WGU

D207 Performance Assessment

EXPLORATORY DATA ANALYSIS

Instructor: Dr. David Gagner

Student name: Mina Saad

Student ID: 010001047

Email: msaad16@wgu.edu

D207 Performance Assessment

A1 - QUESTION FOR ANALYSIS:

How does the Timely Response survey question affect the customer to Churn?

A2 - BENEFIT FROM ANALYSIS:

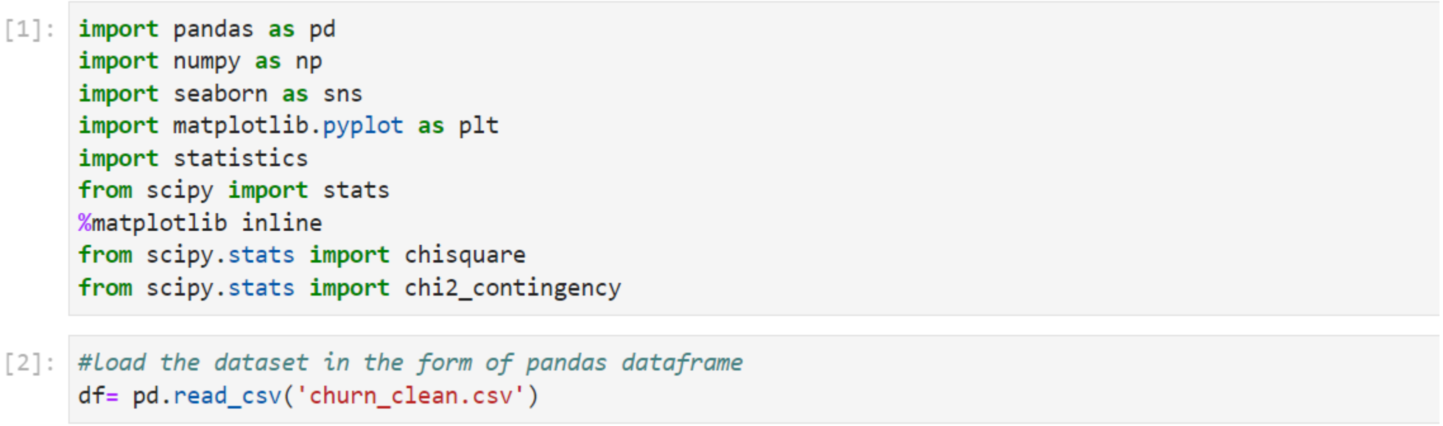
Stakeholders will know from this analysis how likely the customers to churn based on the customer’s answers to the survey question about the “Response Time” they received. This will help Stakeholders to allocate more resources and more efficient tools to meet the customer satisfaction.

A3 - DATA IDENTIFICATION:

The relevant data needed for this analysis would be the [Churn] column and [Item1] column from the survey questions which represent Timely response.

B1 - CODE

Chi Square technique will be used



```
[1]: import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
     import statistics
     from scipy import stats
     %matplotlib inline
     from scipy.stats import chisquare
     from scipy.stats import chi2_contingency
```

```
[2]: #Load the dataset in the form of pandas dataframe
     df= pd.read_csv('churn_clean.csv')
```

