D207
EXPLORATORY DATA ANALYSIS

Performance Assessment Task

WGU - MSDA

Exploratory Data Analysis using the Cleaned Churn Dataset

Richard Flores
rflo147@wgu.edu
Student ID: 006771163

Table of Contents

Part I

Part II

Part III

Part IV

Part V

Part VI

Part I

A1. Question for Analysis

For this assignment, I have opted in favor of using the Telecommunications Churn Database. As stated in the databases' accompanying file Data Cleaning Churn Data Consideration and Dictionary, "Customer 'churn' is defined as the percentage of customers who stopped using a provider's product or service". The objective of this assignment is to provide meaningful and clean data to assess telecommunications organizational need of the understanding of churn as the document goes on to further state, "It costs 10 times more to acquire a new customer than to retain an existing one." (Larose & Larose, 2019) The raw data used in this assignment will come from the provided file 'churn_raw_data.csv'. The question answered in this assignment will be:

1. What are the characteristics of customers who choose to continue telecommunication services with the company and what are common characteristics of customers who terminate services and contribute to churn?

A2. Benefit from Analysis

A thorough analysis of the cleaned churn dataset can allow great insight for stakeholders as to why customers choose to discontinue services with the telecommunications company. A statistical and visual explanation of common trends among customers in the churn pool can immensely help in saving the company time and more importantly money leading to greater revenue in the future.

A3. Data Identification

Successfully answering the question for analysis defined in A1 will require the use of several dependent variables and information discovered from the initial data cleaning. Analyzing the "churn" variable is of paramount importance and is a binary variable that consists of either "Yes" or "No" responses. Data relevant to answering the analysis question and exploratory data analysis include:

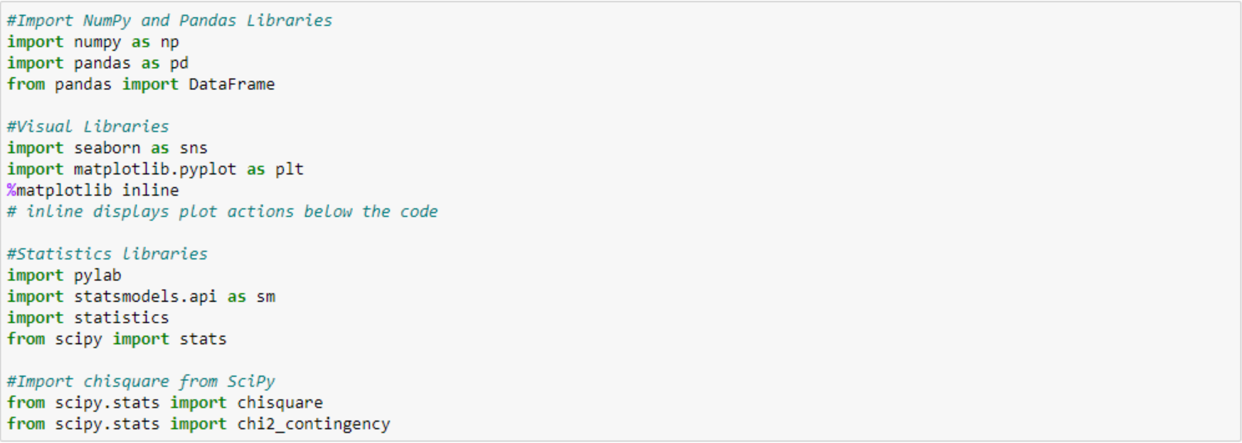
Tenure: Number of months the customer has stayed with the provider
MonthlyCharge: The amount charged to the customer monthly.
Bandwidth_GB_Year: The average amount of data used, in GB, in a year by the customer

Information collected from the PCA analysis will also prove useful in the exploratory analysis and includes the following information from customer surveys: "Timely response, Timely fixes, Timely replacements, Reliability, Options, Respectful response, Courteous exchange, Evidence of active listening."

