**Logistic Regression: Customer Churn**

Heidi L. Petersen

College of Information Technology, Western Governors University

D208: Predictive Modeling

Dr. William Sewell

June 4, 2021

**Objectives of Analysis**

**The main objectives of this analysis will be to discover which variables most customers discontinuing service with the telecommunications company and the specific indicator variables that suggest a customer will leave the company. The target variable will be churn, a binary variable with “yes” and “no” as possible values.**

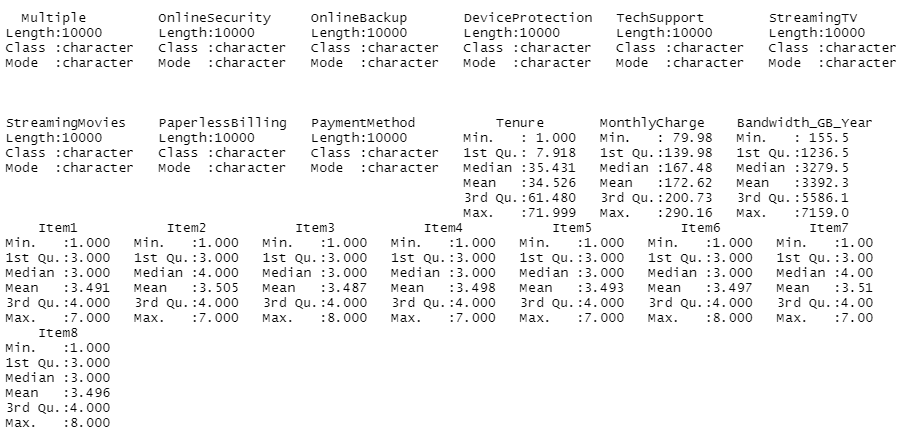
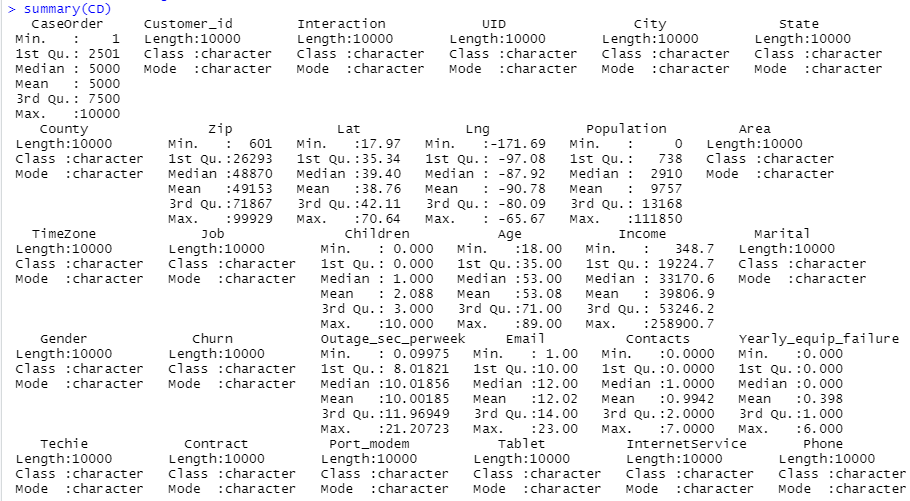
**The analysis will use logistic regression. This technique uses many predicator variables to predict the probability of outcome of a specific response variable. It is an appropriate choice for this analysis because we have a binary outcome; In this case, whether the customer will switch providers. The assumptions of logistic regression include that there exists a linear relationship between the independent and dependent variables, the independent variables do not have a high correlation, and the residuals are normally distributed with a mean of zero.**

I chose R to process this data. R is a good resource for statistical analysis. It has the features needed to achieve the goals of this particular analysis quickly and efficiently. I mostly chose R for its ease. SAS would have been my choice if the data set had been much larger, and I needed to worry about RAM. Python is more of a generalized program where R is specifically designed for statistics.

**Cleaning the Data Set:**

The goals of data preparation are to drop irrelevant data, check for missing or unexpected values, transform variables, and do an initial overview of all the variables in the data set.

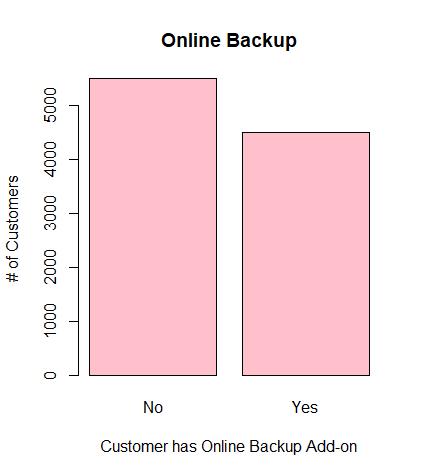
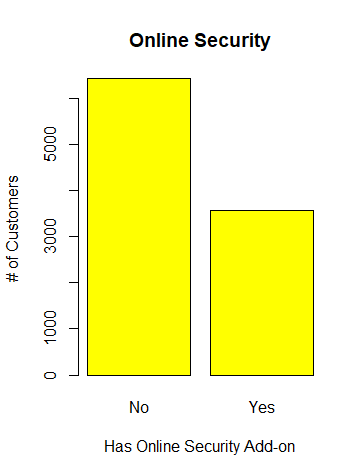
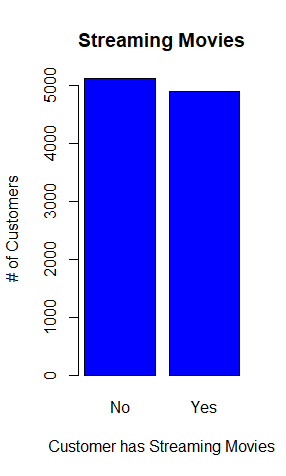
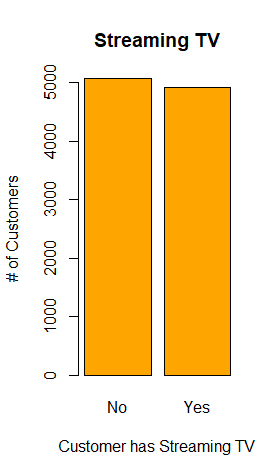
I imported the data set into R and changed the data into a data frame. I summarized the data frame to get an idea of what I would be working with. There are 10,000 observations and 50 variables contained in the set.

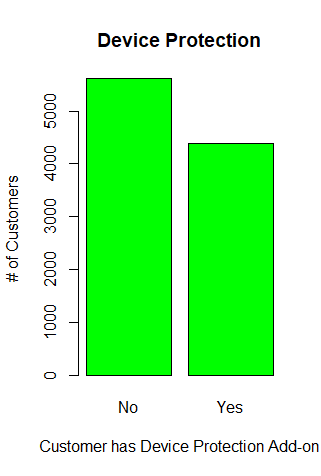
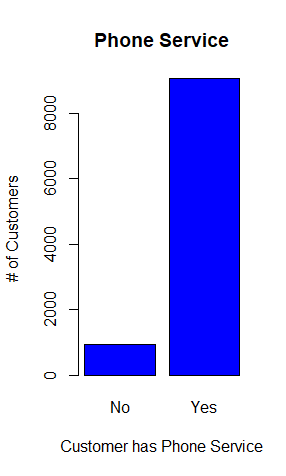
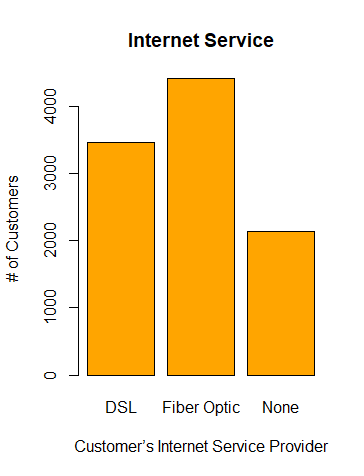
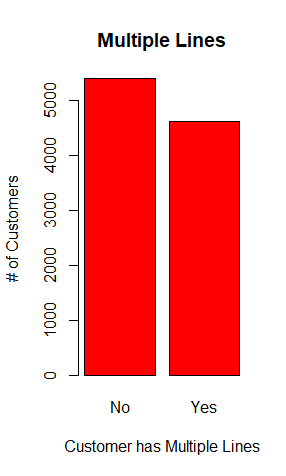


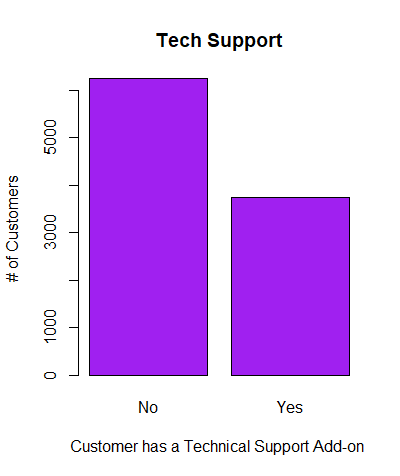
I first dropped CaseOrder, Customer\_id, Interatction, and UID because they are not relevant to the analysis. City, State, County, Zip, Lat, Lng, Population, TimeZone can be covered by Area so I dropped those as well. I dropped the Job column because that info is captured in Income, and it could lead to overfitting by including specific occupations. Monthly charges and bandwidth are clearly correlated to the other columns about the telecommunication usage so I dropped those columns. There were not enough details to know if the survey questions would be relevant or helpful. The survey could be biased based on when it was sent out or how it was distributed so it was not part of the analysis. I changed age to age group, income to income levels, outage seconds per week into outage groups and tenure to tenure groups so that my variables were categorical and more uniform.

**Univariate Analysis**

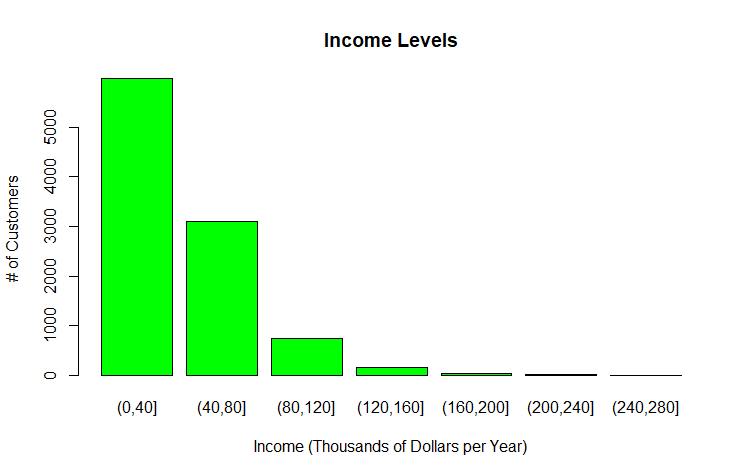
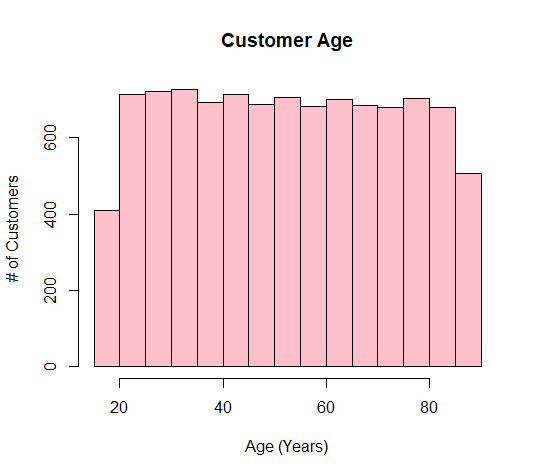
Included in the data, was information on the services each customer utilized. These included streaming TV, streaming movies, online security, online backup, device protection, multiple lines, internet service, tech support, and phone service. The service used more than not by customers, besides internet, was phone service. The internet chosen most was fiber optic. The rest of the services were not in the majority of usage. Streaming Tv and movies were close but still more customers opted out of the service.

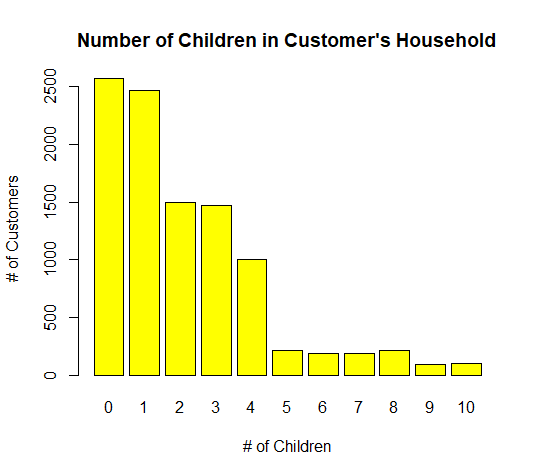
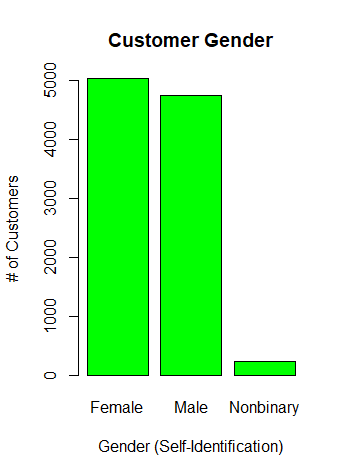
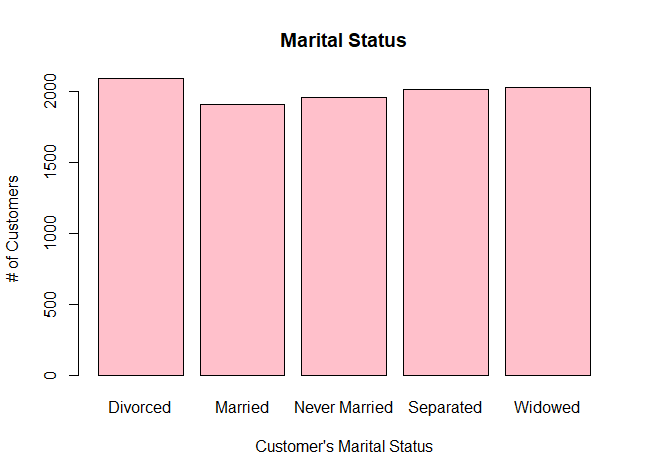
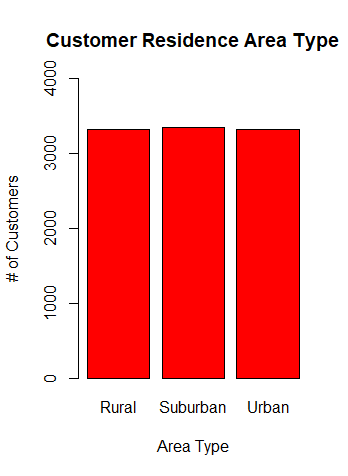


