Customer Retention

Christopher Flowers

College of Information Technology, Western Governors University

C744 Data Analytics and Mining II

Mr. Smith

July 19, 2021

***I: Tool Selection***  
Execute data extraction from the “Customer Data” web link using data mining software (Python, R, or SAS). Provide a screen shot of the code you have written and its successful application with a copy of all the extracted data.

1. Describe the benefits of using the tool you have chosen (Python, R, or SAS) for extracting data in this scenario.

I chose to use the programming language R for the extracting data for this scenario. R has several benefits to using it and the biggest is its versatility. R has the ability to collect and store a lot of data but also has the ability to run any statistic and graphical analysis needed. Also it is easier to clean and prep the data than python. The visual representation in R is native compared to in python having to install additional packages. The statistical advantages are given to R because of its ability to run deep statistical analysis and heavy statistical models, which can lead to better observation of the data, all this with very few lines of code. For example, by just using the summary method it categories columns to give the count for each string value and give the min, max, mean and median.

1. Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the scenario and are represented in the available data.

Does customer contract type depend on the customers age?

Is churn affected by the loyalty of the customer?

Does paperless billings affect the churn?

Does having internet service affect the churn?

What variables give the greatest gauge of what customer will leave? The goal are to figure out what categories need to be focused on in order to mitigate the amount of customers lost. Also what categories have the highest effect on the churn.

1. Select a descriptive method *and* a nondescriptive method (i.e., predictive, classification, or probabilistic techniques) you will use to analyze the data, and explain how the methods you have selected are appropriate for the objectives or goals you have defined.

The descriptive method picked to use is Multi Corresponded Analysis because it deals with qualitative data versus the principle component analysis which deals with quantitative data. This along with FactorMineR to help determine which categories have the most effect on the churn and by helping highlight the categories which have the least effect on churn. The nondescriptive method chose will be logistic regression because of its binomial response variables in how it predicts. This also helps with giving a probability of the churn outcome.

***II: Data Exploration and Preparation***  
Clean the data you have extracted and save as .xls or .xlsx format for submission. Be sure to address all necessary formatting, converting, and missing da

1. Describe the target variable in the data and indicate the specific type of data the target variable is using, including examples that support your claims.

The target variable in the data is the churn variable with a categorial data type. The categories are either ‘Yes’ or ‘No’.

1. Describe an independent predictor variable in the data and indicate the specific type of data being described. Use examples from the data set that support your claims.

Monthly charges, tenure, total charges are the independent predictor variables because they have a quantitative value. In the first observation the tenure is one with a monthly charge and total charge value of 29.5.

1. Propose the goal in manipulation of the data and define your data preparation aims.

The goal in manipulating the data is to have the cleanest data possible by removing unneeded columns and missing or duplicated data. Also after the data has been cleaned we will split the data into two different data sets.

1. Define the statistical identity of the data, including the essential criteria and phenomenon to be predicted.

Can you define the three quantitative, continuous independent variables? Monthly charges, tenure, total charges

What are the 16 categorical, qualitative independent variables? All columns minus churn which have values of ‘Yes’ or ‘No’ are categorical and qualitative independent values.