Part II

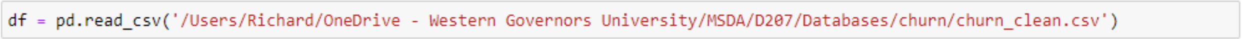
B1. Code

Import Libraries



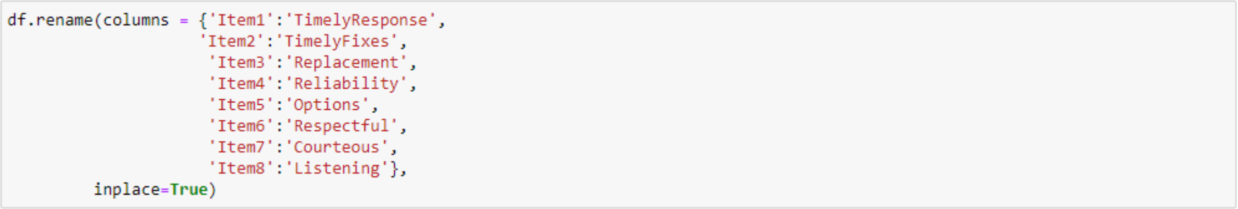
```
#Import NumPy and Pandas Libraries
import numpy as np
import pandas as pd
from pandas import DataFrame

#Visual Libraries
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
# inline displays plot actions below the code

#Statistics libraries
import pylab
import statsmodels.api as sm
import statistics
from scipy import stats

#Import chisquare from SciPy
from scipy.stats import chisquare
from scipy.stats import chi2_contingency
```

Import cleaned dataset into Pandas DataFrame



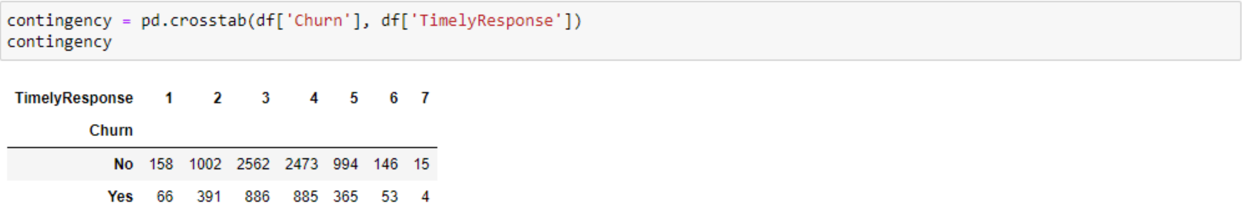
```
df = pd.read_csv('/Users/Richard/OneDrive - Western Governors University/MSDA/D207/Databases/churn/churn_clean.csv')
```

Change the names of Customer Survey features for clarity



```
df.rename(columns = {'Item1':'TimelyResponse',
                    'Item2':'TimelyFixes',
                     'Item3':'Replacement',
                     'Item4':'Reliability',
                     'Item5':'Options',
                     'Item6':'Respectful',
                     'Item7':'Courteous',
                     'Item8':'Listening'},
         inplace=True)
```

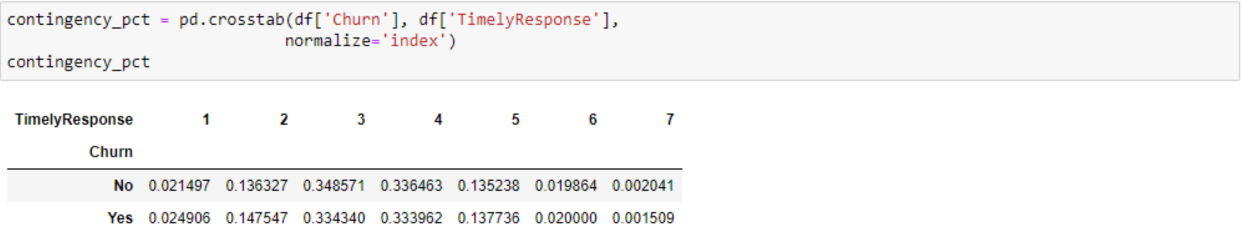
Cross tabulation of Churn & TimelyResponse



```
contingency = pd.crosstab(df['Churn'], df['TimelyResponse'])
contingency
```

