D207: Exploratory Data Analysis Task 1

Justin Huynh

Student ID: 012229514

M.S. Data Analytics

A1. Question for Analysis

My research question for this project is “Does a relationship exist between people who were never married and those with a longer contract?”

A2. Benefit from Analysis

The primary benefit of this analysis for stakeholders is identifying whether there is a statistically significant association between the marital status and contract type of the customers. By using the chi-square test, we can better understand if there is a non-random relationship between these two variables, which could suggest what marketing strategy makes sense for which type of customers based on their marital status. While the chi-square test doesn’t differentiate between the specific classes of the categorical variables, it can establish whether there is an association between them. If there is, further analysis could be conducted to determine the relationship between the customers’ marital statuses and contract types.

A3. Data Identification

The relevant variables for my research question from this data set are as follows:

- Marital (quantitative), example: Never Married
- Contract (qualitative), example: One Year

B1: Code

A chi-square test was performed to evaluate the two qualitative variables for our research question.

```
In [10]: # create a contingency table for marital status to contract length
         table = pd.crosstab(df['Marital'], df['Contract'])
         print(table)
```

```
Contract       Month-to-month  One year  Two Year
Marital
Divorced                 1166       425       501
Married                  1067       399       445
Never Married            1046       416       494
Separated                1085       422       507
Widowed                  1092       440       495
```

```
In [11]: chi = stats.chi2_contingency(table)
         print(chi)
         print(f"The p-value is {chi[1]:.3}.")
```

```
Chi2ContingencyResult(statistic=5.123017148617485, pvalue=0.7443506356112823, dof=8, expected_freq=array([[1141.395
2,  439.7384,  510.8664],
       [1042.6416,  401.6922,  466.6662],
       [1067.1936,  411.1512,  477.6552],
       [1098.8384,  423.3428,  491.8188],
       [1105.9312,  426.0754,  494.9934]]))
The p-value is 0.744.
```

See code attached in WGU_D207_Task_1.ipynb.

B2: Output

The output from running the code resulted in a p-value of 0.744. Using the typical significance level of 0.05, we can establish that there is no significant association between marital status and contract length as 0.744 > 0.05, so we do not reject the null hypothesis. We are able to see the actual counts of the variables from the data set as well, which helped establish the p-value when running the test.

