A.  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).

Which of the customers of the telecommunications company is at a high risk of churn?

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

Stakeholders will benefit from this analysis by being able to understand and identify customers that are churning. Furthermore, stakeholders will then be able to come up with a plan, marketing, or changing initial service that will better benefit the customer and reduce churn rates.

3.  Identify *all* of the data in your data set that are relevant to answering your question in part A1

The Churn column, which is a binary categorical column, that involves the inputs of “Yes” and “No”, is possibly the most relevant data in the set. Income, Tenure, Monthly Charge, Bandwidth\_GB\_Year columns which are continuous numerical data, are also relevant information that can be used to answer question in A1. Furthermore, 8 survey questions which all contain discrete numerical data, and provides relevance to answering part A1. The survey question provides the ordinal numerical data that the customers use to rate from 1 – 8, where 1 is most important and 8 is the least important in the survey questions.

* Churn: Whether a customer discontinued service within the last month
* Income: Annual income of the customer reported at the time of sign up
* Tenure: Number of months the customer has stayed with provider
* Monthly Charge: The amount charged to the customer monthly
* Bandwidth\_GB\_Year: Average amount of data used in a year by a customer
* Item1: Timely response
* Item2: Timely fixes
* Item3: Timely replacements
* Item4: Reliability
* Item5: Options
* Item6: Respectful response
* Item7: Courteous exchange
* Item8 Evidence of active listening

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

Python – Chi-Square

**CODE:**

**Text

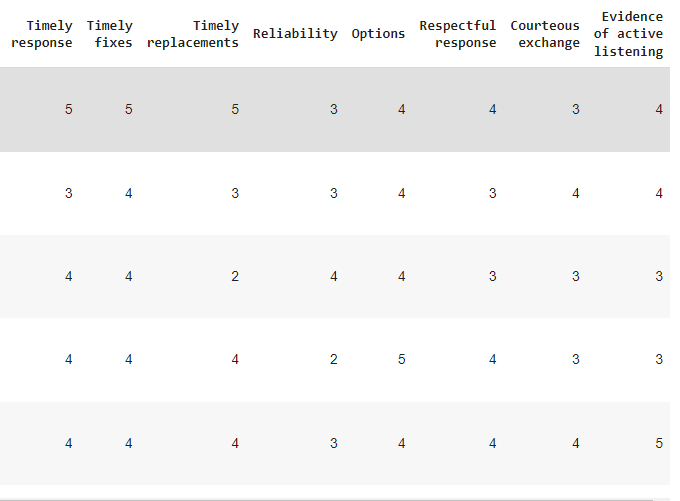
Description automatically generatedGraphical user interface, text, application

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidenceA picture containing text, screenshot, indoor, several

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedText

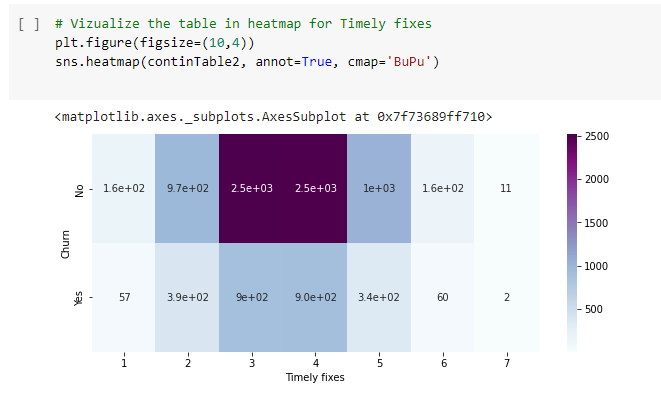
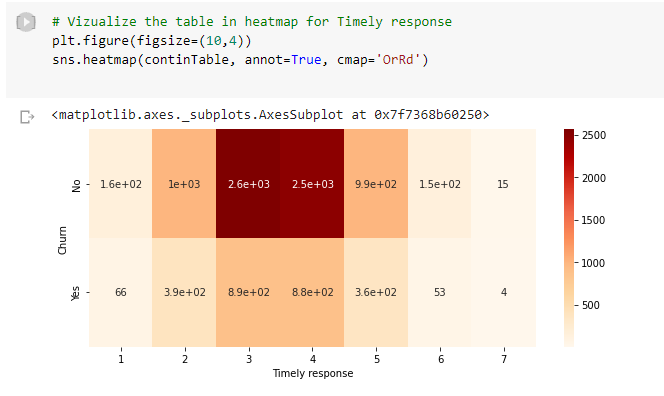
Description automatically generatedA picture containing company name

Description automatically generatedTable

Description automatically generatedGraphical user interface, application

Description automatically generatedA picture containing text

Description automatically generatedA picture containing calendar

Description automatically generated**

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

Graphical user interface, text, application

Description automatically generated

3.  Justify why you chose this analysis technique.

The reason I chose chi-square as an analysis technique to answer the question, “which of the customers of the telecommunications company is at a high risk of churn?”, was because we can look at two categorical variables that appear to be related. Also, will be able to understand the relationship between them. Furthermore, as long a chi-square meets two assumptions, we can use it for testing.