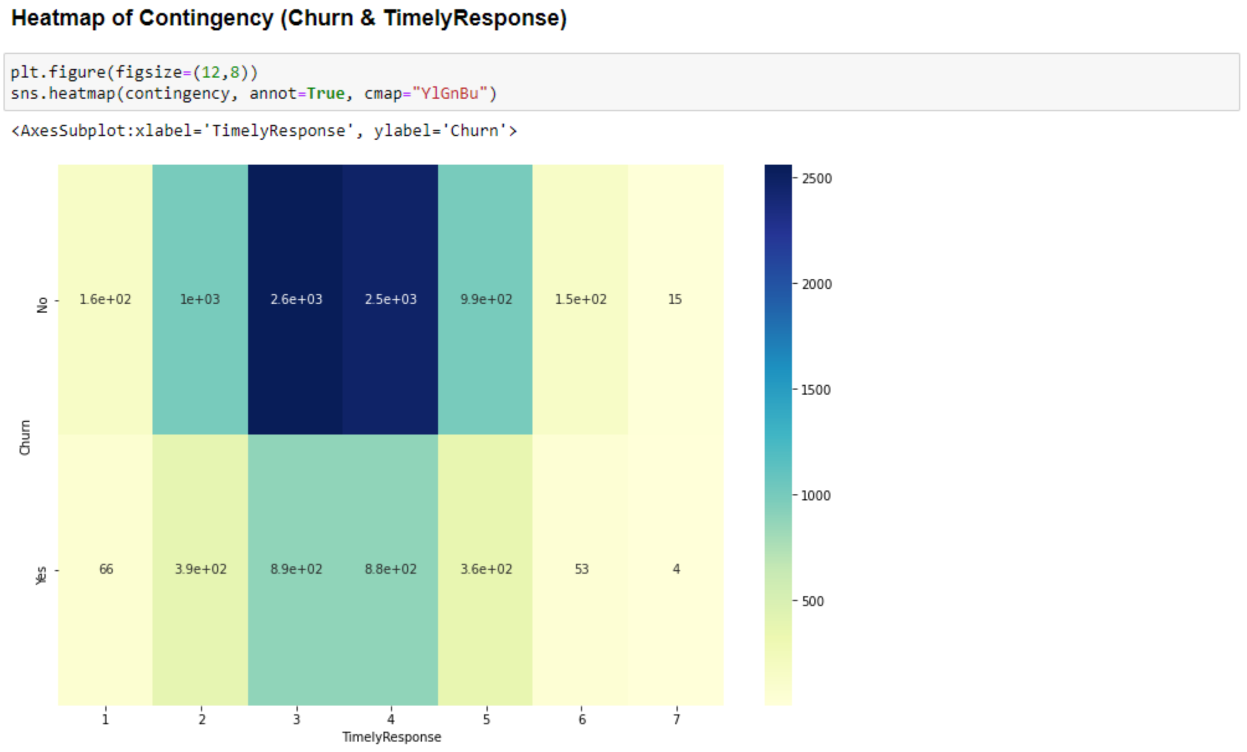
TimelyResponse	1	2	3	4	5	6	7
Churn							
No	158	1002	2562	2473	994	146	15
Yes	66	391	886	885	365	53	4

Cross Tabulation of Churn & TimelyResponse



```
contingency_pct = pd.crosstab(df['Churn'], df['TimelyResponse'],
                              normalize='index')
contingency_pct
```

TimelyResponse	1	2	3	4	5	6	7
Churn							
No	0.021497	0.136327	0.348571	0.336463	0.135238	0.019864	0.002041
Yes	0.024906	0.147547	0.334340	0.333962	0.137736	0.020000	0.001509



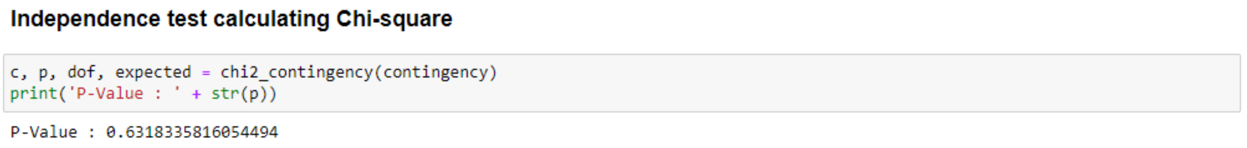
Heatmap of Contingency (Churn & TimelyResponse)

```
plt.figure(figsize=(12,8))
sns.heatmap(contingency, annot=True, cmap="YlGnBu")
```

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

(Chi-Square Test Datacamp 2021)

B2. Output



Independence test calculating Chi-square

```
c, p, dof, expected = chi2_contingency(contingency)
print('P-Value : ' + str(p))
```

```
P-Value : 0.6318335816054494
```

