D207 – Exploratory Data Analysis

Morrell J. Parrish

Western Governors University

Table of Contents

[A. Research Question 3](#_Toc102469342)

[A2. Benefits of Analysis 3](#_Toc102469343)

[A3. Identifying Data 3](#_Toc102469344)

[B. Data Analysis Description 3](#_Toc102469345)

[B2. Output and Results of Analysis 8](#_Toc102469346)

[B3. Justification of Analytical Technique 12](#_Toc102469347)

[C. Univariate Statistics 12](#_Toc102469348)

[C2. Visual of Findings 12](#_Toc102469349)

[D. Bivariate Statistics 14](#_Toc102469350)

[D2. Visual of Findings 14](#_Toc102469351)

[E. Summary 16](#_Toc102469352)

[F. Panopto video recording 16](#_Toc102469353)

[References 17](#_Toc102469354)

D207 – Exploratory Data Analysis

# A. Research Question

During this course of research, we will explore and identify which customers are at a greater risk for churn; does the internet service type, area and customer service skills affect whether customers churn or not?

## A2. Benefits of Analysis

The benefits of this analysis will provide businesses with a better understanding of the services they provide for specific areas, as well as their customer service skills and customer satisfaction ratings; for example, if the analysis reveals a dependency or correlation between areas of service, customer service skills, satisfaction ratings, and customers who churn, businesses can delve into improving their customer service skills, which will raise their customer satisfaction ratings and will over all help retain customers.

## A3. Identifying Data

The data we will be using will consists of the following items: caseorder, internet service

type, churn, and the survey response categories 1 through 8 (timely responses, timely fixes, timely replacement, reliability, options, respectful responses, courteous exchange, and evidence of active listening).

# B. Data Analysis Description

*# Import the following Libraries*import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
from scipy.stats import chi2\_contingency  
from scipy.stats import chi2  
  
*# Read Churn data*df = pd.read\_csv("churn\_clean.csv",  
 usecols=["CaseOrder", "InternetService", "Churn", "Area", "Multiple",  
 "Item1", "Item2", "Item3", "Item4", "Item5",  
 "Item6", "Item7", "Item8"], index\_col="CaseOrder")  
  
*# Selecting a random sample from our Churn data*data = df.sample(n=100)  
print(data)  
  
*# Print description of sample data set*print(data.describe())  
  
*# Sample dataset column info*data.info()  
  
*# Creating a Contingency table for Churn and InternetService variables*contingency\_table = pd.crosstab(data['Churn'], data['InternetService'], margins=True, margins\_name="Total")  
  
print(contingency\_table)  
  
*# Perform Chi-Square Test on Churn and InternetService variables*stat, p, dof, expected = chi2\_contingency(contingency\_table)  
print("stat = ", stat)  
print("p-value = ", p)  
print("Degrees of Freedom =", dof)  
print("Expected = ", expected)  
  
*# Calculate alpha and critical values and interpret test-statistic*prob = 0.95  
critical = chi2.ppf(prob, dof)  
print('significance=%.3f, p=%.3f' % (1 - prob, p))  
print("critical value = ", critical)  
  
*# Test result summary*print('probability=%.3f, critical=%.3f, stat=%.3f' % (prob, critical, stat))  
if abs(stat) >= critical:  
 print('Dependent (reject H0)')  
else:  
 print('Independent (fail to reject H0)')  
  
*# Interpret p-value*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)')

*# Creating a Contingency table for Area and Churn variables*contingency\_table\_2 = pd.crosstab(data['Churn'], data['Area'], margins=False)  
  
print(contingency\_table\_2)  
  
*# Perform Chi-Square Test on Churn and Area variables*

stat, p, dof, expected = chi2\_contingency(contingency\_table\_2)  
print("stat = ", stat)  
print("p-value = ", p)  
print("Degrees of Freedom =", dof)  
print("Expected = ", expected)  
  