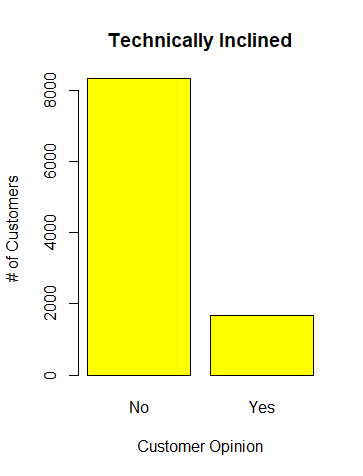
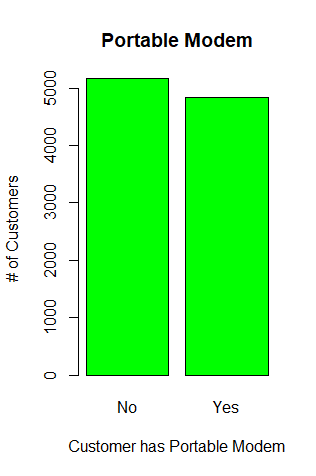
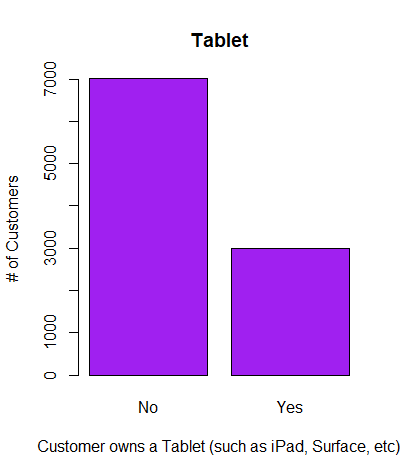
 



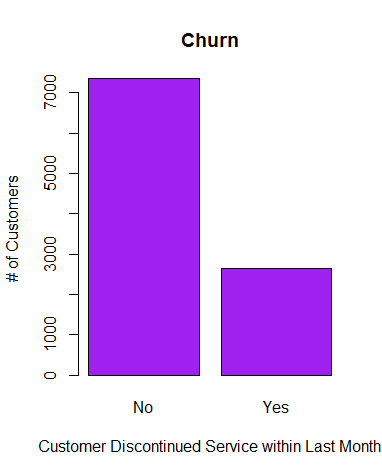
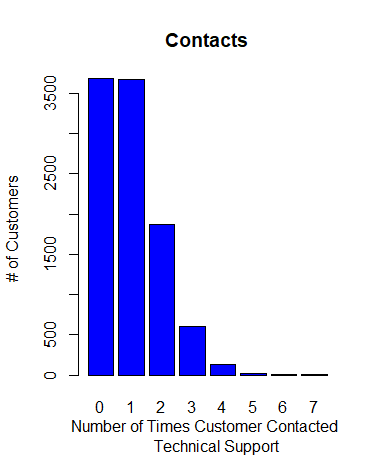
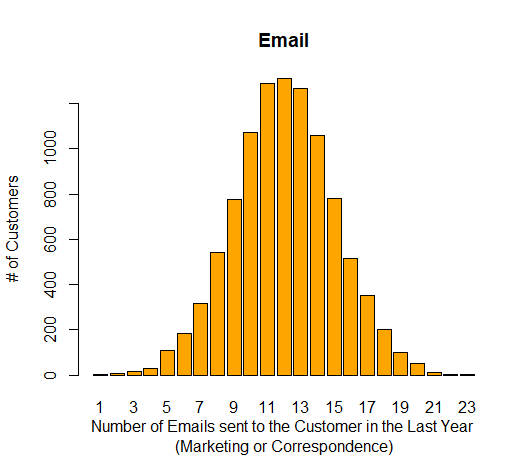
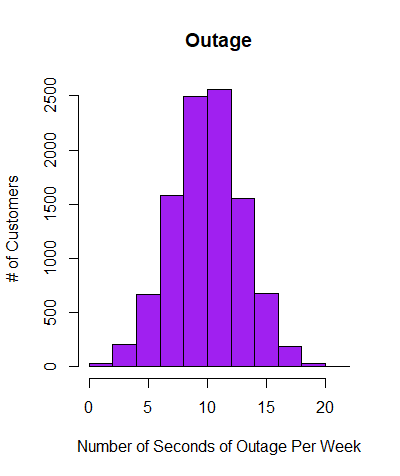
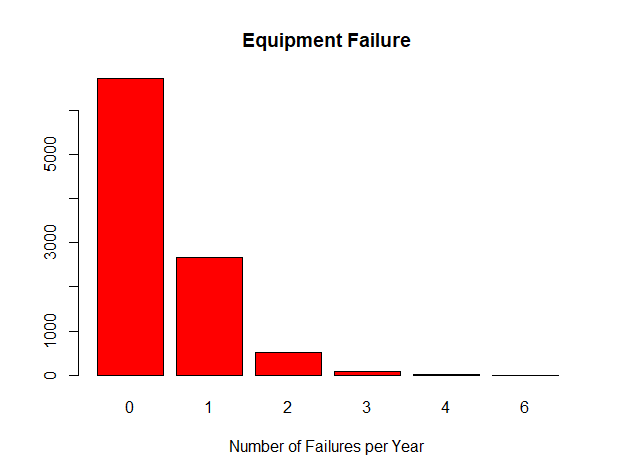
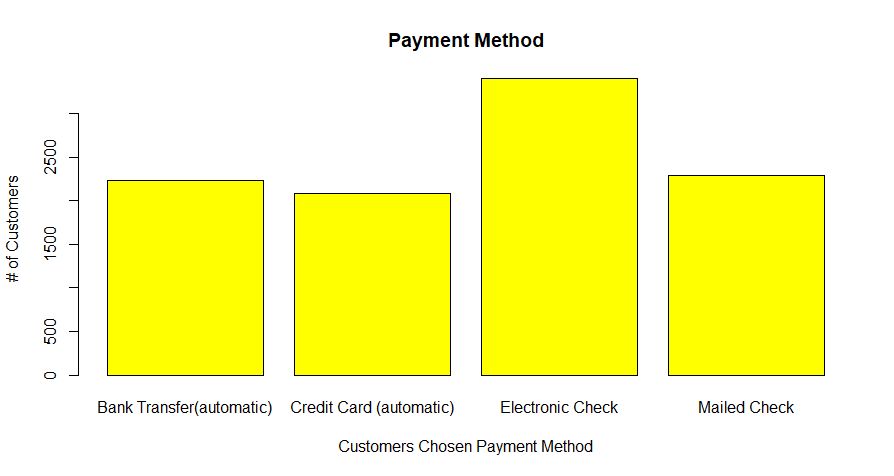
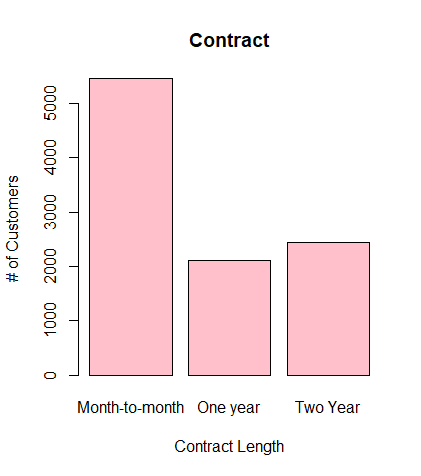
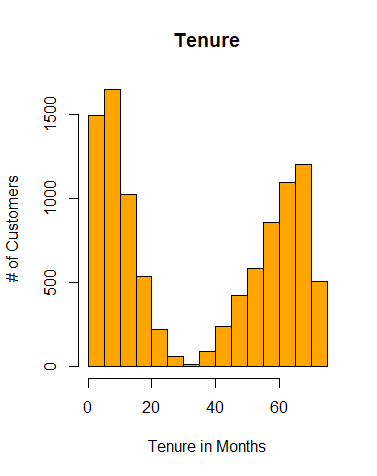
The data also included additional information on each customer. These categories included age, marital status, gender, number of children in the household, income, area, whether they own a tablet or portable modem, and if the customer felt they were technically inclined (most felt they were not).