[3]:
```
#Print few rows from the imported data
print(df.head())
```

```
   CaseOrder Customer_id                           Interaction  \
0          1     K409198  aa90260b-4141-4a24-8e36-b04ce1f4f77b
1          2     S120509  fb76459f-c047-4a9d-8af9-e0f7d4ac2524
2          3     K191035  344d114c-3736-4be5-98f7-c72c281e2d35
3          4      D90850  abfa2b40-2d43-4994-b15a-989b8c79e311
4          5     K662701  68a861fd-0d20-4e51-a587-8a90407ee574

                                UID         City State                 County  \
0  e885b299883d4f9fb18e39c75155d990  Point Baker    AK  Prince of Wales-Hyder
1  f2de8bef964785f41a2959829830fb8a  West Branch    MI                 Ogemaw
2  f1784cfa9f6d92ae816197eb175d3c71      Yamhill    OR                Yamhill
3  dc8a365077241bb5cd5ccd305136b05e      Del Mar    CA              San Diego
4  aabb64a116e83fdc4befc1fbab1663f9    Needville    TX              Fort Bend

     Zip       Lat        Lng  ...  MonthlyCharge Bandwidth_GB_Year Item1  \
0  99927  56.25100 -133.37571  ...     172.455519        904.536110     5
1  48661  44.32893  -84.24080  ...     242.632554        800.982766     3
2  97148  45.35589 -123.24657  ...     159.947583       2054.706961     4
3  92014  32.96687 -117.24798  ...     119.956840       2164.579412     4
4  77461  29.38012  -95.80673  ...     149.948316        271.493436     4

  Item2  Item3  Item4  Item5 Item6 Item7 Item8
0     5      5      3      4     4     3     4
1     4      3      3      4     3     4     4
2     4      2      4      4     3     3     3
3     4      4      2      5     4     3     3
4     4      4      3      4     4     4     5
```

[5 rows x 50 columns]

[4]:
```
#Rename survey responses column names
df.rename(columns = {'Item1':'Timely_Responses','Item2':'Timely_Fixes',
                           'Item3':'Timely_Replacements','Item4':'Reliability',
                           'Item5':'Options','Item6':'Respectful_responses',
                           'Item7':'Courteous_exchange','Item8':'Active_listening'},
                inplace=True)
```

[5]:
```
#Show columns after updates
df.columns
```

[5]:
```
Index(['CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State',
       'County', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'TimeZone', 'Job',
       'Children', 'Age', 'Income', 'Marital', 'Gender', 'Churn',
       'Outage_sec_perweek', 'Email', 'Contacts', 'Yearly_equip_failure',
       'Techie', 'Contract', 'Port_modem', 'Tablet', 'InternetService',
       'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup',
       'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies',
       'PaperlessBilling', 'PaymentMethod', 'Tenure', 'MonthlyCharge',
       'Bandwidth_GB_Year', 'Timely_Responses', 'Timely_Fixes',
       'Timely_Replacements', 'Reliability', 'Options', 'Respectful_responses',
       'Courteous_exchange', 'Active_listening'],
      dtype='object')
```

[6]:

```
chi_Responses= pd.crosstab(df['Churn'], df['Timely_Responses'])
print(chi_Responses)
```

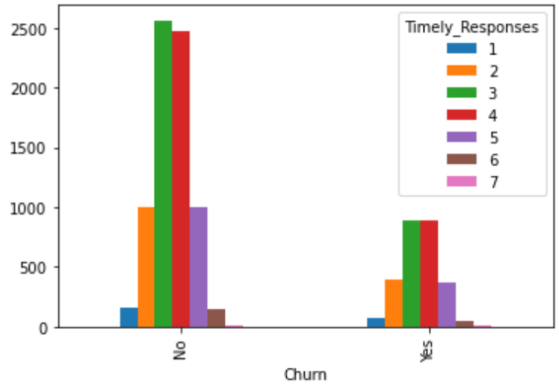
```
Timely_Responses    1     2     3     4    5    6   7
Churn
No                158  1002  2562  2473  994  146  15
Yes                66   391   886   885  365   53   4
```

