**PERFORMANCE ASSESSMENT:**

**D207 – EXPLORATORY DATA ANALYSIS**

Michael Lawson

Western Governors University

D207 – Exploratory Data Analysis

David Gagner

12/30/21

Michael C. Lawson | Student ID: #001479557 | Program: MS Data Analytics (5/1/21) |

Program Mentor: Emil Stoica | 859-265-1888 (EST) | [mlaws77@wgu.edu](mailto:mlaws77@wgu.edu)

1. **Describe one real-world organizational situation in the data.**
2. **Provide a question relevant to the chosen data set:**

Does the variable “age” influence whether a patient will be readmitted to the hospital in the “medical\_clean.csv” data set?

1. **Explain how stakeholders could benefit from the analysis of the data**.

Finding the variables that are significant to patient readmission could help formulate predictions and help the stakeholders make the correct changes to reduce readmits. The hospital will save money if fines for readmitted patients are avoided. The savings would allow the money saved to be used in more beneficial areas of the hospital like research or financial assistance for patients.

1. **Identify the data in your data set relevant to answering the question.**

The nominal categorical variable “ReAdmis” is the dependent variable for the question because it splits the data between readmitted patients and first-time patients using two categories, “Yes” and “No”. The numeric ratio variable “Age” is also relevant to answering the question.

1. **Describe the data analysis:**
2. **Write the code to run the analysis in Python.**

*#import full packages or separate tools needed from packages*

**import** numpy **as** np

**import** pandas **as** pd

**import** statistics

**import** matplotlib

**import** matplotlib.pyplot **as** plt

**import** scipy.stats **as** stats

**import** seaborn **as** sns

*#Import data from hard drive*

med **=** pd**.**read\_csv(r"C:\Users\mlaws\OneDrive - Western Governors University\Documents\WGU\D207\medical\_clean.csv", skiprows**=**0, delimiter**=**",")

*#Set random seed*

np**.**random**.**seed(123)

*#Sample size*

N**=**25

*#Take a sample of 25 readmitted patients and 25 inital visit patients from the data set*

Yes **=** med[med['ReAdmis']**==**'Yes']**.**sample(25)

No **=** med[med['ReAdmis']**==**'No']**.**sample(25)

*#Separate the "Age" variable from the samples*

a **=** Yes['Age']

b **=** No['Age']

*#Histogram to visualize data before running t-test*

sns**.**histplot(data**=**Yes, x**=**'Age', kde**=True**, color**=**"Green")

sns**.**histplot(data**=**No, x**=**'Age', kde**=True**, color**=**"Purple")

*#Boxplots to visualize data before running t-test*

mb **=** sns**.**boxplot(x**=**'Age', y**=**'ReAdmis', data**=**Yes, color**=**'#99c2a2')

mb **=** sns**.**swarmplot(x**=**'Age', y**=**'ReAdmis', data**=**Yes, color**=**'#7d0013')

plt**.**show()

mb **=** sns**.**boxplot(x**=**'Age', y**=**'ReAdmis', data**=**No, color**=**'Tan')

mb **=** sns**.**swarmplot(x**=**'Age', y**=**'ReAdmis', data**=**No, color**=**'Red')

plt**.**show()

*#Calculate variance to be used for standard deviation*

var\_a **=** statistics**.**variance(a)

var\_b **=** statistics**.**variance(b)

*#Standard Deviation*

s **=** np**.**sqrt((var\_a**+**var\_b)**/**2)

*#t-statistics*

t **=** (statistics**.**mean(b)**-**statistics**.**mean(a))**/**(s**\***np**.**sqrt(2**/**N))

*#Degrees of Freedom*

df **=** (2**\***N)**-**2

print(df)

*#P-Value*

p **=** 1**-**stats**.**t**.**cdf(t,df**=**df)

print("t= "**+**str(t))

print("p= "**+**str(2**\***p))

1. **Provide the output and results of calculations from the analysis performed.**

T-test results:

T-Value: 0.11022014988113599

P-Value: 0.9126943347750718

Boxplot and Histogram to visualize data sample:

Chart, histogram

Description automatically generated Chart, box and whisker chart

Description automatically generated

**H0:** Age does not influence readmission rates.

**H1: “**Age” does have an influence on “ReAdmis”

The P-Value of 0.91 is much larger than the alpha limit 0.05. The high p-value and low t-value reveal the analysis failed to reject the null hypothesis that age does not significantly influence readmission rates.