B3. Justification

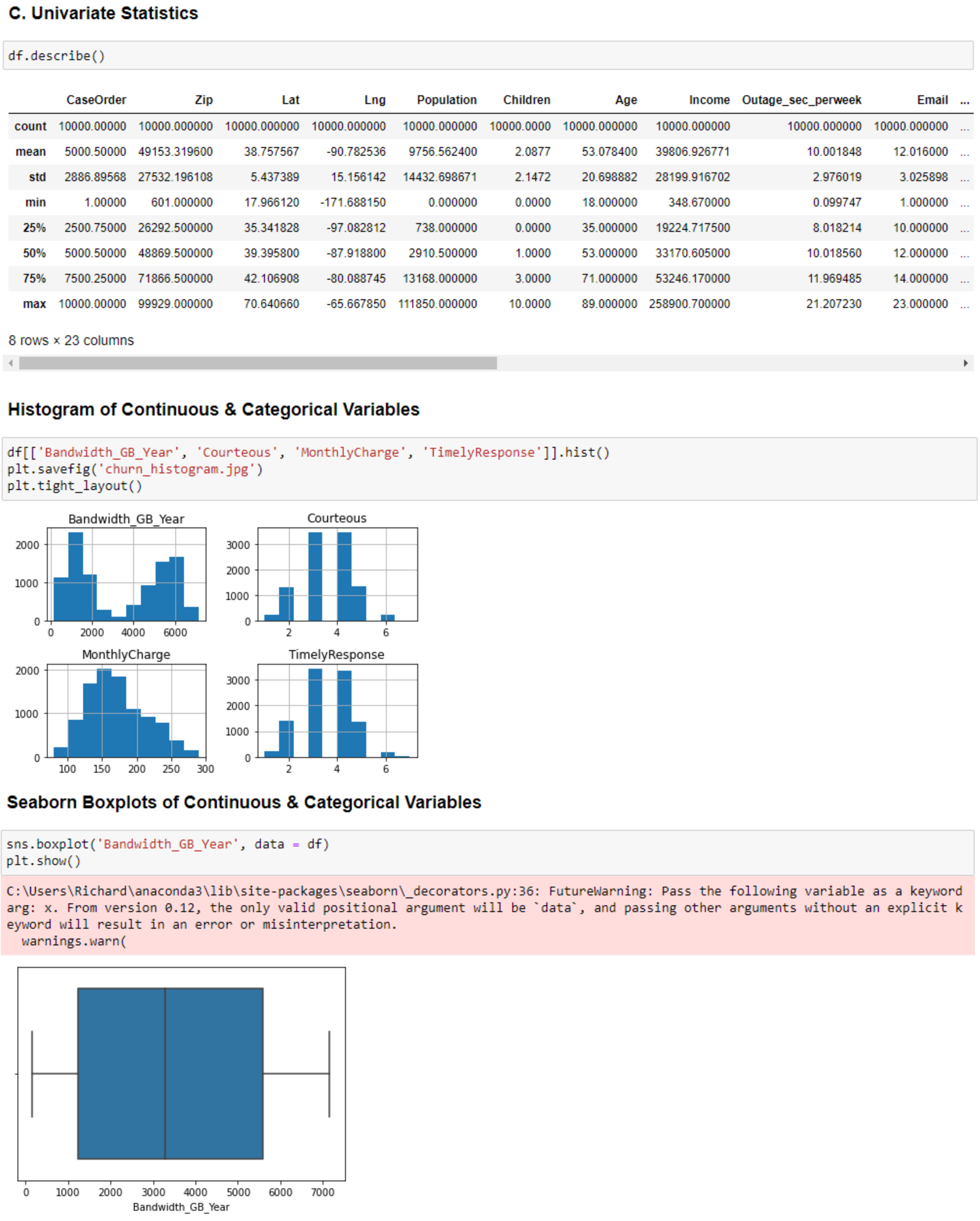
For the analysis, we have the option of choosing from a chi-square test, t-test, or ANOVA. The goal for this exploratory analysis is to define characteristics of customers who contribute to churn and loss of revenue for the telecom company. Churn is a dependent binomial variable with a “Yes” or “No” response. Therefore, as the chi-square test is non-parametric it can be used to find the squared difference among actual and expected data values and diving the result by the expected values. ‘Item1’ is the ordinal level for categorical variables (Larose 2021).

Part III

C1. Univariate Statistics - Visual of Findings

Continuous Variables: Monthly Charge, Bandwidth_GB_Year

Categorical Variables: TimelyResponse, Courteous (Chi-Square Test 2021)



C. Univariate Statistics

```
df.describe()
```

	CaseOrder	Zip	Lat	Lng	Population	Children	Age	Income	Outage_sec_perweek	Email	...
count	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.0000	10000.000000	10000.000000	10000.000000	10000.000000	...
mean	5000.50000	49153.319600	38.757567	-90.782536	9756.562400	2.0877	53.078400	39806.926771	10.001848	12.016000	...
std	2886.89568	27532.196108	5.437389	15.156142	14432.698671	2.1472	20.698882	28199.916702	2.976019	3.025898	...
min	1.00000	601.000000	17.966120	-171.688150	0.000000	0.0000	18.000000	348.670000	0.099747	1.000000	...
25%	2500.75000	26292.500000	35.341828	-97.082812	738.000000	0.0000	35.000000	19224.717500	8.018214	10.000000	...
50%	5000.50000	48869.500000	39.395800	-87.918800	2910.500000	1.0000	53.000000	33170.605000	10.018560	12.000000	...
75%	7500.25000	71866.500000	42.106908	-80.088745	13168.000000	3.0000	71.000000	53246.170000	11.969485	14.000000	...
max	10000.00000	99929.000000	70.640660	-65.667850	111850.000000	10.0000	89.000000	258900.700000	21.207230	23.000000	...