[7]:

```
chi_Responses.plot(kind='bar', stacked=False)
```

[7]:

```
<AxesSubplot:xlabel='Churn'>
```

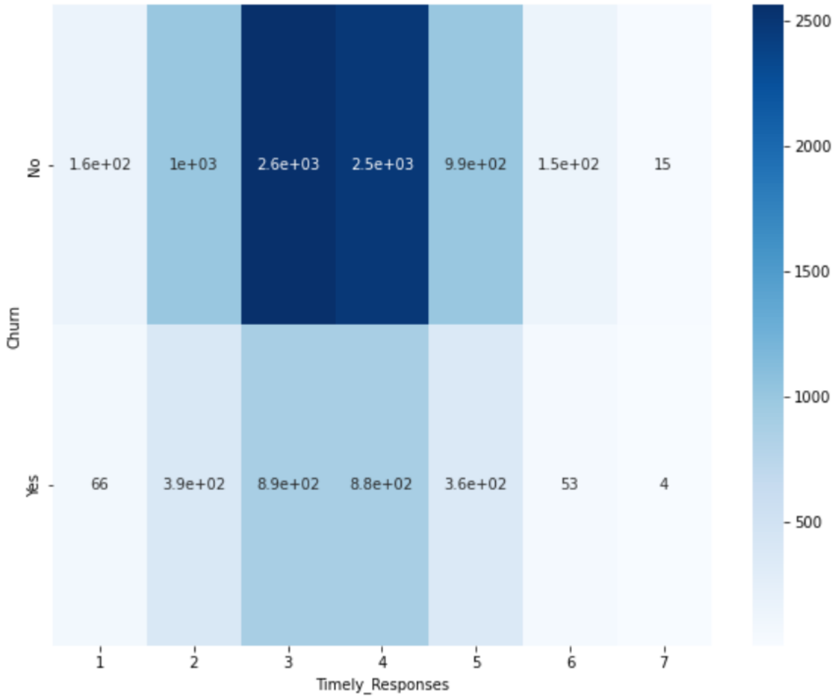


[8]:

```
plt.figure(figsize=(10,8))
sns.heatmap(chi_Responses, annot=True, cmap='Blues')
```

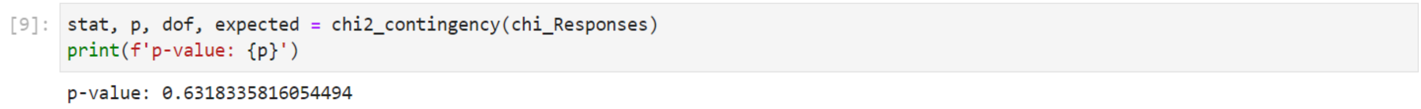
[8]:

```
<AxesSubplot:xlabel='Timely_Responses', ylabel='Churn'>
```



B2 – OUTPUT

Use chi-square to test the independence



```
[9]: stat, p, dof, expected = chi2_contingency(chi_Responses)
     print(f'p-value: {p}')

     p-value: 0.6318335816054494
```

B3 - JUSTIFICATION

Chi-square technique chosen to test the dependency between two categorical variables [Churn] and [Timely_Response] to determine whether the two categorical variables are likely related or not.

C - UNIVARIATE STATISTICS:

Continuous variables:

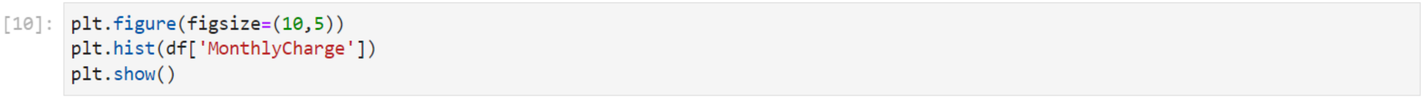
- Monthly Charge
- Income

Categorical variables:

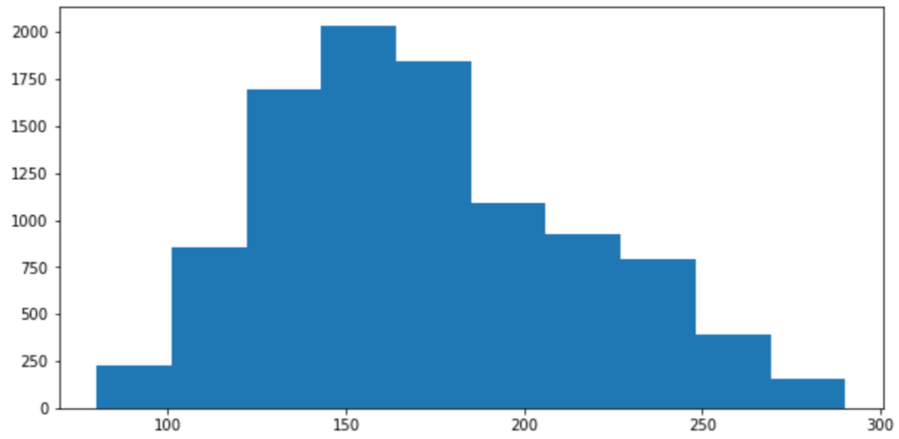
- Internet Service
- Contract

C1:VISUAL OF FINDINGS:

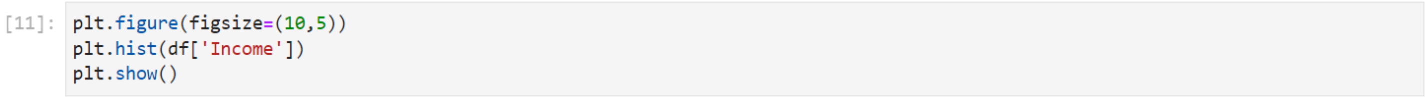
- Continuous Variable (Monthly Charge)



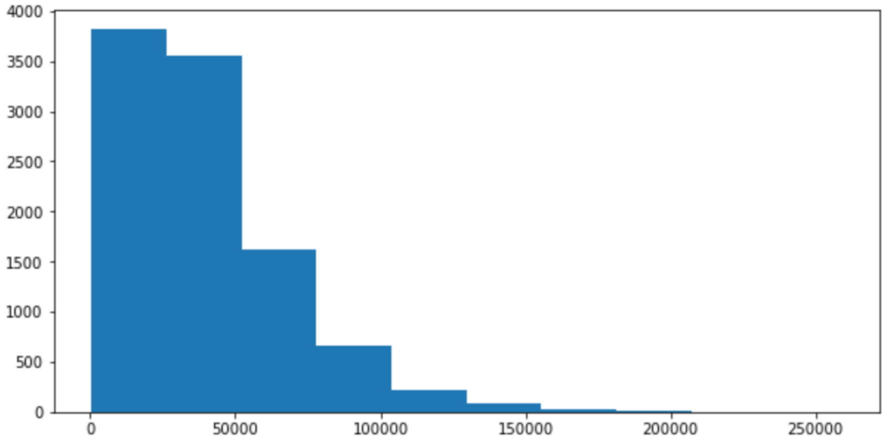
```
[10]: plt.figure(figsize=(10,5))
      plt.hist(df['MonthlyCharge'])
      plt.show()
```



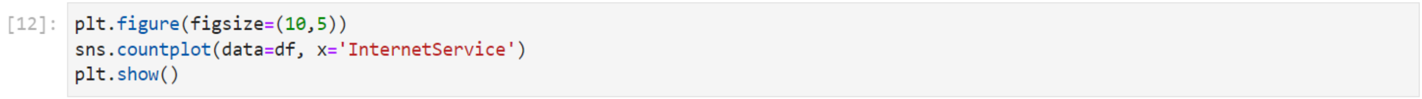
- Continuous Variable (Income)



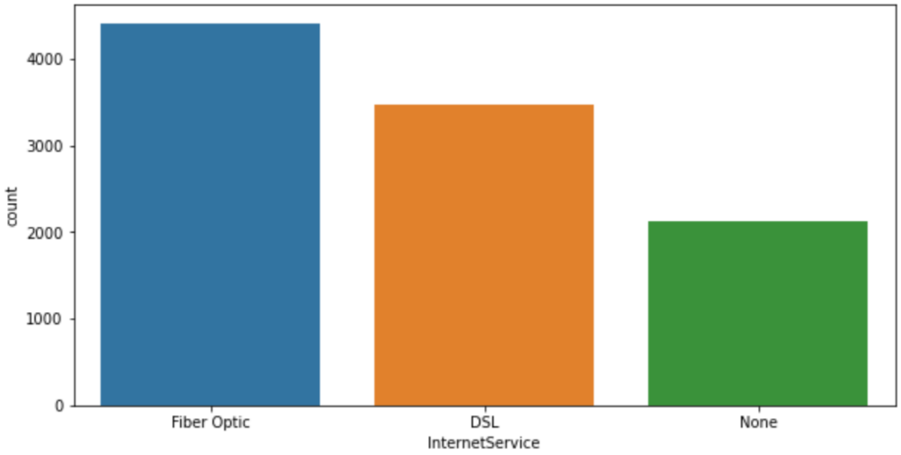
```
[11]: plt.figure(figsize=(10,5))
      plt.hist(df['Income'])
      plt.show()
```



