Exploratory Data Analysis (OEM2) – D207

Jeffrey Williams

Western Governors University

Student #: 001173968

Date: 02/03/2023

Mentor: Jared Knepp

**Table of Contents**

**Part A1.** Question For Analysis3

**Part A2.** Benefit From Analysis3

**Part A3.** Data Identification3

**Part B1.** Code**3-8**

**Part B2.** Output3-8

**Part B3.** Justification9

**Part C.** Univariate Statistics**9**

**Part C1.** Visual of Findings9-10

**Part D.** Bivariate Statistics10

**Part D1.** Visual of Findings10

**Part E1.** Results of Analysis**11**

**Part E2.** Limitations of Analysis11

**Part E3.** Recommended Course of Action11

**Part F.** Video11

**Part** **G& H.** Sources for Third-Party Code & Paper Sources…………………………………………………**12**

**Part A1. Question for Analysis**

Out of the three types of community (urban, rural, and suburban), do any of the three charge more and do patients spend more time during their visit based off time spent in the hospital and if any charges vary.

**Part A2. Benefit From Analysis**

Stakeholders profit from hospital patient stays and the organization total profit. By focusing on each region, the stakeholders will have a better logistical understanding of which community needs profits more opposed to the less profiting regions. This analysis will provide additional insight on which communities have longer patient stays and provide metrics for total incidents and length of stay in relation to medical problems.

**Part A3. Data Identification**

To answer the question I will need to examine the “Area” column to identify where the patient are from (urban, rural, or suburban). I will also be examining “TotalCharge”, “Additional\_charges”, and “Initial\_days in order to determine the total time patients spent in the hospital along with how much the patients were charged.

**Part B1. Code & B2. Output**

To determine if any of the variables differentiate between Urban, Rural, or Suburban, I will be using the “t-test” technique. I will need to pair up the regions so I can analyze both the p and v values. - *The statistical analysis t-test explained for beginners and experts: (2020, April 11)*

Listed below is both the results and code of both the calculations and analysis conducted by utilizing the “t-test” technique:

Text

Description automatically generated with medium confidence

The technique utilized indicates no significant difference between the three different regions analyzed. The test resulted that out of the three, only “Urban/Rural” differentiates with a p-value of 0.258 and a t-value of 1.13. With “Urban/Rural” having a p-value of 0.258, indicates no real difference between the three regions. “Suburban/Urban” had a t-value of 0.55 with a p-value of 0.58 and finally the “Rural/Suburban” had a t-value of 0.58 with a p-value of 0.56.

**Complete Code in order that it was utilized: -** *Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley*

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Part B3. Justification**

I eliminated both the “ANOVA” and “chi-square” techniques because my focus was on continuous variables and categorical (total cost and regions of where the patients reside). The “chi-square” would need two categories to conduct an analysis (which I did not have) and “ANOVA” determines variance of a data set by using statistics to analyze the means of more than two groups, both techniques would not have answered the question I posed in the same way the “T-Test” did. - *P-Value, T-test, Chi-Square test, ANOVA, When to use Which Strategy (2020, March 17)*

**Part C. Univariate Statistics**

The continuous variables chosen to conduct my analysis was “Number of days in the hospital” and “Number of days in the hospital”. The two categorical variables are “Complication risk” and “Area”. - *Histograms: (2022, November 27)*

**Part C1. Visual of Findings**

I conducted a univariate analysis using a histogram (displayed below). With the histogram tool, I evaluated all the variables against themselves. - *Histograms: (2022, November 27)*

Chart

Description automatically generated

* The first “Area” histogram displayed above shows no deviation between urban, rural, and suburban.
* The second “Complication\_risk” histogram shows little deviation between the three Areas.
* The third “Initial\_days” histogram indicates the patients spend both very little time and a lot of time in the hospital based the inversed bell curve shown.
* The fourth “Total\_cost” histogram shows 1 and a half bell curves, around $14500 and $45000.

**Part D. Bivariate Statistics & D1. Visual of Findings**

The two continuous variables chosen for the distribution are “Number of days in the hospital” and “Number of days in the hospital”. The two categorical variables I used for bivariate statistics are “Complication risk” and “Area”. Using the heatmap visual tool I can show the relationship between each of the variables. - *Overview of Multivariate Analysis | what is Multivariate Analysis and Model Building Process (2022, December 02)*

Chart

Description automatically generated

By using the heatmap visual tool shown above, I compared relationships between all the factors. The closer the number is to 1, the more related the 2 factors are. The only related factors are with “Total\_cost” and “Initial\_days”. This indicates a rational linear relationship because it is common knowledge the more days spent in the hospital for care will always cost more for the patients. *-All About Heatmaps (2020, December 24)*

**Part E1. Results of Analysis**

Determining if the three different community regions “Urban”, “Suburban”, and “Rural” patients show any significant difference between them and the charges paid was the original hypothesis. With the full analysis performed, it was determined that no significant variance between the three community regions in terms of cost. The R-values came back as 0.58, 0.56, and 0.26, concluded no significant variation for both area of residency and the costs received. With the values so far off concluded that the null hypothesis was accurate because aspects of cost and area appear to be completely unrelated to each other. In order for the alternative hypothesis to be correct, we would see variance between the area of residency and cost sustained at the hospital. Since the null hypothesis is correct, we known the alternative hypothesis in incorrect.

**Part E2. Limitations of Analysis**

Limitations that occurred during my analysis was adding all the costs together in order to have a total amount of costs charged. By doing so, my examination of the question ignores other possible discrepancies. If I perhaps focused on and compared “Urban”, “Rural”, and “Suburban” and just the “additional charges” or the “total amount of time” in the hospital, I could have possibly found some sort of discrepancy, but the analysis ignores the categories independently and focuses on everything in its totality. The “t-test” used shows that the null hypothesis was correct because it displays only a relationship between initial days and total cost, but because both the P and R-values were so far out concluded a low level of confidence with this result. My hypothesis shows limitation with the bivariate analysis between total costs, initial days, complication risk, and area .If the chi-square test technique was used instead of the T-test, it may have sorted out inconsistencies that are present in the alternative hypothesis.

**Part E3. Recommended Course of Action**

A course of action I would recommend is to focus on breaking down the total coast by each variable that equates the total amount value. By focusing on the length of stay for the patient, the cost per day, and additional charges autonomously to figure out if the average for the total cost would be the same because of those independent aspects or if no discrepancy for “Areas” on any of the variables that make up “total\_cost”.

**Part F. Video**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=07232f36-8845-4740-8bfa-afb40055acd7>

**G-Sources for Third-Party Code & H-Sources**

*The statistical analysis t-test explained for beginners and experts (2020, April 11)*

*https://towardsdatascience.com/the-statistical-analysis-t-test-explained-for-beginners-and-experts-fd0e358bbb62*

*P-Value, T-test, Chi-Square test, ANOVA, when to use Which Strategy (2020, March 17)*

*https://medium.datadriveninvestor.com/p-value-t-test-chi-square-test-anova-when-to-use-which-strategy-32907734aa0e*

*Histograms: (2022, November 27)*

*https://corporatefinanceinstitute.com/resources/excel/histogram/*

*Overview of Multivariate Analysis | what is Multivariate Analysis and Model Building Process (2022, December 02)*

*https://www.mygreatlearning.com/blog/introduction-to-multivariate-analysis/*

*All About Heatmaps (2020, December 24)*

*https://towardsdatascience.com/all-about-heatmaps-bb7d97f099d7*

*ColSums: Form Row and Column Sums and Means (2019, December 31)*

*https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/colSums*

*{Bibliography}*

*Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley*