SMU®

Data Science Machine Learning DS7331.402

Amber Burnett
Lance Dacy
Shawn Jung
Jeremy Otsap

Lab 1

January 26, 2020

Business Understanding

The goal of this project is to evaluate the Telco industry's ability to retain existing customers as well as pinpoint variables to predict if a person might leave the provider ("Churn"). In addition to the customer's propensity to leave; trends will be discovered that allow the firm to proactively install retention programs aimed at slowing the attrition rate of their customers.

According to an article dated April of 2019; DemandJump discovered that the estimated cost of new customer acquisition is approximately five times higher than retaining an existing customer. AnnexCloud discovered that only a third of Telco customers switch carriers due to lower prices. Factors such as dissatisfaction with quality of service, advancing technology and media features, competitors having better cellular coverage, and poorly implemented loyalty programs are among the remaining contributing factors that lead to customer attrition.

The goal is to analyze data in order to discover trends that might help providers retain customers instead of acquiring new ones at the rate they attrit. Happy and healthy customers can affect the bottom line in a positive way for less cost than acquiring new customers. The business question is as follows:

> *"Can we pinpoint positive indicators that lead us to recommend various programs to our Telco client that potentially prevent attrition; ultimately affecting the bottom line in a positive way even though they are increasing spend for the retainment programs"?*

Purpose of the Dataset

The data set acquired from Kaggle is a random sampling of approximately 7,000 anonymized customers from an unnamed Telco provider. It contains 20 features that describe such characteristics as demographics, account information, and subscribed services. The original purpose of the dataset was " To predict tenure and loyalty of the customers".

The data set includes information about:

- Customers who left within the last month – the column is called churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Define and Measure the Outcomes

InData Labs provides some guidance on common tools in Data Science. These tools are available to help define and measure outcomes. Usually the problems in defining measures fall into two categories that will be explored throughout this project:

- Classification problems
- Regression problems

Classification problems revolve around being able to predict what category a feature falls into within the dataset. The common tools for classification problems are:

- Percent correction classification (PCC): measures overall accuracy. Every error has the same weight.
- Confusion matrix: also measures accuracy but distinguished between errors, i.e false positives, false negatives and correct predictions.
- Area Under the ROC Curve (AUC – ROC): is one of the most widely used metrics for evaluation. Popular because it ranks the positive predictions higher than the negative. Also, ROC curve it is independent of the change in proportion of responders.
- Lift and Gain charts: both charts measure the effectiveness of a model by calculating the ratio between the results obtained with and without the predictive model. In other words, these metrics examine if using predictive models has any positive effects or not.

Regression problems revolve around predicting a quantity within the dataset. The common tools to measure outcomes for regression are:

- R-squared: indicate how many variables compared to the total variables the model predicted. R-squared does not take into consideration any biases that might be present in the data. Therefore, a good model might have a low R-squared value, or a model that does not fit the data might have a high R-squared value.
- Average error: the numerical difference between the predicted value and the actual value.
- Mean Square Error (MSE): good to use if you have a lot of outliers in the data.
- Median error: the average of all difference between the predicted and the actual values.
- Average absolute error: similar to the average error, only you use the absolute value of the difference to balance out the outliers in the data.
- Median absolute error: represents the average of the absolute differences between prediction and actual observation. All individual differences have equal weight, and big outliers can therefore affect the final evaluation of the model.

Effectiveness of a Prediction Algorithm

As the project progresses the data set purpose is to help the firm retain existing customers, therefore the effectiveness of the model will be based on the model's ability to predict outcomes via regression variables or classification variables.

Given that only a third of customers switch carriers due to lower prices; more and more factors such as dissatisfaction with quality of service, advancing technology and media features, competitors having better cellular coverage, and poorly implemented loyalty programs are all contributing to customer attrition.

By analyzing all relevant customer data in this dataset, we will effectively understand the performance of each of the models that are tested by selecting metrics that truly measure how well each model achieve the overall business goal of customer retention.

The effectiveness of the algorithm will be scored by how well the model can have the same predictive ability across many different datasets (testing and training sets at minimum). The results will need to be comparable, measurable, and reproducible. The process by which this is determined will evolve as the project moves forward.

Essentially, the goal is to predict whether a customer will "Churn" based on the selected categories of the model. This will allow the providers to be pro-active in creating promotional programs aimed at the probably of churned customers as well as target the variables that have the most positive impact on customer loyalty (tenure).

Data Understanding

The initial pass of data inspection indicates that the data is wellformed and provides ample data size for predictive analysis. Given that most people have a device of some kind, our own team can relate to some of the indicators that might lead a customer to change providers. This knowledge is helpful, but also could be harmful if bias is introduced.

Key insights of this data related to the customer's churn indicators are as follows:

- 45% of the users that church are paying via electronic check
- 43% of the users that churn are under a month-to-month contract
- 42% of the users that church have fiber optic internet
- 42% of the users church if they do not have online security features
- 42% of the users church that opt out of TechSupport for their devices
- 42% of the senior citizen users tend to churn
- 40% of the users that church do not opt for OnlineBackup capabilities
- 39% of the users that church opt out of the DeviceProtection programs
- The longer the customers remain with the provider the less chance they churn
- The higher monthly charges tend to lead to a higher churn rate

Data Description

The following fields are contained in the dataset along with a description of the field, values, and potential scale of the variable.

- customerID: Unique alpha-numeric string to anonymously represent an individual customer
- gender: Categorical String value to represent customer's gender (Male or Female)
- SeniorCitizen: Boolean int value to show whether the customer is a senior citizen or not (1, 0)
- Partner: Boolean string value showing whether the customer has a partner or not (Yes, No)
- Dependents: Boolean string value showing whether the customer has dependents or not (Yes, No)
- tenure: Numeric value showing number of months the customer has stayed with the company
- PhoneService: Boolean string value showing whether the customer has a phone service or not (Yes, No)
- MultipleLines: Categorical string value that shows if the customer has multiple lines or not (Yes, No, No phone service)
- InternetService: Categorical string value that shows the customer's internet service provider (DSL, Fiber optic, No)

- OnlineSecurity: Categorical string value showing whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup: Categorical string showing whether the customer has online backup or not (Yes, No, No internet service)
- DeviceProtection: Categorical string showing whether the customer has device protection or not (Yes, No, No internet service)
- TechSupport: Categorical string showing whether the customer has tech support or not (Yes, No, No internet service)
- StreamingTV: Categorical string showing whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies: Categorical string showing whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract: Categorical string that represents the contract term (Month-to-month, One year, Two year)
- PaperlessBilling: Boolean string showing whether the customer has paperless billing or not (Yes, No)
- PaymentMethod: Categorical string that shows the customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges: Numeric value showing the amount charged to the customer each month
- Churn: Boolean string showing whether or not the customer 'churned' or terminated services (Yes or No)

Data Quality

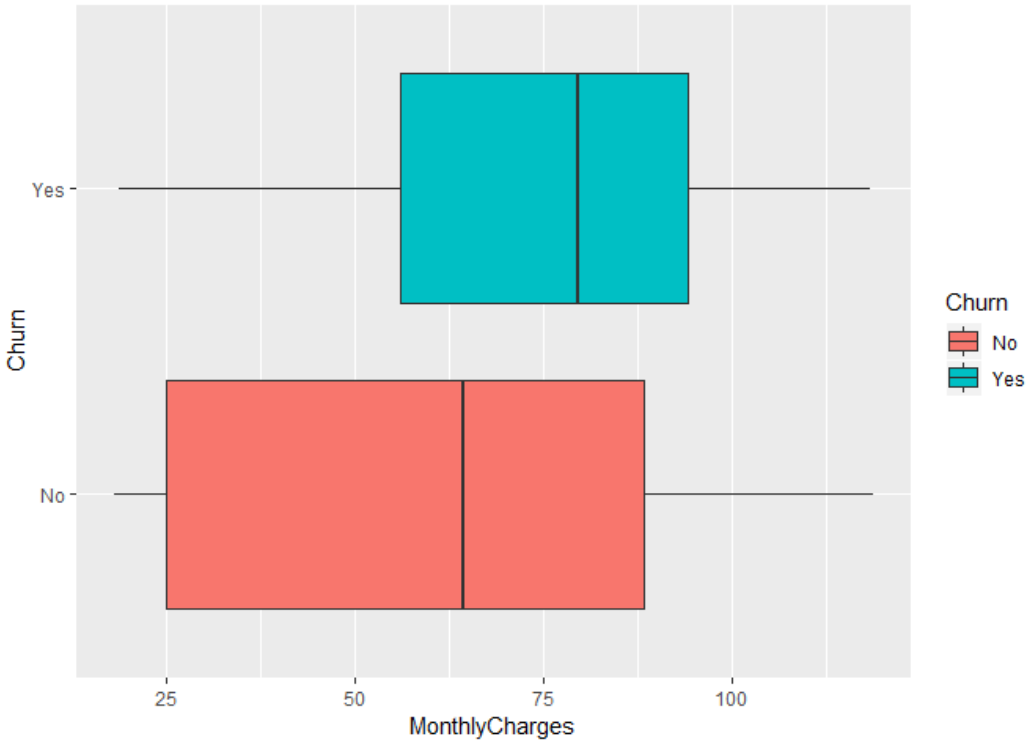
Naturally, with any large dataset, it is prudent to evaluate missing values and determine imputation strategies that allow statistical analysis to remain sound. This dataset had very few missing data points:

- MonthlyCharges has 11 missing values. Given it is such a small percentage of the total ~7,000 values, it should not affect our analysis. At this point, we are simply excluding these 11 observations from the initial data exploration and visualization exercises.
- SeniorCitizen is given as numeric, when for all practical purposes it is a categorical factor. (i.e. a customer cannot be 0.77 of a senior citizen; they either are under 65 or they are not). We will simply convert from an integer to a categorical factor.

To allow for visual reference for each categorical variable, Box Plots for each "Churn" category were used to assess distribution in an effort to determine whether this data can be used for predictive analysis.

The Box Plot for "Churn" and "Monthly Charges" (Figure 1) shows a significant difference in the distribution. Validating this with a Two-Sample t-Test for Equal Means, will likely show the two distributions differ significantly.

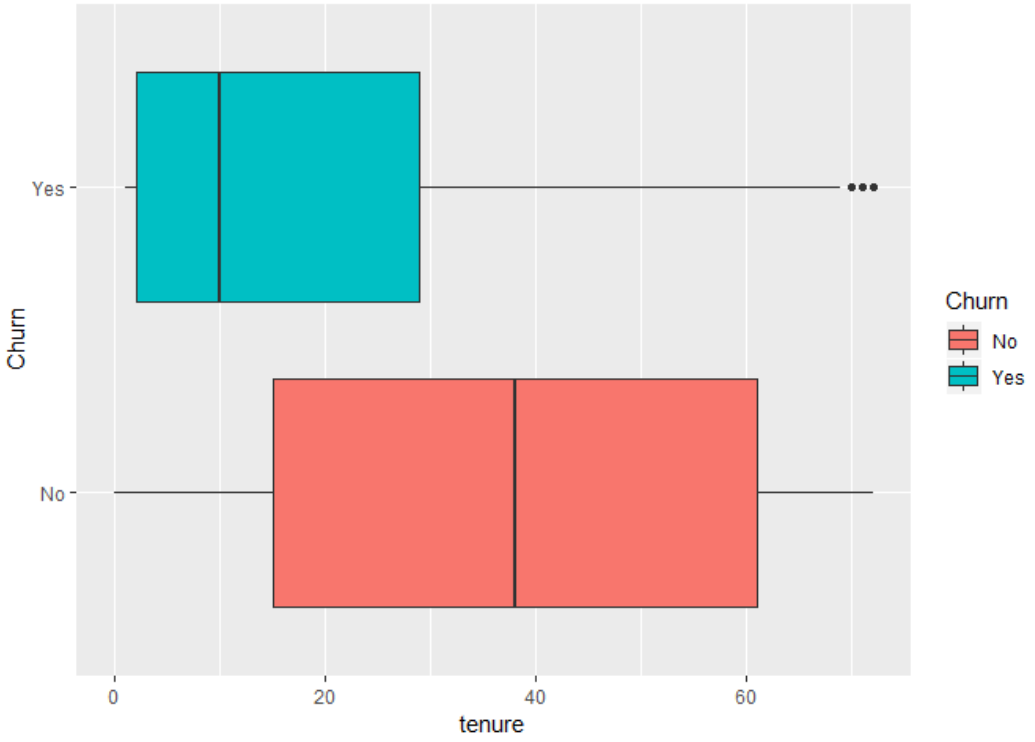
Figure 1: Box Plot for Monthly Charges and Churn



Since the data are just observations, causality cannot be inferred. However, given that the data are a random sampling; generalizations can be made. The plot indicates there is some correlation between "Monthly Charges" and "Churn", which can be tested with Logistic Regression and/or Linear Discriminate Analysis.

Upon investigating "Tenure" against "Churn" (Figure 2), a Welch's t-test would likely show significant deviations in the two different distributions. This is fairly intuitive since customers who "Churn" usually do so within the first one to two years of their service contract.

Figure 2: Box Plot for Tenure and Churn



Moving more toward the advanced visuals to round out assumptions, the data was split into 2 separate data frames; numerical and categorical. Both will contain the response variable "Churn" which allows for appropriate visualizations based on each of variables considered for analysis.

Simple Statistics

As part of the data exploration exercise, a few simple summary statistics provide insight into the type of data that has been provided (Figure 3).

Figure 3: Summary Statistics

Tenure

Min	1st Qu.	Median	Mean	3rd Qu.	Max
$0.00	$9.00	$29.00	$32.37	$55.00	$72.00

Monthly Charges

Min	1st Qu.	Median	Mean	3rd Qu.	Max
$18.25	$35.50	$70.35	$64.76	$89.85	$118.75

Total Charges

Min	1st Qu.	Median	Mean	3rd Qu.	Max
$18.80	$401.40	$1,397.50	$2,283.30	$3,794.70	$8,684.80

Using these statistics as a starting point, it is apparent that “Tenure" is slightly skewed to the right with the Mean hovering around $32.00. 50% of the data ranges from approximately $9.00-$55.00.

“Monthly Charges” show that the Median is greater than the Mean, indicating data is skewed to the left, with a long tail of low “Monthly Charges” pulling the Mean down more than the Median.

“Total Charges” is skewed to the right, indicating there is an abundant amount of extreme “Total Charges” in the data. The Mean of the “Total Charges” hovers around $2,000.00 with 50% of the data ranging from approximately $400.00-$4,000.00.

Visualizations

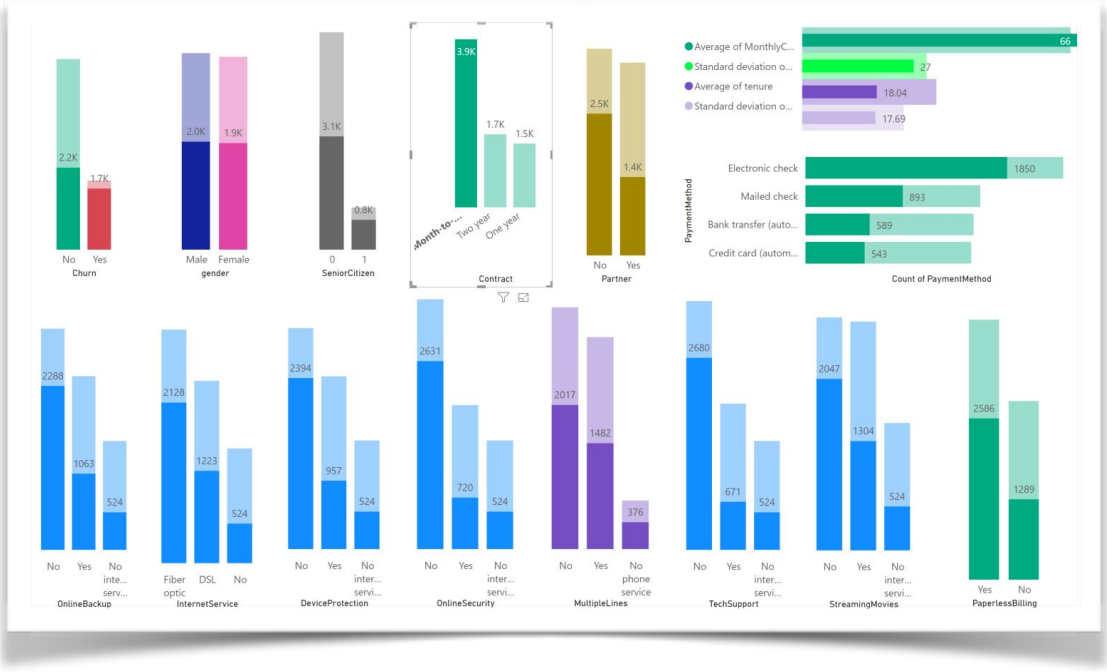
Given the number of variables in the data set, it is prudent to reduce the ones used for regression down to as few as possible. Keeping the model simple will help prevent "over-fitting". Visualizing the important variables will help us ensure the model can be kept within margins for regression.

The variables will be introduced in this section to allow for inspection of them individually. Exploration of the relationships as well as comments on the interesting ones as the journey to regression continues can be found in the "Exploring Relationships" section. Thus far the following five variables are to be considered in the model:

- "Contract"
- "PaymentMethod"
- "MultipleLines"
- "TechSupport"
- "Total Charges"

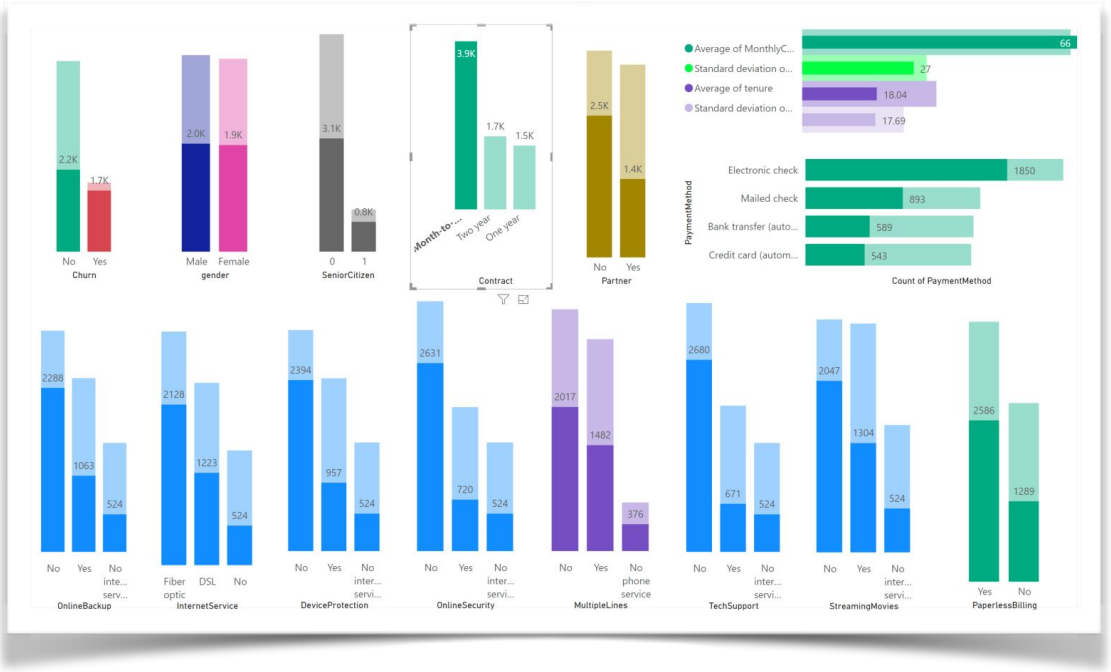
Monthly Contract

Using the visuals below, a deduction can be made that the ratio of "Churn" (Yes v. No) is disproportionate (having almost a 1 to 1 relationship). The 1 year and 2 year are more closely aligned to the total Yes vs. No behavior.



Payment Method

In addition to the "MonthlyContract" relationship, it is important to note that most of the contracts are paid with eCheck.



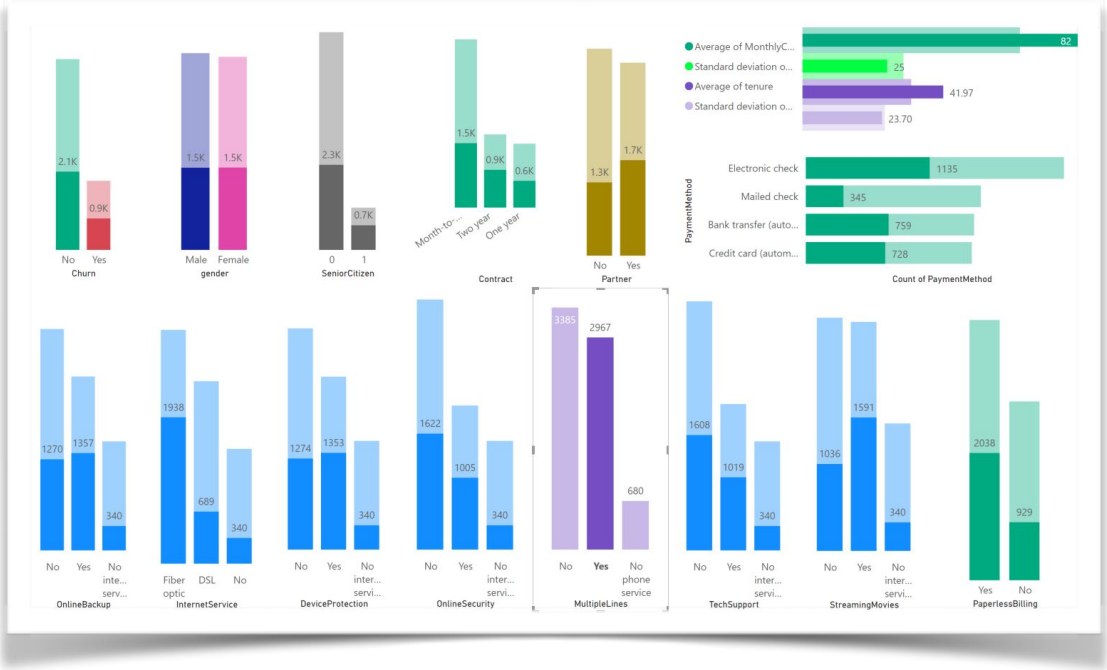
This can be a signifiant indicator to the ease of changing providers. One would think that having automatic payments would be a reason to stay with a provider (given the time and setup required), the data shows that this not necessarily true. While most of the data represented suggests that a majority of paying customers use eCheck, it also indicates that those are the most likely to leave the provider.

Is this because they are more technically savvy? Is it because they demand more of the provider? This variable has much to be explored to see if it can be used to deduce variations in the "Churn". It can be also deduced that given that the majority of contracts are paid with eCheck it naturally has the higher rate of "Churn" by coincidence to the statistics.

Multiple Lines

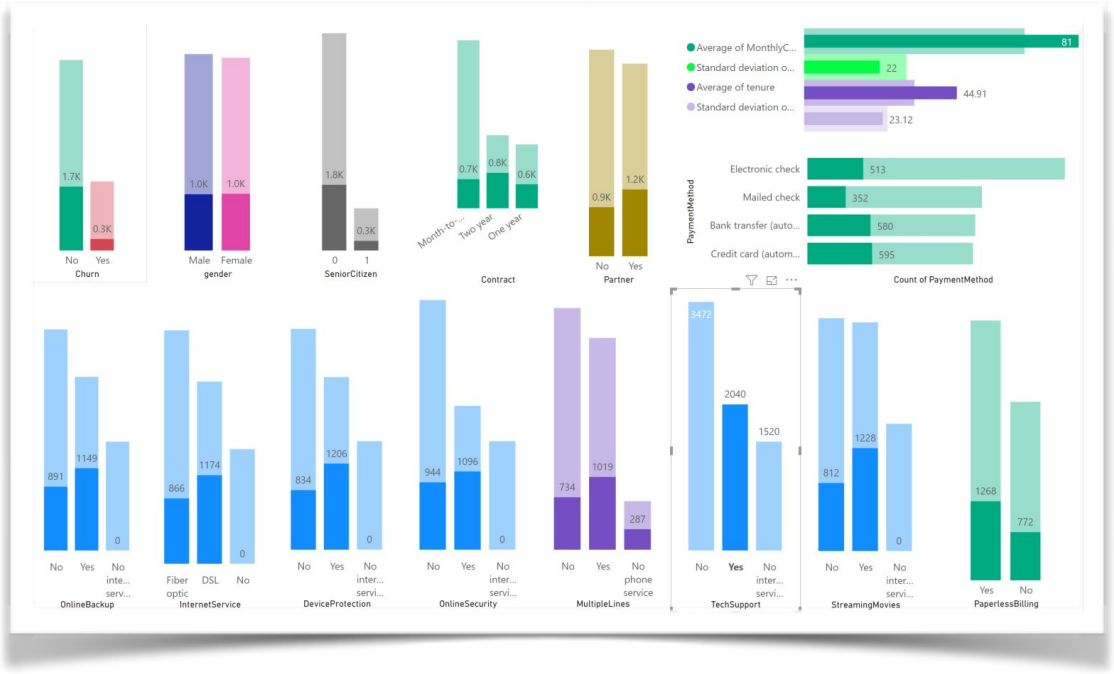
This metric is interesting in the sense that most families have more than one line on the contract. It would appear that the more families that use the same contract, the harder it is for them to change providers. Of the 2,967 customers who have multiple phone lines, 340 of them do not have Internet capabilities. This would indicate an opportunity to market services like Skype or WebEx or even advanced VoIP services and video conferencing.

These customers also tend to have a higher average monthly spend and higher average tenure. Also of note is the ratio for churn (~2 to 1), is significantly higher than the entire dataset's churn ration (~3 to 1). This data point could lead to the conclusion that the customers spend their money more productively by getting modern features (VoIP and video conferencing) for the same money, and multiple lines are no longer needed. Inference can be made as well that this leads to better loyalty, better service, and the customers could have migrated off of an aging technology like analog phone lines.