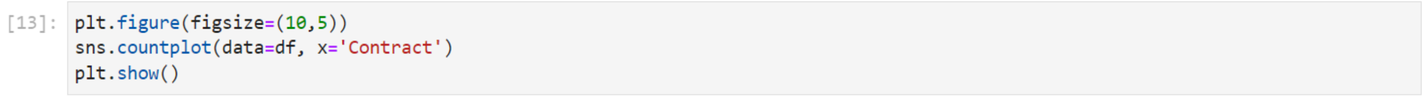
- Categorical variables (InternetService)



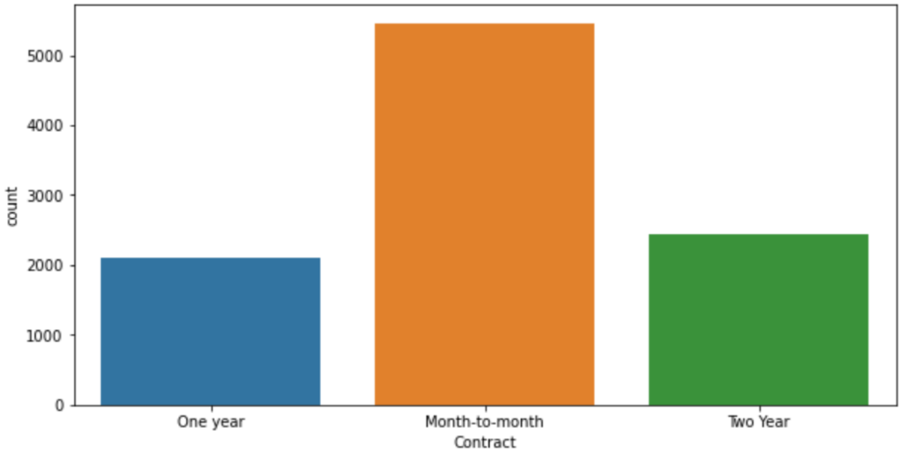
```
[12]: plt.figure(figsize=(10,5))
      sns.countplot(data=df, x='InternetService')
      plt.show()
```



- Categorical variables (Contract)



```
[13]: plt.figure(figsize=(10,5))
      sns.countplot(data=df, x='Contract')
      plt.show()
```



D - BIVARIATE STATISTICS:

Continuous variables:

