**Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:**

1. Provide **one** question that is relevant to your chosen data set. You will answer this question later in the task through an analysis of the cleaned data, using one of the following techniques: chi-square, t-test, or analysis of variance (ANOVA).

What variable is the most important predictor of customer churn?

My hypothesis is that higher monthly charge will have the greatest impact on the customer churn.

2. Explain how stakeholders in the organization could benefit from an analysis of the data.

Stakeholders would value this data as it would give them insight into the specific features they should look into to reduce the overall customer churn.

3. Identify *all* of the data in your data set that are relevant to answering your question in part

The variables that are relevant to answering my question in this dataset are:

* Churn: yes or no variable showing if the customer decided to continue or discontinue their service
* Tenure: the length of time in months that the customer has stayed with the provider
* Outage\_sec\_perweek: the amount of time that an outage occurs in a customers neighborhood, counted in seconds
* Yearly\_Equip\_Failure: the number of times that a customer has had an issue or had to have their equipment replaced in the last year
* MonthlyCharge: The average amount that a customer is charged monthly by their provider

Bandwidth\_GB\_Year: The average amount of data that is used by a customer per year, counted in Gigabytes

Also the survey results are relevant to the question above:

* Item1: Responding to customer issues in a timely manner
* Item2: Fixing issues in a timely manner
* Item3: Timely replacements for damaged equipment
* Item4: Reliability with solving issues
* Item5: Plentiful Options
* Item6: Respectful responses to customer complaints
* Item7: Courteous exchange between customer and employee
* Item8: Evidence of active listening

To answer the question I am going to use the Monthly Charge, Churn, Tenure, and timely response variables.

**B. Describe the data analysis by doing the following:**

1. Using one of the following techniques, write code (in either Python or R) to run the analysis of the data set:

• chi-square

• t-test

• ANOVA

2. Provide the output and the results of *any* calculations from the analysis you performed.

Shown in attached jupyter notebook

3. Justify why you chose this analysis technique.

I chose this analysis technique mainly due to the variables that I ended up focusing on being numerical. I know it would be useful to compare the two groups of data that I ended up creating, the yes’s vs the no’s in relation to the monthly charge amount. I wanted to see if the results that were received could have happened by chance.

C.  **Identify the distribution of two continuous variables and two categorical variables using univariate statistics from your cleaned and prepared data.**

1. Represent your findings in Part C, visually as part of your submission.

Shown in the attached jupyter notebook

*Note: To draw a graph or visualization, you may use one or a combination of the following:*

*- A spreadsheet program, such as Excel (\*.xls)*

*- A graphics program, such as Paint (\*.jpeg, \*.gif)*

*- A word-processing program, such as Word (\*.rtf)*

*- A scanned hand-drawn graph (\*.jpeg, \*.gif)*

**D. Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from your cleaned and prepared data.**

1. Represent your findings in Part D, visually as part of your submission.

Shown in attached jupyter notebook

*Note: To draw a graph or visualization, you may use one or a combination of the following:*

*- A spreadsheet program, such as Excel (\*.xls)*

*- A graphics program, such as Paint (\*.jpeg, \*.gif)*

*- A word-processing program, such as Word (\*.rtf)*

*- A scanned hand-drawn graph (\*.jpeg, \*.gif)*

**E. Summarize the implications of your data analysis by doing the following:**

1. Discuss the results of the hypothesis test.

The data shows that there is a pretty strong correlation between monthly charge and Churn with higher monthly charge values tending to lead towards increased churn. Looking at the T-statistic (40.3) lets us know that the difference between the churn and monthly charge isn’t necessarily due to chance. The P-value (0.0) shows that the difference is statistically significant.

2. Discuss the limitations of your data analysis.

- There are factors we cannot consider like the location that this study took place. Internet companies can reach a wide range of places and costs + factors can be different as places vary.

- This dataset is a frozen picture in time, and it was not specified (unless I missed it) the timeframe in which this was recorded. Different lengths of time can yield many different results.

3. Recommend a course of action based on your results.

I would suggest a deep dive into the specific things that affect the monthly charge. I know that the last thing a company wants is to reduce their costs too much but there could be some other factors within or alongside the monthly charge that further skews the customers to churn away from the company. Altering those could help reduce a good portion of the churn.

**F. 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.**

*Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.*

*To submit your recording, upload it to the Panopto drop box titled “Exploratory Data Analysis – OEM2 \ D207.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.*

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

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

GfG. “Python - Seaborn.Pairplot() Method.” *GeeksforGeeks*, 11 Nov. 2022, www.geeksforgeeks.org/python-seaborn-pairplot-method/. Accessed 24 Mar. 2024.

“T-Test, ANOVA and Chi Squared Test Made Easy.” *YouTube*, 5 Apr. 2022, www.youtube.com/watch?v=ijeEYFnS2v4&t=16s. Accessed 03 Apr. 2024.

**I. Demonstrate professional communication in the content and presentation of your submission.**