**Exploratory Data Analysis**

**By: Richard Kemonou**

**Date: August 2023**

A-1- Research Question

This research paper intends to evaluate our ability to analyze complex data sets and statistically explore the relationship between variables inside the data sets using different programming languages. We opted to analyze the churn data CSV file using the python programming language. We chose the churn data set because of our familiarity with it since we have used the same data set in our previous assignments and the python programming language because we are more familiar with it than the R programming language. After a study of the data set, we decided to analyze whether the variables Item1 (Timely response), Item6 (Respectful response), and Item8 (Active listening) have a dependency on consumers’ churn decision-making.

Analyzing whether the consumers’ satisfaction survey ratings related to timely response, respectful response and active listening have a dependency on churn is relevant for the telecommunication industry because consumers rate highly the quality of services provided to them and stakeholders can use the insights from this analysis to set up targeted consumers retention marketing strategies; given that in the telecommunication industry, consumers retention is as much important as new consumers acquisition.

We will be performing the chi-square statistical analysis on the data set because the consumers’ churn variable we are studying is a nominal one, which means it is a categorical variable that only takes the values of yes or no.

A-2- The chi-square analysis of the data set will enable us to understand whether the variables Item1 (Timely response), Item6 (Respectful response), and Item8 (Active listening) have a dependency on consumers’ churn decision-making. The data set depicts many variables ranging from consumers’ age to their service satisfaction survey ratings and this analysis can find out whether any of the three variables discussed above contribute to consumers’ churn decision-making. Doing so can help stakeholders set up adequate consumers retention marketing strategies; given that in the telecommunication industry, consumers retention is as important as new consumers acquisition.

A-3- This paper examines whether the variables Item1 (Timely response), Item6 (Respectful response), and Item8 (Active listening) have a dependency on consumers’ churn decision-making. The churn data set provided to us details 10,000 consumers data from a telecommunication firm.

In our research, the dependent variable depicts consumers who responded “Yes” or “No” to the question of whether they discontinued their telecommunication service in the last month, which is a qualitative variable.

In order to analyze this dependent variable, we picked three independent variables from the data set. We picked these three independent variables based on the results of the principal component analysis we run in our D-206 Data Cleaning Course. Back then, we identified 11 variables as being significant to the consumers' churn decision-making. These 11 variables are: Tenure, MonthlyCharge, Bandwidth\_GB\_Year, item1, item2, item3, item4, item5, item6, item7, and item8.

The Tenure column indicates the number of months the consumers spent with the telecommunication service provider (example: 17). The MonthlyCharge column indicates the monthly charges billed to the consumers, computed as per customer average (example: 148.18). The Bandwidth\_GB\_Year column describes the consumers’ average GB in data used per year (example: 800.779). The item1 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is a timely response (example: 1, 8, 9). The item2 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important are timely fixes (example: 1, 8, 9). The item3 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important are timely replacements (example: 1, 8, 9). The item4 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is reliability (example: 1, 8, 9). The item5 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important are options (example: 1, 8, 9). The item6 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is a respectful response (example: 1, 8, 9). The item7 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is a courteous exchange (example: 1, 8, 9). The item8 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is active listening (example: 1, 8, 9).

the Tenure, MonthlyCharge and Bandwidth\_GB\_Year columns depict continuous numerical data while the last eight consumers’ satisfaction survey responses depict discrete numerical data.

The three independent variables we retained for this study are as follows:

1. Item1 (Timely response)
2. Item6 (Respectful response)
3. Item8 (Active listening)

B-1-

***Import necessary python libraries***

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

%matplotlib inline

import seaborn as sns

from scipy import stats

from scipy.stats import chi2

from scipy.stats import chi2\_contingency

***#### Import the cleaned data set***

df = pd.read\_csv(r'C:\Users\richa\OneDrive\Desktop\d207\churn\_clean.csv')

***#####Rename survey response columns with names that better describes their attributes***

df.rename(columns = {'Item1':'Timely\_response', 'Item2':'Timely\_fixes',

'Item3':'Timely\_replacements', 'Item4':'Reliability',

'Item5':'options', 'Item6':'Respectful\_response',

'Item7':'Courteous\_exchange', 'Item8':'Active\_listening'}, inplace=True)

print(df)

***Provide descriptive statistics for the chi-square variables***

dataframe = df[['Timely\_response', 'Respectful\_response', 'Active\_listening']]

dataframe.describe()