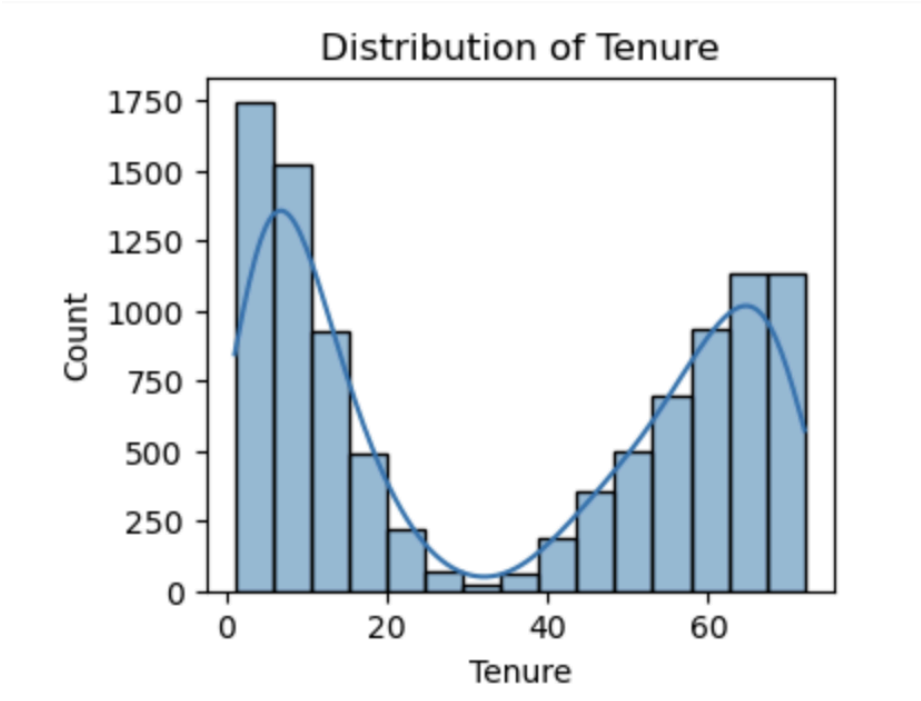
B3: Justification

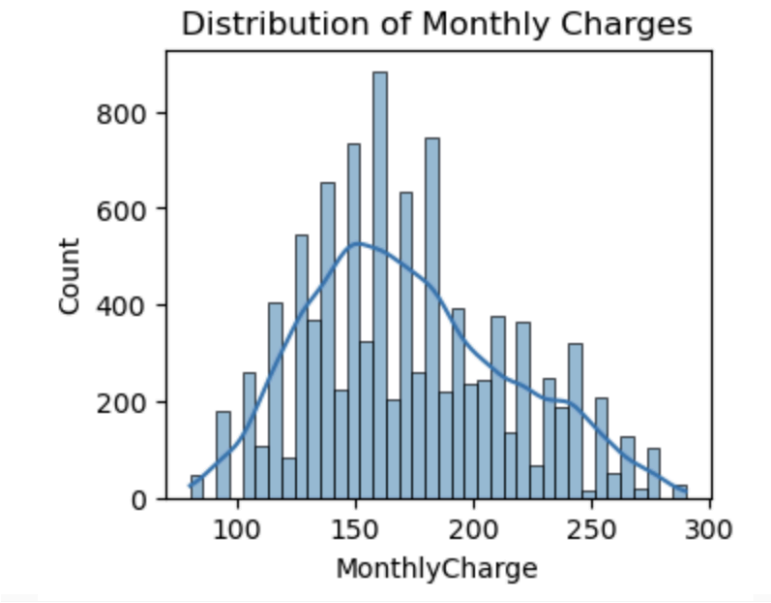
The chi-square test was selected due to its ability to test the association between two categorical variables: Marital and Contract. Since both variables are qualitative, a chi-square test is most suitable for this analysis. Using Python's scipy.stats library, this allows for a clear reporting on the test statistic, p-value, and frequencies when performing the chi-square test of independence. Based on these given points, I'd determined that a chi-square test was most appropriate for this task.

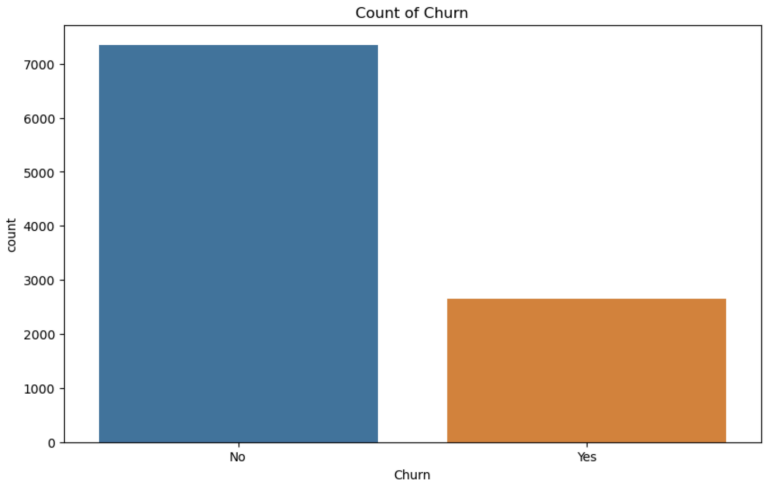
C: Univariate Statistics

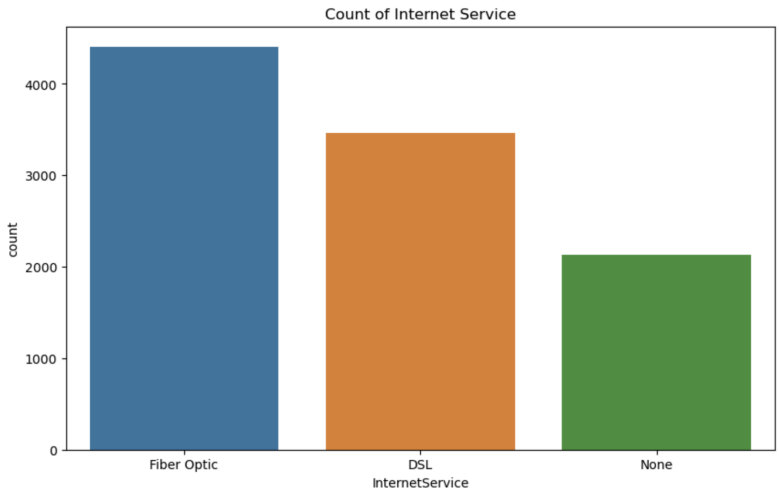
The distribution of the continuous variable ‘Tenure’ is slightly right-skewed, indicating most customers have a shorter tenure but also a long tail of customers with very high tenure as well. The mean tenure is 34 months, the median is 35 months, and the standard deviation is 24 months. The continuous variable ‘MonthlyCharge’ seems relatively normal but with a slight left-skew, with very few customers on both the high and low end of the graph. The mean monthly charge is $172, the median is $167, and the standard deviation is $30. The categorical ‘Churn’ variable has only ‘yes’ or ‘no’ values that indicate whether a customer has churned or not. It seems that around 27% of the customers have churned while 73% have not churned. The categorical ‘InternetService’ variable shows whether customers have DSL, Fiber Optic, or no service. The bar graph indicates that about 44% use Fiber Optic, 34% use DSL, and 21% have no internet service.