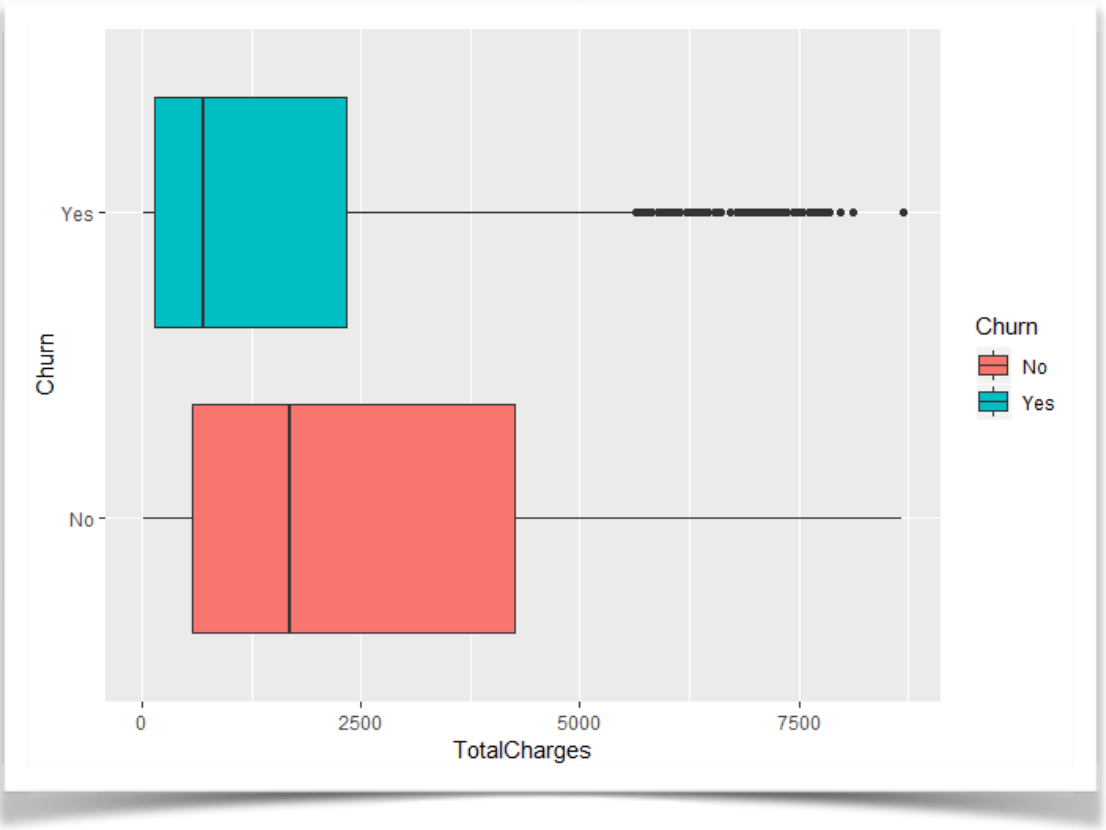
Technical Support

Most providers allow for technical support in their contracts as an option. The population is diverse in people that are used to technology and can figure out most of the options on the phone while others struggle significantly which leads to poor outlook overall on the provider. It appears in the data that customers that opt for technical support tend to have a higher tenure rate with the provider.



Total Charges

Naturally customers are likely to look for the lowest cost solutions in just about any area of their life for discretionary spending. Interestingly that the data provided seems to describe the behavior that if a customer's total chargers are lower, they are more likely to leave the provider. In addition to this phenomena is the idea that the range of total charges is smaller with those who leave the provider.



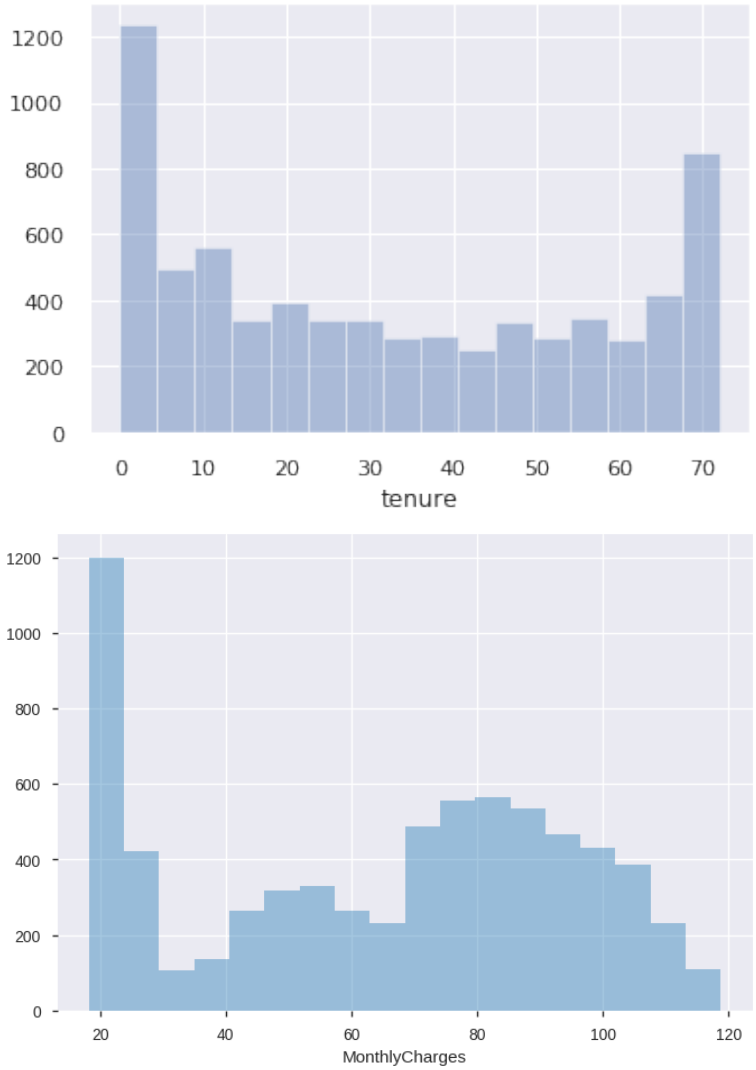
Exploring Relationships

In this section, relationships among the variables will be explored; especially feature related to the target variable ("tenure").

Histogram

The first question in the dataset appeared to be around the probability that "tenure", "MonthlyContracts", and "TotalCharges" can be transformed into categorical values. Their histograms below (Figure 4) indicate a lot of short-term customers and long-term customers at the sides. This is an indication that defining a stratified category variable over "tenure" might be necessary.

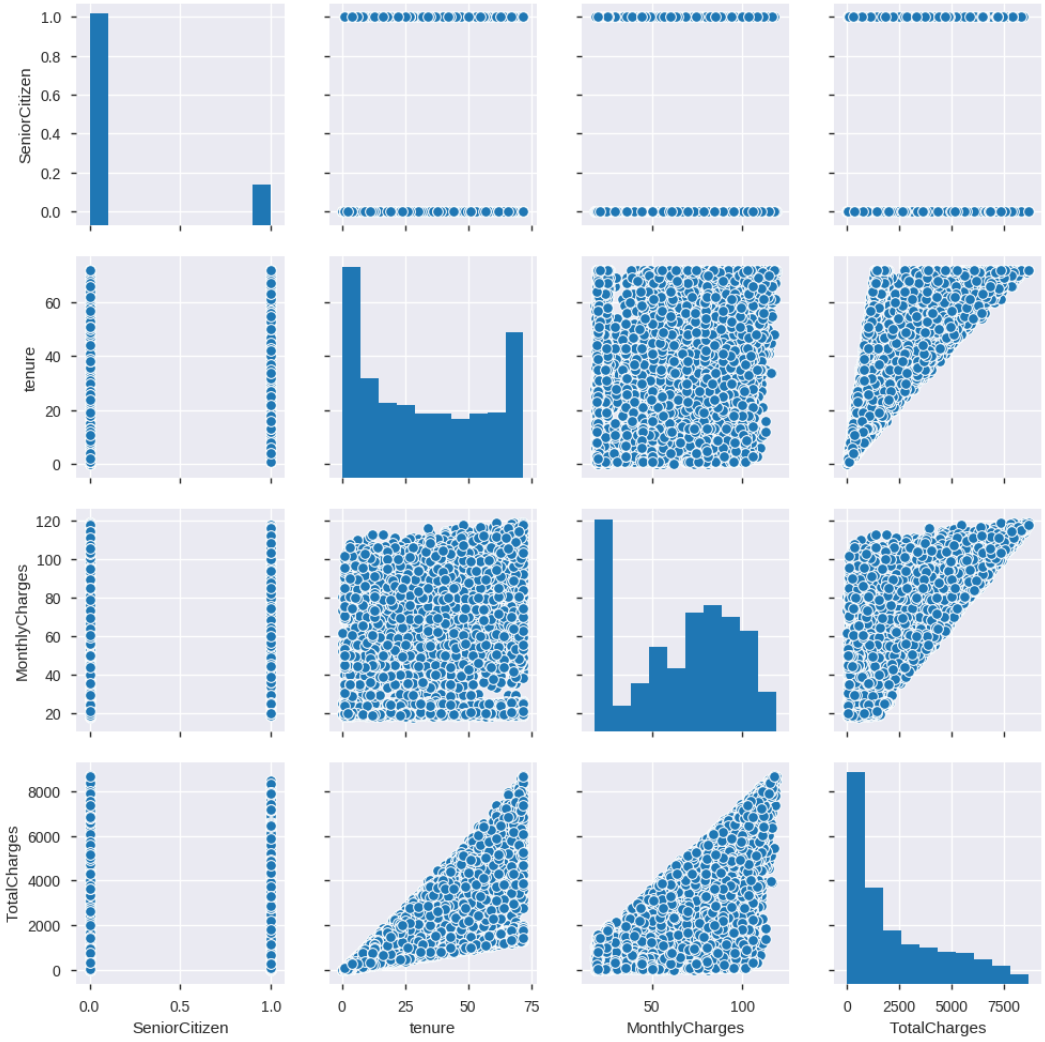
Figure 4: Initial Tenure Relationship



Scatter Plot

In order to determine (visually) the initial target relationship; a scatter plot matrix will allow for the beginning stages of a multiple regression analysis. Figure 5 below provides insight into this starting point of deciding relationships.

Figure 5: Scatter Plot Matrix



In the table below (Figure 6), month-to-month contracts have the highest percentages of customers who left while two-year contracts have the highest percentages of customers who stay.

Figure 6: Summary of Contracts and Churn

Contract Types vs. Churn

Churn	Month to Month	One Year	Two Year
No	2,220	1,307	1,647
Yes	1,655	166	48

It is also interesting to note the relationship between Internet services and their tenure (Figure 7). DSL service has a higher percentage of customers who stay with the provider than those with Fiber Optic. In addition, people with no Internet service stay longer than people with Internet service.

Figure 7: Internet Service Types

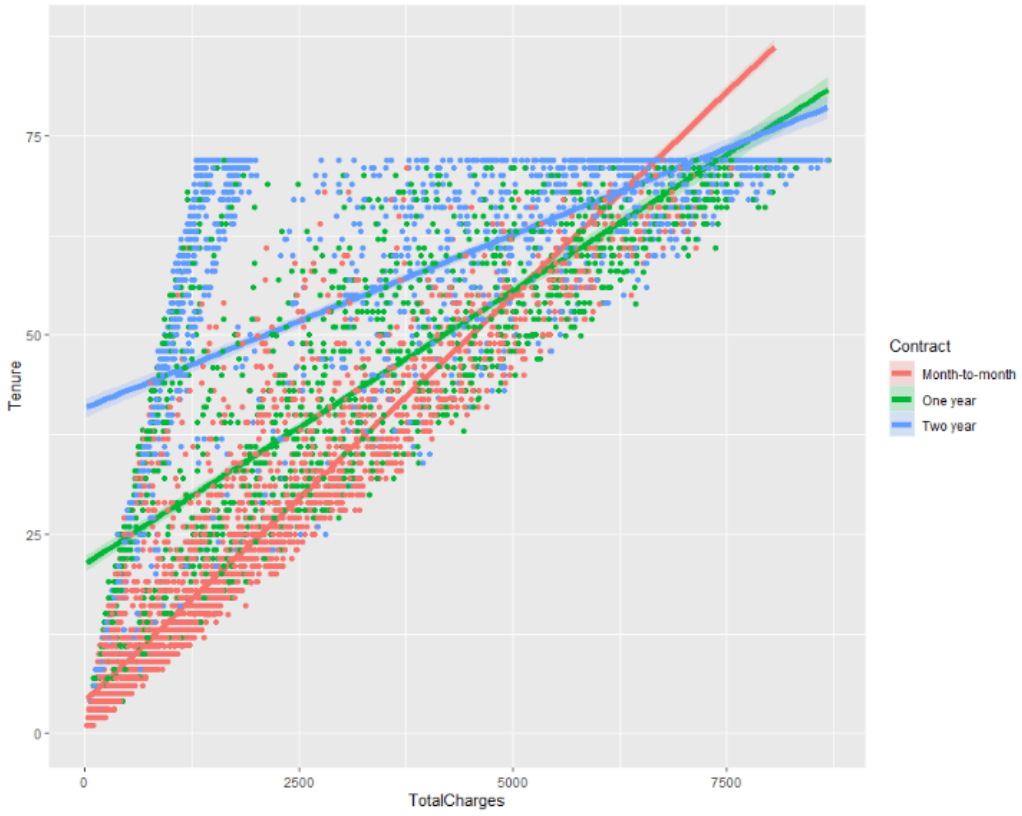
Internet Service vs. Churn

Churn	DSL	Fiber Optic	None
No	1,962	1,799	1,413
Yes	459	1,297	113

Tenure and total charges appear to have positive relationship (Figure 8). The correlation varies by different types of contracts. Holding the total charges between 0 and $5,000, the tenure is longer for two-year contracts than a one-year or month-to-month contract.

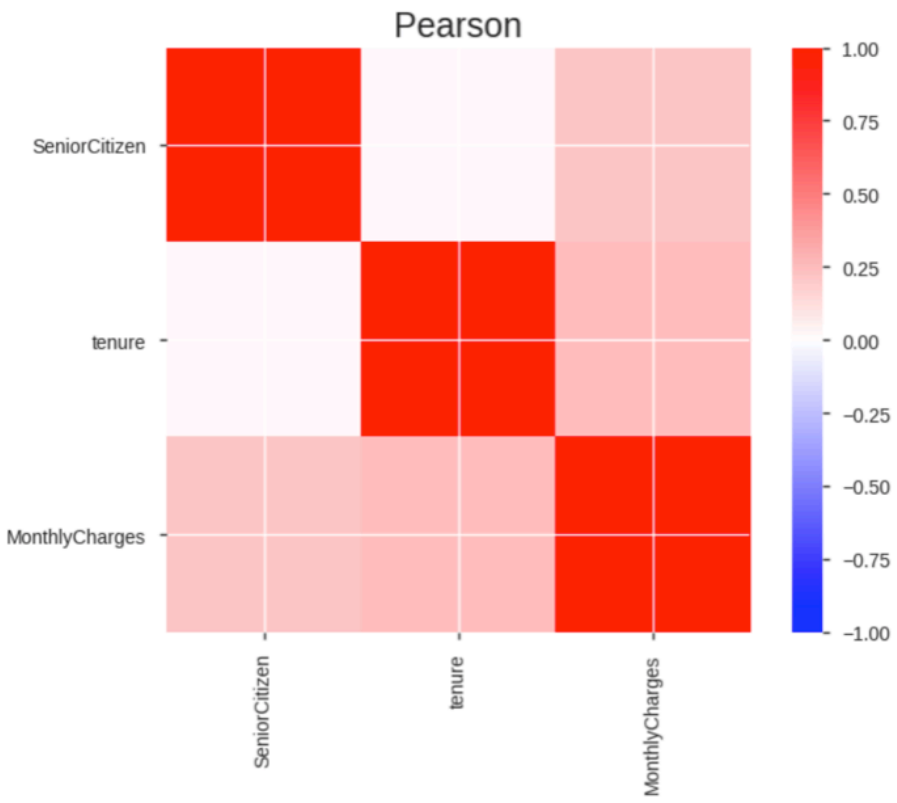
When total charges are more than $5,000 or $7,500, this direction can be reversed. This leads to an interesting trend in the relationship that needs to be explored further.

Figure 8: Initial Tenure Relationship



When looking through various correlations, Senior Citizen also continued to manifest itself as a statistically significant variable. Using a Pearson correlation coefficient (Figure 9), the variables “SeniorCitizen”, “tenure”, and “MonthlyCharges” seem to be relevant.

Figure 9: Pearson Correlation Coefficient



A Pearson correlation coefficient ranges from −1 to 1, making it a bit easier to decipher relationships. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases.

A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables. The Pearson chart above (Figure 8) allows for visually seeing the possibility of the correlation on a gradient scale.

Interesting Relationships

The dependent variable is whether a customer stays or leave which is indicated by the variable churn. There were many interesting relationships between factors and class.

- For example, one-year and two-year contracts account for higher percentages of loyal customers
- People without multiple lines and also used mailed checks have a higher probability of loyalty
- People that pay via electronic check have higher probability to leave the provider
- People who have tech support and no internet service are more likely to be loyal
- People who pay more have a higher probability of being loyal and have a longer tenure

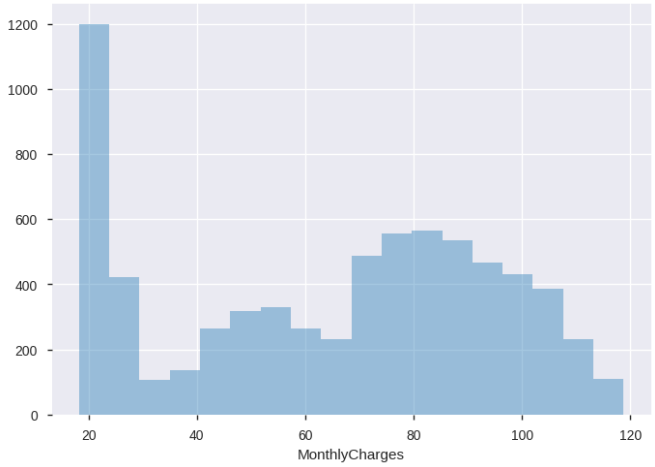
Other Features to Consider

During the first pass of a large data set there are many things to be considered. Most of the summary analysis has been provided; however, there are other features that will need to be considered as this data is explored and modeled.

There should be a consideration to transform the numeric "MonthlyCharges" feature into a stratified categorical feature. The descriptive statistics of this variable (Figure 10) shows mild skewness or lower Kurtosis, but we see homogeneous low-charge group at the left side of the histogram.

It is assumed that the churn behavior of such low-charge group can be different from other mid-to-high charge groups, and confirm the hypothesis as the exploration continues.

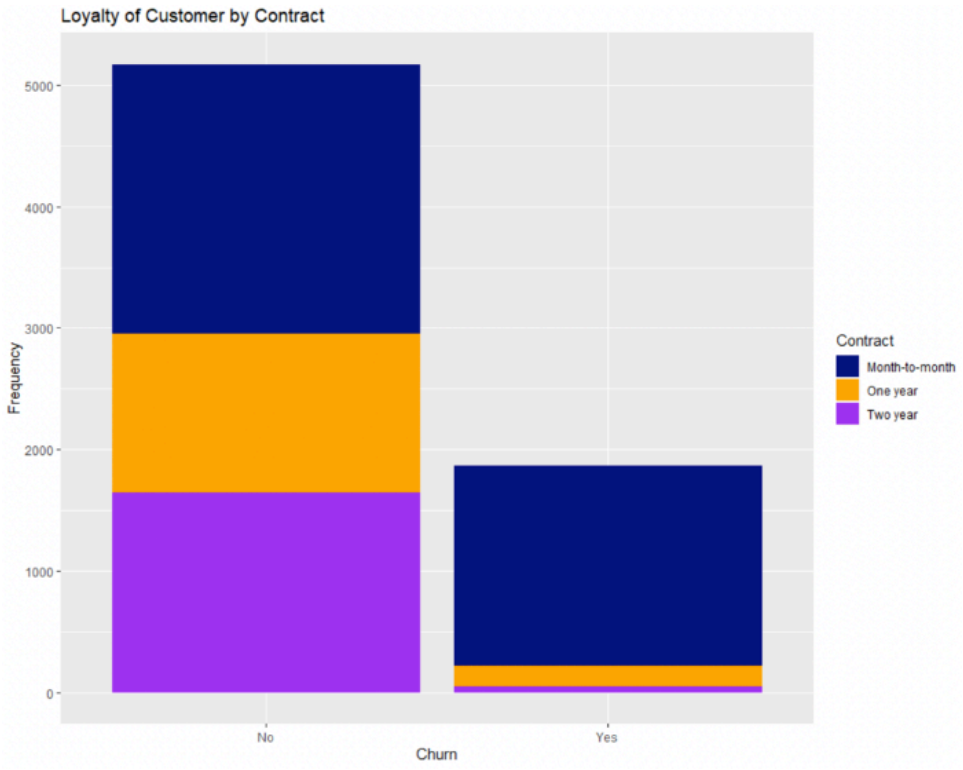
Figure 10: Loyalty of Customer with Technical Support



Type	Value
Standard Deviation	30.09
Coefficient of Variation	0.46463
Kurtosis	-1.2573
Mean	64.762
MAD	26.222
Skewness	-0.22052
Sum	456,120
Variance	905.41

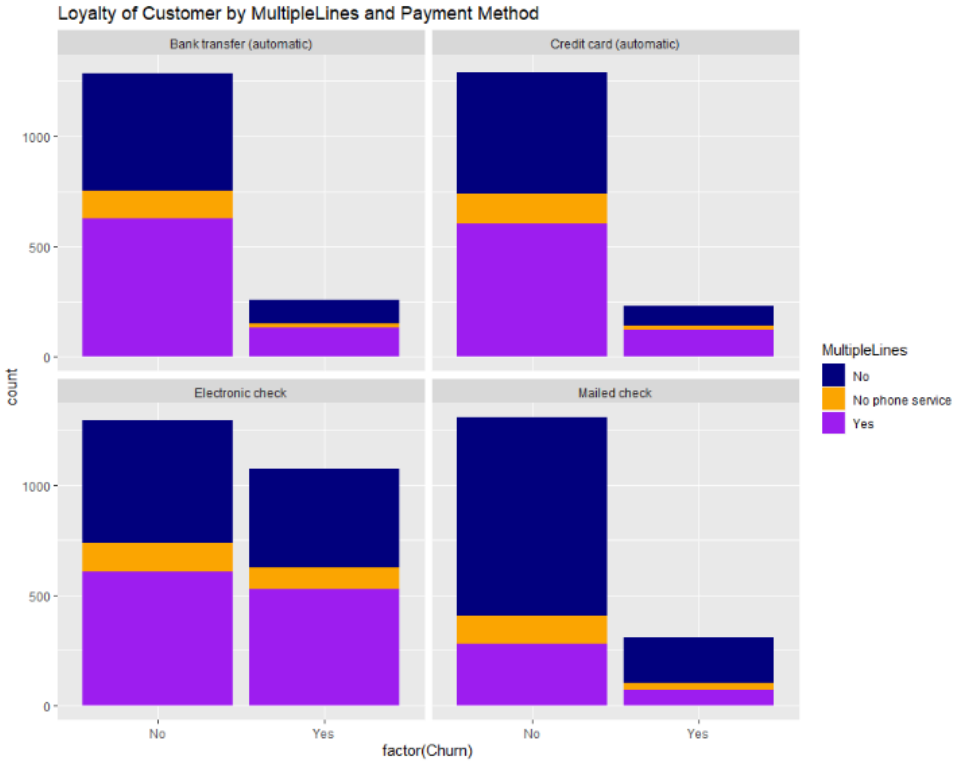
In Figure 11 below, 1 year and 2 year contracts accounted for a higher percentage of customers who remained with the provider (“tenure").

Figure 11: Loyalty of Customers by Contract Length



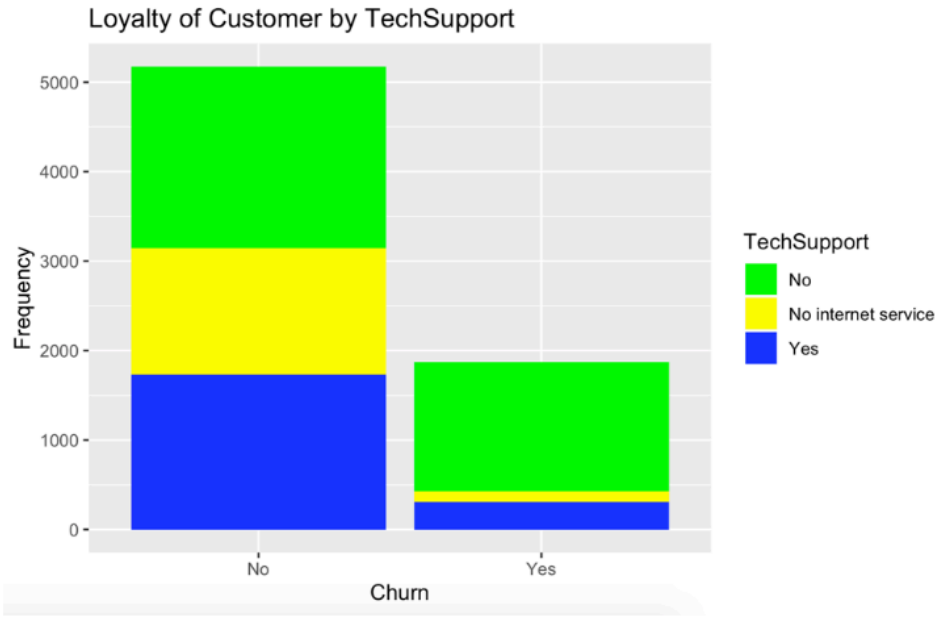
In Figure 12 below, it is visible that people without multiples lines that also used “Mailed Checks” have a higher probability to remain with the provider (“tenure”). People who pay with eCheck have a higher probability to changing providers. This was also presented in the “Visualizations” section under “Payment Methods”.

Figure 12: Loyalty of Customers by Multiple Lines and Payment Methods



As “tenure” is further explored (Figure 13), people who have no Internet capabilities on their device, but opt for technical support are more likely to remain with the provider.

Figure 13: Loyalty of Customer with Technical Support



Appendix

In order to provide visualizations into the data as well as summary and advanced statistical modeling, many tools are used. These tools naturally required programming skills and data manipulation skills. In order to make this work as reproducible as possible, the following links to coding notebooks are provided.

Code Link: Initial Discovery
Code Link: Relationship Discovery
Code Link: Visuals (outside of PowerBI)