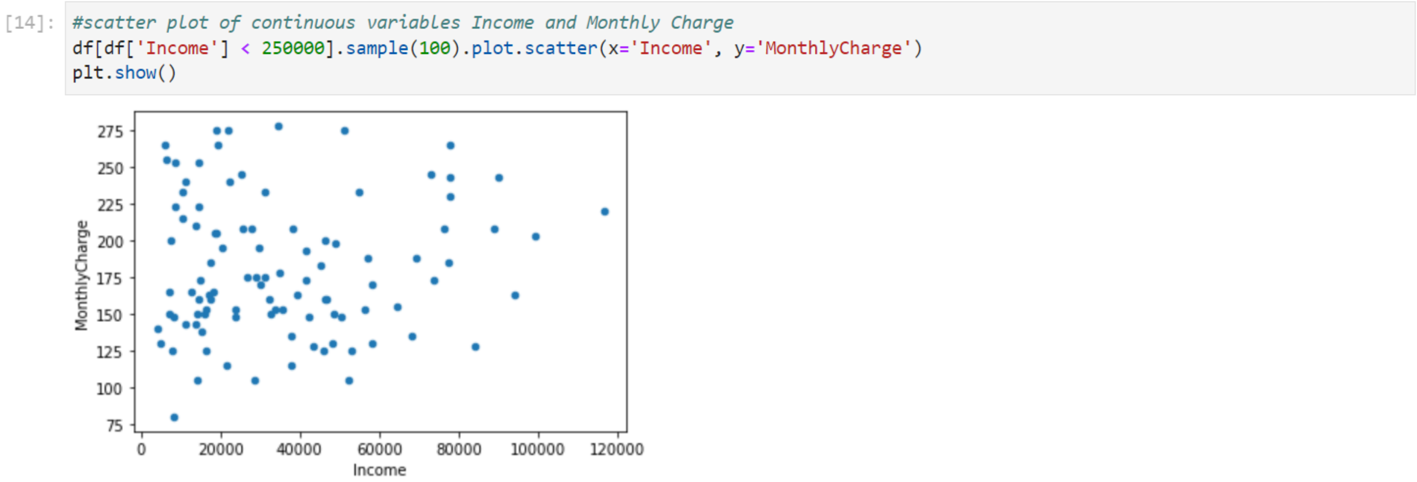
- Monthly Charge
- Income

Categorical variables:

- Churn
- Multiple

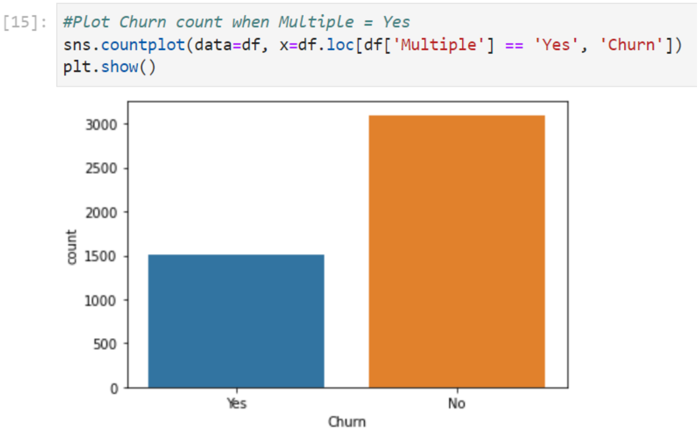
D1 - VISUAL OF FINDINGS:

- Scatter plot for continuous variables Income and Monthly Charge.

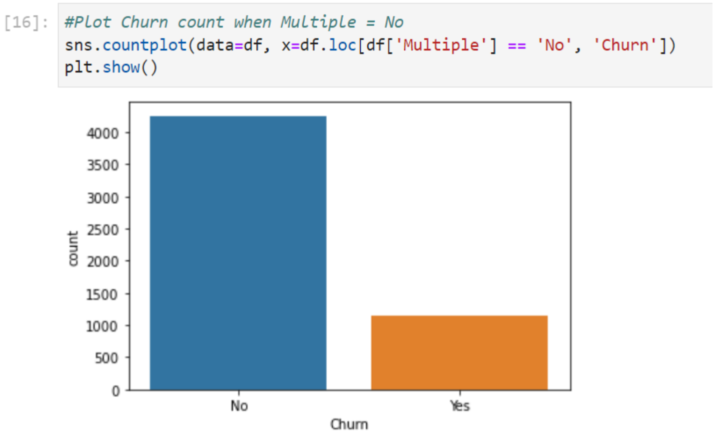


```
[14]: #scatter plot of continuous variables Income and Monthly Charge
      df[df['Income'] < 250000].sample(100).plot.scatter(x='Income', y='MonthlyCharge')
      plt.show()
```

