D207 Exploratory Data Analysis: Performance Assessment

Jill Rivera

Western Governors University

Table of Contents

[I. Variables and Research Question 2](#_Toc101689450)

[II. Statistical Analysis 6](#_Toc101689451)

[III. Univariate Statistics 9](#_Toc101689452)

[IV. Bivariate Statistics 11](#_Toc101689453)

[V. Results 14](#_Toc101689454)

[VI. References 15](#_Toc101689455)

# **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 in order 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. This company would like to analyze their data set to find the variables that are most likely associated with a customer leaving the company.
   2. The research question that will be analyzed for this purpose will be: What variables are able to predict likelihood of a customer to churn away from the company?
2. Independent Variables:
   1. For this analysis, the variables associated with churn needed to be isolated. The variables that were irrelevant to churn, or had a high cardinality, were dropped from the data set. The variables that were dropped were: CaseOrder, Customer\_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Children, Marital, and PaymentMethod. These 17 variables have no effect on the customer churn rate or have more than three data categories in a categorical variable. Eliminating this data allows the statistical tests and visualizations to better depict data related to churn. Following that data drop, a correlation matrix was produced to show the relation between the variables. The data set was then further reduced to 20 variables. The following 13 variables were eliminated from the data set: Options, Reliability, Email, Contract, Contacts, Port\_modem, TechSupport, Techie, Tablet, Phone, Age, InternetService and Income. This will eliminate those variables with low or no correlation to each other as well as reduce the number of variables to increase the accuracy of the ChiSquared test.
   2. To determine the association between variables, a correlation heatmap was created and is shown below. From this, it was determined that timely response, timely fixes and timely replacements are all highly correlated. This graph also shows that tenure and bandwidth are not correlated at all. This analysis shows the variables that when improved, could affect the rate of customer churn.

**Chart, scatter chart

Description automatically generated**

1. Data Set Analysis:
   1. The given data set contains telecommunication customer data. The data set contains 10,000 customer records (rows), and 48 variables (columns). 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.

After cleaning the data set to account for missing, duplicate and irrelevant variables, the data set is reduced to 20 variables. The variables along with their type and an example are listed in section b. The remaining variables consist of mainly categorical variables that do not follow a normal distribution pattern. This status will determine the statistical method used for analysis.

* 1. Table of the Data Set:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** | **Example** |
| Gender | Object | Customer self-identification as male, female, or nonbinary. | Male |
| Churn | Object | Whether the customer discontinued service within the last month. | No |
| Outage\_sec\_perweek | Float64 | Average number of seconds per week of system outages in the customer’s neighborhood. | 6.97256609 |
| Yearly\_equip\_failure | Int64 | The number of times customer’s equipment failed and had to be reset/replaced in the past year. | 1 |
| Multiple | Object | Whether the customer has multiple lines. | No |
| OnlineSecurity | Object | Whether the customer has an online security add-on service. | Yes |
| OnlineBackup | Object | Whether the customer has an online backup add-on service. | Yes |
| DeviceProtection | Object | Whether the customer has device protection add-on service. | No |
| StreamingTV | Object | Whether the customer has streaming TV service. | No |
| StreamingMovies | Object | Whether the customer has streaming movies service. | Yes |
| PaperlessBilling | Object | Whether the customer has paperless billing service. | Yes |
| Tenure | Float64 | Number of months the customer has been with the provider | 6.79551295 |
| MonthlyCharge | Float64 | Average amount customer is charged monthly | 171.449762 |
| Bandwidth\_GB\_Year | Float64 | Average amount of data used yearly in GB | 904.53611 |
| TimelyResponses | Int64 | Survey Response: Timely response | 5 |
| TimelyFixes | Int64 | Survey Response: Timely fixes | 5 |
| TimelyReplacements | Int64 | Survey Response: Timely replacements | 5 |
| RespectfulResponse | Int64 | Survey Response: Respectful response | 3 |
| CourteousExchange | Int64 | Survey Response: Courteous exchange | 4 |
| ActiveListening | Int64 | Survey Response: Evidence of active listening | 4 |

# **Statistical Analysis**

1. Statistical analysis of data is used to test the difference between the mean of the population versus the mean of a sample size of the data. This process assumes a null hypothesis that there is no statistical difference between these calculations, and an alternate hypothesis that there is a statistical difference. The test is then performed resulting in a p-value, which is this calculated difference between the two means. The lower the p-value, the greater the statistical significance of the observed difference. The p-value is compared to the alpha value, which is the p-value limit that is acceptable. Typically, that alpha value is 0.05, which means that you would have a 95% certainty that the difference in variation of the variable cannot occur by random chance. If the p-value is greater than the alpha value, the null hypothesis is accepted and the alternate hypothesis is rejected. The opposite occurs when the p-value is lower than the alpha value.