A table with numbers and text

Description automatically generated

***########## ANALYZE CHURN AND ACTIVE LISTENING***

***Create Contingency table for Churn and Active Listening***

contingency = pd.crosstab(df['Churn'], df['Active\_listening'])

contingency

contingency\_pct = pd.crosstab(df['Churn'], df['Active\_listening'], normalize='index')

contingency\_pct

plt.figure(figsize=(12,8))

sns.heatmap(contingency, annot=True, cmap="YlGnBu")

***####### Chi-square test of independence***

c, p, dof, expected = chi2\_contingency(contingency)

print('p-value = ' + str(p))

B-2- Please find the following screenshots, the PDF attachment, and the Panopto video

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A close-up of a logo

Description automatically generated

B-3- In this paper, we aimed at analyzing whether the variables Item1 (Timely response), Item6 (Respectful response), and Item8 (Active listening) have a dependency on consumers’ churn decision-making. Our dependent variable "Churn" is a binomial categorical variable that takes either the value Yes or No. Therefore, we opted to use the chi-square statistical testing because it is a non-parametric test. We chose three independent variables from our data set because we anticipated that these variables are significant to the consumers' churn decision-making. These independent variables are the consumers’ satisfaction survey ratings depicting timely response, respectful response, and active listening.

Through our chi-square analysis, we computed different statistical indicators to investigate whether any of these three independent variables we retained is significantly dependent on the consumers' churn decision-making. Stakeholders can use the insights from this analysis to set up targeted consumers retention marketing strategies; given that in the telecommunication industry, consumers retention is as much important as new consumers acquisition.

C- According to Shah (2021), Univariate statistics explore each variable in a data set, separately. Histograms and box plots represent some of the most commonly used univariate statistics methods to represent the data visually.

In this paper, we chose two continuous and two categorical variables from the churn data set. The two continuous variables we chose are MonthlyCharge and Bandwidth\_GB\_Year. The two categorical variables we chose are item1 (Timely response) and item8 (Active listening). The item1 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is a timely response (example: 1, 8, 9). The item8 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is active listening (example: 1, 8, 9).

***Provide descriptive statistics for the univariate variables***

dataframe = df[['MonthlyCharge', 'Bandwidth\_GB\_Year', 'Timely\_response', 'Active\_listening']]

dataframe.describe()

A table with numbers and text

Description automatically generated

***VISUALIZING CONTINUOUS VARIABLES***

***We create Seaborn boxplots for the continuous variables***

sns.boxplot(x=df['MonthlyCharge'])

A screenshot of a graph

Description automatically generated

The above boxplot visualizes the distribution of MonthlyCharge based on its minimum value (79.97), first quartile (140), median (172), third quartile (200), and maximum value (290)

sns.boxplot(x=df['Bandwidth\_GB\_Year'])

A blue rectangular object with numbers

Description automatically generated with medium confidence

The above boxplot visualizes the distribution of Bandwidth\_GB\_Year based on its minimum value (156), first quartile (1236), median (3392), third quartile (5586), and maximum value (7159)

***We generate SCATTER PLOT for MonthlyCharge***

Since for data set is too large for a scatter plot, we randomly generated a sample data of size 100 from the original data set

import random

datasample = df.head(100)

print(datasample)

***We generated Scatter plot for MonthlyCharge***

plt.scatter(datasample.index, datasample['MonthlyCharge'])

plt.show()

A graph with blue dots

Description automatically generated

The scatter plot for MonthlyCharge shows that the data points are spread out and don't display any trend, therefore there is no correlation.

***We generated Scatter plot for Bandwidth\_GB\_Year***

plt.scatter(datasample.index, datasample['Bandwidth\_GB\_Year'])

plt.show()

A graph with blue dots

Description automatically generated

The scatter plot for Bandwidth\_GB\_Year shows that the data points are spread out and don't display any trend, therefore there is no correlation.

***We generate histogram for MonthlyCharge***

plt.hist(datasample['MonthlyCharge'])

A graph with numbers and lines

Description automatically generated with medium confidence

The histogram for MonthlyCharge shows the number of consumers who spend specific amount monthly on telecommunication services; for instance, the above graph shows that 20 consumers in our data sample spend between 150 and 165 monthly on telecommunication services.

***We generate histogram for Bandwidth\_GB\_Year***

plt.hist(datasample['Bandwidth\_GB\_Year'])

A blue graph with numbers

Description automatically generated

The histogram for Bandwidth\_GB\_Year shows the number of consumers with specific average GB in data used per year; for instance, the above graph shows that 17.5 consumers in our data sample consume between 1000 and 1250 GB in data per year.