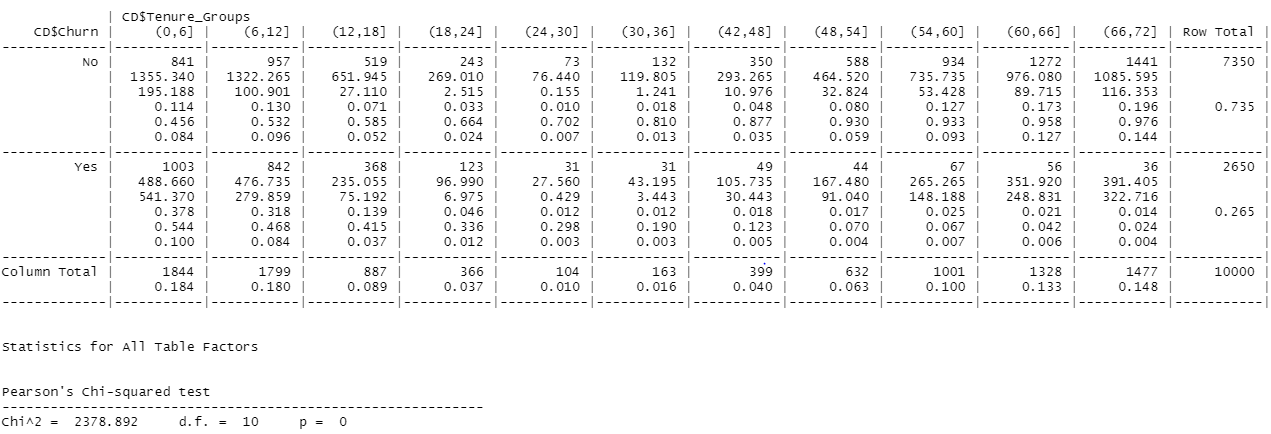
The data includes information about the customers experience with the company: length of time they have been a customer, the type of contract, their chosen payment method, their use of paperless billing, how many seconds per week they experienced outage, how many times they had equipment failure per year, how many emails they received from the company, how many times they contacted tech support, and whether they had left the company in the last month.



**Bivariate Analysis**

I created a crossplot with each variable against churn. I also graphed each variable with the percentage of each that churned to better visualize the relationship. I have listed them as follows in order by smallest p value to largest. The variables towards the top have the highest correlation with switching providers in the last 30 days.

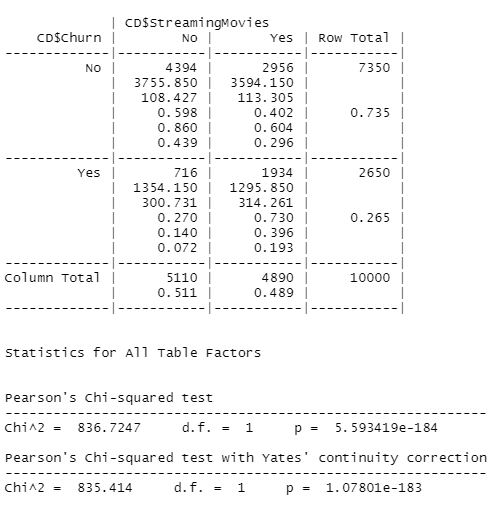
Tenure\_Groups



Chart, bar chart, histogram

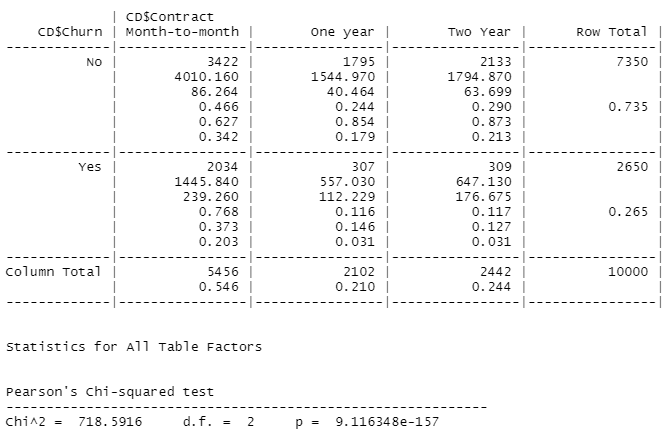
Description automatically generated

Streaming Movies

 Chart, bar chart

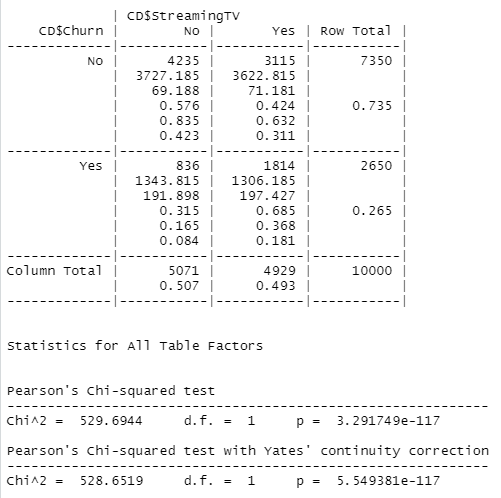
Description automatically generated

Contract

 Chart, bar chart

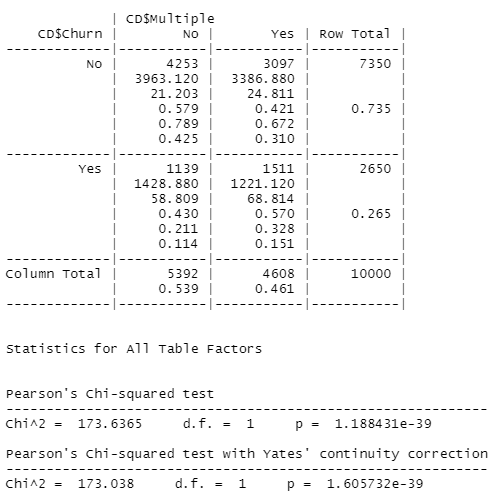
Description automatically generated

Streaming TV

 Chart, bar chart

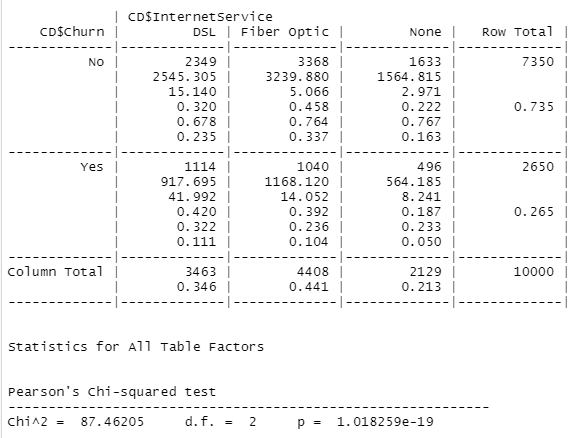
Description automatically generated

Multiple Lines

 Chart, bar chart

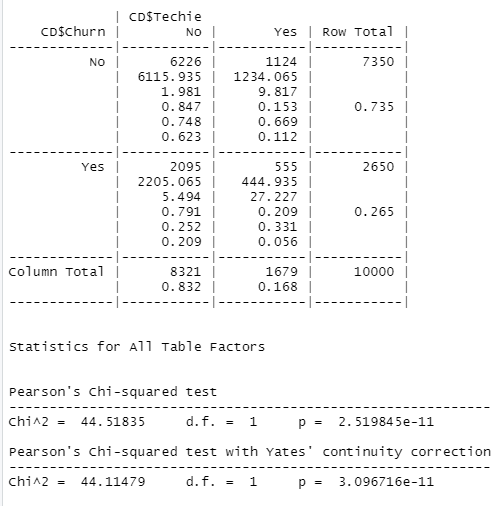
Description automatically generated

InternetService

 Chart, bar chart

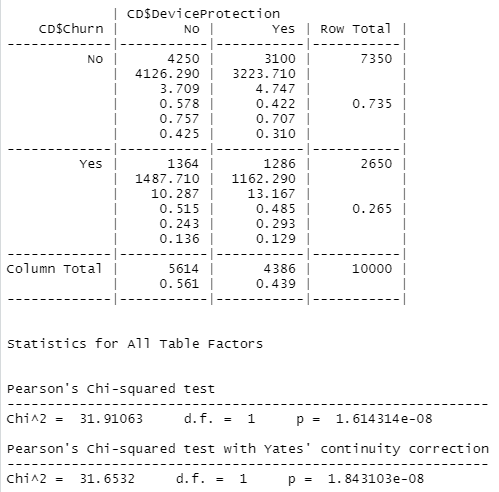
Description automatically generated

Techie

 Chart, bar chart

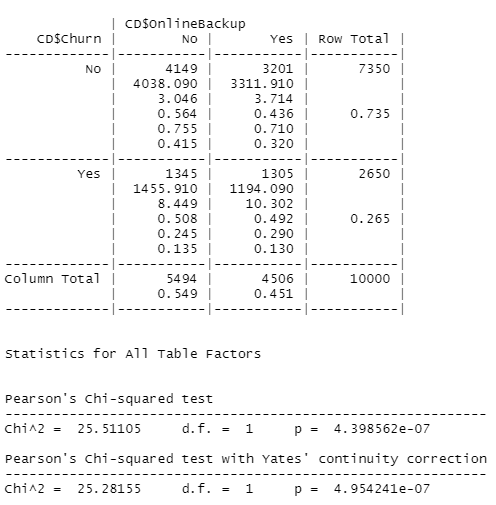
Description automatically generated

Device Protection

 Chart, bar chart

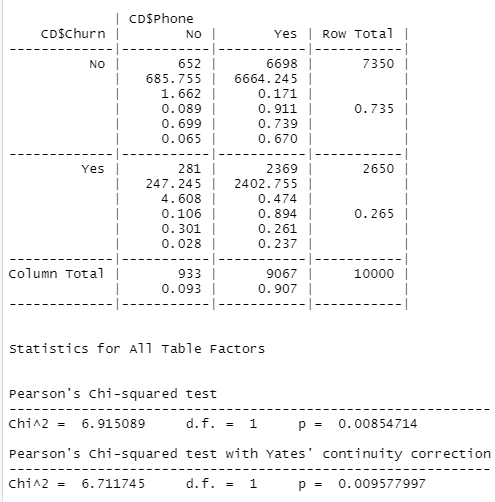
Description automatically generated

Online Backup

 Chart, bar chart

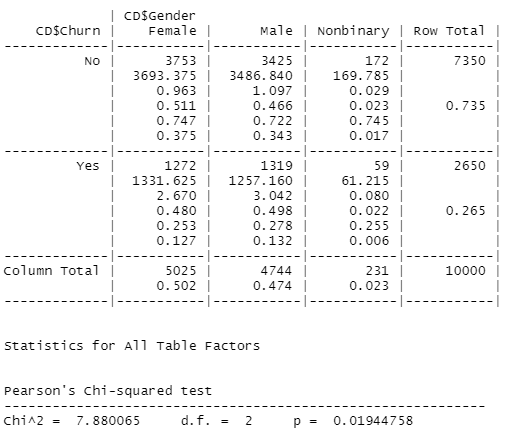
Description automatically generated

Phone

 Chart, bar chart

