D207_Performance_Assessment

September 22nd,2021

Table of Contents

Performance Evaluation of Exploratory Data Analysis, D207

Name: Rekha Alle
Student ID: 000778673
Course: Masters Data Analytics
Date: 09/22/2021
Program Mentor: David Gagner
Contact: 3854282643
Email: David.gagner@wgu.edu

Introduction to Exploratory Data Analysis (EDA)

Artificial intelligence, Big-data, machine centric decision making, and machine learning are all buzzwords in today's world. All these strategies or methodologies necessitate a significant amount of statistical and analytical effort, as well as a sample dataset and model training. Data analytics is concerned with several aspects of the data/business decision pipeline, starting with data collecting and ending with data conclusion and report development. However, there are a few things to keep in mind when designing and deploying an analytical model. Some of them are – the data has been acquired in its entirety and is free of anomalies and outliers. The next stage in the data analysis process would be to search for early trends in the collected historical data and conduct preliminary analysis

Exploratory Data Analysis (EDA) is a series of stages and processes that involves spotting data anomalies, checking assumptions with statistics, central tendency, and testing hypothesis in its most basic form. We also look at how to portray data in a graphical way for easier comprehension and how to start making sense of the information acquired.

A: Analytical Question for Research

Exploratory data analysis, as shown in the course, explores the skills and techniques required to do parametric or non-parametric hypothesis testing, identify the distribution of acquired data, compute central tendency, and visually plot the data. I've decided to use the churn data set for this assessment. In business, churn is defined as the percentage of customers who cancel their subscriptions or whose accounts do not need to be renewed after a certain period has gone. Churn can be defined as the number of people who discontinue using a product.

In this D207- assessment, I will analyze a churn dataset based on real-world questions such as: Do parameters such as streaming TV, streaming movies and having several lines have a dependency on churn? This will help us better understand the customer response and reduce churn rate over time. For the sake of this examination, we shall employ the Chi-Square test as a technique. In the category level the variable is measured at the nominal level, the chi-square statistic is utilized. The Chi-Square statistic is used when the variable is measured at the nominal (also known as category) level.[1]

A1. Analytical Question:

Is there a dependency between churn and other factors like streaming TV, having several lines, and streaming movies?

A2. Advantages/Benefits from Analysis:

collaborators in the organization will gain from knowing, with some certainty, which customers are most likely to churn, as this will provide weight to selling enhanced services to consumers with these traits and previous user experiences.

We will be able to answer several questions with a high level of confidence by using the Chi-Square test on this sample dataset, such as: Does any of the streaming services or having numerous lines affect the likelihood of retaining a customer

A3. Identification of the data:

"Churn," is the controlled variable with only two values, which is binary categorical, "Yes" or "No," is most significant to our decision-making process. The data columns "Tenure" has continuous numerical (the number of months the customer has been with the provider), "Monthly Charge" (the average monthly charge to the customer), and "Bandwidth GB Year" were identified to be relevant while cleaning the data (the average annual amount of data utilized by customer, in GB.). Finally, discrete numerical data derived from consumer survey responses on various customer service elements is very helpful in the decision-making process. Customers has ordinal numerical data by ranking 8 customers in the surveys. This analysis is conducted based on the limited data points as per the churn raw data file.

Following are Variables of customer service factors based on customer survey range between 8 to 1 (8= Most important, 1 = least important):

- Active_Listening
- Courteous_exchange
- Respectful_Response
- Options
- Reliability
- Timely_Replacements
- Timely_Fixes
- Timely_Responses

B: Code Analysis

B1. Describe the Chi-Square testing process:

1. Explanation of Approach:

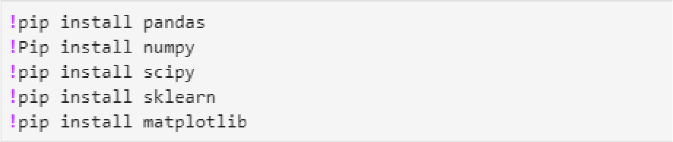
For this assessment we are using Chi-Square testing used to see if there's a link between categorical variables. The Chi-Square test's null hypothesis is that there is no link between the categorical variables in the churn; they are independent

2. Tools will be used:

I'll use Python because I have some experience with it, having spent Artificial Intelligence and ML independently over the last year before starting this master's degree, and because of its ability to do a lot of things "out of the box" features (Poulson, 2016, section 2). Python has become widely used in the data science field because of its clear, intuitive, and understandable syntax. In addition, I find Jupyter notebooks to be a useful way to execute code visually, thanks to its appealing single-document markdown style, ability to display code and graphic visualization results, and crystal-clear running documentation for future reference. Rather than building code from the ground up, a Python installation, packages and libraries will supply custom created code to perform complex data science tasks.
NumPy used to work with arrays,
Pandas used to load datasets,
Matplotlib used to plot charts,
Scikit-learn used for machine learning model classes,
SciPy used for mathematical problems, specifically linear algebra transformations, and
Seaborn used for a high-level interface and appealing visualizations.
Using the Pandas library and its accompanying "read csv" function to transform our data as a dataframe is a quick, exact example of loading a dataset and constructing a variable efficiently: imported pandas as pd, df(dataframe) = pd.read csv('ChurnData.csv')