8 rows × 23 columns

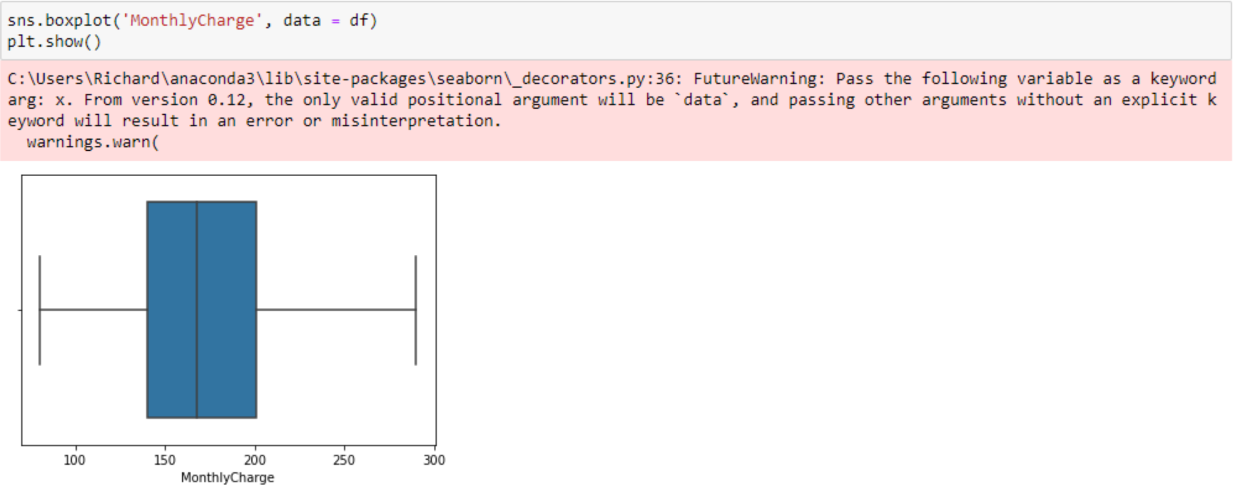
Histogram of Continuous & Categorical Variables

```
df[['Bandwidth_GB_Year', 'Courteous', 'MonthlyCharge', 'TimelyResponse']].hist()
plt.savefig('churn_histogram.jpg')
plt.tight_layout()
```

Seaborn Boxplots of Continuous & Categorical Variables

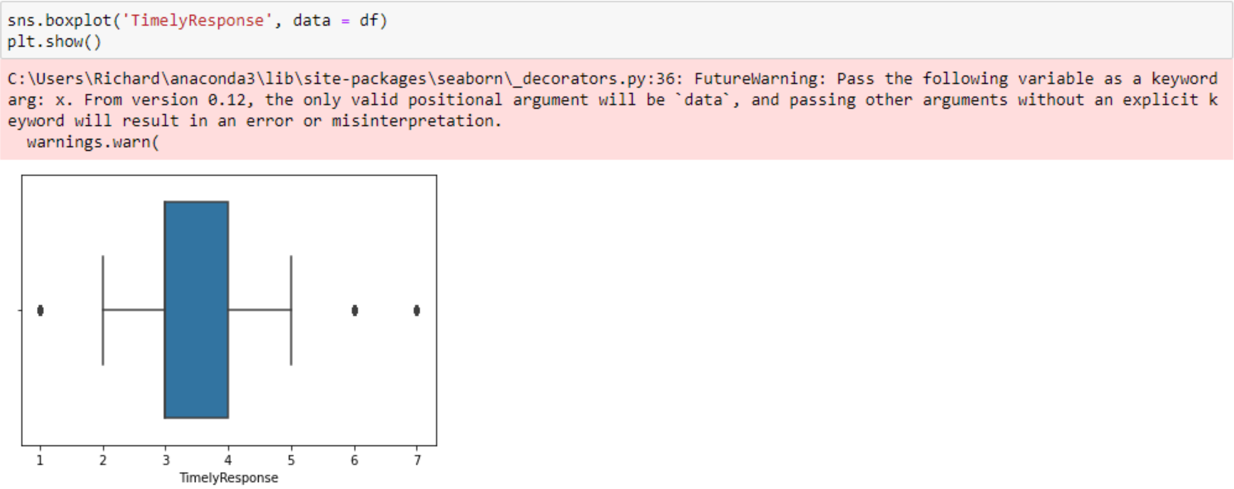
```
sns.boxplot('Bandwidth_GB_Year', data = df)
plt.show()
```

```
C:\Users\Richard\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword
arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit k
eyword will result in an error or misinterpretation.
  warnings.warn(
```