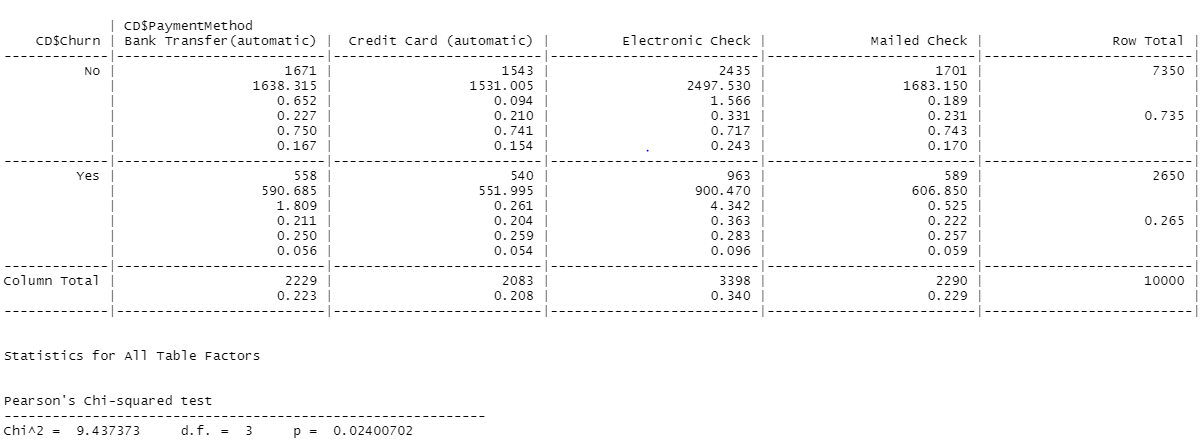
Description automatically generated

Gender

 Chart, bar chart

Description automatically generated

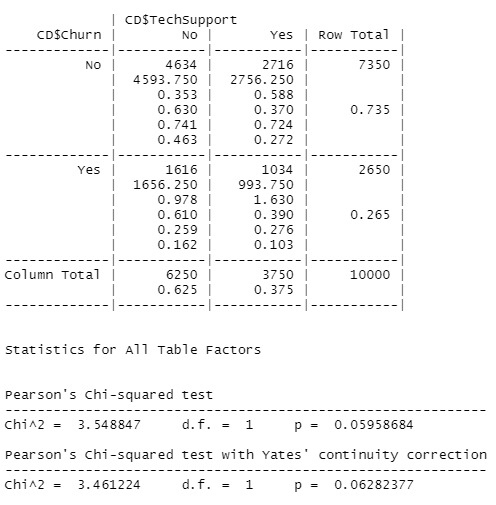
Payment Method



Chart, bar chart

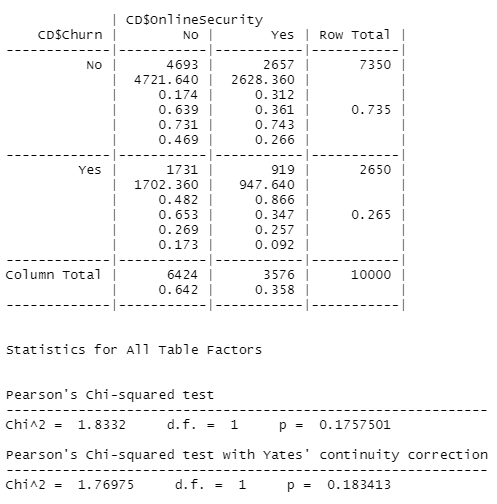
Description automatically generated

Technical Support

 Chart, bar chart

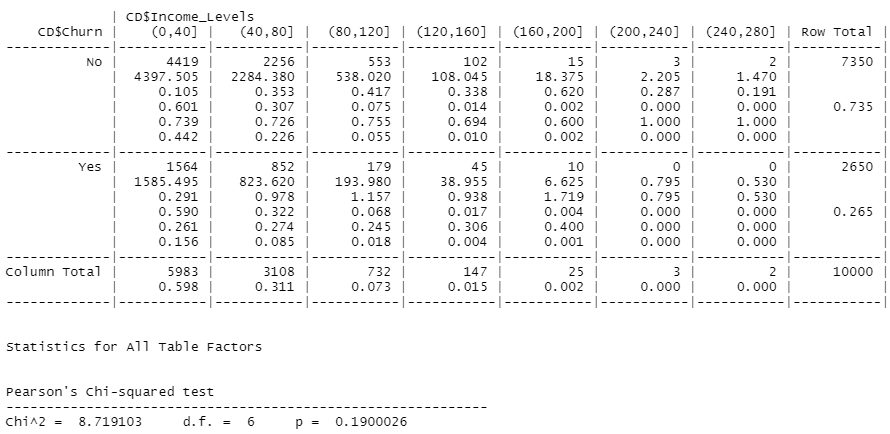
Description automatically generated

Online Security

 Chart, bar chart

Description automatically generated

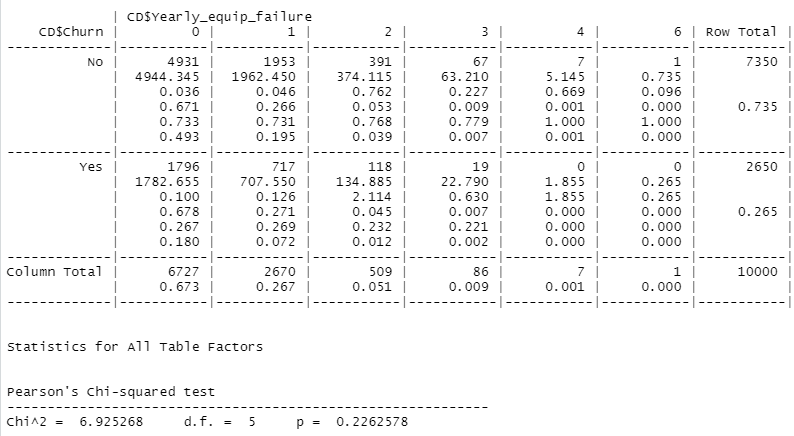
Income\_Levels



Chart, bar chart, histogram

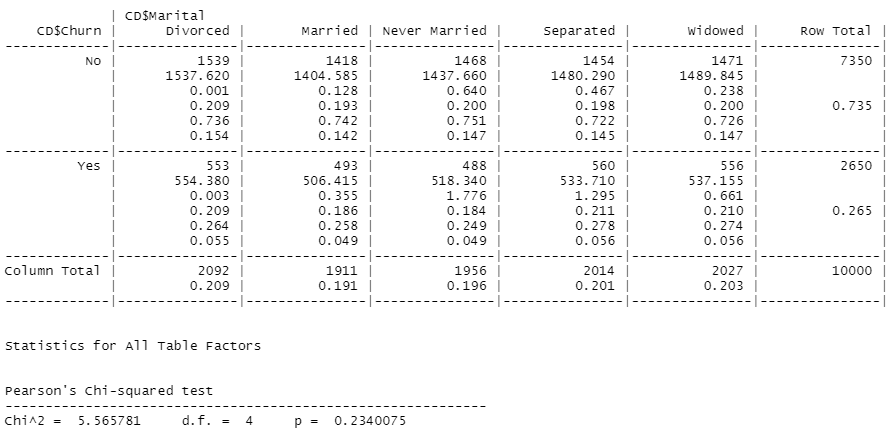
Description automatically generated

Yearly\_Equipment\_Failure

Chart, bar chart

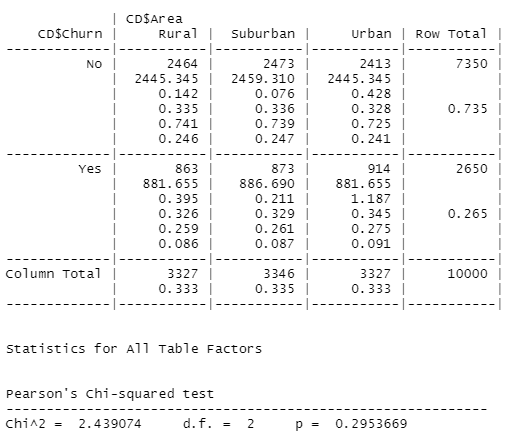
Description automatically generated

Marital Status

Chart, bar chart, histogram

Description automatically generated

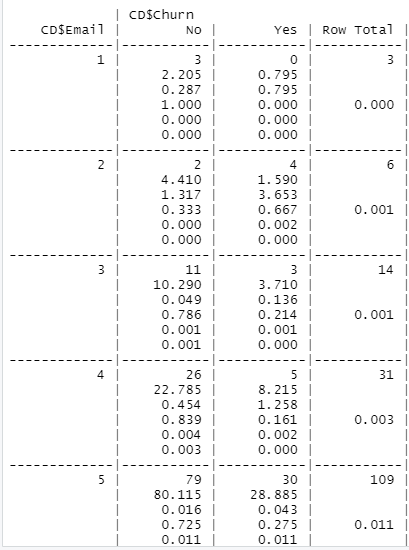
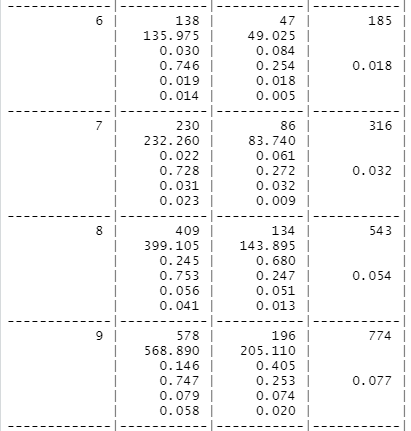
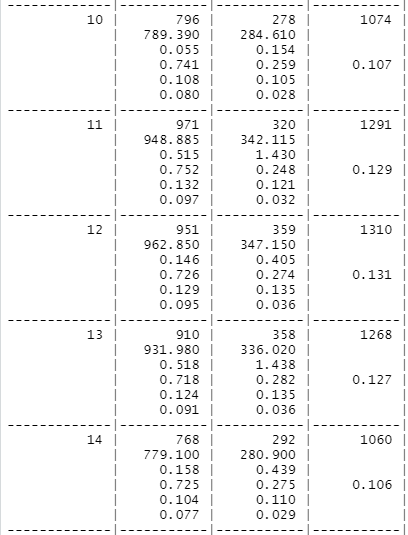
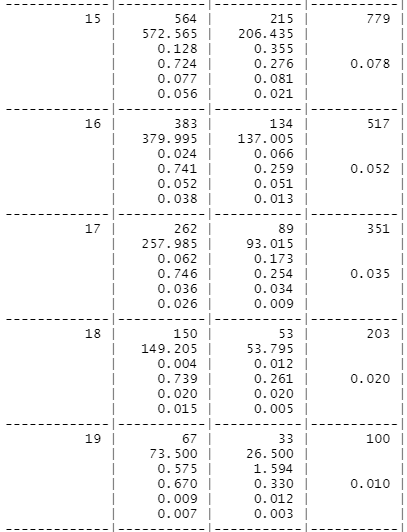
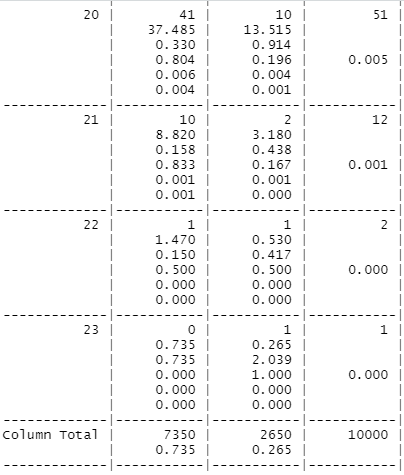
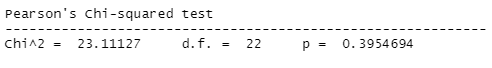
Customer Area

