WGU D207 PA

Andrew Shrestha

Requirement A

1. Question: Are we able to identify and quantify the risk at which customers will churn on their service provider?

2. Benefits to Stakeholders

The information obtained by this analysis will greatly benefit the stakeholders as they will have concrete figures/statistics to guide their decisions on which variables are of the greatest significance to the churn results. This is extremely important and useful for a company's profit as they are able to optimize the allocation of their time and resources on improving customer retention, which will lead to further business wide growth.

3. Identifying Relevant Data

The dependent variable in this data set will be the "Churn" outcome which takes a binary option of either "Yes" or "No". This is the most important variable in the dataset as it is the outcome that will be focused on in this analysis.

Other relevant data in this data set will consist of the numerical variables of: "Age" (Age of customer as reported in sign-up information) , "Income" (Annual income of customer as reported in sign-up information), "Tenure" (Number of months a customer has stayed with the provider), "MonthlyCharge" (Amount charged to customer monthly), and "Bandwidth_GB_Year" (Average amount of Bandwidth used, in GB, in a year) as these have been identified after observing the cleaned data, to be the most applicable.

Finally, we would also like to include the survey question responses as the voice of opinions directly from customers will also play a significant role in our churn analysis. These variables are also Numeric in nature and will be contained in the range of 1-8 where 1 is the most important and 8 is the least important. The 8 survey questions will be as follows: "Item 1" (Timely Response), "Item 2" (Timely Fixes), "Item 3" (Timely Replacements), "Item 4" (Reliability), "Item 5" (Options), "Item 6" (Respectful Response), "Item 7" (Courteous Exchange), "Item 8" (Evidence of Active Listening).

Requirement B

1. Describe the Data Analysis

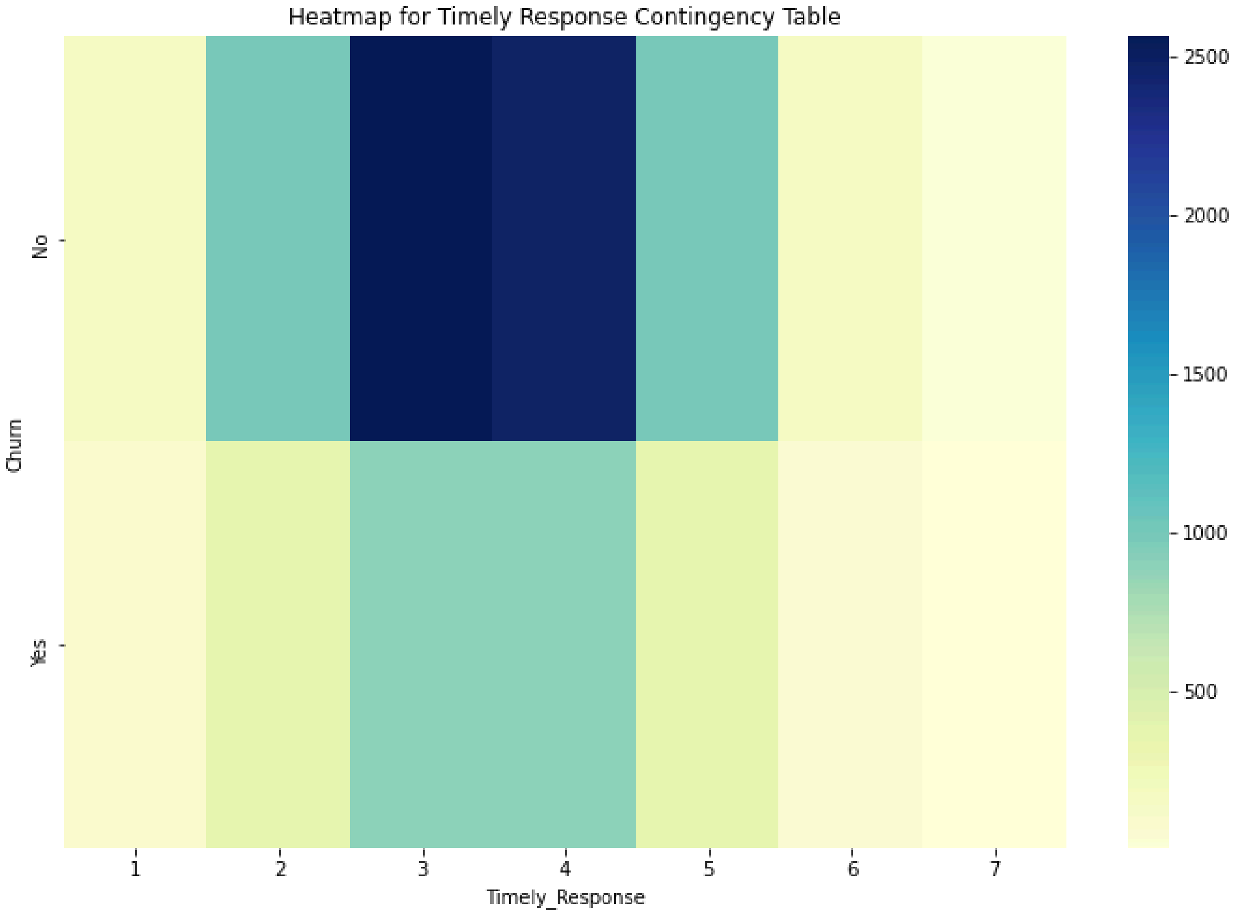
The analysis will be done using the Chi-Square technique in Python. The reasoning behind choosing to use the Chi-Square test is the excellent capability of handling data distributions, ease of computation, and providing detailed statistics **McHugh, M.(2013)**