3. Enter the Code:

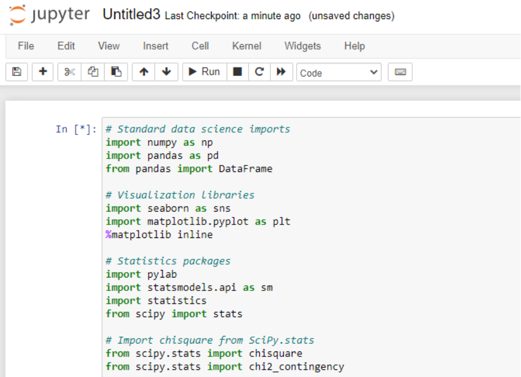
3.1 Install the required packages:



```
!pip install pandas
!Pip install numpy
!pip install scipy
!pip install sklearn
!pip install matplotlib
```

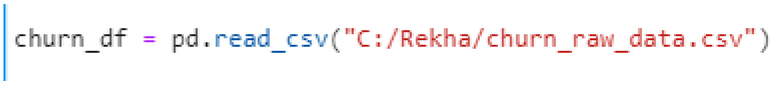
```
Requirement already satisfied: pandas in c:\users\kaila\anaconda3\lib\site-packages (1.2.4)
Requirement already satisfied: pytz>=2017.3 in c:\users\kaila\anaconda3\lib\site-packages (from pandas)
(2021.1)
Requirement already satisfied: numpy>=1.16.5 in c:\users\kaila\anaconda3\lib\site-packages (from pandas)
(1.20.1)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\kaila\anaconda3\lib\site-packages (from
pandas) (2.8.1)
Requirement already satisfied: six>=1.5 in c:\users\kaila\anaconda3\lib\site-packages (from python-
dateutil>=2.7.3->pandas) (1.15.0)
Requirement already satisfied: numpy in c:\users\kaila\anaconda3\lib\site-packages (1.20.1)
Requirement already satisfied: scipy in c:\users\kaila\anaconda3\lib\site-packages (1.6.2)
Requirement already satisfied: numpy<1.23.0,>=1.16.5 in c:\users\kaila\anaconda3\lib\site-packages (from
scipy) (1.20.1)
Requirement already satisfied: sklearn in c:\users\kaila\anaconda3\lib\site-packages (0.0)
Requirement already satisfied: scikit-learn in c:\users\kaila\anaconda3\lib\site-packages (from sklearn)
(0.24.1)
Requirement already satisfied: joblib>=0.11 in c:\users\kaila\anaconda3\lib\site-packages (from scikit-
learn->sklearn) (1.0.1)
Requirement already satisfied: scipy>=0.19.1 in c:\users\kaila\anaconda3\lib\site-packages (from scikit-
learn->sklearn) (1.6.2)
Requirement already satisfied: numpy>=1.13.3 in c:\users\kaila\anaconda3\lib\site-packages (from scikit-
learn->sklearn) (1.20.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\kaila\anaconda3\lib\site-packages (from
scikit-learn->sklearn) (2.1.0)
Requirement already satisfied: matplotlib in c:\users\kaila\anaconda3\lib\site-packages (3.3.4)
Requirement already satisfied: numpy>=1.15 in c:\users\kaila\anaconda3\lib\site-packages (from
matplotlib) (1.20.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\kaila\anaconda3\lib\site-packages (from
matplotlib) (8.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\kaila\anaconda3\lib\site-packages (from
matplotlib) (0.10.0)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\kaila\anaconda3\lib\site-packages (from
matplotlib) (2.8.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\users\kaila\anaconda3\lib\site-
packages (from matplotlib) (2.4.7)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\kaila\anaconda3\lib\site-packages (from
matplotlib) (1.3.1)
Requirement already satisfied: six in c:\users\kaila\anaconda3\lib\site-packages (from cycler>=0.10-
>matplotlib) (1.15.0)
```