1. **Justify the analysis technique.**

An independent sample two-tail t-test was chosen because a nominal categorical variable, “ReAdmis”, was tested against a numeric ratio variable, “Age”. Two sample sets of 25 patients were created to find the mean age of the category ‘Yes’ and the mean age of the category ‘No’ in the column labeled “ReAdmis”. The two sample means were compared, and the data is structured independently. The “Decision Tree” in the “Episode 4 PowerPoint” created by Dr. Sewell was used to find that an independent sample t-test was the appropriate test given these observations.

1. **Identify the univariate distribution of two continuous variables. Identify the univariate distribution of two categorical.**
2. **Represent the findings in Part C, visually.**

**Continuous Variables (Univariate):**

1. “VitD\_levels”

Chart, histogram

Description automatically generated

1. “Additional\_charges”

Chart, histogram

Description automatically generated

**Categorical Variables (Univariate):**

1. “Gender” Chart, histogram

   Description automatically generated
2. “Services” Chart, histogram

   Description automatically generated

1. **Identify the bivariate distribution of two continuous variables. Identify the bivariate distribution of two categorical variables.**
2. **Represent the findings in Part D, visually.**

**Continuous Variables (Bivariate)**

“Initial\_days”, “TotalCharge”

**Chart, scatter chart

Description automatically generated**

**Categorical Variables (Bivariate)**

“Marital”, “Gender”

**Chart, histogram

Description automatically generated**

1. **Summarize the implications of the analysis by:**
2. **Discuss the results of the hypothesis test.**

The alternative hypothesis was that “Age” influences the “ReAdmis” variable; conversely, the null hypothesis was that “Age” has no effect on “ReAdmis”. A sample of 25 patients in the category “Yes of the “ReAdmis” column was taken from the population. A sample of 25 patients in the category “No” of the “ReAdmis” column was also taken. Each sample has 24 degrees of freedom for a total of 48 degrees of freedom. A t-test table was used to find that the t-critical value is 2.943 at an alpha level of 0.05 for a two tailed t-test (MedCalc Software Ltd).

A bivariate boxplot and histogram were created to visualize the distribution of the data. The boxplot and histogram gave no clear difference in the variables. The t-value was 0.1102 which indicates that the data is in the normal range because it falls below the t-critical value of 2.943. The p-Value was 0.9127. The p-value suggests that the samples are well within normal; therefore**, the analysis has failed to reject the null-hypothesis.** Age does not have a significant effect on readmission.

1. **Discuss the limitations of your data analysis.**

Only 50 patients were selected from 10,000 and tested, so there is a chance that the sample does not accurately represent the entire dataset. It would be better to take several samples. T-tests are also limited to data with a normal distribution. There is a possibility the random 50 patients chosen just happened to be similar, and a larger sample size or more samples would reveal different results.

1. **Recommend a course of action based on the results.**

The hospital should focus on variables other than “Age” to predict if a patient will be readmitted. Based on these results, the hospital should allow for more analysis to find if any other variables do influence the rate of patient readmission. The hospital can safely assume that a patient’s age will not have an effect on whether they are readmitted.

1. **Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the tool(s) used.**

[*https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=91804c2b-b265-401e-8a76-ae0e01782dc0*](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=91804c2b-b265-401e-8a76-ae0e01782dc0)

**References**

1. **Reference web sources used to acquire segments of third-party code to support the analysis.**

*Python: t-test, one-way ANOVA, categorical features*. Mark Keith (November 9, 2020) <https://www.youtube.com/watch?v=u3Hwt_jbbTc&list=PLe9UEU4oeAuV7RtCbL76hca5ELO_IELk4&index=6>

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

*D207 Exploratory Data Analysis Webinar, Episode 4*. Dr. William Sewell.

<https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/EUq-_jd08c5LtMHuOhkmEPUBWlr9UjsaTE9IRIo2q3KXLw?e=4S4Kw0>

*D207 Exploratory Data Analysis*. Datacamp. <https://app.datacamp.com/learn/custom-tracks/custom-d207-exploratory-data-analysis>

1. **Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.**

*T-distribution Table (two tailed)*. MedCalc Software Ltd. <https://www.medcalc.org/manual/t-distribution-table.php>