2. Code for Chi-Square Analysis

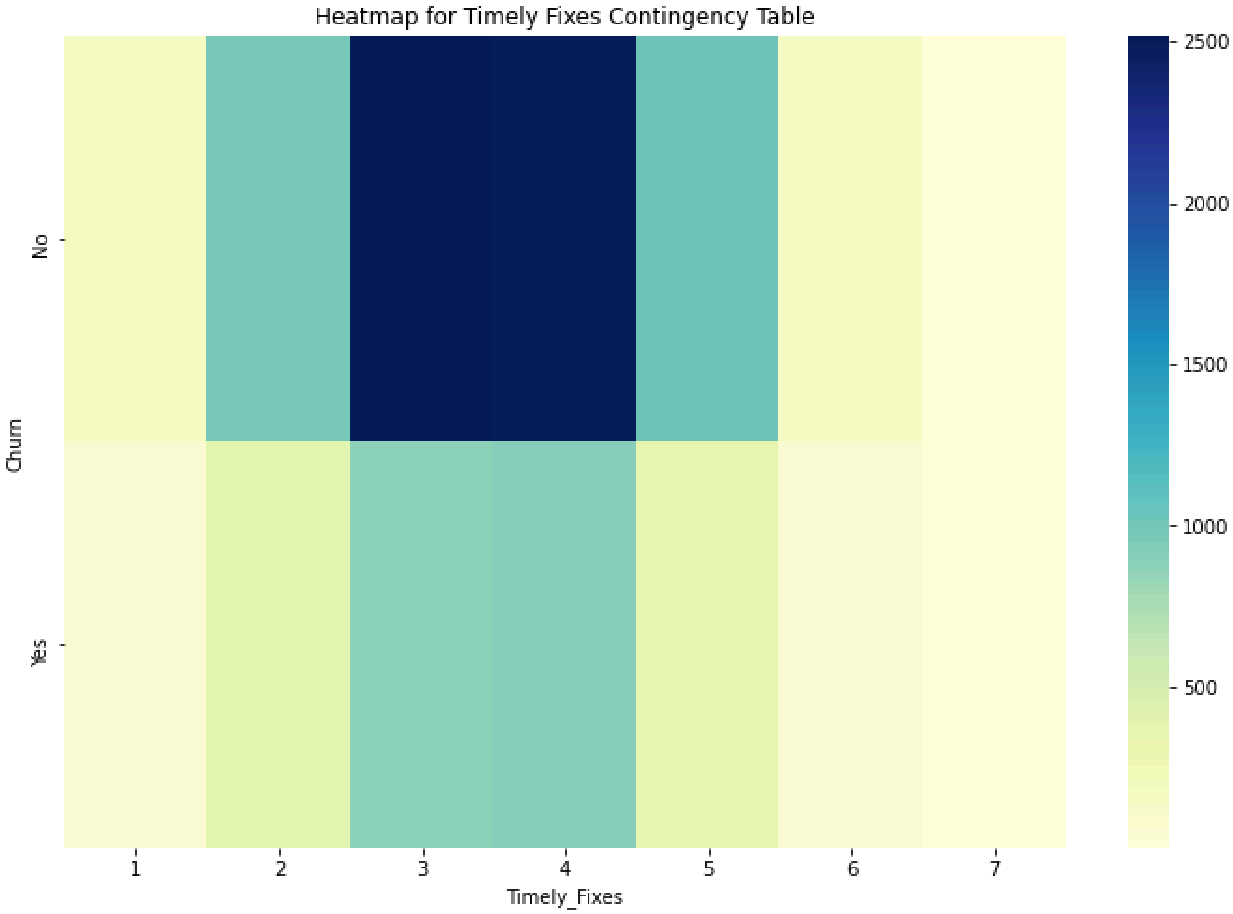
```
#Import Necessary Libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

import pylab
import statsmodels.api as sm
import statistics
from scipy import stats

from scipy.stats import chisquare
from scipy.stats import chi2_contingency
from scipy.stats import chi2


#Load the Provided Cleaned Data Set
df = pd.read_csv(r"C:\Users\andre\OneDrive\Desktop\churn_clean_207.csv")

#Gain an overview of the data
df.describe

df.info()

df.dtypes


# We want to address the survey Items and rename them to more meaningful
lables, thus making it easier to interperate
# and keep track
df.rename(columns = {'Item1':'Timely_Response', 'Item2':'Timely_Fixes',
'Item3':'Timely_Replacements','Item4':'Reliability',

'Item5':'Options','Item6':'Respectful_Response','Item7':'Courteous_Exchange',
                     'Item8':'Evidence_Of_Active_Listening'}, inplace=True)
```

```
# Now we create a contingency table. We do this as to observe if there is a
correlation between two categorical variables as
#pertained to one another

Data_Contingency_Timely_Response = pd.crosstab(
df['Churn'],df['Timely_Response'])
Data_Contingency_Timely_Response

# Going in more detail, we look at the percentage distribution of the Churn
outcomes for each response
Data_Contingency_Timely_Response_Norm = pd.crosstab(
df['Churn'],df['Timely_Response'], normalize='index')
Data_Contingency_Timely_Response_Norm

Data_Contingency_Timely_Fixes = pd.crosstab( df['Churn'],df['Timely_Fixes'])
Data_Contingency_Timely_Fixes

Data_Contingency_Timely_Fixes_Norm = pd.crosstab(
df['Churn'],df['Timely_Fixes'], normalize ='index')
Data_Contingency_Timely_Fixes_Norm

Data_Contingency_Timely_Replacements = pd.crosstab(
df['Churn'],df['Timely_Replacements'])
Data_Contingency_Timely_Replacements

Data_Contingency_Timely_Replacements_Norm = pd.crosstab(
df['Churn'],df['Timely_Replacements'], normalize='index')
Data_Contingency_Timely_Replacements_Norm

Data_Contingency_Reliability = pd.crosstab( df['Churn'],df['Reliability'])
Data_Contingency_Reliability

Data_Contingency_Reliability_Norm = pd.crosstab( df['Churn'],df['Reliability'],
normalize='index')
Data_Contingency_Reliability_Norm

Data_Contingency_Options = pd.crosstab( df['Churn'],df['Options'])
Data_Contingency_Options

Data_Contingency_Options_Norm = pd.crosstab( df['Churn'],df['Options'],
normalize='index')
Data_Contingency_Options_Norm

Data_Contingency_Respectful_Response = pd.crosstab(
df['Churn'],df['Respectful_Response'])
Data_Contingency_Respectful_Response

Data_Contingency_Respectful_Response_Norm = pd.crosstab(
df['Churn'],df['Respectful_Response'], normalize ='index')
Data_Contingency_Respectful_Response_Norm
```

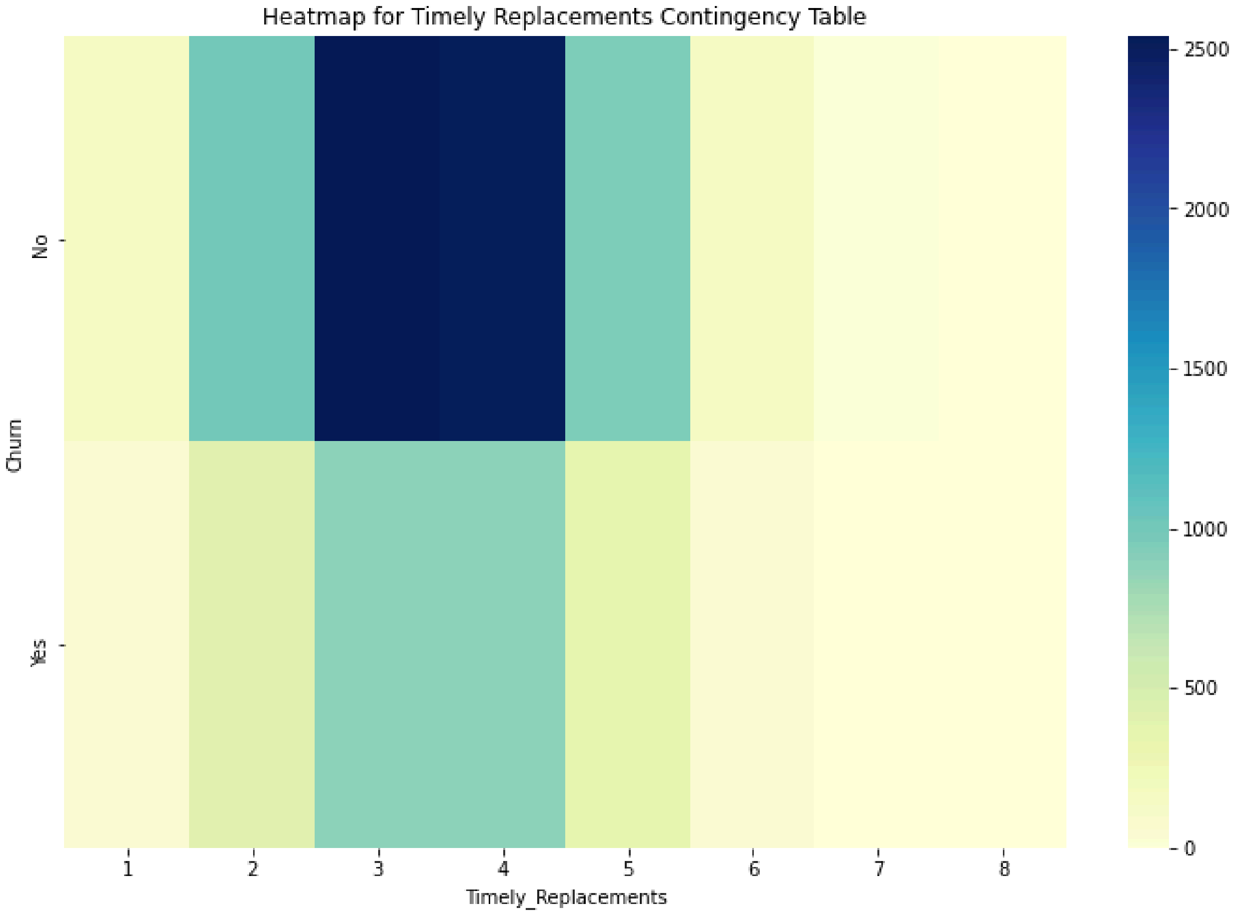
```
Data_Contingency_Courteous_Exchange = pd.crosstab(
df['Churn'],df['Courteous_Exchange'])
Data_Contingency_Courteous_Exchange

Data_Contingency_Courteous_Exchange_Norm = pd.crosstab(
df['Churn'],df['Courteous_Exchange'], normalize='index')
Data_Contingency_Courteous_Exchange_Norm

Data_Contingency_Evidence_Of_Active_Listening = pd.crosstab(
df['Churn'],df['Evidence_Of_Active_Listening'])
Data_Contingency_Evidence_Of_Active_Listening

Data_Contingency_Evidence_Of_Active_Listening_Norm = pd.crosstab(
df['Churn'],df['Evidence_Of_Active_Listening'],normalize='index')
Data_Contingency_Evidence_Of_Active_Listening_Norm


#Obtain a visualization of these contingency tables via Heatmap
plt.figure(figsize =(12,8))
sns.heatmap(Data_Contingency_Timely_Response, cmap="YlGnBu")
plt.title('Heatmap for Timely Response Contingency Table', fontsize=12)
```



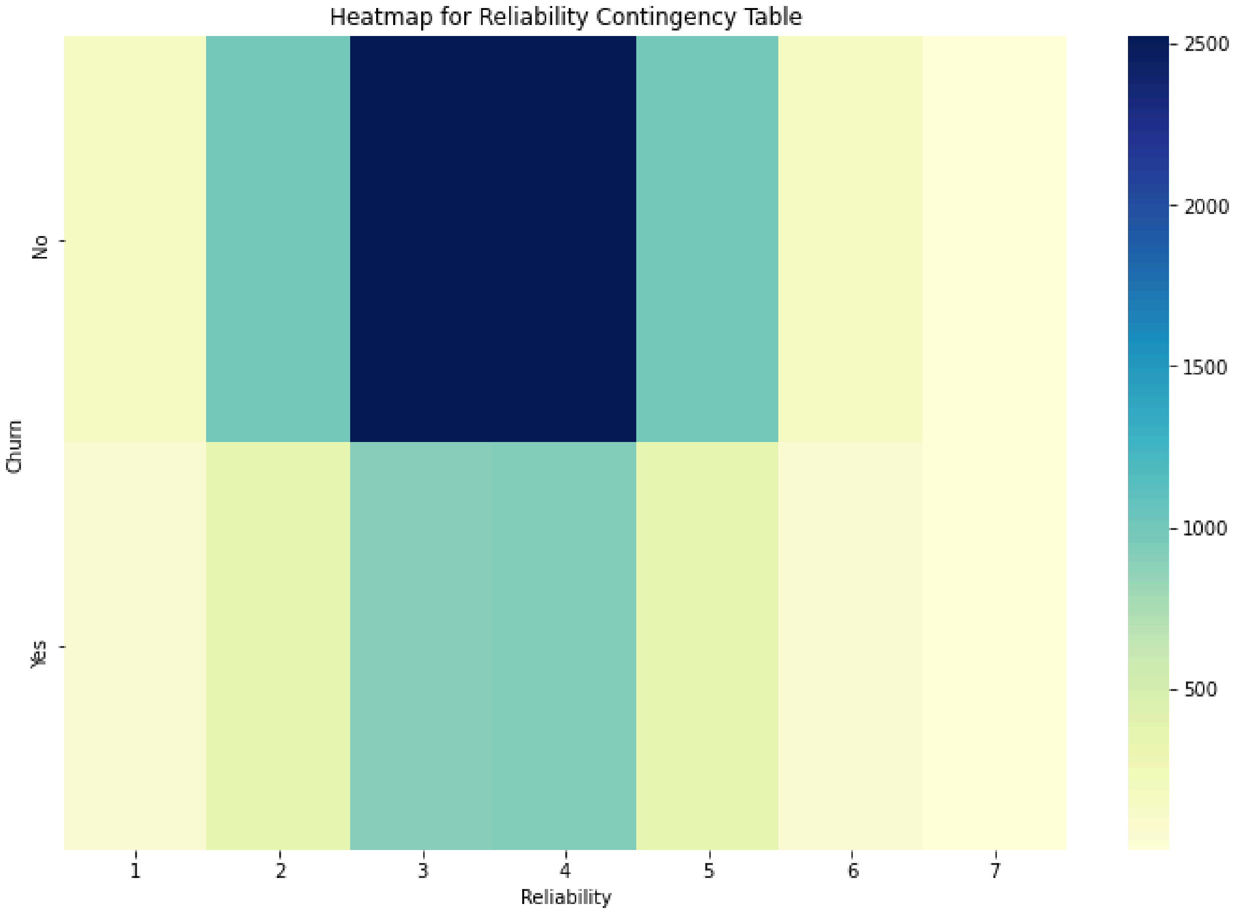
```
plt.figure(figsize =(12,8))
sns.heatmap(Data_Contingency_Timely_Fixes, cmap="YlGnBu")
plt.title('Heatmap for Timely Fixes Contingency Table', fontsize=12)
```



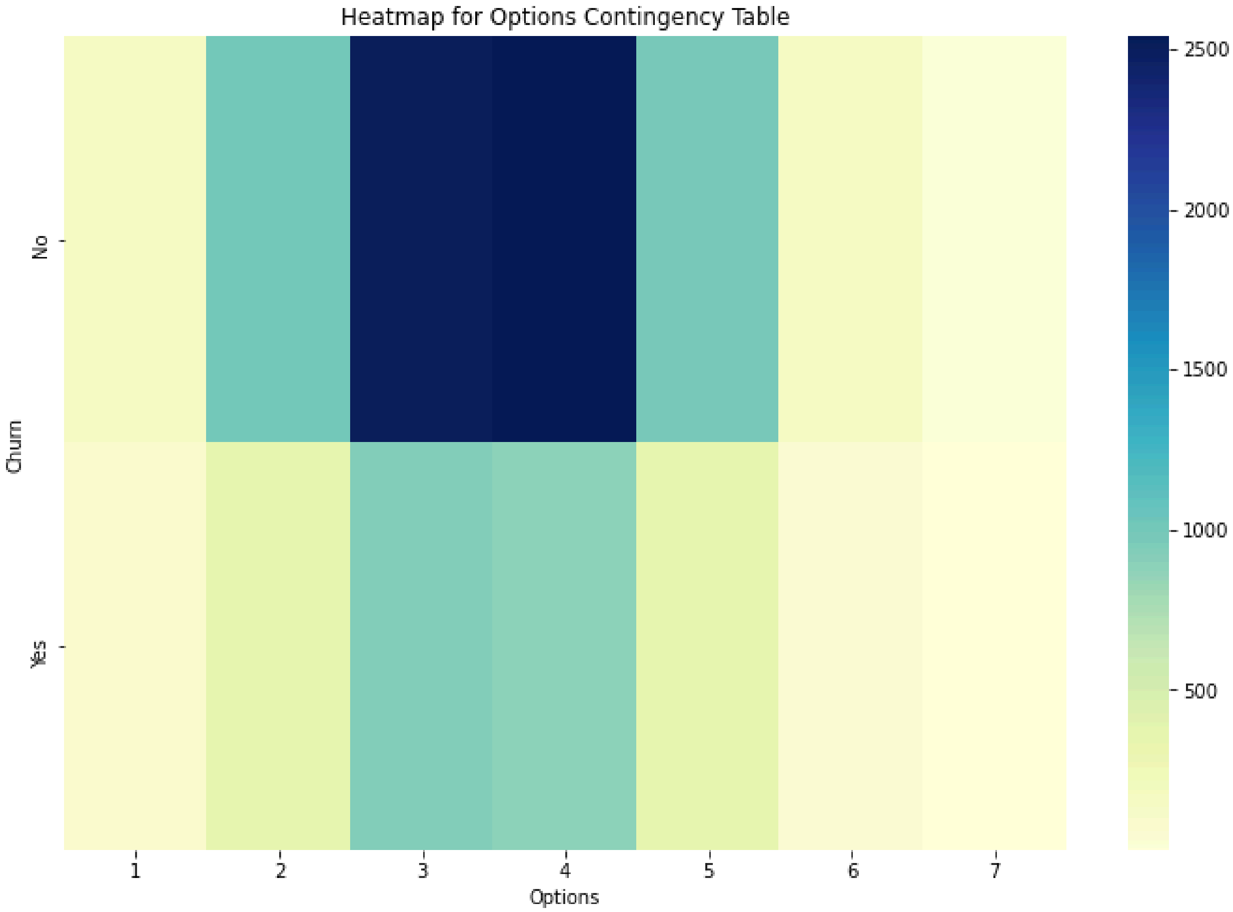
```
plt.figure(figsize =(12,8))
sns.heatmap(Data_Contingency_Timely_Replacements, cmap="YlGnBu")
plt.title('Heatmap for Timely Replacements Contingency Table', fontsize=12)
```



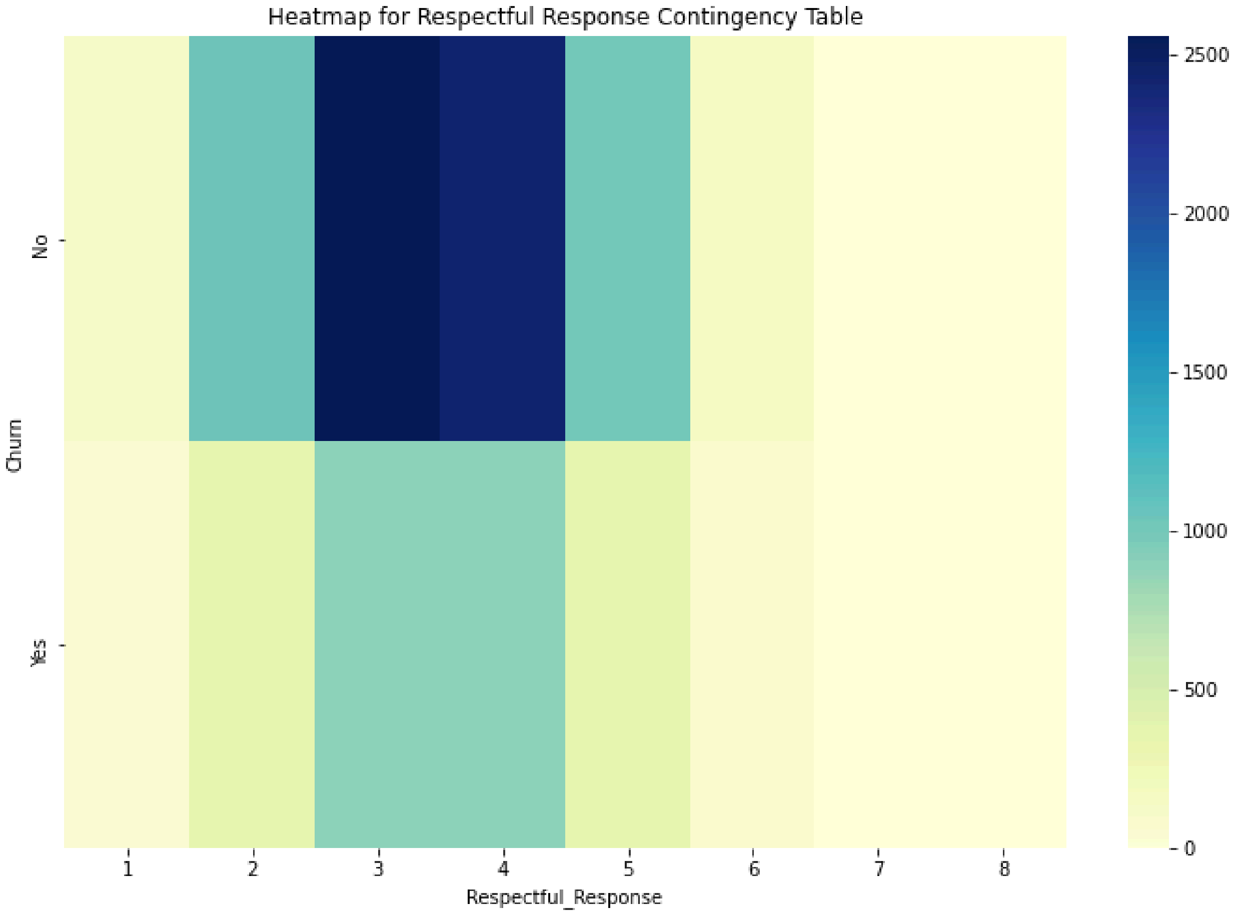
```
plt.figure(figsize =(12,8))
sns.heatmap(Data_Contingency_Reliability, cmap="YlGnBu")
plt.title('Heatmap for Reliability Contingency Table', fontsize=12)
```



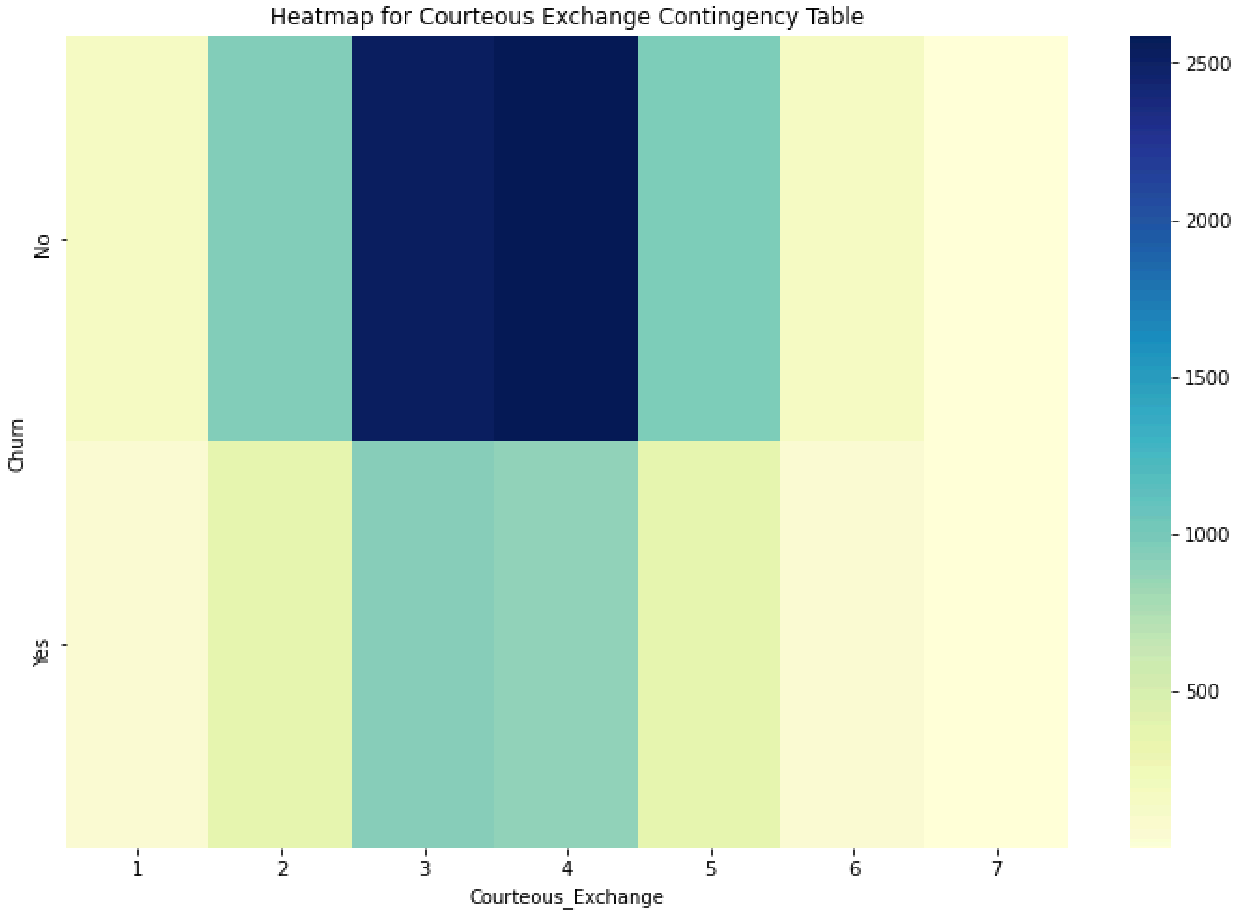
```
plt.figure(figsize =(12,8))
sns.heatmap(Data_Contingency_Options, cmap="YlGnBu")
plt.title('Heatmap for Options Contingency Table', fontsize=12)
```



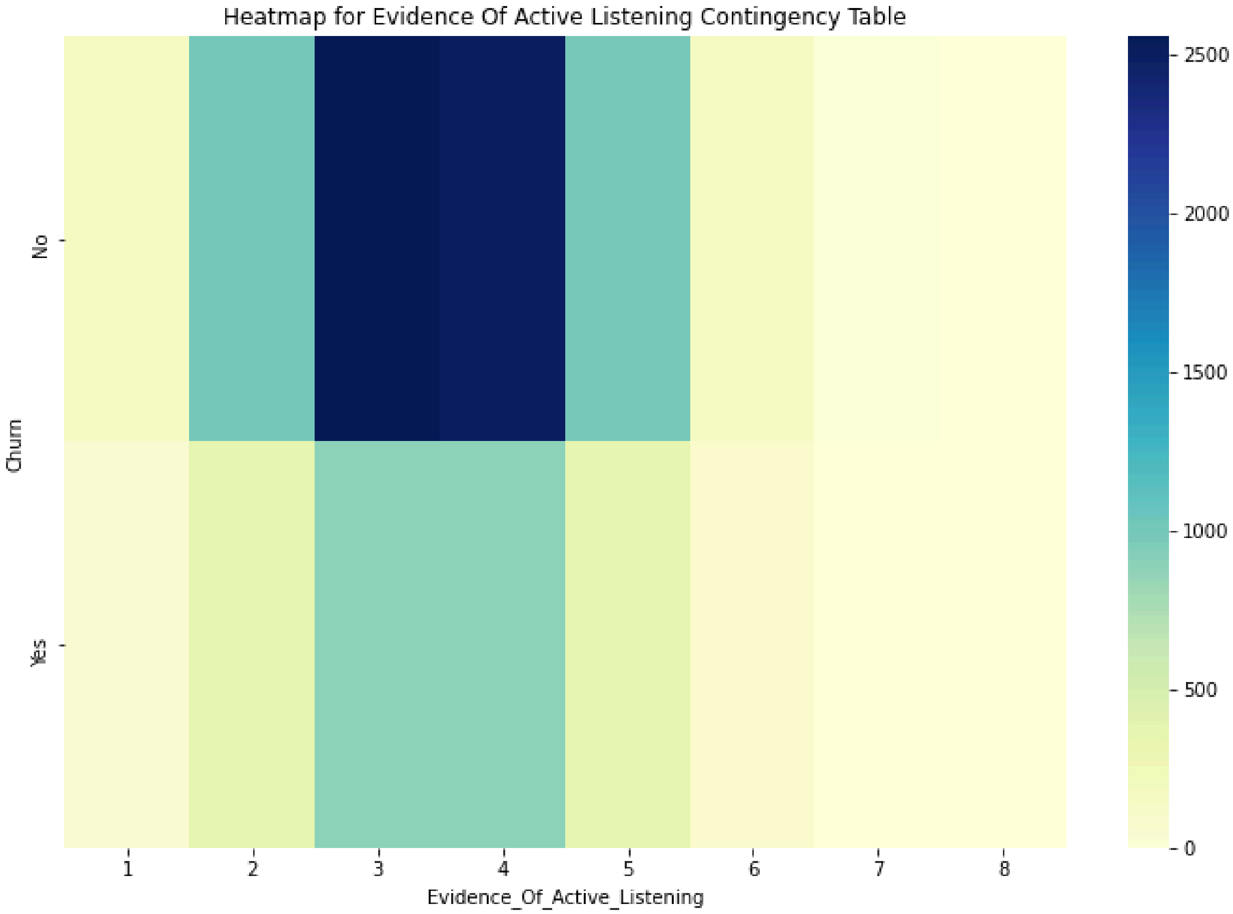
```
plt.figure(figsize =(12,8))
sns.heatmap(Data_Contingency_Respectful_Response, cmap="YlGnBu")
plt.title('Heatmap for Respectful Response Contingency Table', fontsize=12)
```



```
plt.figure(figsize =(12,8))
sns.heatmap(Data_Contingency_Courteous_Exchange, cmap="YlGnBu")
plt.title('Heatmap for Courteous Exchange Contingency Table', fontsize=12)
```



```
plt.figure(figsize =(12,8))
sns.heatmap(Data_Contingency_Evidence_Of_Active_Listening, cmap="YlGnBu")
plt.title('Heatmap for Evidence Of Active Listening Contingency Table',
fontsize=12)
```



3. Output

```
#Utilize the Chi-Square Test for Independence on each of the Contingency
Tables.
print(Data_Contingency_Timely_Response)
stat, p, dof, expected_Timely_Response=
chi2_contingency(Data_Contingency_Timely_Response)
print('dof=%d' % dof)
print(expected_Timely_Response)


#Interpreting the P-Value for the Timely Response Contingency Table. We will
using the significance level of 5%
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
print(Data_Contingency_Timely_Fixes)
stat, p, dof, expected_Timely_Fixes=
chi2_contingency(Data_Contingency_Timely_Fixes)
print('dof=%d' % dof)
print(expected_Timely_Fixes)
```

```
#Interpreting the P-Value for the Timely Fixes Contingency Table. We will using
the significance level of 5%
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
print(Data_Contingency_Timely_Replacements)
stat, p, dof, expected_Timely_Replacements=
chi2_contingency(Data_Contingency_Timely_Replacements)
print('dof=%d' % dof)
print(expected_Timely_Replacements)
```

```
#Interpreting the P-Value for the Timely Replacements Contingency Table. We
will using the significance level of 5%
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
print(Data_Contingency_Reliability)
stat, p, dof, expected_Reliability=
chi2_contingency(Data_Contingency_Reliability)
print('dof=%d' % dof)
print(expected_Reliability)
```

```
#Interpreting the P-Value for the Reliability Contingency Table. We will using
the significance level of 5%
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
print(Data_Contingency_Options)
stat, p, dof, expected_Options= chi2_contingency(Data_Contingency_Options)
print('dof=%d' % dof)
print(expected_Options)
```

```
#Interpreting the P-Value for the Options Contingency Table. We will using the
significance level of 5%
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
print(Data_Contingency_Respectful_Response)
stat, p, dof, expected_Respectful_Response=
chi2_contingency(Data_Contingency_Respectful_Response)
print('dof=%d' % dof)
print(expected_Respectful_Response)
```

```
#Interpreting the P-Value for the Respectful Response Contingency Table. We
will using the significance level of 5%
alpha = 1.0 - prob
```

```
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
print(Data_Contingency_Courteous_Exchange)
stat, p, dof, expected_Courteous_Exchange=
chi2_contingency(Data_Contingency_Courteous_Exchange)
print('dof=%d' % dof)
print(expected_Courteous_Exchange)


#Interpreting the P-Value for the Courteous Exchange Contingency Table. We will
using the significance level of 5%
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

```
print(Data_Contingency_Evidence_Of_Active_Listening)
stat, p, dof, expected_Evidence_Of_Active_Listening=
chi2_contingency(Data_Contingency_Evidence_Of_Active_Listening)
print('dof=%d' % dof)
print(expected_Evidence_Of_Active_Listening)