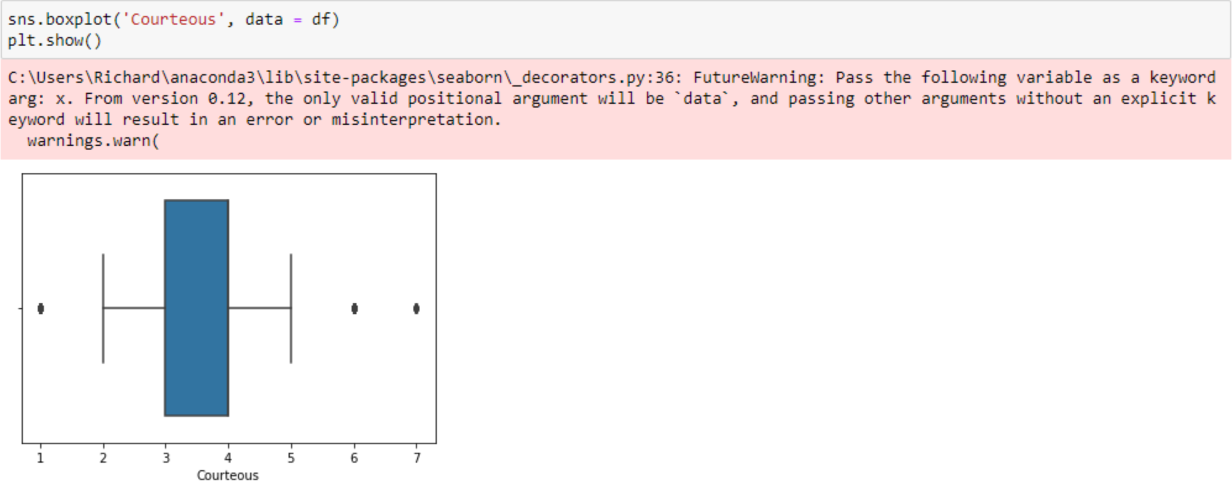
```
sns.boxplot('MonthlyCharge', data = df)
plt.show()
```

```
C:\Users\Richard\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword
arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit k
eyword will result in an error or misinterpretation.
  warnings.warn(
```



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

```
C:\Users\Richard\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword
arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit k
eyword will result in an error or misinterpretation.
  warnings.warn(
```



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

```
C:\Users\Richard\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword
arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit k
eyword will result in an error or misinterpretation.
  warnings.warn(
```

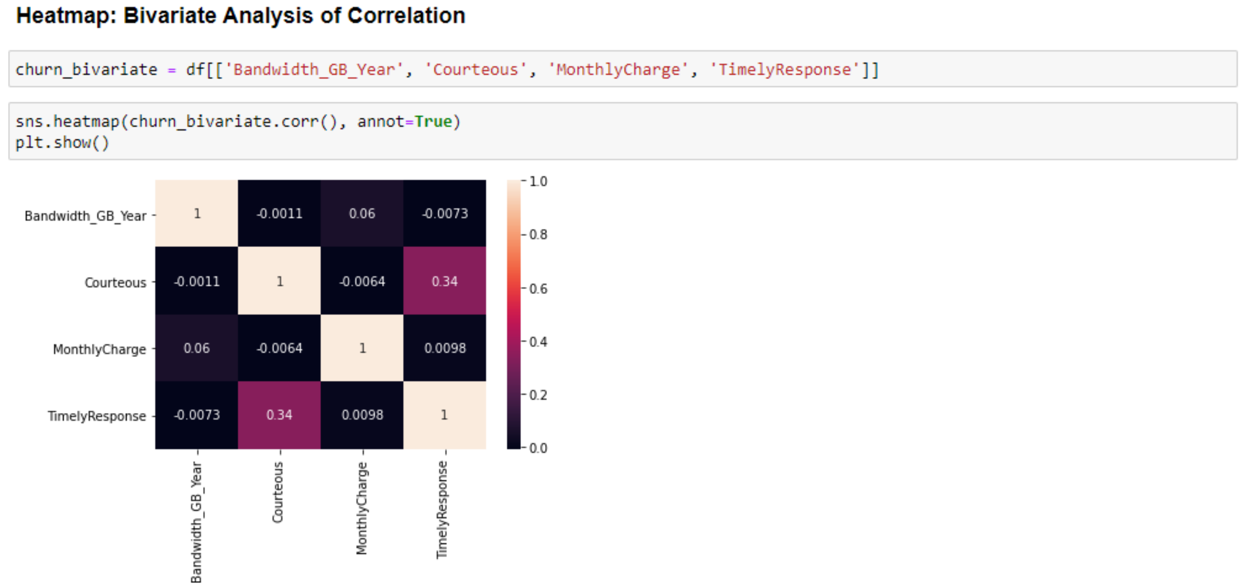
(Exploratory Data Analysis 2021)