3.2 Include standard imports all the required references:



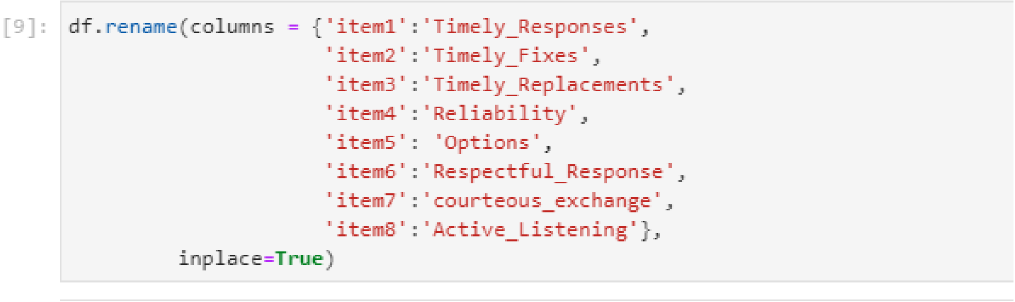
```
In [*]: # Standard data science imports
        import numpy as np
        import pandas as pd
        from pandas import DataFrame

        # Visualization libraries
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline

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

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

3.3 Using pandas read the data from raw data file:



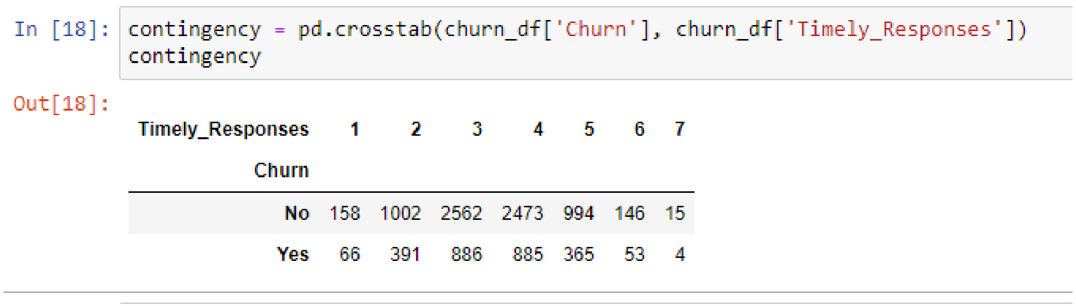
```
churn_df = pd.read_csv("C:/Rekha/churn_raw_data.csv")
```

3.4 Change the names of the last eight survey columns to better describe the variables:



```
[9]: df.rename(columns = {'item1':'Timely_Responses',
                          'item2':'Timely_Fixes',
                          'item3':'Timely_Replacements',
                          'item4':'Reliability',
                          'item5': 'Options',
                          'item6':'Respectful_Response',
                          'item7':'courteous_exchange',
                          'item8':'Active_Listening'},
               inplace=True)
```

3.5 Data shows frequency between Churn and Timely Responses:

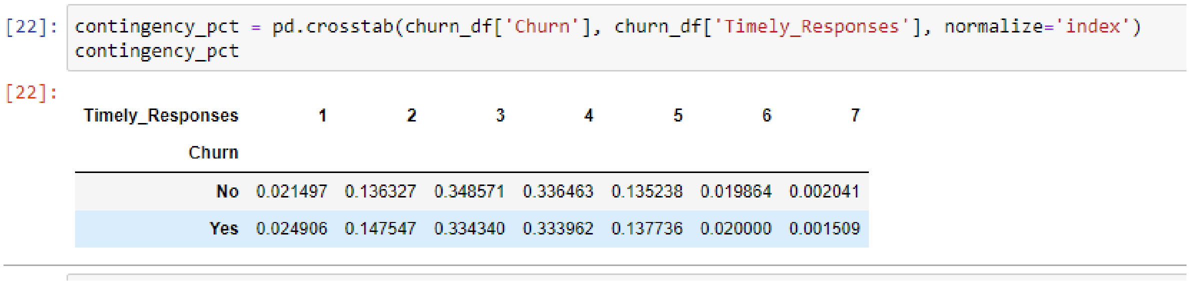


In [18]:
```
contingency = pd.crosstab(churn_df['Churn'], churn_df['Timely_Responses'])
contingency
```

Out[18]:

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

3.6 To List the normalized contingency of Churn and Timely Responses:

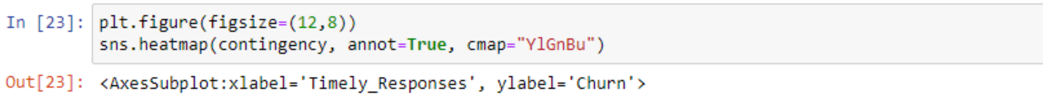


[22]:
```
contingency_pct = pd.crosstab(churn_df['Churn'], churn_df['Timely_Responses'], normalize='index')
contingency_pct
```

[22]:

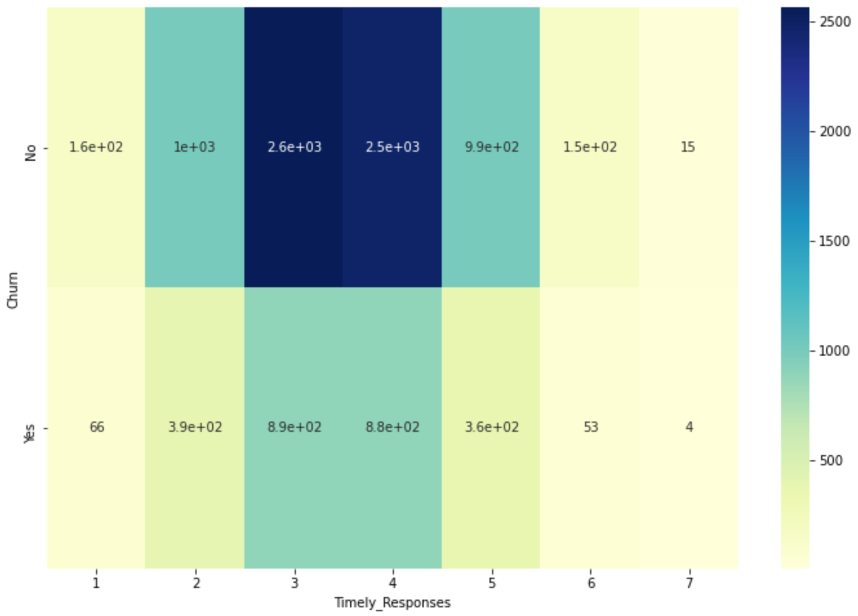
Timely_Responses	1	2	3	4	5	6	7
Churn							
No	0.021497	0.136327	0.348571	0.336463	0.135238	0.019864	0.002041
Yes	0.024906	0.147547	0.334340	0.333962	0.137736	0.020000	0.001509

3.7 Heatmap:



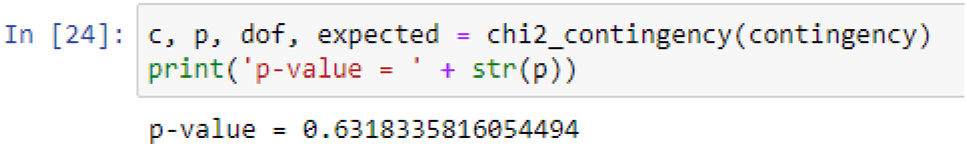
In [23]:
```
plt.figure(figsize=(12,8))
sns.heatmap(contingency, annot=True, cmap="YlGnBu")
```

Out[23]:
```
<AxesSubplot:xlabel='Timely_Responses', ylabel='Churn'>
```



B2: Output/Return Analysis

1.1 Independence test of Chi-Square Analysis:



```
In [24]: c, p, dof, expected = chi2_contingency(contingency)
         print('p-value = ' + str(p))

         p-value = 0.6318335816054494
```

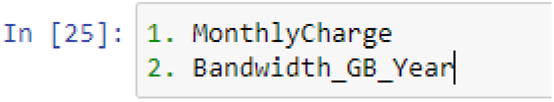
P-Value using Chi-Square testing analysis ²

B3: Foundation of Analysis

We're looking at churn from a telecom firm, the key analysis ("whether customers stay or quit the company?") in our study. "**Churn**" is a categorical dependent variable with a binomial distribution. For this "yes/no" target variable use non-parametric test, we will utilize chi-square testing. "**Timely_Responses**," our other categorical variable, is at the ordinal level.

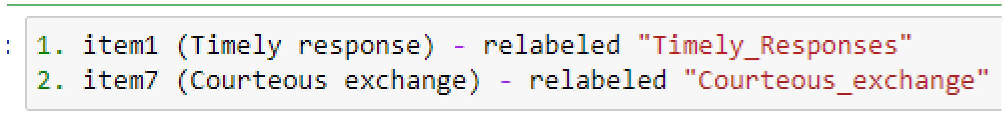
C: Statistics using Only One Variable

1.1 These are the continuous variables for this analysis:



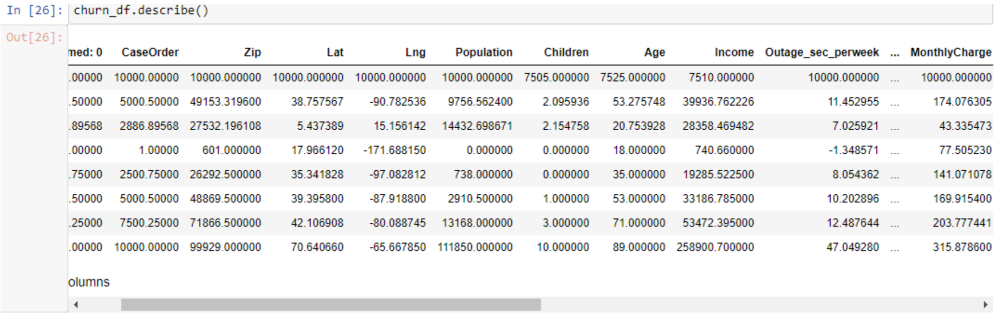
```
In [25]: 1. MonthlyCharge
         2. Bandwidth_GB_Year
```

2.1 These are the Ordinal variables for the analysis:



```
: 1. item1 (Timely response) - relabeled "Timely_Responses"
  2. item7 (Courteous exchange) - relabeled "Courteous_exchange"
```

3.1 Listing the Data Frame:



```
In [26]: churn_df.describe()
```

Out[26]:

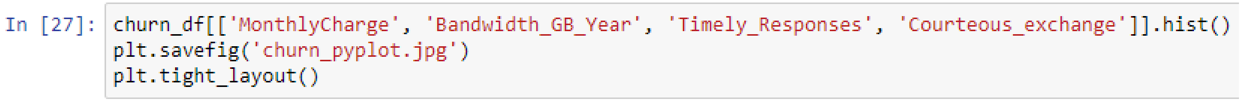
med: 0	CaseOrder	Zip	Lat	Lng	Population	Children	Age	Income	Outage_sec_perweek	...	MonthlyCharge
.00000	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	7505.000000	7525.000000	7510.000000	10000.000000	...	10000.000000
.50000	5000.50000	49153.319600	38.757567	-90.782536	9756.562400	2.095936	53.275748	39936.762226	11.452955	...	174.076305
.89568	2886.89568	27532.196108	5.437389	15.156142	14432.698671	2.154758	20.753928	28358.469482	7.025921	...	43.335473
.00000	1.00000	601.000000	17.966120	-171.688150	0.000000	0.000000	18.000000	740.660000	-1.348571	...	77.505230
.75000	2500.75000	26292.500000	35.341828	-97.082812	738.000000	0.000000	35.000000	19285.522500	8.054362	...	141.071078
.50000	5000.50000	48869.500000	39.395800	-87.918800	2910.500000	1.000000	53.000000	33186.785000	10.202896	...	169.915400
.25000	7500.25000	71866.500000	42.106908	-80.088745	13168.000000	3.000000	71.000000	53472.395000	12.487644	...	203.777441
.00000	10000.00000	99929.000000	70.640660	-65.667850	111850.000000	10.000000	89.000000	258900.700000	47.049280	...	315.878600

olumns

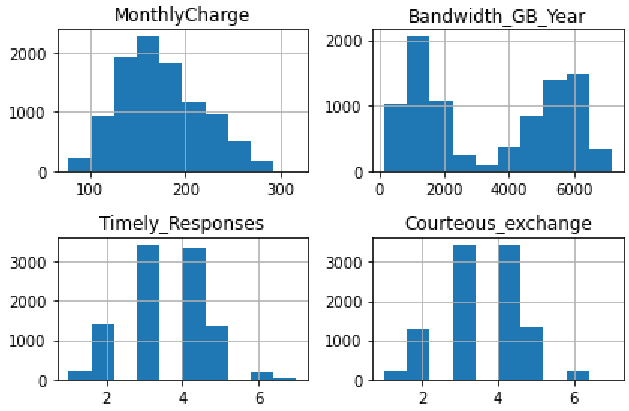
C1: Visualization of the Results

1.1 Data Visualization & Anomaly Detection:

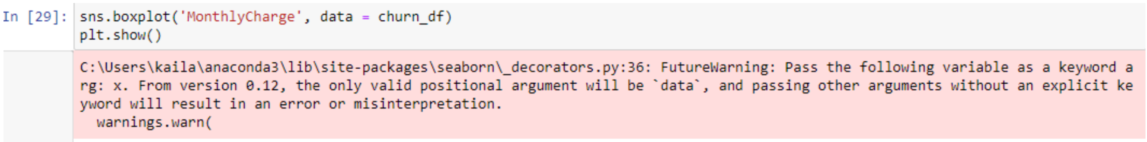
2.1 Make histograms of critical variables:



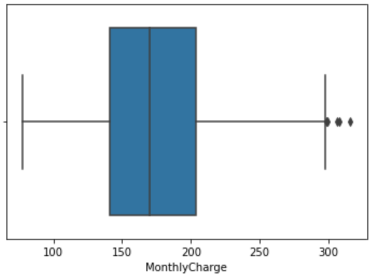
```
In [27]: churn_df[['MonthlyCharge', 'Bandwidth_GB_Year', 'Timely_Responses', 'Courteous_exchange']].hist()
         plt.savefig('churn_pyplot.jpg')
         plt.tight_layout()
```

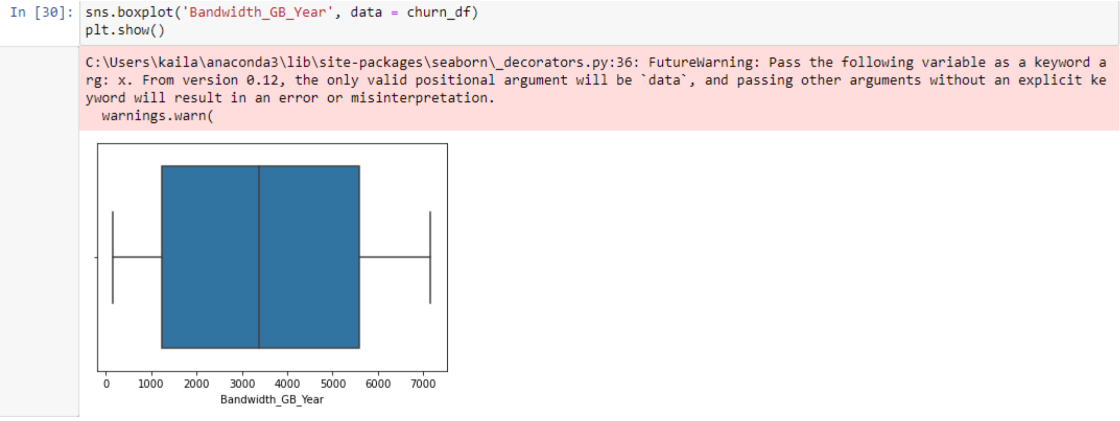


3.1 Let's have a look at the Seaborn boxplot charge and bandwidth:



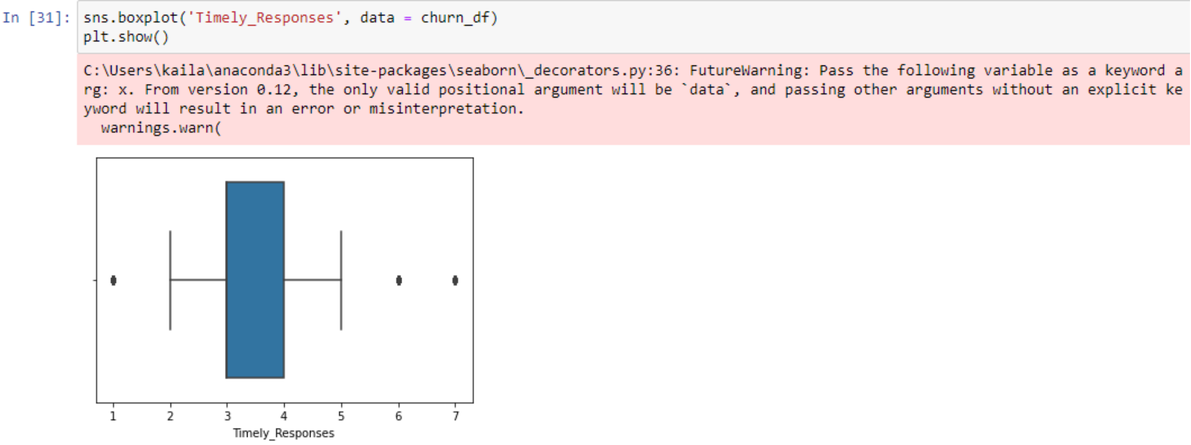
```
In [29]: sns.boxplot('MonthlyCharge', data = churn_df)
         plt.show()