C1: Visual of Findings




Distribution of Monthly Charges
Count
MonthlyCharge
0
200
400
600
800
100
150
200
250
300


Count of Churn
count
Churn
No
Yes
0
1000
2000
3000
4000
5000
6000
7000

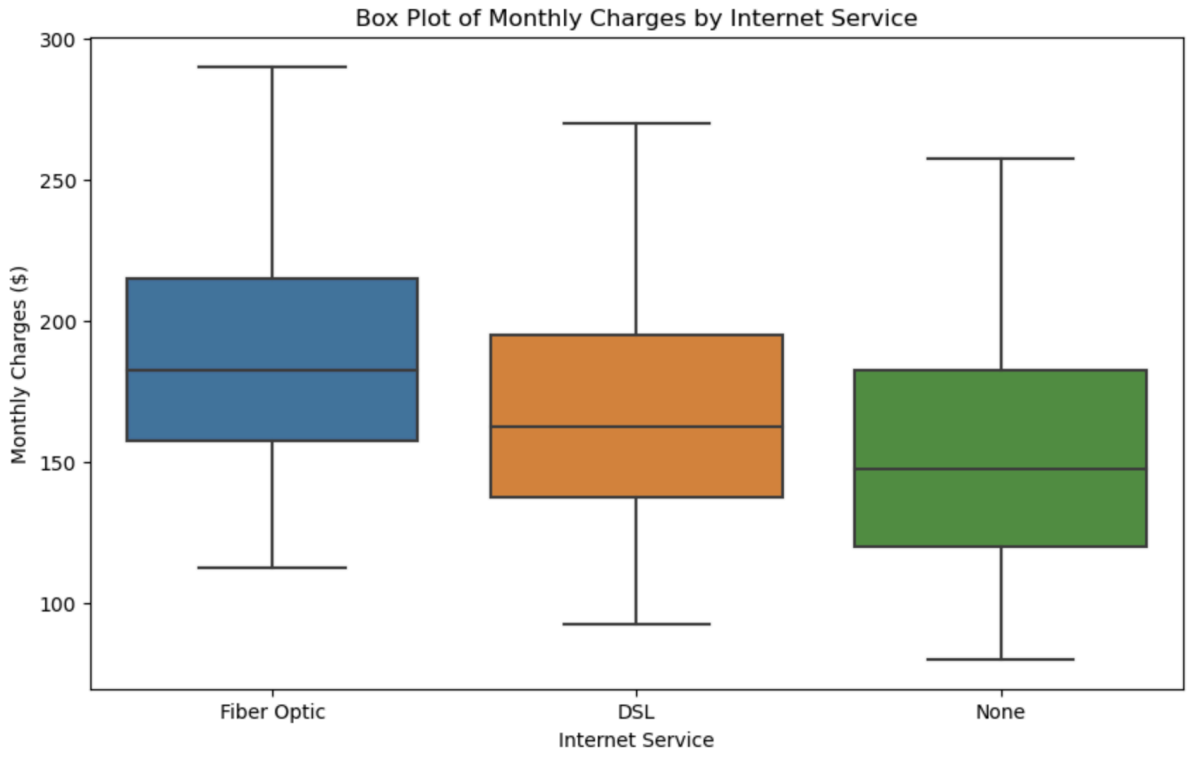


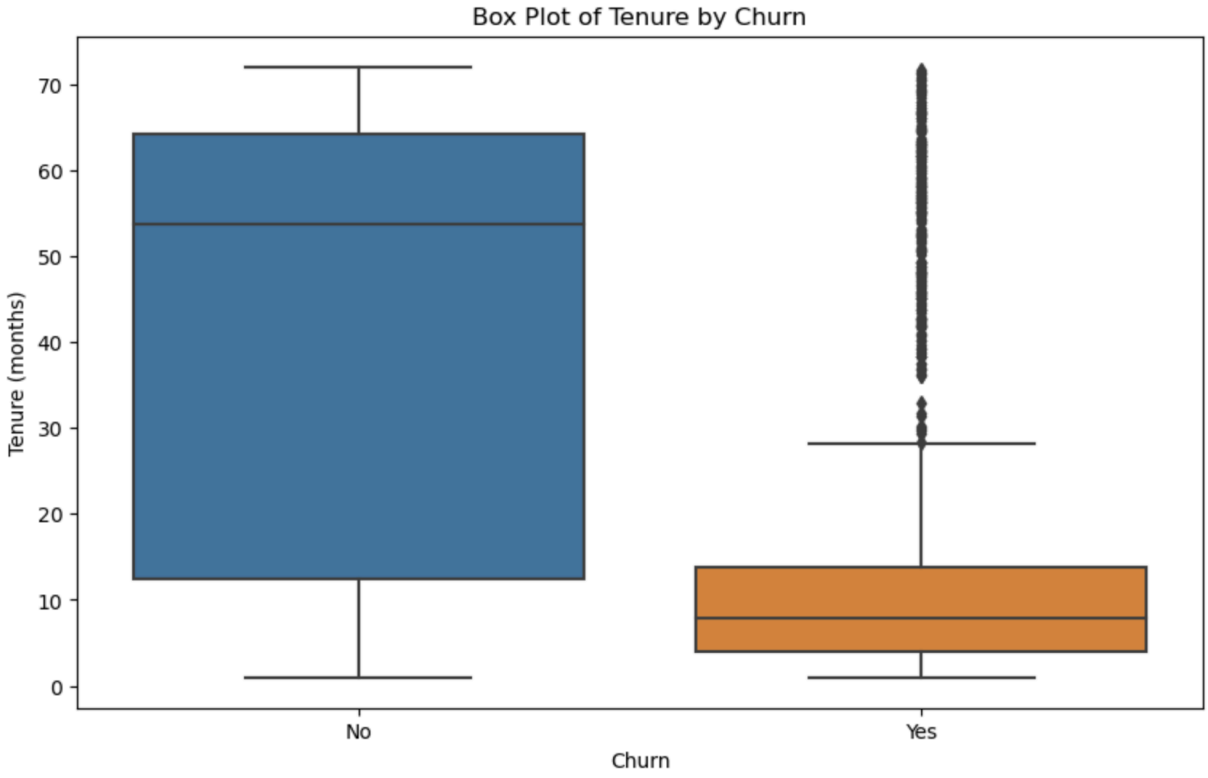
See code attached in WGU_D207_Task_1.ipynb.

D: Bivariate Statistics

Using bivariate statistics, we'll be looking at the relationship between the continuous variable 'Tenure' and the categorical variable 'Churn' and comparing them in our analysis. We first used a box plot to observe that those who don't churn stay for longer periods, with the median time being around 37 months. In contrast, those who have churned stayed for a median time of about 9 months. We next used a box plot to look at the relationship between the continuous variable 'MonthlyCharge' and the categorical variable 'InternetService'. We can observe that customers with Fiber Optic have the highest monthly charges, with the median being $180, those with DSL have a median charge of $162, and those with no service have a median of $150.

D1: Visual of Findings





See code attached in WGU_D207_Task_1.ipynb.

E1: Results of Analysis

In the chi-square test of independence for our research question, we examined if there was a relationship between marital status and the length of contract that the customers chose. The results showed that the p-value was 0.744, which is greater than the typical alpha level of 0.05. Due to this observation, we fail to reject the null hypothesis that marital status and contract length are independent of each other. The results imply that there is not a significant association between marital status and the type of contract the customers sign up for.

E2: Limitations of Analysis