***VISUALIZING CATEGORICAL VARIABLES***

***Generating BAR CHARTS FOR Timely\_response AND Active\_listening***

datasample = df.head(100)

Df = pd.DataFrame(datasample)

colors = ['green','blue','purple','brown','teal', 'red', 'yellow', 'pink']

plt.bar(Df['Timely\_response'], Df['Active\_listening'], color=colors)

plt.title('Timely\_response Vs Active\_listening', fontsize=14)

plt.xlabel('Timely\_response', fontsize=14)

plt.ylabel('Active\_listening', fontsize=14)

plt.grid(True)

plt.show()

A graph with different colored bars

Description automatically generated

D- According to Wolfram Research (2011), Bivariate statistics analyze the data on each of two variables, where each value of one of the variables is paired with a value of the other variable. Heat maps and scatter plots represent some of the most commonly used bivariate statistics methods to represent the data visually.

In this paper, we chose two continuous and two categorical variables from the churn data set. The two continuous variables we chose are MonthlyCharge and Bandwidth\_GB\_Year. The two categorical variables we chose are item7 and item8. The item7 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is a courteous exchange (example: 1, 8, 9). The item8 column indicates on a scale of 1 ‘Most Important’ – 8 ‘Least Important’, how important is active listening (example: 1, 8, 9).

***Provide descriptive statistics for the bivariate variables***

dataframe = df[['MonthlyCharge', 'Bandwidth\_GB\_Year', 'Courteous\_exchange', 'Active\_listening']]

dataframe.describe()

A table with numbers and text

Description automatically generated

***# Create dataframe for heatmap bivariate analysis of correlation***

df\_bivariate = df[['MonthlyCharge', 'Bandwidth\_GB\_Year', 'Courteous\_exchange', 'Active\_listening']]

sns.heatmap(df\_bivariate.corr(), annot=True)

plt.show()

A screenshot of a computer

Description automatically generated

***Create a scatter plot of continuous variables MonthlyCharge & Bandwidth\_GB\_Year***

df\_bivariate[df\_bivariate['MonthlyCharge'] < 300].sample(100).plot.scatter(x='MonthlyCharge', y='Bandwidth\_GB\_Year')

A graph of blue dots

Description automatically generated

The scatter plot for Bandwidth\_GB\_Year and MonthlyCharge shows that the data points are spread out and don't display any trend, therefore there is no correlation.

***Create a scatter plot of categorical variables Courteous\_exchange & Active\_listening***

df\_bivariate[df\_bivariate['Courteous\_exchange'] < 7].sample(100).plot.scatter(x='Courteous\_exchange', y='Active\_listening')

A graph of blue dots

Description automatically generated

The scatter plot for Active listening and Courteous exchange shows that the data points are spread out and don't display any trend, therefore there is no correlation.

df\_bivariate[df\_bivariate['MonthlyCharge'] < 300].plot.hexbin(x='MonthlyCharge', y='Bandwidth\_GB\_Year', gridsize=15)

A graph of a number of green hexagons

Description automatically generated

E-1-In the section C1, we performed chi-square test on Item8 (Active listening) in order to analyze how significant this variable is related to the consumers' churn decision-making. We performed the chi-square test with two hypotheses: The null hypothesis and the alternative hypothesis. The null hypothesis states that Item8 (Active listening) is significantly related to the dependent variable while the alternative hypothesis states that Item8 (Active listening) is not significantly related to the dependent variable. The confidence interval we retained for this study is 95 %; which means, in each of these two hypotheses, we have 5 % chance of being wrong. A confidence level of 95 % indicates an alpha value of 0.05. Neyman and Pearson (1928) defined alpha value as the probability of rejecting the null hypothesis when it is true.

According to the output results, in the case of churn and Item8 (Active listening), the computed p value (0.974) is greater than the alpha value, that means we fail to reject the null hypothesis. Consumers' churn decision-making is dependent on Item8 (Active listening).

E-2-Our computed p-value is higher than the alpha value. This is probably an indication of data quality and accuracy issues. The researcher needs to be involved in the data collection process and the choice of the variables to be retained for data collection.

E-3-From our analysis, we found out that the independent variable Item8 (Active listening) is significantly dependent on the consumers' churn decision-making. We recommend that stakeholders improve the quality of the active listening variable in consumers' retention marketing strategies.

F- Panopto video

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=af6ea331-03bb-4709-a294-b05d011269dd>

G- Web Source References

Wolfram Research (2011) <https://mathworld.wolfram.com/Bivariate.html>

Shah, Khushi., (2021) Exploratory Analysis Using Univariate, Bivariate, and Multivariate Analysis Techniques. <https://www.analyticsvidhya.com/blog/2021/04/exploratory-analysis-using-univariate-bivariate-and-multivariate-analysis-techniques/>

H- References

Neyman, J. and Pearson, E.S., (1928) On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Biometrika, 20A, 175-240.