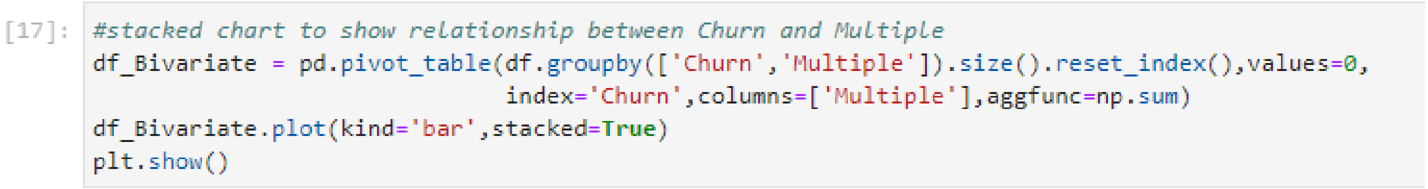
- Show categorical variables the Churn count when Multiple = Yes/No



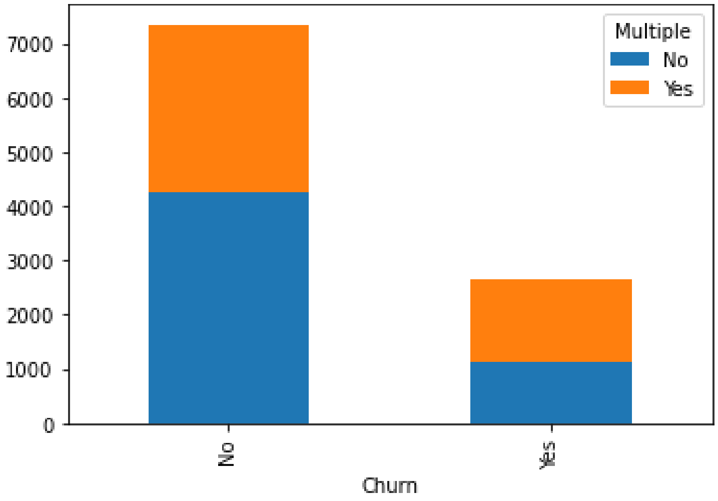
```
[15]: #Plot Churn count when Multiple = Yes
      sns.countplot(data=df, x=df.loc[df['Multiple'] == 'Yes', 'Churn'])
      plt.show()
```



```
[16]: #Plot Churn count when Multiple = No
      sns.countplot(data=df, x=df.loc[df['Multiple'] == 'No', 'Churn'])
      plt.show()
```



```
[17]: #stacked chart to show relationship between Churn and Multiple
      df_Bivariate = pd.pivot_table(df.groupby(['Churn','Multiple']).size().reset_index(),values=0,
                                    index='Churn',columns=['Multiple'],aggfunc=np.sum)
      df_Bivariate.plot(kind='bar',stacked=True)
      plt.show()
```



E1 - RESULTS OF ANALYSIS:

The p-value result from the chi-square test = 0.6318335816054494 and with 0.05 alpha value, we cannot reject the null hypothesis. Given this result, there is no relationship between the response time survey question results and the customer decision whether to churn or not.

E2 - LIMITATIONS OF ANALYSIS:

- With the consideration of the high p-value the effect of the independent variable might exist, but the limitation of the sample data isn’t enough to gather meaningful information.
- Further analysis and gathering more data are required.

E3 - RECOMMENDED COURSE OF ACTION:

As the response to the customer in a timely manner seems to be important but the result of the analysis indicates the need for more data exploration and to continue to analyze different variables to find other insights that can help decision-makers to take the right decision.

F – VIDEO:

Mon Feb 28 2022 4:18:36 PM (panopto.com)

G - SOURCES FOR THIRD-PARTY CODE:

- Bivariate plotting with pandas. Kaggle. Bivariate plotting with pandas | Kaggle
- Python: Correlation and P-value Concepts Python: Correlation and P-value Concepts - YouTube
- Chi Square Test | How to do Bivariate Analysis of Categorical Variables Chi Square Test | How to do Bivariate Analysis of Categorical Categorical Variables - YouTube

H - SOURCES:

- No additional sources or in-text citation were used.