There are several limitations with our analysis, one being the granularity of the variables. We only have several categories for the ‘Marital’ and ‘Contract’ variables but what if there are more categories (i.e. 3 years, 4 years, etc.) that make our variables not as broad? Another limitation is causality. While we found no association between the variables, this doesn’t mean that other variables don’t influence the contract type the customers purchase. Lastly, one more limitation could be timeline factors. We only had a snapshot of a particular time with this data set, which doesn’t account for changes in consumer behavior over time.

E3: Recommended Course of Action

Due to having no association between ‘Marital’ and ‘Contract’ variables, a recommended action for the organization would be to tailor their marketing strategies using other factors such as age, income, gender, etc. Another action that could be done is to take customer feedback to better understand why certain demographics purchase certain contract types. The organization could also improve their data collection method

to include more variables that could potentially influence contract choice. Further research and periodic updates to the data set is recommended to capture any changes in customer behavior.

F: Panopto Video

The URL link for the Panopto video can be found here. It will also be submitted in the Performance Assessment task submission.

G: Sources of Third Party Code

1. "Contingency Table Functions." Retrieved from https://docs.scipy.org/doc/scipy/reference/stats.contingency.html
2. "Simple Scatter Plots." Retrieved from https://jakevdp.github.io/PythonDataScienceHandbook/04.02-simple-scatter-plots.html
3. "Create a Python Heatmap with Seaborn." Retrieved from https://absentdata.com/python-graphs/create-a-heat-map-with-seaborn/

H: Web Sources

1. WGU Courseware. Retrieved from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bccf2cb6-39e2-4a53-8744-ad1900e9aa91
2. WGU Courseware. Retrieved from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=958f4551-c08f-4611-ab2a-b109003ab67a