**Performance Assessment: Exploratory Data Analysis (OEM2)**

Bader Ale

Department of Information Technology, Western Governors University

D207 Exploratory Data Analysis

Professor William Sewell

Table of Contents

[Part A: 3](#_Toc137294816)

[Research Question 3](#_Toc137294817)

[Question Importance 3](#_Toc137294818)

[Data Identification 3](#_Toc137294819)

[Part B: 3](#_Toc137294820)

[Technique Used 3](#_Toc137294821)

[Part C: 4](#_Toc137294822)

[Univariate Statistics 4](#_Toc137294823)

# Part A:

## Research Question

For this performance assessment, my research question is: Is there a significant difference in readmission based on the number of office visits?

## Question Importance

This research question is particularly valuable because it can help the hospital administration, investors and stakeholders determine if the amount of visits the physician makes to the patient affects the readmission rate. In this way, readmission rates can be decreased, potentially saving the facility money and resources.

## Data Identification

In order to answer the research question, the right variables have to be chosen. In this case *ReAdmis* and *Doc\_visits.* Per the Data Dictionary, *ReAdmis* is defined as whether or not the patient was readmitted back to the hospital within thirty days following their original discharge while *Doc\_visits*  is defined as the number of times the primary physician visited the patient in the hospital during their admittance.

# Part B:

## Technique Used

The technique chosen for this analysis was the t-test and more specifically a two-sample t-test. This technique was chosen because the research question concentrated on two variables, *ReAdmis* and *Doc\_visits; ReAdmis* is a categorical variable with two levels (Yes and No) while *Doc\_visits* is a discrete quantitative variable. Two groups will be further created – *ReAdmis\_yes* and *ReAdmis\_no*. The independent samples (two-sample) t-test will calculate the means of the two groups, *Readmis\_yes* and *Readmis\_no* and using the formula will generate both a t-statistic and a p-value.

A screenshot of a computer program

Description automatically generated with low confidence

Figure 1: Identifying t-test Variables.

A screenshot of the analysis output is shown below:

A picture containing text, font, line, white

Description automatically generated

Figure 2: t-test Output

# Part C:

## Univariate Statistics

For the univariate analysis section, we will be looking at variables *Income* and *VitD\_levels* for the continuous variables and *Gender*  and *Area* for our categorical variables.

Density plots were used for both continuous variables and histograms were used for the categorical variables, as shown below.

A picture containing text, screenshot, diagram, plot

Description automatically generated A picture containing diagram, text, plot, screenshot

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generatedA picture containing text, screenshot, rectangle, diagram

Description automatically generated

For the continuous variables, Income exhibited a positively skewed distribution while Vitamin D level was normally distributed. Moreover, Gender and Area were uniformly distributed.

## Bivariate Statistics

Next, bivariate analysis was performed; Income and total charge was compared and gender versus area. Below are the screenshots of the graphs.

A picture containing screenshot, text

Description automatically generated A picture containing text, screenshot, diagram, plot

Description automatically generated

We can see that there were more total charges below and above $5000 than any other quantity. Moreover, there were some who earned more than $150,000 per year that were charged less than the $5000 and conversely some who earned less than $50,000 per year who were charged around $9,000. For our gender versus area graph, we can see that overall, there were more females than males hospitalized. The code below shows the value counts for the gender variable:

A screenshot of a computer

Description automatically generated with medium confidence

# Part E:

## Discussion

After running the t-test, we can see that our p-value and t-statistic are 0.9804363 and 0.02452273 respectively. We can safely conclude that, because of the results shown, there is no statistical significance between the number of visits from a doctor during a patient’s hospitalization and the rate of readmission thirty days following discharge. We can therefore accept the null hypothesis which states there is no significant statistical significance between the number of doctor visits and the rate of readmission.

There are limitations to our data. For one, the data cannot account for the various other reasons a patient might be readmitted; for example, when the patient leaves the hospital once cannot know for certain if the individual followed all the instructions given to them in order to not be readmitted and to get better. Moreover, we cannot know for sure if the individual was readmitted if it was for the same reason.

All things considered, it is recommended not to pursue further action regarding doctors visiting the patients, more than they are, as the analysis showed that it had no statistical significance in readmission rates. Further analysis should be performed in order to determine the correlation between other variables and the readmission rates. In other words, no action needs to be taken regarding doctor visits as it had no impact on readmission.