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





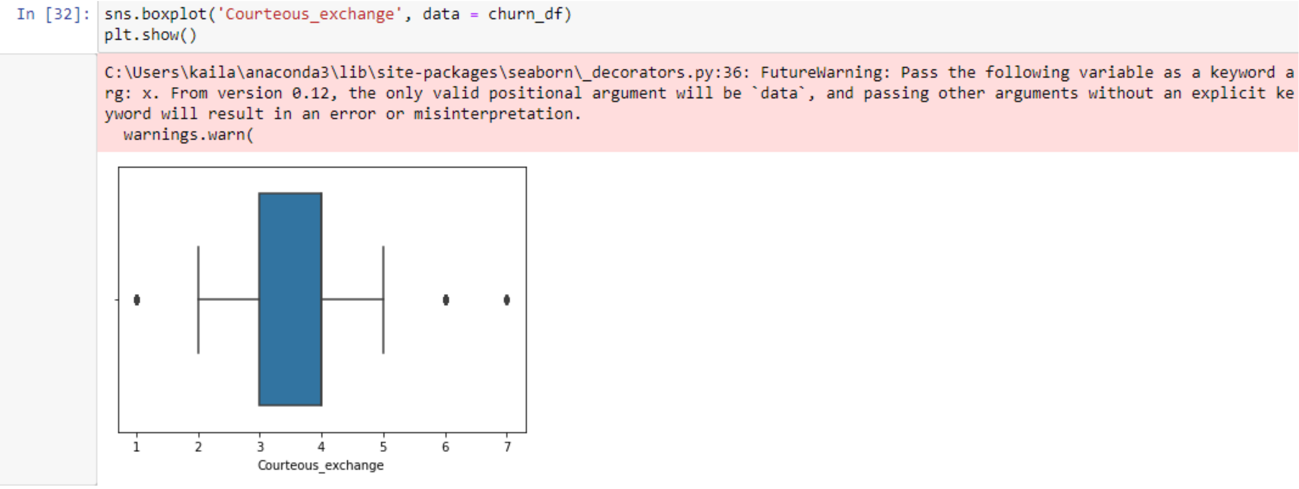
```
In [30]: sns.boxplot('Bandwidth_GB_Year', data = churn_df)
         plt.show()
```

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



```
In [31]: sns.boxplot('Timely_Responses', data = churn_df)
         plt.show()
```

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

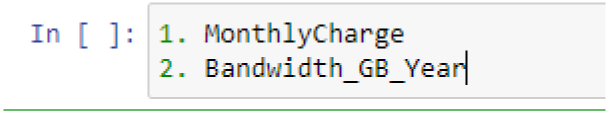


```
In [32]: sns.boxplot('Courteous_exchange', data = churn_df)
         plt.show()
```

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

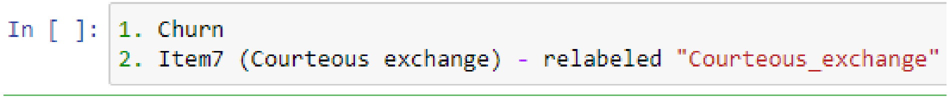
D: Bivariate Statistics are statistics that have two variables.

