D207 Exploratory Data Analysis: Performance Assessment

Jill Rivera

Western Governors University

Table of Contents

[I. Variables and Research Question 3](#_Toc103200839)

[II. Statistical Analysis 5](#_Toc103200840)

[III. Univariate Statistics 8](#_Toc103200841)

[IV. Bivariate Statistics 13](#_Toc103200842)

[V. Results 15](#_Toc103200843)

[VI. References 17](#_Toc103200844)

[VII. Table of Figures 17](#_Toc103200845)

# **Variables and Research Question**

1. Research Question:
   1. This telecommunications company is specifically wanting to reduce customer churn. Churn is a term used to describe a customer that has discontinued service with a company to obtain the same service with another company. This corporation has determined that it costs 10 times the amount of money to get a new customer, then keep an existing one. With an average rate of churn in telecommunications companies being around 25 percent, reducing this rate will save the company a lot of money. Specifically, this analysis will examine the relationship between the churn and gender variables.
   2. The research question that will be analyzed will be: Does gender influence the churn rate for the company?
   3. Answering this question will benefit the stakeholders of the company and give valuable insight into the data that they have. This will allow a decision of keeping or dropping certain features in the data set to improve the churn model.
2. Variables:
   1. For this analysis, the target variable is Churn. The target variable will be analyzed against the variable of Gender to determine their relationship. This will be the predictor dependent variable for the analysis. Churn is a categorical variable with Yes/No responses determining if a customer has churned in the last 30 days. The Gender variable is a categorical variable as well which is a self-identified value of male, female, or nonbinary.
   2. The data set consists of 10,000 customer records with 49 variables. This translates to a data frame of 10,000 rows and 49 columns as a portion is seen in Figure 1.

**Graphical user interface, text, application, email

Description automatically generated**

Figure : First 5 rows of the data set.

The variables consist of customer demographic information (gender, marital status, etc.), services the customer has or doesn’t (phone, tech support, paperless billing, etc.), and eight rating responses to a customer satisfaction survey. With this analysis evaluating churn, other variables that would likely be associated with churn will be analyzed. In addition to the target and dependent variable, this analysis will evaluate the following variables:

* + 1. Gender – Categorical variable with customer self-identification as either male, female, or nonbinary.
    2. Churn – Categorical variable that states yes or no if the customer has discontinued service with the provider in the last 30 days.
    3. Tenure – Continuous numerical variable with a value of the number of months that a customer has had service with the provider.
    4. Bandwidth\_GB\_Year – A continuous numerical variable that averages the amount of data in GB used by the customer yearly.
    5. Timely Response – Customer satisfaction survey response of 1 to 7 on how timely the response from the provider was to their inquiry.
    6. Timely Fixes - Customer satisfaction survey response of 1 to 7 on how timely the fix of the customer’s problem was.

# **Statistical Analysis**

1. For this analysis, Python 3 programming language in a Jupyter Notebook will be used to calculate and visualize the data. The package libraries of Pandas, Numpy, Seaborn, Matplotlib and SciPy Stats will be used to perform the research.
2. The test chosen to analyze this data set is the ChiSquare Independence Test. This method was picked due to the target variable of Churn and dependent variable of Gender, being categorical variables. This analysis method assumes a null hypothesis that the two variables are independent, and an alternate hypothesis that they are dependent. The results of this test will tell if the two variables are dependent or independent of each other. If they are independent, then they are not related, and the null hypothesis of our test must be accepted. If the two variables are dependent, then they are related to each other, and the alternate hypothesis is accepted. Setting a significance level of 0.05 will determine a significant relationship with a 95% certainty by comparing the distributions of each data point.
3. A Q-Q plot will first be constructed. The Q in Q-Q plot stands for quantile, which means that this diagram is a probability plot in which two probability distributions are plotted graphically with their quantiles as the plots. This type of graph when constructed with a 45-degree line will show the variation of the data from the standard deviation. A Q-Q plot of the churn variable was constructed below.

**Chart, line chart

Description automatically generated**

Figure : Q-Q Plot of the Churn Variable

This graph shows that the churn variable does not follow the normal standard deviation.

A Chi Squared test is performed on the churn and gender variables to see if we are to accept or reject the null hypothesis as discussed previously.