The two assumptions are:

1. “Your two variables should be measured at an ordinal or nominal level (i.e., categorical data).” (“Your two variables should be measured”, September 23, 2021)
2. “Your two variables should consist of two or more categorical, independent group. Example independent variables that meet this criterion include gender (2 groups: Males and Females), ethnicity (e.g., 3 groups: Caucasian, African American and Hispanic), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth.” (“Your two variables should consist of”, September 23, 2021)

Categorical variables:

* Churn: binary variable/ binomial categorical dependent variable “yes/no”
* Timely fixes: Ordinal level variable

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

Two continuous variables:

* Income
* MonthlyCharge

Two categorical variables:

* Timely response: Item1
* Timely fixes: Item2

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

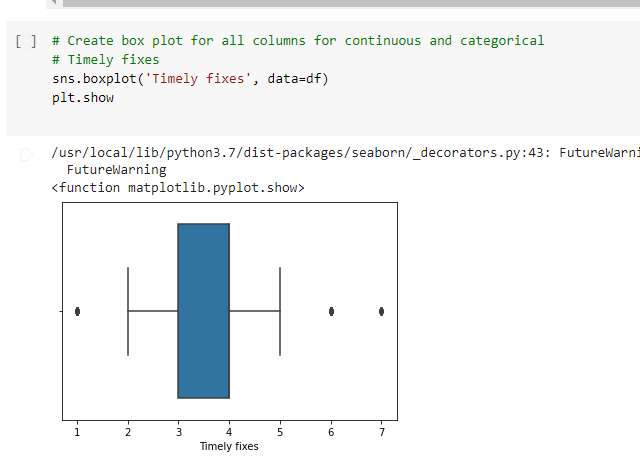
Text

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart

Description automatically generatedChart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generated

*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.

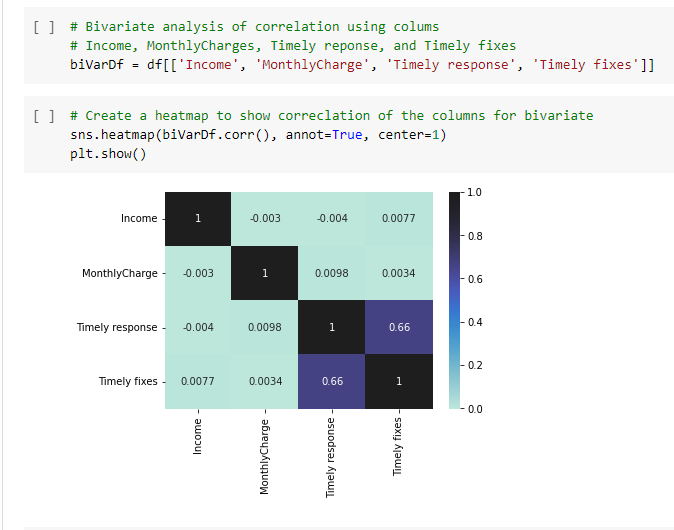
Two continuous variables:

* Income
* MonthlyCharge

Two categorical variables:

* Churn
* Timely fixes: Item2

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

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart

Description automatically generated with medium confidence

*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.

As I look at the high p-value of “0.6318335816054494”, this value shows how likely the data will occur by random chance, for example, that null hypothesis is true. This allows us to know if we should reject the null hypothesis if p-value is less than 0.05, because the smaller the p-value the more statistically significant and stronger the evidence to reject the null hypothesis. But in this case, we are greater than the 0.05, so it is not statistically significant and indicates that there is a strong evidence of null hypothesis. So, we must retain that null and reject the alternative hypothesis.

Text

Description automatically generated

1. Discuss the limitations of your data analysis.

We can clearly see a high p-value of “0.6318335816054494”, which means that we should look further into the data. Also, we may need to gather better and more significant data. Furthermore, we could investigate different columns and try for better results for testing. Otherwise, we will continue to struggle with proper information, because of the limited dataset and the gathering of meaningful data.

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

The recommended course of action is to consider the survey responses at a greater length to help resolve customer churn. Improving these areas will create a better experience with the customer and better results with the company.

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.

Anon, Chi-square test for association using SPSS statistics. *Chi-Square Test for Association using SPSS Statistics - Procedure, assumptions and reporting the output*. Available at: [Accessed September 23, 2021].

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