 Chart, bar chart

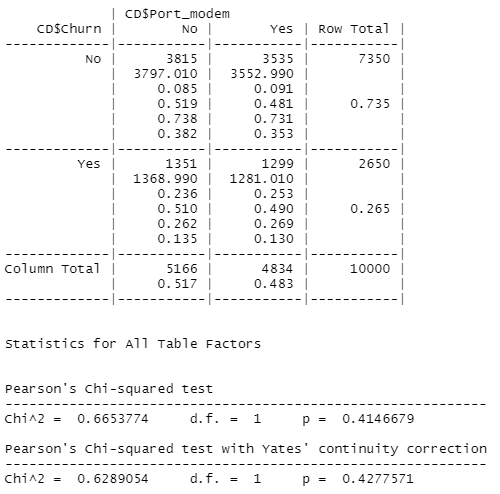
Description automatically generated

Email

Chart, histogram

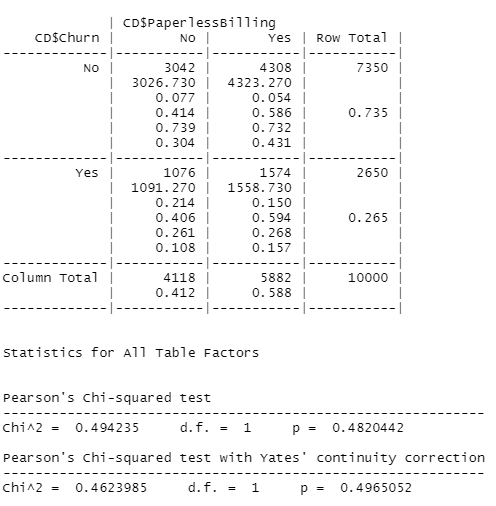
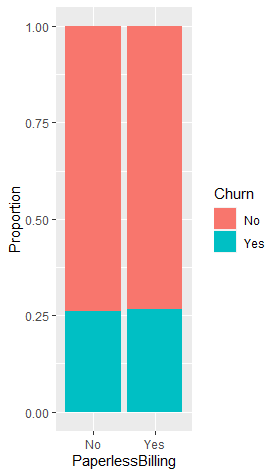
Description automatically generated     

Portable\_Modem

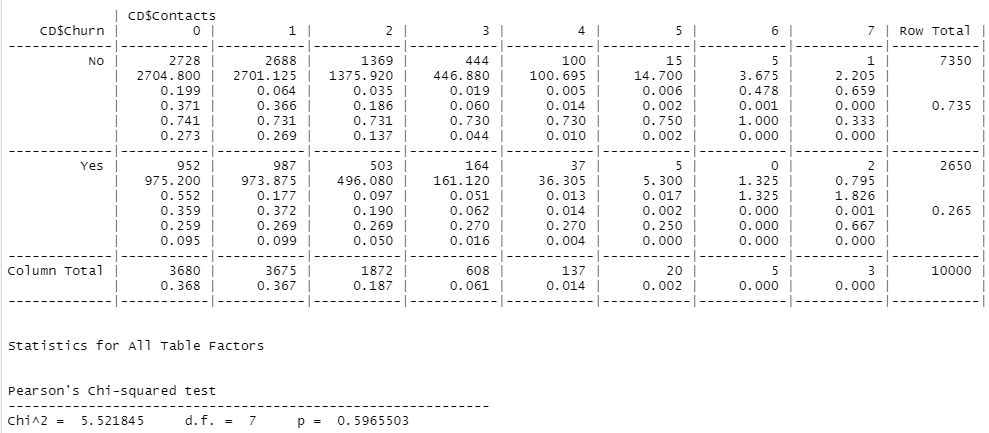
 Chart, bar chart

Description automatically generated

Paperless Billing

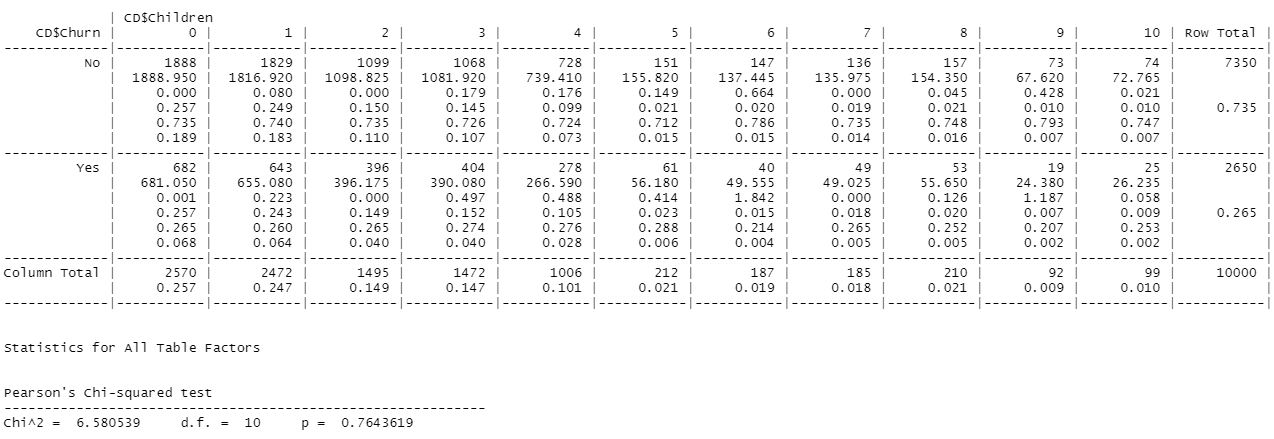
Number of Times Customer Contacted Technical Support