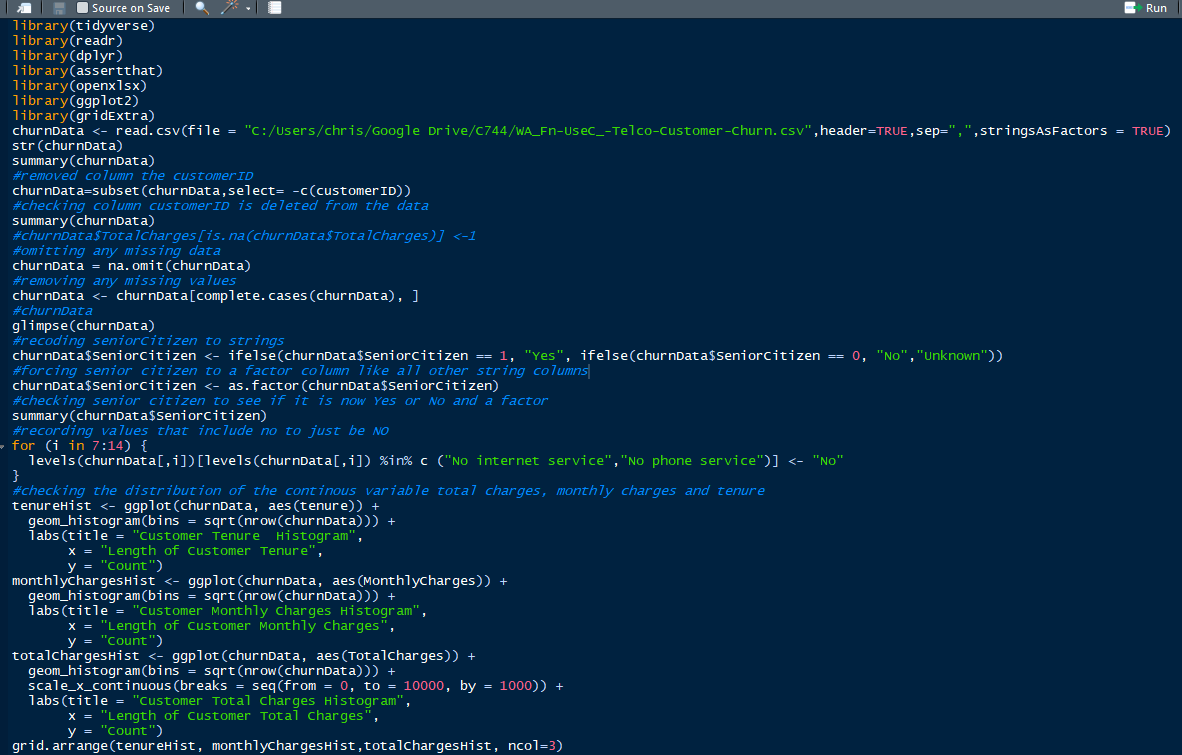
What is the one categorical dependent variable (target)? The categorical dependent variable is churn and it is the only dependent variable in the dataset.

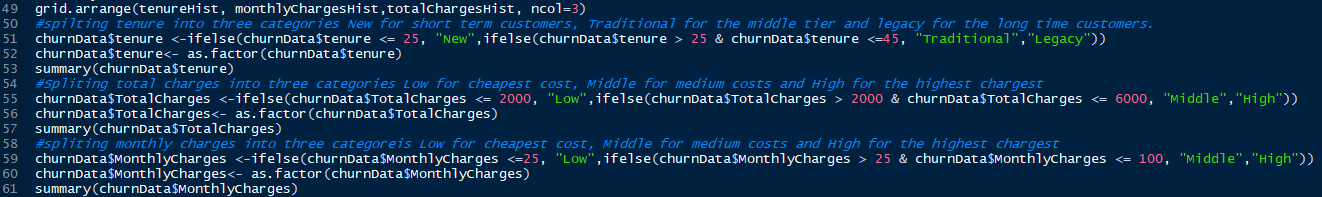
Is there an identifier for all of the individuals records in the dataset? There are key values in the data set to set individual records about in the dataset.

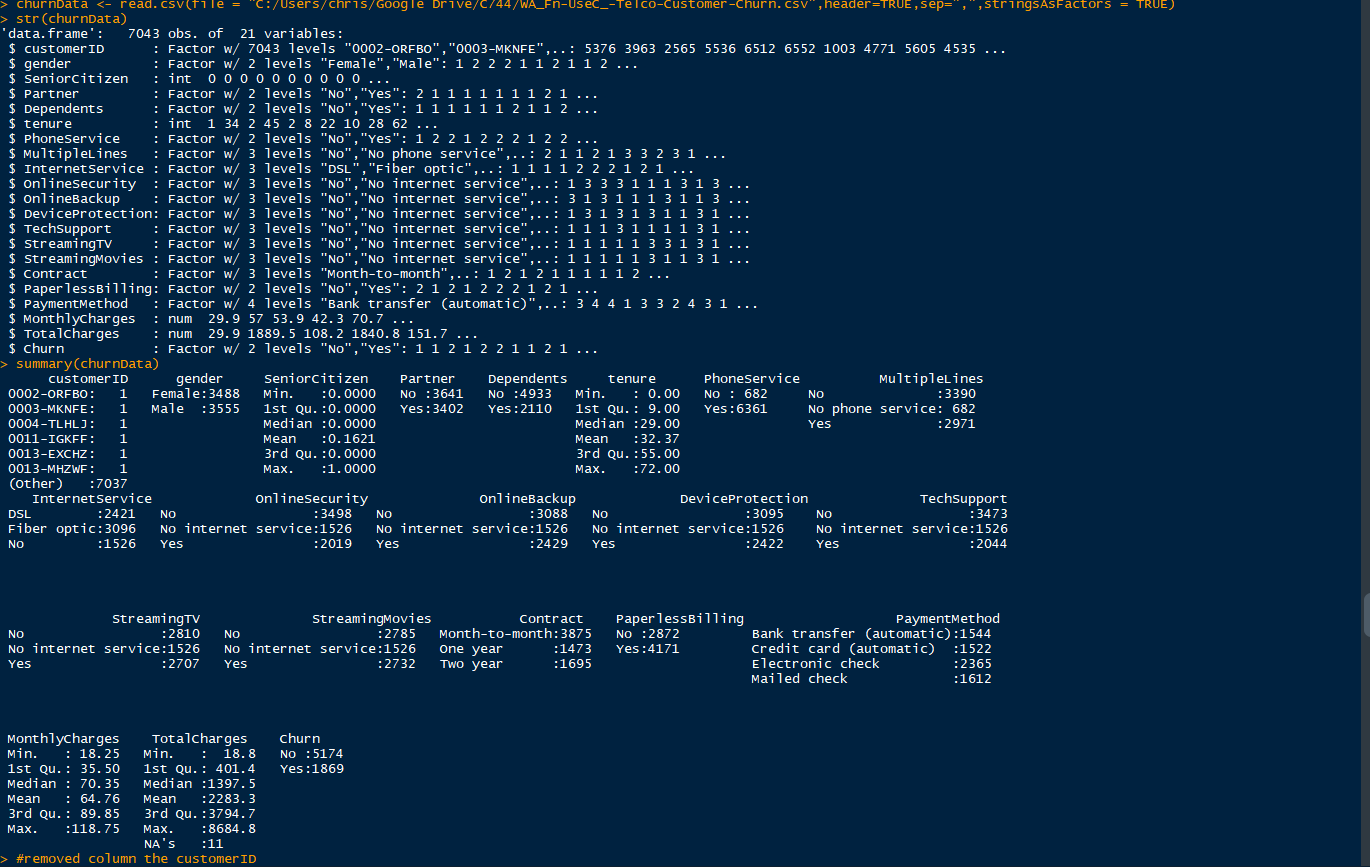
What is the essential criteria to be predicted? Essential criteria to predict how each variable is distributed.

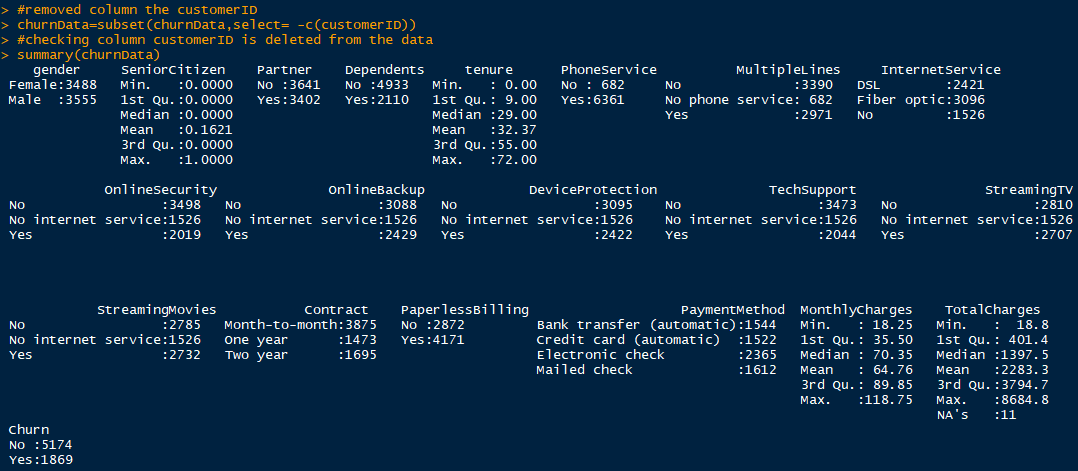
Explain the steps used to clean the data and how you addressed any anomalies or missing data.

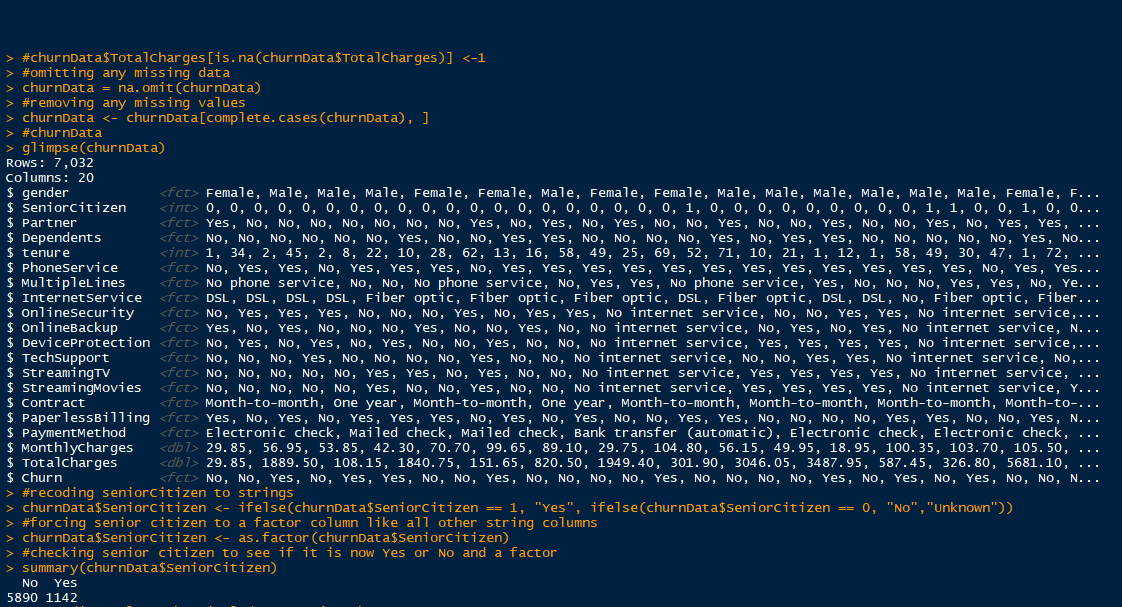
First thing in cleaning the data was to remove the customerID column because it is not need to determine the likely of who would leave the company. We need to check for missing data to do that we will remove it by using an omit function on the data and use the complete.cases function in order to remove any missing values in the data. Next thing is to change values of columns that are not in the right data type and the first column to do this to is the senior citizens. Then any column that has a value which includes ‘No phone service’ or ‘No internet service’ to just be ‘No’. All columns need to have a qualitative value for Multiple correspondence Analysis, in order to do this is to convert three columns. The three columns being converted are tenure, monthly charges and total charges and all columns will split into three categories. Tenure was split into categories called New, Traditional and Legacy. Total Charges and Monthly Charges were spilt into categories called low, middle and high. Now the data is cleaned and ready to be spilt into two data sets.