*# Calculate alpha and critical values and interpret test-statistic*prob = 0.95  
critical = chi2.ppf(prob, dof)  
print('significance=%.3f, p=%.3f' % (1 - prob, p))  
print("critical value = ", critical)  
  
*# Test result summary*print('probability=%.3f, critical=%.3f, stat=%.3f' % (prob, critical, stat))  
if abs(stat) >= critical:  
 print('Dependent (reject H0)')  
else:  
 print('Independent (fail to reject H0)')  
  
*# Interpret p-value*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)')  
  
*# Creating a Contingency table for Churn and Multiple variables*contingency\_table\_3 = pd.crosstab(data['Churn'], data['Multiple'], margins=False)  
  
print(contingency\_table\_3)  
  
*# Creating dataframe for our Univariate Statistics*df2 = pd.read\_csv("churn\_clean.csv", usecols=["MonthlyCharge", "Bandwidth\_GB\_Year"])

*# Create histograms for MonthlyCharge and Bandwidth\_GB\_Year*print(df2)  
  
histogram = df2.hist()  
  
plt.show()  
  
*# Creating histograms for Item1 and Item2*df3 = pd.read\_csv("churn\_clean.csv", usecols=["Item1", "Item2"])  
  
print(df3)  
  
*# Rename Item1 and Item2*dictionary = {'Item1': 'Timely\_Responses', 'Item2': 'Timely\_Fixes'}  
  
df3.rename(columns=dictionary, inplace=True)  
  
hist = df3.hist()  
  
plt.show()  
  
*# Creating Boxplots for Variables*df4 = pd.read\_csv("churn\_clean.csv", usecols=["MonthlyCharge", "Bandwidth\_GB\_Year", "Item1", "Item2"])  
  
sns.boxplot(y=df4["MonthlyCharge"])  
  
plt.show()  
  
sns.boxplot(y=df4["Bandwidth\_GB\_Year"])  
  
plt.show()  
  
sns.boxplot(y=df4["Item1"])  
  
plt.ylabel("Timely\_Responses")  
  
plt.show()  
  
sns.boxplot(y=df4["Item2"])  
  
plt.ylabel("Timely\_Fixes")  
  
plt.show()  
  
*# Randomly selecting data from our Churn dataset*df5 = pd.read\_csv("churn\_clean.csv", usecols=["MonthlyCharge", "Bandwidth\_GB\_Year", "Area", "Item1", "Item2", "Item3", "Item4", "Item5", "Item6", "Item7", "Item8"])  
  
data3 = df5.sample(n=100)  
  
print(data3)  
  
*# Create a scatter plot of continuous variables MonthlyCharge & Bandwidth\_GB\_Year*x = data3['MonthlyCharge']  
y = data3['Bandwidth\_GB\_Year']  
  
plt.scatter(x, y, c='red')  
plt.scatter(x, y, c='blue')  
  
plt.xlabel("MonthlyCharge")  
plt.ylabel("Bandwidth\_GBYear")  
  
plt.plot(np.unique(x), np.poly1d(np.polyfit(x, y, 1))(np.unique(x)), color='red')  
  
plt.show()  
  
*# Correlation Matrix*

sns.heatmap(df5.corr(), linewidths=.3, annot=True)  
plt.show()