#Interpreting the P-Value for the Evidence of Active Listening Contingency
Table. We will using the significance level of 5%
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
        print('Dependent (reject H0)')
else:
        print('Independent (fail to reject H0)')
```

4. Justification for Analysis Technique

The use of Chi-Square Analysis was done due to the fact that we were able to compare the independence of the categorical dependent variable “Churn” with other categorical survey variables **P&SS(2021)**: Timely_Response, Timely_Fixes, Timely_Replacement, Reliability, Options, Respectful_Response, Courteous_Response, and Evidence_Of_Active_Listening.

This is perfect for answering out initial question with regards to Churn as the Chi-Square will illuminate any relations between the selected categorical variables obtained by the survey directly to the Churn results.

Requirement C

Identify distribution of two continuous and categorical variables via Univariate Statistics

The two continuous variable distribution that are going to be used will be as follows:

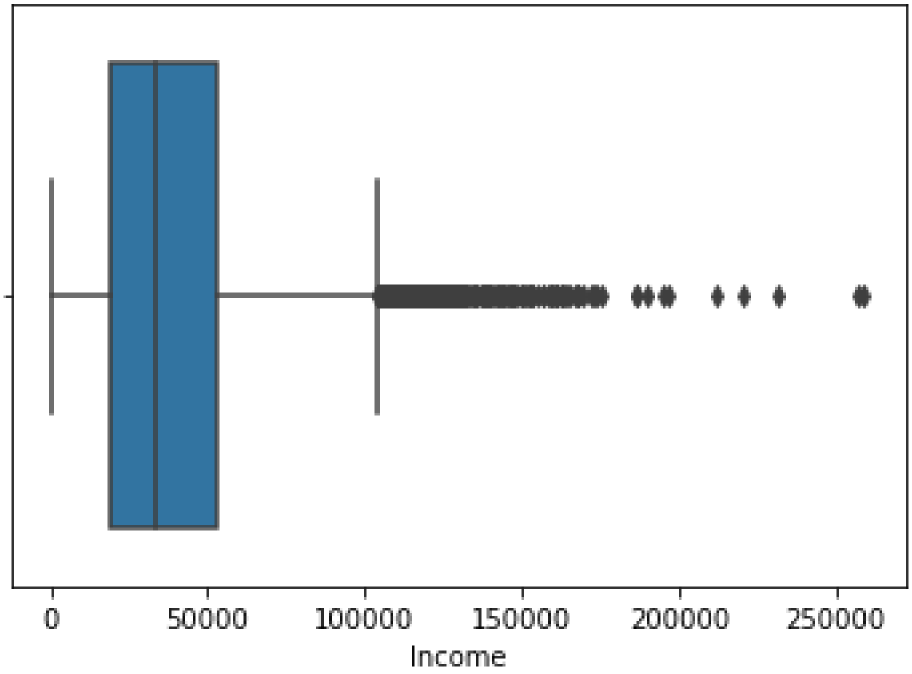
a. Income
b. MonthlyCharge

The two categorical variable distribution that are going to be used will be as follows:

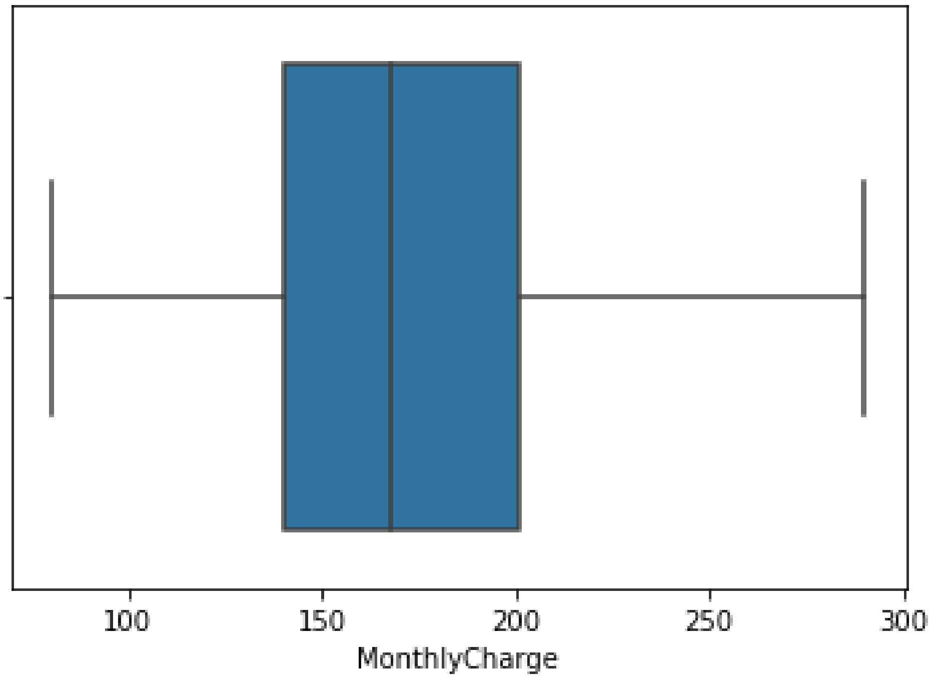
a. Timely_Fixes
b. Timely_Replacements

1. Visual Representation

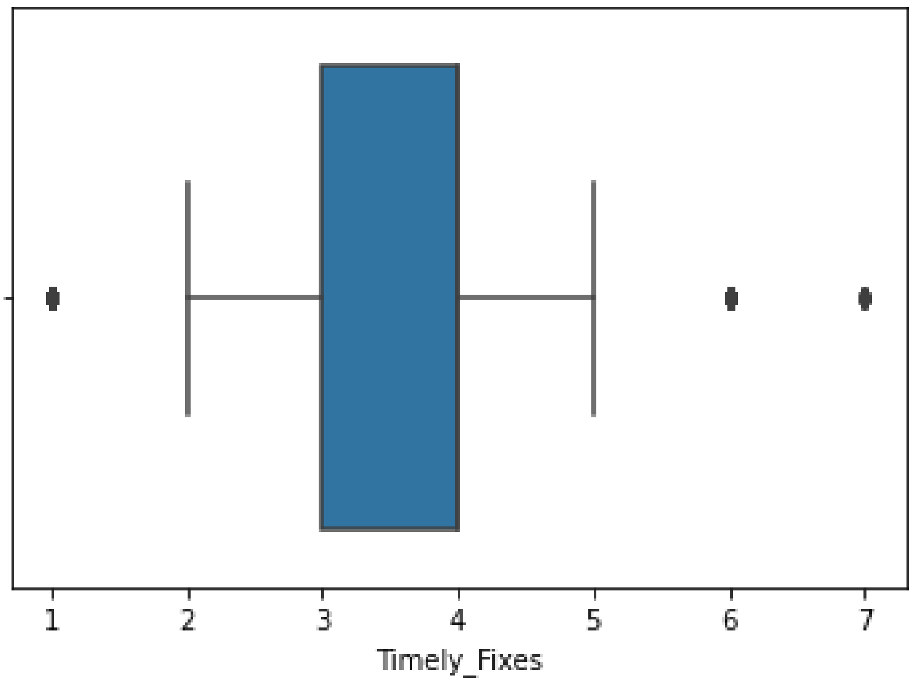
```
#Univariate Statistics for our Chosen two continuous and categorical variables
in the below dataset
df.describe()


#Box Plots of the choosen continuous and categorical variables: Income,
MonthlyCharge, Timely_Fixes, Timely_Replacements via seaborn
sns.boxplot('Income', data=df)
plt.show()
```



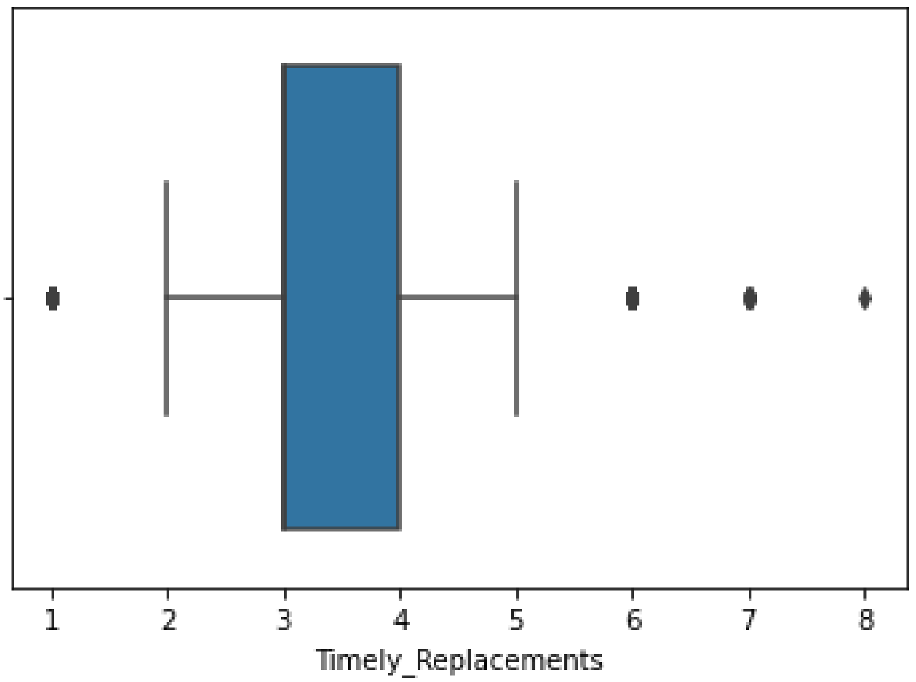
```
sns.boxplot('MonthlyCharge', data=df)
plt.show()
```



