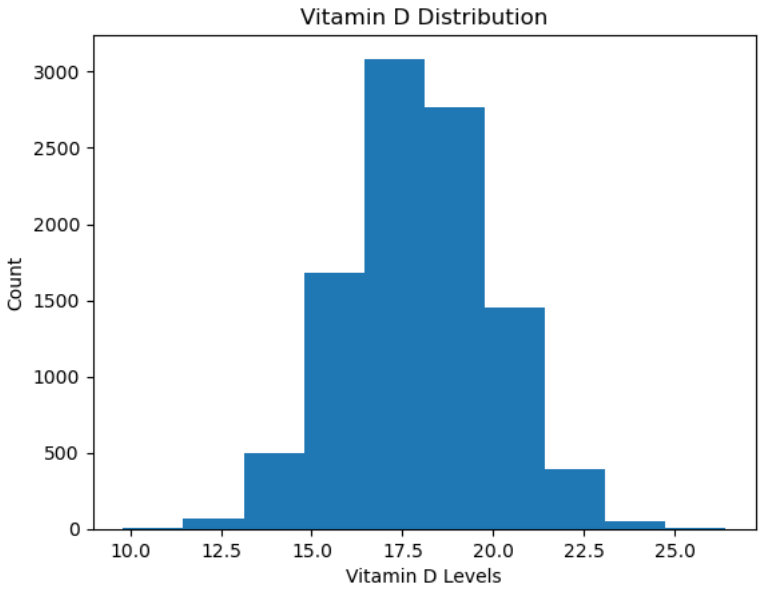
1. **Univariate Distributions**

For the univariate analysis, we are requested to choose two continuous variables and two categorical variables and create visualizations that represent their respective distributions. For the continuous variables I chose Income, which describes the income for each patient, and VitD_levels which describes the results for vitamin D levels after testing for each patient. For the categorical variables, I chose HighBlood, which describes if the patient has a history of high blood pressure, and Stroke, which describes if the patient has a history of strokes. Each of the categorical variables have results such as “yes” or “no”.

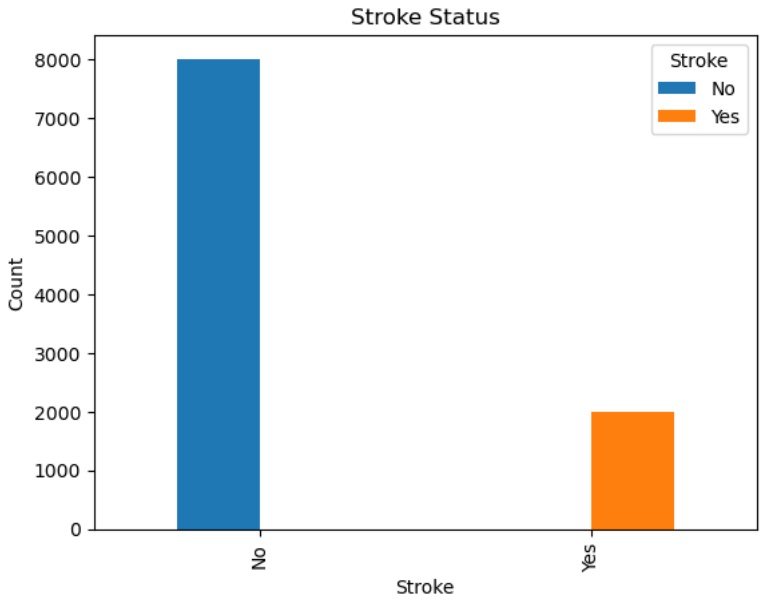
For the continuous variables, I chose to create histograms. The income variable presents a right skewed distribution while the VitD_levels variable presents a normal distribution. For the categorical variables, I chose to create bar charts for the visualizations that represent each category in a different color. Screen shots are attached for reference of each visualization tool.