Chart, bar chart, histogram

Description automatically generated

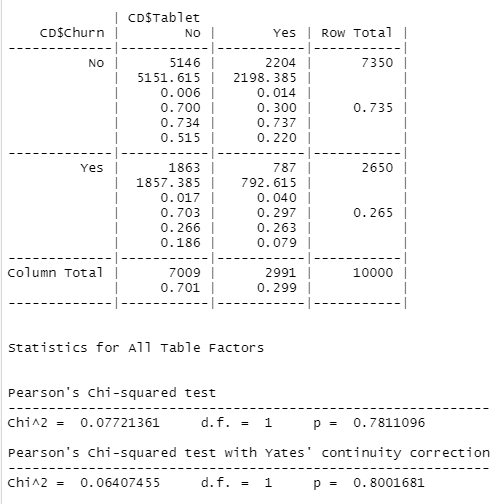
Number of Children in Household



Chart, bar chart, histogram

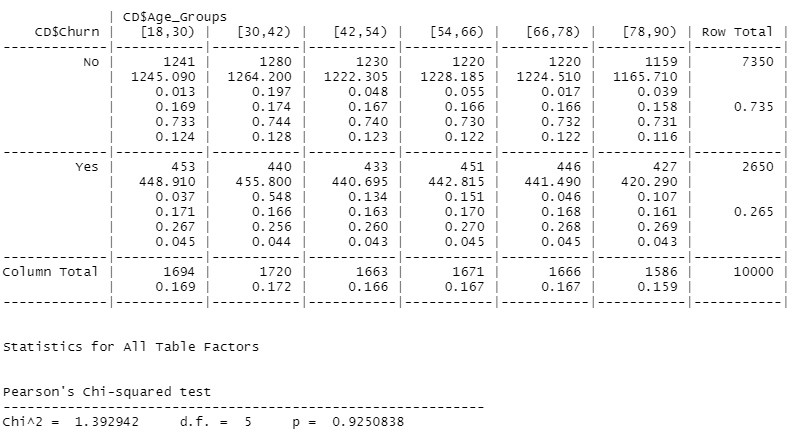
Description automatically generated

Tablet

 Chart, bar chart

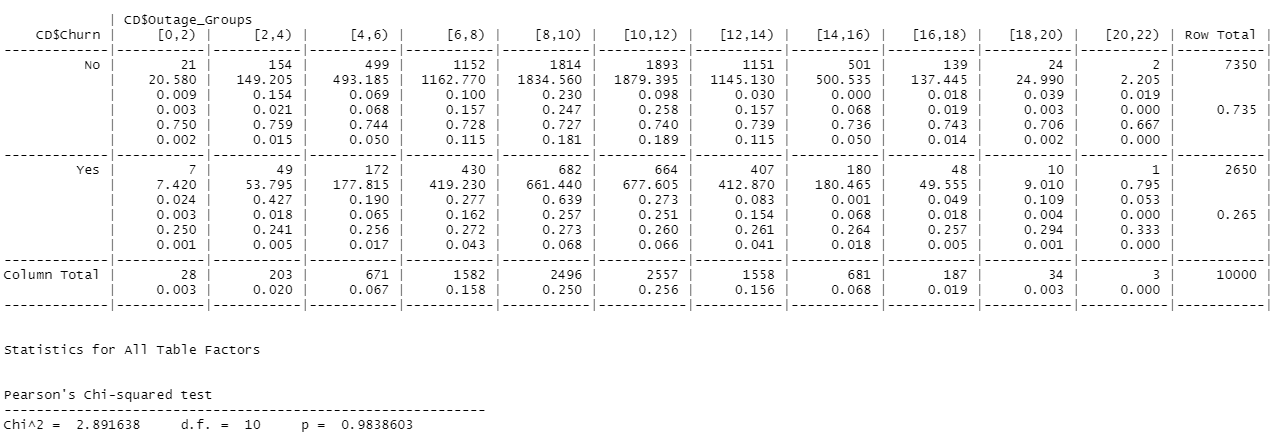
Description automatically generated

Age\_Groups

Chart, bar chart, histogram

Description automatically generated

Outage\_Groups

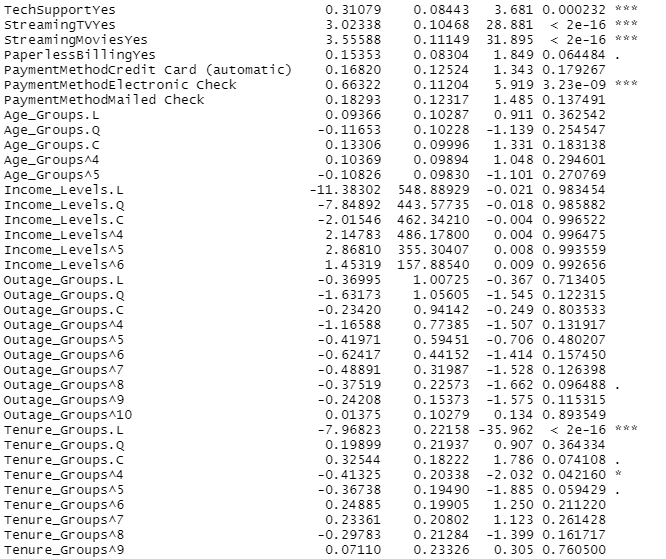
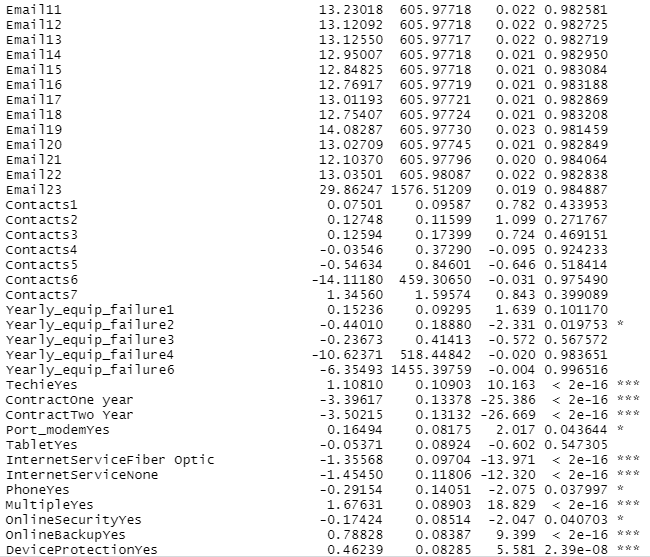
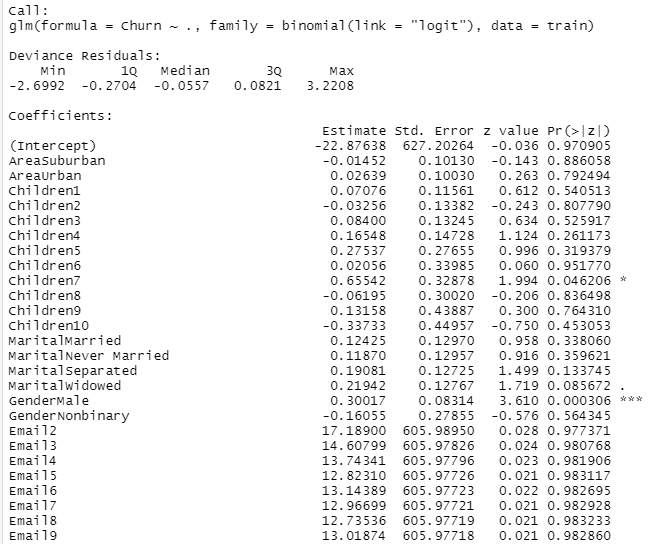


Chart, bar chart, histogram

Description automatically generated

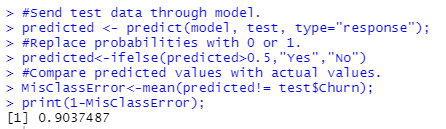
**Gross Logistic Regression:**

I divided the data into a training set (9013 observations) and a testing set (987 observations). I used the training set to develop my model. I then summarized the model created:



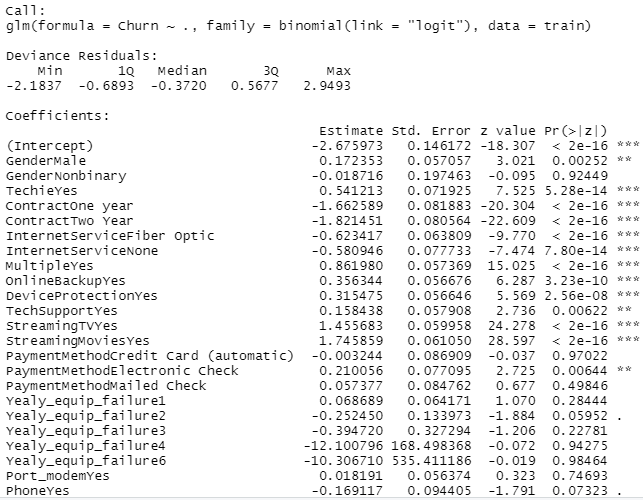
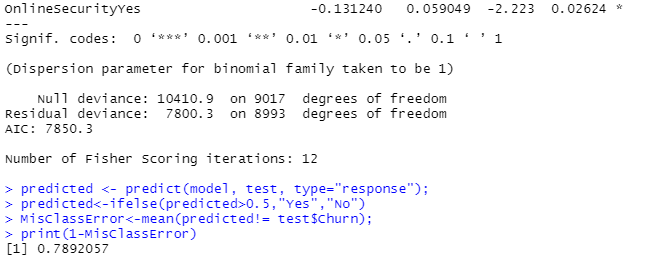
This indicates that customers with a one or two year contract and who have fiber optic or no internet service are less likely to churn. Customers who are male, consider themselves technically inclined, have multiple lines, have online backup, device protection, have tech support, stream tv and movies, and pay by electronic check are more likely to churn.

I then ran the testing data through the model so I could check its accuracy. I compared the predicted results using the model versus the actual data in the testing set. Based on the results, my model has about a 90% accuracy.



**Reduced Logistic Regression:**

I selected variables with sinficance greater than 0.05 (Gender, yearly equipment failure, Techie, Contract, Port\_Modem, Internet Service, Phone, Multiple Lines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaymentMethod) and created a new logistic model. This reduced my variables from 26 to 15. The model still has about 80% accuracy but with fewer variables.

The telecommunications company should be aware of the variables that indicate a customer may leave. They could identify the customers most likely to churn and offer increased incentive for those specific customers. They could also offer incentive that corresponds to the indicator variables.

**Part VI: Demonstration**

G.  Provide a Panopto video recording that includes all of the following elements:

•  a demonstration of the functionality of the code used for the analysis

•  an identification of the version of the programming environment

•  a comparison of the **two** logistic regression models you used in your analysis

•  an interpretation of the coefficients

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.

To submit your recording, upload it to the Panopto drop box titled “Logistic Regression Modeling – NBM2 | D208.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.