## B2. Output and Results of Analysis

Graphical user interface, text, application, email

Description automatically generated

Figure 1: Input

Table

Description automatically generated

Figure 1: Output

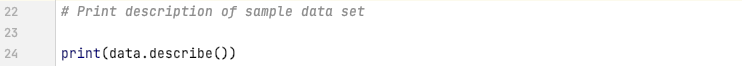


Figure 2: Input

Table, Excel

Description automatically generated

Figure 2: Output



Figure 3: Input

Table

Description automatically generated

Figure 3: Output

A picture containing text

Description automatically generated

Figure 4: Input

A picture containing text

Description automatically generated

Figure 4: Output

Graphical user interface, text, application, email

Description automatically generated

Figure 5: Input

Background pattern

Description automatically generated with medium confidence

Figure 5: Output

Graphical user interface, text, application

Description automatically generated

Figure 6: Input

Text

Description automatically generated with medium confidence

Figure 6: Output

A picture containing diagram

Description automatically generated

Figure 7: Input

A picture containing table

Description automatically generated

Figure 7: Output

Graphical user interface, text, application, email

Description automatically generated

Figure 8: Input

Text

Description automatically generated with medium confidence

Figure 8: Output

Graphical user interface, text, application

Description automatically generated

Figure 9: Input

Text

Description automatically generated

Figure 9: Output

Graphical user interface, text, website

Description automatically generated

Figure 10: Input

Background pattern

Description automatically generated with low confidence

Figure 10: Output

Text

Description automatically generated with low confidence

Figure 11: Input

Table

Description automatically generated

Figure 11: Output

## B3. Justification of Analytical Technique

A chi-square test is a statistical test which is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration; the categorical variable we are analyzing is churn rate, we want to determine if a relationship exists between our variable and if a relationship is proven to exist what affect does it have on churn rates (Chi Square, nd).

# C. Univariate Statistics

To perform our univariate statistics, we will be using “***monthlycharge****”* and “***bandwidth\_gb\_year****”* for our continuous variables and *“****item 1****”* (timley\_responses) and “***item 2***” (timely\_fixes) as our categorical variables from our churn dataset.

## C2. Visual of Findings

Graphical user interface, text, application

Description automatically generated

Figure 12: Input - Creating Histograms for Monthly Charge, Bandwidth, Item 1 and 2

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

*Figure 12: Output*

Graphical user interface, text, application

Description automatically generated

Figure 13: Boxplots for our Variables

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

*Figure 13: Output - Monthly Charges. Figure 13: Output - Bandwith\_GB\_Year*

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

*Figure 13: Output - Timely Responses Figure 13: Output - Timely Fixes*

# D. Bivariate Statistics

To perform our bivariate statistics, we will be using “***monthlycharge****”* and “***bandwidth\_gb\_year****”* for our continuous variables and *“****item 1****”* (timley\_responses) and “***item 2***” (timely\_fixes) as our categorical variables from our churn dataset. I will be using heatmaps and scatterplots to perform the bivariate statistics; the result of the scatterplot shows that there isn’t a strong relationship between or variables; however, our correlation matrix shows that there is some correlation between Item 1 (***timely\_responses***) an Item 2 (***timely\_fixes***).

## D2. Visual of Findings

Graphical user interface, text, application, email

Description automatically generated

Figure 14: Input - Scatterplot Monthly Charge and Bandwidth

Chart, scatter chart

Description automatically generated

Figure 14: Output

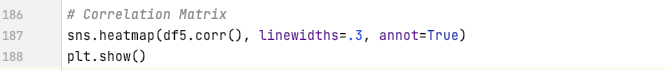


Figure 15: Input - Heatmap (Correlation Matrix)