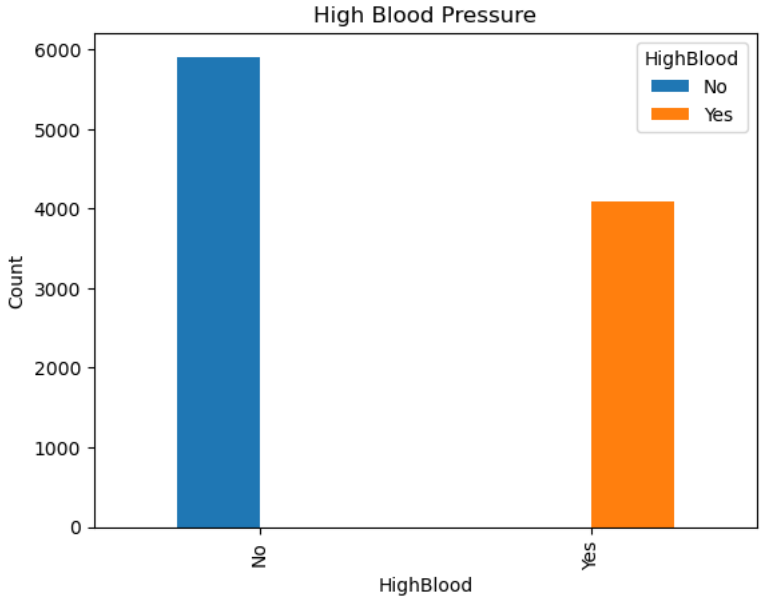
```
#Create histogram for continuous variables
    plt.hist(mdf.Income)
    plt.title('Income Distribution')
    plt.xlabel('Income')
    plt.ylabel('Count')
    plt.show()

    plt.hist(mdf.VitD_levels)
    plt.title('Vit D Levels Distribution')
    plt.xlabel('Vitamin D Levels')
    plt.ylabel('Count')
    plt.show()
```






```
#Create bar chart for categorical variables
crosstab=pd.crosstab(index=mdf['Stroke'], columns=mdf['Stroke'])
crosstab.plot(kind='bar')
plt.ylabel('Count')
plt.title('Stroke Status')
plt.show()

crosstab= pd.crosstab(index=mdf['HighBlood'], columns=mdf['HighBlood'])
crosstab.plot(kind='bar')
plt.ylabel('Count')
plt.title('High Blood Pressure')
plt.show()
```