The test chosen to analyze this data set is the ChiSquare test. This method was picked due to the data set being comprised of categorical variables being in the majority. The variables also do not follow a normal distribution pattern, so T-Test and ANOVA testing will not result in accurate analysis.

1. For this analysis, 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**

This graph shows that the churn variable does not follow the normal standard deviation. A chisquared test is performed on the churn variable to see if we are to accept or reject the null hypothesis as discussed previously.

**Graphical user interface, text, application

Description automatically generated**

This complex analysis is achieved in only one line of Python code using the SciPy Stats Chisquared command. The results shown are the p-value and the test score for the chisquared analysis. (scipy.stats.chisquare, 2022) For the churn variable, the p-value came back as a 1.0. This means that we will need to accept the null hypopthesis of there being no statistical difference between the sample and population data sets of the churn variable. It can be said with a certainty that there is no difference, and that the data is a good model to determine churn rate of customers.

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. 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 command. 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. (Univariate Analysis, 2022) By analyzing a single dependent variable, 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 the Matplotlib library to create a histogram of the variable. A histogram visually summarizes the data point showing the patterns of distribution and statistics easily. 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 monthly charge and tenure were selected to study for the continuous variables. Categorical variables are discrete numerical values with set values. Timely response and timely fixes are the categorical variables that will be shown in this analysis. The following figure shows the four variables for analysis.

Chart, box and whisker chart

Description automatically generated

Analysis of the histograms can show the median value, quartile values and the distribution pattern of the variable. Monthly charge can be seen to have a median value of about $155, a 25% quartile below $100, a 75% quartile above $250, and a normal distribution pattern. Tenure shows a bimodal distribution pattern, indicating that there are two values that occur most frequently. These values would be around 5 and 50 months. These values suggest that there are several new and long-term customers, but little customers with tenure in between those values. Timely response and timely fixes visualize an intermittent normal distribution pattern with a median value of 3.5 each. It is also shown that 1 and 5 are the values of the 25% and 75% quartiles, respectively, of timely response and timely fixes. The result of this analysis proves that the majority of these variables follow a normal distribution pattern and that tenure could be an indication of determining churn rate of customers.

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. This shows a comparison of the two variables and how their distributions relate to each other. (Bivariate Analysis, 2022) In section B, two categorical and two continuous variables are reviewed using this method. A box plot was created for each pair of variables using Matplotlib’s library in Python.
2. The continuous and categorical variables are plotted into box plots below for analysis. The continuous variables chosen were outage seconds per week and tenure.

**Chart

Description automatically generated**

From the box plot, the range of each variable can be viewed easily as well as any outlying factors that are present in the data. No outlying variables are present in the tenure variable; however, the outage variable has several outlying factors outside of the range. The range for the outage variable is between 8 and 12 seconds of down time per week. The tenure variable has a big range from 8 to 61 months of service. The median for customer tenure is clearly 35 months from the graph. This analysis can determine that these variables are not related to each other or dependent of each other. Next, the categorical values of timely replacements and timely fixes are created for the categorical variables.

**Chart, line chart

Description automatically generated**

Again, the range, median and outlying factors for each variable are shown easily in the box plot. Both variables have identicalstatistical measurements with a median value of 3.5, a range of 3 to 4 and outlying factors at 1, 6, 7 and 8. This analysis shows that these variables are closely related and could be dependent of each other.

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 ChiSquared test was to either accept or reject the null hypothesis. The null hypothesis stated that there is no statistical difference in the population mean or variance in the data with the sample size tested. The null hypothesis was accepted in this analysis since the p-value was calculated to be more than 0.05, which means that it can be said with more than a 95% certainty that the data is modeled accurately.
2. 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.
3. This telecommunications company is attempting to reduce the churn rate of their customers due to it costing more to obtain a new customer. The course of action for the company to get this result would be to evaluate the variables closely associated to churn statistically. Following this analysis, these variables were determined to be timely response, timely replacements, and timely fixes. The customer churn rate will improve if these variables are improved. They can be improved through reducing the time that it takes to replace faulty equipment and the time that it takes to respond to problems or questions that a customer has.

# **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/.