1.1 These are two continuous variables for this analysis:



```
In [ ]: 1. MonthlyCharge
        2. Bandwidth_GB_Year
```

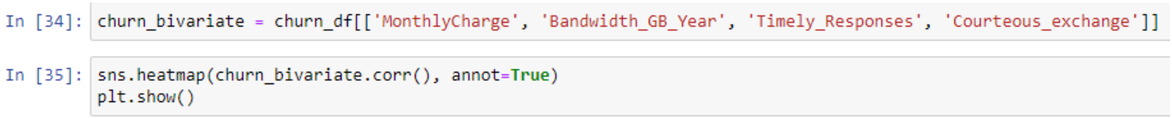
2.1 These are the Ordinal variables for the analysis:



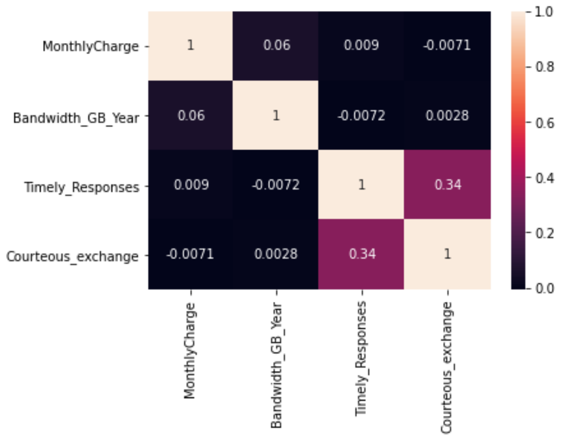
```
In [ ]: 1. Churn
        2. Item7 (Courteous exchange) - relabeled "Courteous_exchange"
```

D1: Visualization of the Results

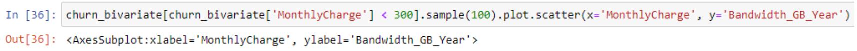
1.1 Create a data frame for a heatmap bivariate correlation analysis.:



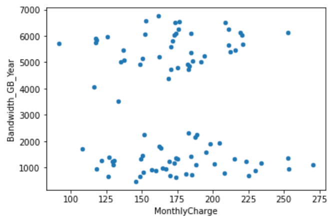
```
In [34]: churn_bivariate = churn_df[['MonthlyCharge', 'Bandwidth_GB_Year', 'Timely_Responses', 'Courteous_exchange']]

In [35]: sns.heatmap(churn_bivariate.corr(), annot=True)
         plt.show()
```



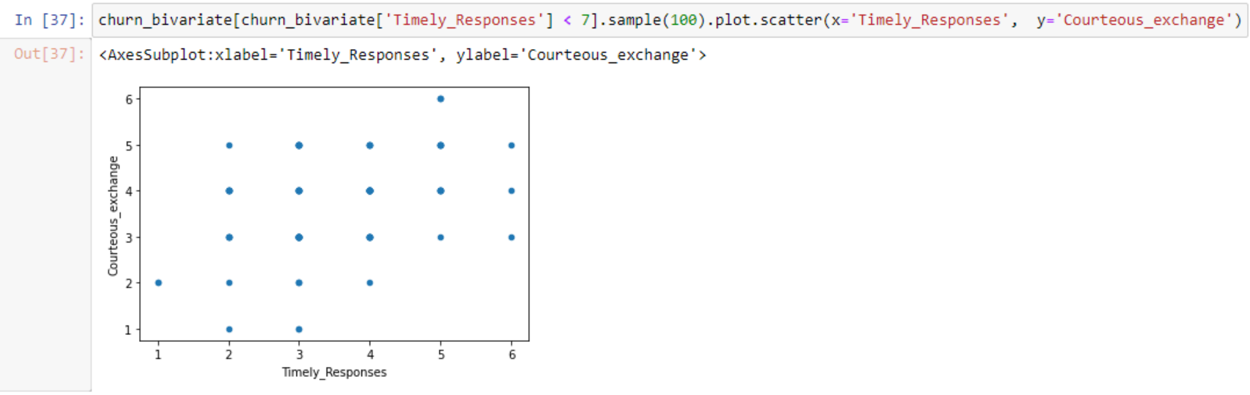
2.1 Using continuous variables, make a scatter plot. Bandwidth GB Year & MonthlyCharge:



```
In [36]: churn_bivariate[churn_bivariate['MonthlyCharge'] < 300].sample(100).plot.scatter(x='MonthlyCharge', y='Bandwidth_GB_Year')
Out[36]: <AxesSubplot:xlabel='MonthlyCharge', ylabel='Bandwidth_GB_Year'>
```