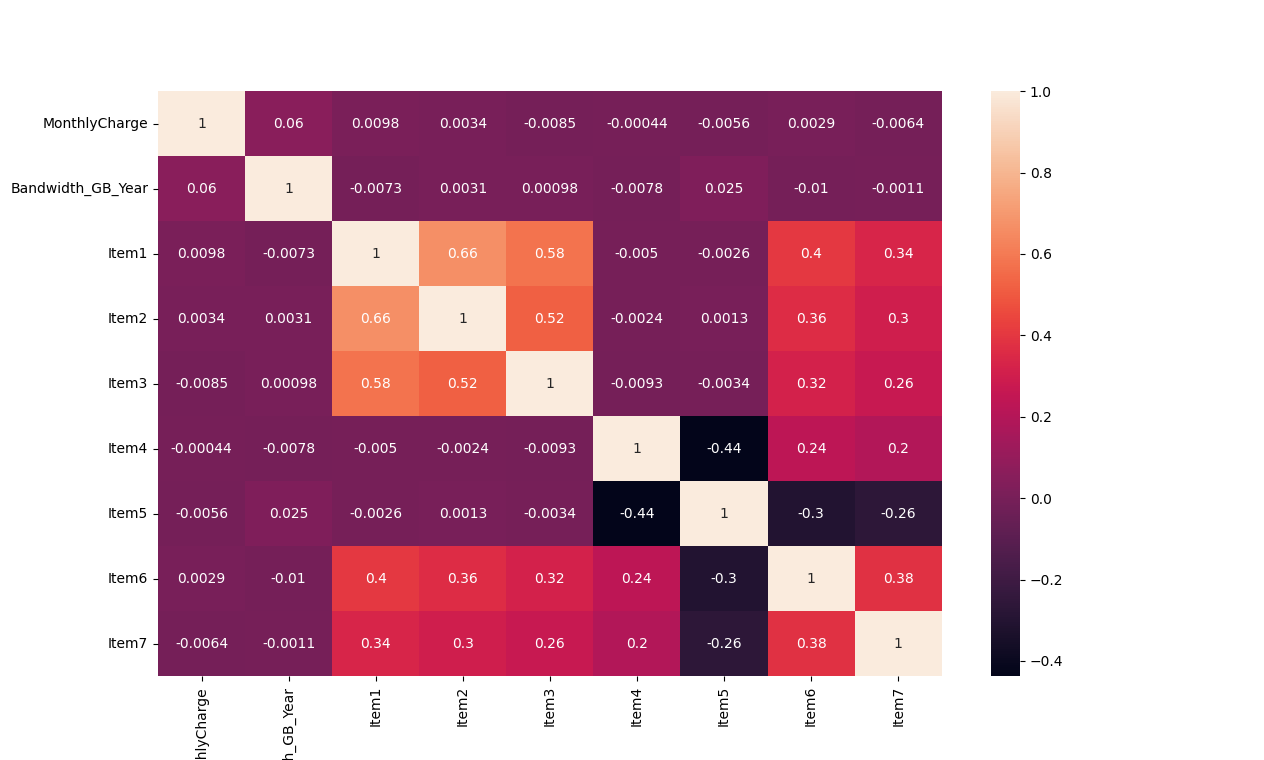


Figure 15: Output

# E. Summary

We performed Chi-Square test on five different variables - ***churn, item 1, item 2, internet service type***, and ***area***; from our analysis we can see that there is no direct correlation between any of our chosen variables and customers who churn. All tests proved that the variables are independent of each other and have no impact on weather a customer churn or not. The dataset may have some limitations because of how unbalanced it is, only a small portioned of customers have churned.

I would recommend that another study be done with more balanced data. I would also focus more on ***tenure, contract,*** ***monthly charge,*** and ***area***. By doing this, this will allow the company to see if these variableshave an impact on customers who churn. Furthermore, it will also allow the company to revamp its contracts length and monthly fees.

# F. Panopto video recording

[Video Link](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e36d2ccf-c985-49d9-b148-ae8a0129ad3d)

References

*Chi Square*. Chi Square | Practical Applications of Statistics in the Social Sciences | University of Southampton. (n.d.). Retrieved May 3, 2022, from https://www.southampton.ac.uk/passs/full\_time\_education/bivariate\_analysis/chi\_square.page

Choueiry, G. (n.d.). *P-value: A simple explanation for non-statisticians*. Quantifying Health. Retrieved May 3, 2022, from https://quantifyinghealth.com/p-value-explanation/

DataCamp. (n.d.). Retrieved May 3, 2022, from https://app.datacamp.com/learn/courses/data-manipulation-with-pandas

OEM2 Task 1: EDA: Exploratory Data Analysis (n.d.). Retrieved May 3, 2022, from https://tasks.wgu.edu/student/000194226/course/23260006/task/2786/overview