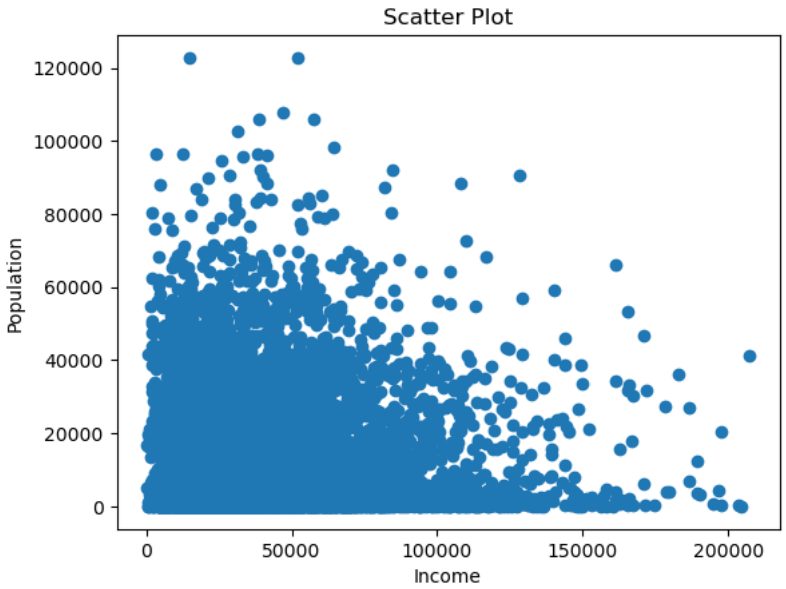


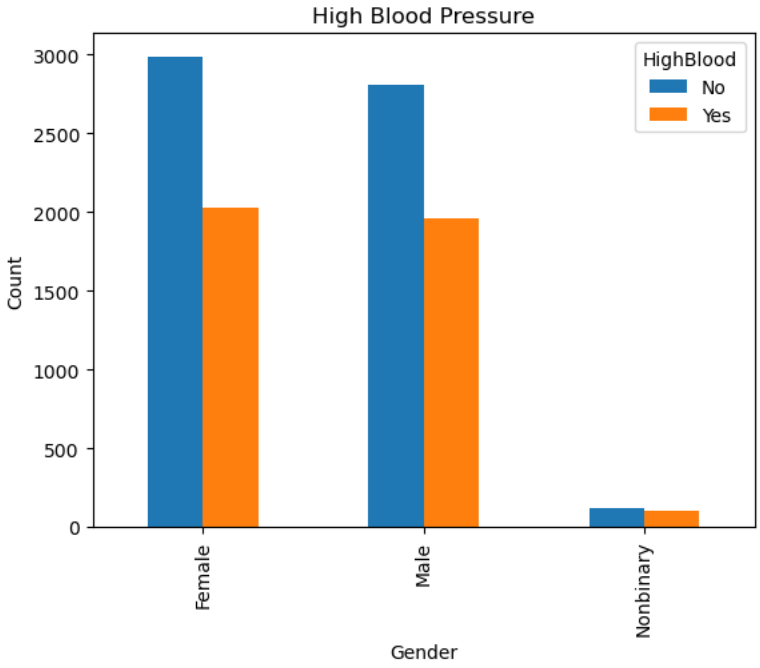
Part D: Bivariate Statistics

1. **Bivariate Distributions**

   The bivariate analysis requested the same instance of choosing two continuous and two categorical variables for analysis. In this scenario, I chose Income and Population as my continuous variables. For my categorical variables, I chose Gender and HighBlood. For my continuous variables, I used a scatterplot for analysis and compared the two against one another in the same scatterplot to visualize distributions.

For the categorical variables, I chose to use a bar graph just as I did for the univariate scenario. However, in this bivariate analysis, I took Gender and HighBlood variables and compared them in a bar graph to visualize how many people had responded with a “yes” or “no” to having high blood pressure, and how many of each gender did or did not have high blood pressure. Graphs are included for visualization and reference.





Part E: Summary of Data Analysis

1. **Results of Hypothesis Test**

   In this exploratory data analysis project, the goal was to ask a question about the data set that could potentially help a healthcare organization. The hypothesis here is that those with a history of a stroke, also have high blood pressure. Both stroke and high blood pressure variables are categorical ones, with stroke being our dependent variable and high blood pressure being our independent variable. So for the test, our H1 is that there is a statistical significance and our H0 is the null hypothesis, or that the variables have no association with one another. The goal is to determine whether we can accept our observed hypothesis and reject the null hypothesis or the opposite.

   Using a chi-square test, we can determine whether or not there is a statistical significance between these two categorical variables. There are two factors that are important in determining whether to accept the H1 or the H0. Those are an alpha value set to 0.05 and determining a p-value. After running my chi-square formula in python, my p-value came out to 1.0462842473664897e-227, well below the set alpha of 0.05. Therefore, it was confirmed that there is a statistical significance/dependency between the variables, and I could reject the null hypothesis.

2. **Limitations**

   I believe that a limitation to the data analysis here is that just because a person said yes to both questions, does not necessarily mean that the high blood pressure caused their stroke. There could be many other risk factors in a patient's life that contribute to the cause of strokes.

3. **Course of Action**

   With the information from this analysis, the hospital could use this data to reduce the risk of strokes in some of their patients with high blood pressure. The goal could be to implement more aggressive treatment measures for those with high blood pressure and educate patients about the risks upon discharge from the hospital. In addition, aside from focusing on just the patient, if there is an increase in the number of patients with strokes and high blood pressure in a specific population, such as the elderly, it could help hospitals become more prepared to handle these types of emergencies.

Part F: Panopto Video

The link to my Panopto video is attached to the submission of this task.

Part G: Third Party Sources

No third-party sources were used for this assessment.

Part H: Sources

No sources used.