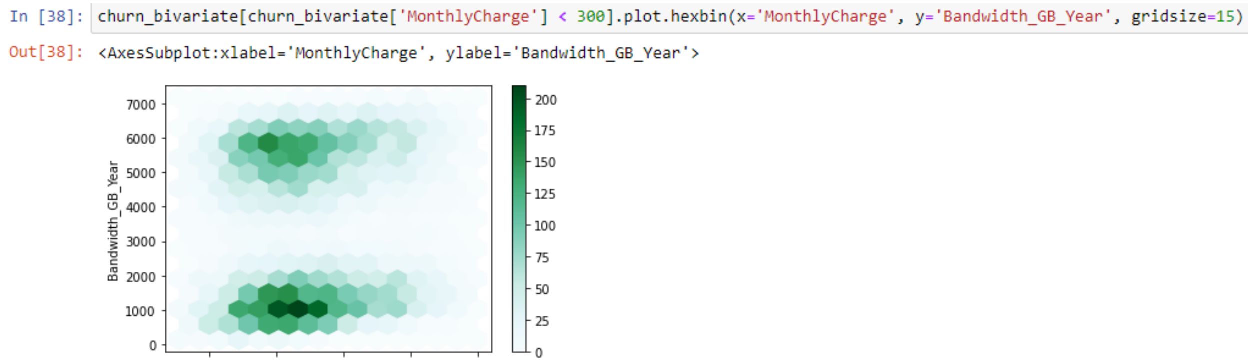


3.1 Using categorical variables, make a scatter plot. Response Time & Courtesy:



```
In [37]: churn_bivariate[churn_bivariate['Timely_Responses'] < 7].sample(100).plot.scatter(x='Timely_Responses',  y='Courteous_exchange')
```

```
Out[37]: <AxesSubplot:xlabel='Timely_Responses', ylabel='Courteous_exchange'>
```



```
In [38]: churn_bivariate[churn_bivariate['MonthlyCharge'] < 300].plot.hexbin(x='MonthlyCharge', y='Bandwidth_GB_Year', gridsize=15)
```

```
Out[38]: <AxesSubplot:xlabel='MonthlyCharge', ylabel='Bandwidth_GB_Year'>
```

E1: Analytical Findings

In this analysis, we can't reject the null hypothesis at a standard significance threshold of alpha = 0.05 with a p-value as huge as our chi-square significance testing output, p-value = 0.6318335816054494. Given the cleaned data, it's unclear whether there's a statistically significant link between survey replies (basically, "How well did we, the telecom firm, take care of you as a customer?") and whether this led to customers leaving the company.

E2: Analysis Limitations:

The sample dataset for this Chi-Square test, as previously indicated, is not genuinely random, but rather the restricted churn data points. It could be deceiving not to totally randomize the sample dataset from the churn dataset.

P-Value is a result of the analysis, but looking at it's such a high value 0.6318335816054494, which clearly indicates that we need to examine much deeper and possibly obtain additional and better data. Our ability to extract relevant and actionable data from this information has been greatly impeded, which is troubling.

E3: Course of Action Suggestion:

Customer churn was found to be unaffected by the gender variable. As a result, we recommend that the "gender" variable be removed/dropped from the dataset before moving on to the next phase in the data analysis cycle.

We can also recommend (with a 95 percent confidence interval) delving deeper into the co-relation and factors for the remaining three variables that influence turnover rate, which could considerably improve the company's business and revenue because of this analysis.

F: Step for Exploratory Data Analysis (EDA)

F1: Step by Step Exploratory Data Analysis (EDA) Summary

Describe the findings:

In this analysis, it is not clear as cleaned data available for statistically significant relationship between the Survey responses and we cannot reject the null values as standard importance of Alpha = 0.05 with a p-value is more with chi-square testing, p-value = 0.6318335816054494. this is basically "how well we are taking care of customer and whether this led to customer departing the company.

Defend your techniques for dealing with each sort of anomaly found in the dataset:

This dataset has so limited, we need to gather more and better data, with Chi-square analysis, it's a huge P-value, p-value = 0.6318335816054494.

Code used to Prevent anomalies:

Please find the attached Video recording.

Cleaned data set copy for EDA:

Please find the attached CSV file.

Constraints of Exploratory Data Analysis steps:

According to chi-square test there is little correlation between variables includes timely action with respect to customer satisfaction. (Timely Responses, Fixes, Replacements, and Respectfulness), we assume that these elements are more weight to reduce the “Churn rate” from its

G. Documentary Evidence

G1. Panopto recording:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1137756c-6246-45f1-a528-adbd0189b62d

G2. Third Party Evidence:

Practiced code: Bivariate plotting with pandas
URL: https://www.kaggle.com

Article: Predict Customer Churn in Python.
URL: https://towardsdatascience.com/

LinkedIn: https://www.linkedin.com/learning/python-statistics-essential-training/introducing-pandas?u=2045532

G3. References:

1. [1] Connelly, L. (2019). Chi-Square Test. MEDSURG Nursing, 28(2), 127.
2. [2] Article: P-Values. (2020). StatsDirect Limited. https://www.statsdirect.com/help/basics/p_values.htm