Part IV

D1. Bivariate Statistics - Visual of Findings

Continuous Variables: MonthlyCharge, Bandwidth_GB_Year

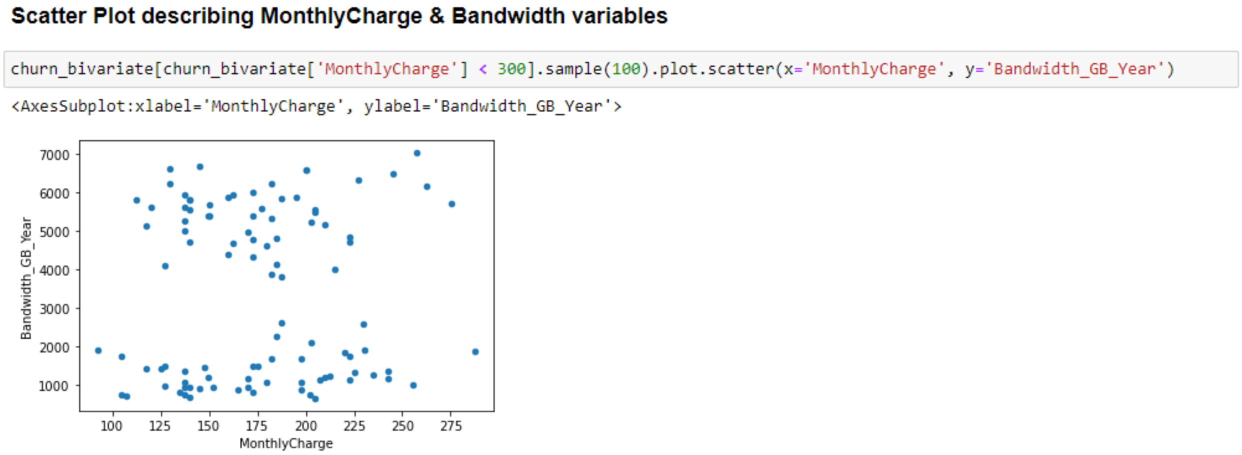
Categorical Variables: Churn (binomial), Courteous (ordinal) (Bivariate Plots 2021)



Heatmap: Bivariate Analysis of Correlation

```
churn_bivariate = df[['Bandwidth_GB_Year', 'Courteous', 'MonthlyCharge', 'TimelyResponse']]
```

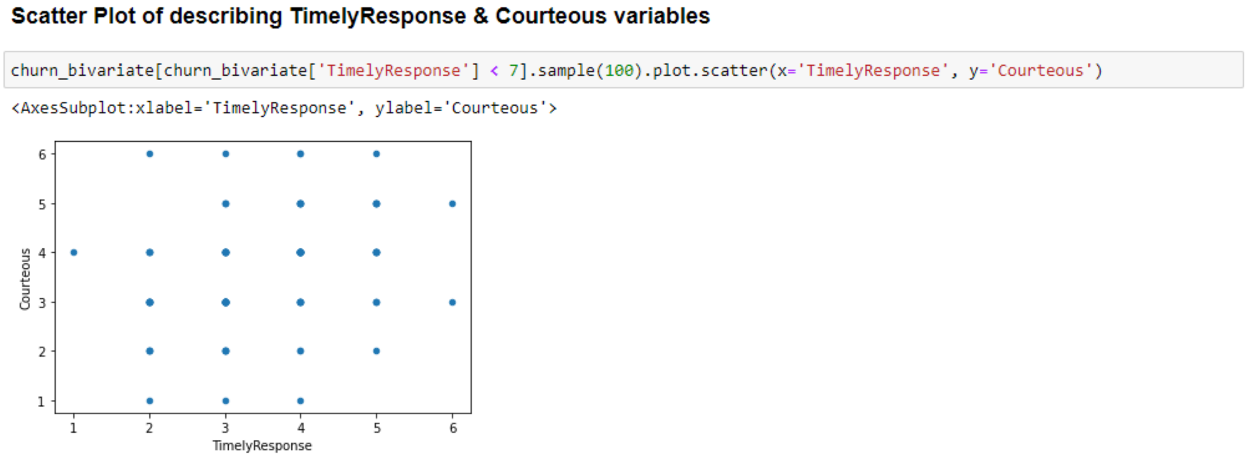
```
sns.heatmap(churn_bivariate.corr(), annot=True)
plt.show()
```