**Table

Description automatically generated**

Figure : Chi Squared Test of Independence (SciPy, 2022)

This complex analysis is achieved with simple Python code using the SciPy Stats Chisquared command. It (McHugh, 2022) will calculate a p-value (probability of independence value) and a chi-square value (numerical result of the formula). The degree of freedom for this analysis is 2, since it is calculated by subtracting 1 from the number of rows in the analysis (3). First, a table of the two variables is created with the crosstab command and the table is then passed into the chi2 command. The results shown are the p-value and the Chi Squared score for the analysis. For the churn variable, the p-value came back as a 0.01945. This means that we will need to reject the null hypothesis and accept the alternative hypothesis stating that the gender and churn variables are dependent of each other. It can then be determined that gender does influence the churn rate for this telecommunications company.

1. The ChiSquared statistical method is a versatile testing method that does not require the variable data to follow a normal distribution pattern, as it is non-parametric. It has flexibility when it comes to handling data from multiple groups and analyze the fit of each variable (McHugh, 2022). The disadvantage is that the number of variables should be less than 20 so that the data is not interpreted incorrectly. This method is also completed in Python using very little code through the SciPy Stats Chisquare chi2 command (SciPy, 2022). This simple, yet powerful command will obtain the statistical variance and p-value to confirm or deny your null hypothesis.

# **Univariate Statistics**

1. A simple form of data analysis is univariate statistics, where a single variable is analyzed. The term univariate statistics is a statistical analysis that includes a single dependent variable (Canova, Luigi Cortinovis, & Ambrogi, 2017). Through this analysis, the statistical data of the variable can be seen in a graphical format. The distribution pattern can be assessed, and a determination can be made on its relation to other variables. This variable can be easily visualized using Python and plotnine in a scatter plot. Two continuous and two categorical variables are reviewed in this manner in section B.
2. Continuous variables are those which are numerical or quantitative and are not set numbers, they continue. The variables Tenure and Bandwidth\_GB\_Year were selected to study for the continuous variables against the Churn variable. Those graphs are shown as follows in Figures 4 and 5 below.Chart, box and whisker chart

   Description automatically generated

Figure : Continuous Univariate Box Plot (Gosset & Wright, 2017)

This box plot shows tenure in relation to churn. We can see that the retained customers have an average tenure of 50 months of service, while those that left averaged only 8 months of service. This tells us that customers are switching providers early on into their service. From this plot we can also see the range of tenure for the retained customers is large (from 9 to 63 months), while the customers who churned have a smaller range (4 months to 13 months). There are also several outlying data points on the customers who discontinued their service after 27 months of service.

Chart

Description automatically generated

Figure : Continuous Univariate Density Plot (Gosset & Wright, 2017)

The density plot of bandwidth used yearly in relation to churn shows that the distribution is irregular. The customers who churned have a right skewed density curve, meaning that the mean is greater than the median. This curve is also multimodal since it has two peaks. The density curve for the retained customers is also multimodal, having two peaks as well. This curve is more left skewed; however, the data is more evenly distributed in comparison to the customers who did churn. From this graph, the average bandwidth usage (around 1200 GB) for customers who churned was much lower than those who were retained (around 4500 GB).

Categorical variables are discrete numerical values with set values. Timely Response and Timely Fixes are the categorical variables that will be analyzed against the Churn variable in this analysis. The graphs can be seen in Figures 6 and 7 below.

Chart

Description automatically generated

Figure : Categorical Univariate Density Plot (Gosset & Wright, 2017)

The density plot of Timely Response in relation to churn shows a normal pattern of distribution for both the retained and churned customers. The responses for this category were the same regardless of the customer being retained or not. This indicates that this variable has no impact on the churn rate. From the graph, we can see the average response was a 3.5 for all customers, and the amount of retained customer responses outnumbered those that were not retained.

Chart, box and whisker chart

Description automatically generated

Figure : Categorical Univariate Box Plot (Gosset & Wright, 2017)

The box plot for Timely Fixes in relation to churn in Figure 7 shows that the range of responses for both the retained customers and those that were not retained were the same. The average response for those who were retained was a 4, while those that were not retained had an average of 3 for a response. This indicates that there were customers who were not satisfied with the amount of time that it took to fix their issue, and that could be a factor into why they decided to leave the provider.

1. Advantages of using univariate analysis on variables are that it is simple, little code to write and visually displays the variable’s statistical measurements. The disadvantages of using this analysis type are that only one variable is analyzed thus limiting the results, and it mainly gives a description of a variable and not how it relates to other variables.