```
sns.boxplot('Timely_Fixes', data=df)
plt.show()
```



```
sns.boxplot('Timely_Replacements', data=df)
plt.show()
```



Requirement D

Identify distribution of two continuous and categorical variables via Bivariate Statistics

The two continuous variable distribution that are going to be used will be as follows:

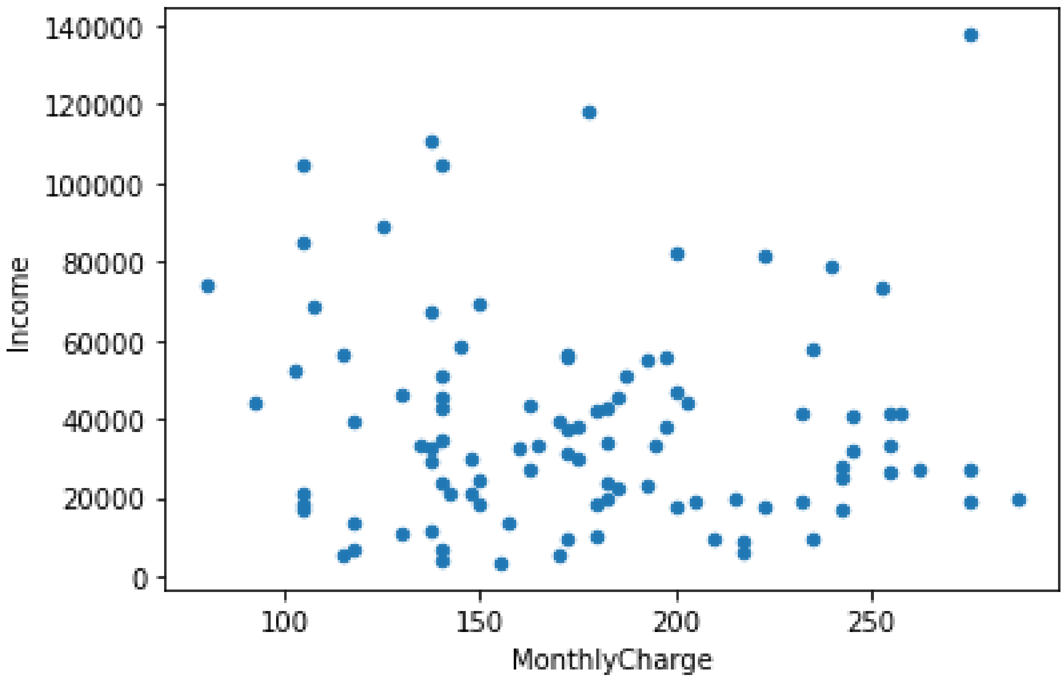
c. Income
d. MonthlyCharge

The two categorical variable distribution that are going to be used will be as follows:

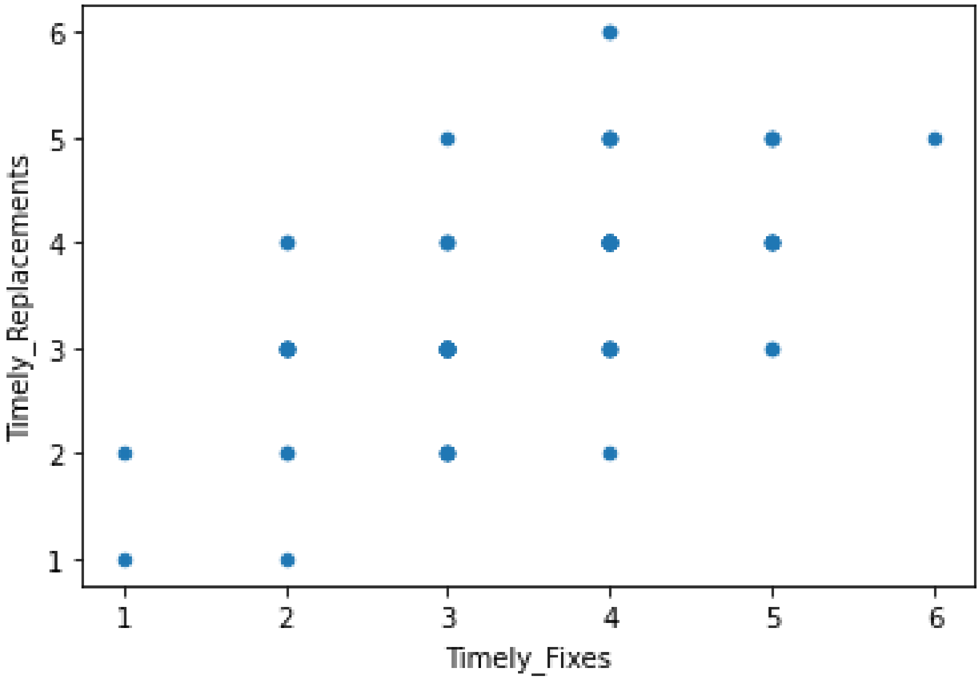
c. Timely_Fixes
d. Timely_Replacements

1. Visual Representation

```
#Bivariate Statistics Scatter Plot of the choosen continuous and categorical
variables: Income, MonthlyCharge, Timely_Fixes, Timely_Replacements via seaborn

Bivariate_Data =
df[['Income','MonthlyCharge','Timely_Fixes','Timely_Replacements']]


# Obtaining Scatter Plot for the Continuous Variables: Income and MonthlyCharge
Bivariate_Data [Bivariate_Data['MonthlyCharge']
<300].sample(100).plot.scatter(x='MonthlyCharge', y='Income')
```



```
# Obtaining Scatter Plot for the Categorical Variables: Timely_Fixes and
Timely_Replacements
Bivariate_Data [Bivariate_Data['Timely_Fixes']
<7].sample(100).plot.scatter(x='Timely_Fixes', y='Timely_Replacements')
```



Requirement E

Implications of the Data Analysis

1. Results of the Hypothesis Test

Through our calculations of the P-Values, we have found that for each of the survey questions, the outcome was ultimately to not reject the null hypothesis. This was implied through the calculation outputs when the values the p-values being greater than the significant value. It is also important to note that this was done using a 95% significance level or 0.05 alpha.

This was a surprising outcome, However this analysis shows there is not a significant relationship between any of the survey questions and the Churn outcomes for the customers.

2. Limitations of the Analysis

One limitation is that there was a collective sample of 10,000 observations. This is not a small amount, however if we were to gather more data and increase the scope, we would be able to find a relationship toward the Churn outcome.

3. Recommended Course of Action based on Results

Based on our results, we clearly see that there is no significant relationship between the customer service oriented survey questions and churn results. Therefore, it may be in the best interest of the company to allocate more time and resources into their pricing models and product features offered instead. The reasoning for this is due to the insights that perhaps customers who decide to churn or not are only interested in the more tangible aspects of the service provider (price, benefits, products, etc) and not so much on the intangible/customer service aspect of the company.

That being said, it would be unwise to completely neglect the customer service aspect in the state survey areas as opportunities to keep improving on.

Requirement F

Panopto Video link:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6b5641ad-5939-4752-8474-ad59000cec5e

Requirement G

Third Party Code References:

Moonbooks(2021). How to plot a contingency table(heatmap) in python using seaborn and matplotlib?. https://moonbooks.org/Articles/How-to-plot-a-contingency-table-heatmap-in-python-using-seaborn-and-matplotlib-/

GeeksforGeeks(2019). Contingency Table in Python. https://www.geeksforgeeks.org/contingency-table-in-python/

Jason Browniee(2015). A Gentle Introduction to the Chi-Squared Test for Machine Learning. https://machinelearningmastery.com/chi-squared-test-for-machine-learning/

Kaggle(2018). Bivariate plotting with Pandas. https://www.kaggle.com/residentmario/bivariate-plotting-with-pandas

Requirement H

Source Code References :

JMP(2021). Chi-Square Test of Independence. https://www.jmp.com/en_ca/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html

Jason Browniee(2015). A Gentle Introduction to the Chi-Squared Test for Machine Learning. https://machinelearningmastery.com/chi-squared-test-for-machine-learning/

McHugh, M.(2013). Chi-Square Test of Independence. https://pubmed.ncbi.nlm.nih.gov/23894860/

Plant & Soil Sciences (2021).When Chi-Square is Appropriate – Strengths/Weaknesses. https://passel2.unl.edu/view/lesson/9beaa382bf7e/14