Scatter Plot describing MonthlyCharge & Bandwidth variables

```
churn_bivariate[churn_bivariate['MonthlyCharge'] < 300].sample(100).plot.scatter(x='MonthlyCharge', y='Bandwidth_GB_Year')
```

```
<AxesSubplot:xlabel='MonthlyCharge', ylabel='Bandwidth_GB_Year'>
```



Scatter Plot of describing TimelyResponse & Courteous variables

```
churn_bivariate[churn_bivariate['TimelyResponse'] < 7].sample(100).plot.scatter(x='TimelyResponse', y='Courteous')
```

```
<AxesSubplot:xlabel='TimelyResponse', ylabel='Courteous'>
```

Part V

E1. Result of Analysis

The exploratory analysis was based on observations from a Chi-Square test. From the chi-square test, I chose a test for independence rather than a test for goodness of fit. Using the independence chi-square test allows us to compare two variables in a contingency table to gain insight into their relationship. The independence test displays whether distributions of categorical variables differ from each other.

The Chi-Square function imported from the SciPy library uses the following formula for calculation:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where the subscript “c’ is degrees of freedom, “O” is the observed value, and “E” is the expected value (Chi-Square Statistics 2021).

For the contingency test we invoked the crosstab function which calculates results using the following formula:

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Where “O” is the observed value, “E” is the expected value, and “I” is the “ith” position in the contingency table (Chi-Square Statistics 2021).

Using a p-value of the same magnitude as the result of the chi-square independence test, the p-value = 0.63183 cannot be rejected in a null hypothesis under the 0.05 alpha level.

Given the available data, it is not clear if there are statistically significant values that prove a correlation between the observed variables in survey responses. We cannot state with confidence the results of the customer satisfaction survey reflect the customers’ decision to discontinue service with the Telcom company.

E2. Limitations of Analysis

The strength of the chi-square test is the ease of statistical computation in Python and R. The independence test measures data on a nominal scale and can be used to identify differences in two groups of variables. (When Chi-square is appropriate 2021).

However, with such a high p-value, it is obvious that we must do more research and acquire more meaningful and insightful data. The churn dataset has a very limited ability to produce actionable results and practical information (Larose 2021).

E3. Recommended Course of Action

The results show that there is no statistically definable correlation between the variables involved in the measurement of customer satisfaction relating to survey responses i.e., Monthly Charge, Bandwidth_GB_Year, Timely Response, and Courtesy affecting churn. There is little to no evidence of a linear relationship among these variables, however, the results visually describe dissatisfaction in these areas may contribute to churn.

While there is no definite correlation between the survey data and customer churn, it is safe to recommend that effort and time spent on increasing performance in these areas and increasing customer satisfaction will contribute to reduced churn and a reduction in loss of profits for the company.

Part VI

F. Panopto Recording

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c00d7dc0-cc47-45af-bf08-adc70073e9ab

G. Sources for Third-Party Code

Chi-Square Test. Python. (n.d.). Retrieved October 22, 2021, from https://campus.datacamp.com/courses/experimental-design-in-python/the-basics-of-statistical-hypothesis-testing?ex=10.

Bivariate plots in Pandas. Python. (n.d.). Retrieved October 22, 2021, from https://campus.datacamp.com/courses/python-for-r-users/plotting-4?ex=3.

Exploratory Data Analysis in python. DataCamp. (n.d.). Retrieved October 22, 2021, from https://www.datacamp.com/courses/exploratory-data-analysis-in-python#!

H. Sources and References

Larose, C. D., & Larose, D. T. (2019). *Data Science using python and R*. Wiley.

Chi-square statistic: How to calculate it / distribution. Statistics How To. (2021, September 30). Retrieved October 22, 2021, from https://www.statisticshowto.com/probability-and-statistics/chi-square/.

When Chi-square is appropriate - strengths/weaknesses. passel. (n.d.). Retrieved October 22, 2021, from https://passel2.unl.edu/view/lesson/9beaa382bf7e/14.