# **Bivariate Statistics**

1. The term bivariate statistics is like the univariate method used before, however now there are two dependent variables that are being graphically analyzed (Trauth, 2007). This shows a comparison of the two variables and how their distributions relate to each other. In section B, two categorical and two continuous variables are reviewed using this method. A scatter plot was created for each pair of variables using plotnine in Python.
2. The continuous and categorical variables are plotted into scatter plots below for analysis in Figures 8 and 9. The continuous variables chosen were Badnwidth\_GB\_Year and Tenure in relation to Churn.

**Chart, scatter chart

Description automatically generated**

Figure : Continuous Bivariate Scatter Plot (Gosset & Wright, 2017)

The scatter plot in Figure 8 depicting tenure in relation to the amount of bandwidth used, shows a strong, positive correlation between the variables. This shows that the two are closely related and that customers who are retained consume more bandwidth than those who leave.

Calendar

Description automatically generated

Figure : Categorical Bivariate Scatter Plot (Gosset & Wright, 2017)

From the scatter plot of the categorical variables in Figure 9, we can see that the data points are not close together. This shows a weak relation between the Timely Fixes and Timely Response variables. The graph shows a positive, yet very weak correlation, that follows the same pattern of distribution.

1. The advantages of using bivariate analysis are that it can determine an empirical relationship between two variables as well as it is simple to complete with only a few lines of code in Python. The disadvantages of this analysis method are that it limits the analysis to the two variables chosen as well as correlation cannot determine causation. While this method can show if the two variables follow the same patterns, but it will not show if they are dependent of each other.

# **Results**

1. The goal of the analysis was to determine if there was a dependent relationship between the gender of customers and the churn rate. The Chi Squared test was used to calculate a p-value, which will aid in either accepting or rejecting the null hypothesis. The null hypothesis stated that the two variables are independent, and an alternate hypothesis that they are dependent. In the analysis, the p-value was calculated to be 0.01945, which is less than the alpha value of 0.05. This means that we reject the null hypothesis and accept the alternative hypothesis of the variables being dependent of each other. This analysis can conclude that gender has an influence over the churn rate for the company.

Through univariate statistics, it was determined that customers who churn from the company use less bandwidth (1200 GB yearly) and only stay with the company for 8 months. Also, timely responses to problems had less of an impact on customers churning than a timely fix to the problem. Those customers who indicated a slower time to fix their problem were more likely to churn.

Bivariate statistics revealed a strong positive correlation between the amount of bandwidth a customer used and the length of their tenure with the company. This shows that the longer a customer stays with the provider, the more bandwidth they consume. This method of analysis also revealed a weak correlation between the timely fixes and responses variables.

1. The major limitation of this data analysis was that the data had to be interpreted and determined to be relevant or not to the analysis by the analyst. Not being able to discuss the data with the person collecting it, leaves it up to interpretation which can be biased. Speaking to the person collecting the data as to which variables are relevant or more important would be a valuable tool that is not available to this study and therefore a limitation. Also, correlation cannot determine causation. Finding a strong correlation between two features, does not mean that they are direct causes of each other.
2. This telecommunications company has more insight into their data and how it can be used to reduce the churn rate of their customers. An improvement in amount of time it takes to fix a customer’s problem would have a positive impact on the churn rate. Those customers with low bandwidth usage are more likely to churn, as well as those who have had service for less than 8 months. It would benefit the company if they were to target those customers and directly improve their satisfaction with their service.

# **References**

*Bivariate Analysis*. (2022, 04 23). Retrieved from Statistics How To: https://www.statisticshowto.com/bivariate-analysis/

*Scipy.stats.chisquare*. (2022, 04 22). Retrieved from SciPy: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html

*Univariate Analysis*. (2022, 04 23). Retrieved from https://www.statology.org/univariate-analysis/.

# **Table of Figures**

[Figure 1: First 5 rows of the data set. 4](#_Toc102905340)

[Figure 2: Q-Q Plot of the Churn Variable 6](#_Toc102905341)

[Figure 3: Chi Squared Test of Independence 7](#_Toc102905342)

[Figure 4: Continuous Univariate Box Plot 9](#_Toc102905343)

[Figure 5: Continuous Univariate Density Plot 10](#_Toc102905344)

[Figure 6: Categorical Univariate Density Plot 11](#_Toc102905345)

[Figure 7: Categorical Univariate Box Plot 12](#_Toc102905346)

[Figure 8: Continuous Bivariate Scatter Plot 13](#_Toc102905347)

[Figure 9: Categorical Bivariate Scatter Plot 14](#_Toc102905348)