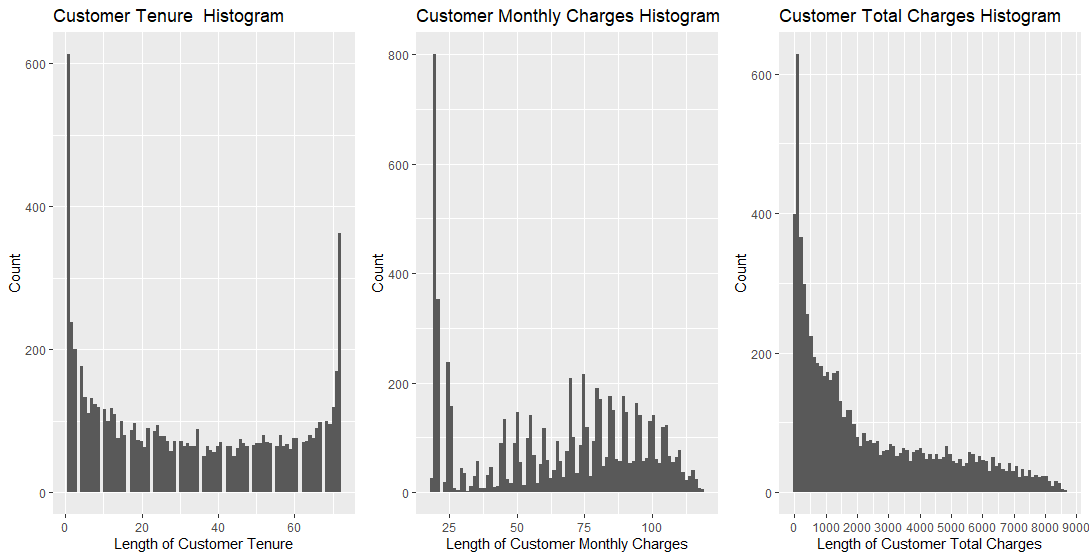


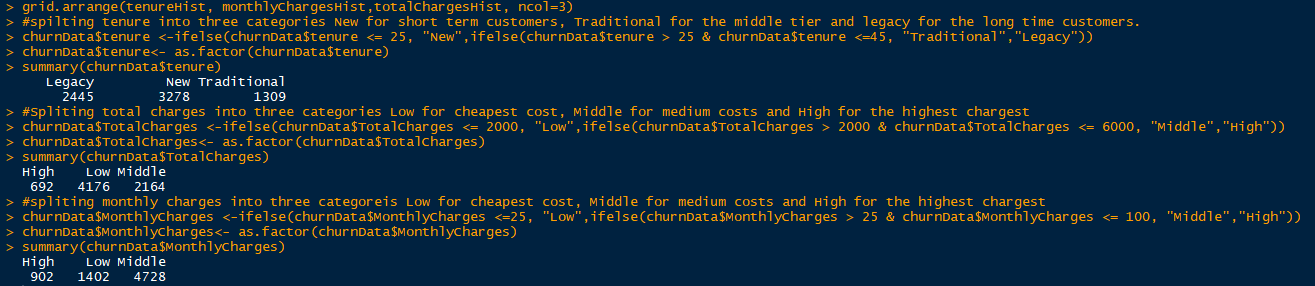










******

***III: Data Analysis***  
For each of the following steps, be sure to clearly indicate each step within your data sheet with a screen shot and annotations in your final submission. All algorithms used need to be clearly identified in the screen shot and submission.

1. Identify the distribution of variables using univariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission.

We have already done a distribution on all the variables in the previous section but this time and show them graphically. The categorical variables used bar charts to show the distributions amongst the subset of each variable. Then histograms were used to show the quantitative variables amongst its subset of each variable.

Summary the charts and identify types of distribution

1. Identify the distribution of variables using bivariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission.

Interpret the data by cross referencing the variables with churn

1. Apply an analytic method and an evaluative method. Annotate the data showing both methods and your findings.

Just apply the methods.

1. Justify the methods you have chosen to analyze your data. Be sure to include details about how the methods you have chosen better represents your findings than other methods.
2. Justify the methods you have chosen to visually present your data. Be sure to include details about how the presentation methods you chose better represents your findings than other presentation methods.

***IV: Data Summary***  
Summarize the findings of your data evaluation. Provide the final findings dataset, including evaluation measures.

1. Explain how your data shows that it was discriminating or not and whether the phenomenon you wanted to detect was present in your findings. Provide specific examples from the data to support your claims.
2. Describe the methods you used for detecting interactions and for selecting the most important predictor variables. Include the specific interactions you detected and the most important predictor variables that you found.
3. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.
4. <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/>
5. <http://www.sthda.com/english/articles/22-principal-component-methods-videos/71-mca-in-r-using-factominer-quick-scripts-and-videos/>
6. <https://www.rdocumentation.org/packages/FactoMineR/versions/2.4/topics/MCA>
7. <https://husson.github.io/MOOC_GB/RMarkdown_MCA_Hobbies.pdf>
8. <https://stackoverflow.com/questions/1249548/side-by-side-plots-with-ggplot2>
9. <https://www.r-bloggers.com/2017/11/predict-customer-churn-logistic-regression-decision-tree-and-random-forest/>
10. <https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/>
11. <http://r-statistics.co/Logistic-Regression-With-R.html>