D206 PA: Data Cleaning

Adam Eyerman

2022-06-22

## Part I: Research Question

**A)** What variables contribute most to the churn rate within the data set for telecommunication services?

**B)** The data set consists of fifty-two individual columns and 10,000 values per column. However, the first column is an unlabeled redundancy or duplication of the second column.

* The dependent variable will be the “churn” column which define a client’s termination of service from the telecommunications company.
* The independent variables include the remaining columns which will help define the possible correlation these aspects have to the contribution of the churn rate.
* Data Types include numeric, character, and integer. Examples for each type would be Outage\_sec\_perweek (numeric), Churn (character), and Email (integer).

#Load data set into R  
Churn\_Raw\_Data <- read.csv('D206\_Churn\_Raw\_Data.csv')  
  
#View data and structure  
View(Churn\_Raw\_Data)  
str(Churn\_Raw\_Data)

## 'data.frame': 10000 obs. of 52 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ CaseOrder : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Customer\_id : chr "K409198" "S120509" "K191035" "D90850" ...  
## $ Interaction : chr "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0f7d4ac2524" "344d114c-3736-4be5-98f7-c72c281e2d35" "abfa2b40-2d43-4994-b15a-989b8c79e311" ...  
## $ City : chr "Point Baker" "West Branch" "Yamhill" "Del Mar" ...  
## $ State : chr "AK" "MI" "OR" "CA" ...  
## $ County : chr "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...  
## $ Zip : int 99927 48661 97148 92014 77461 31030 37847 73109 34771 45237 ...  
## $ Lat : num 56.3 44.3 45.4 33 29.4 ...  
## $ Lng : num -133.4 -84.2 -123.2 -117.2 -95.8 ...  
## $ Population : int 38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...  
## $ Area : chr "Urban" "Urban" "Urban" "Suburban" ...  
## $ Timezone : chr "America/Sitka" "America/Detroit" "America/Los\_Angeles" "America/Los\_Angeles" ...  
## $ Job : chr "Environmental health practitioner" "Programmer, multimedia" "Chief Financial Officer" "Solicitor" ...  
## $ Children : int NA 1 4 1 0 3 0 2 2 NA ...  
## $ Age : int 68 27 50 48 83 83 NA NA 49 86 ...  
## $ Education : chr "Master's Degree" "Regular High School Diploma" "Regular High School Diploma" "Doctorate Degree" ...  
## $ Employment : chr "Part Time" "Retired" "Student" "Retired" ...  
## $ Income : num 28562 21705 NA 18925 40074 ...  
## $ Marital : chr "Widowed" "Married" "Widowed" "Married" ...  
## $ Gender : chr "Male" "Female" "Female" "Male" ...  
## $ Churn : chr "No" "Yes" "No" "No" ...  
## $ Outage\_sec\_perweek : num 6.97 12.01 10.25 15.21 8.96 ...  
## $ Email : int 10 12 9 15 16 15 10 16 20 18 ...  
## $ Contacts : int 0 0 0 2 2 3 0 0 2 1 ...  
## $ Yearly\_equip\_failure: int 1 1 1 0 1 1 1 0 3 0 ...  
## $ Techie : chr "No" "Yes" "Yes" "Yes" ...  
## $ Contract : chr "One year" "Month-to-month" "Two Year" "Two Year" ...  
## $ Port\_modem : chr "Yes" "No" "Yes" "No" ...  
## $ Tablet : chr "Yes" "Yes" "No" "No" ...  
## $ InternetService : chr "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...  
## $ Phone : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ Multiple : chr "No" "Yes" "Yes" "No" ...  
## $ OnlineSecurity : chr "Yes" "Yes" "No" "Yes" ...  
## $ OnlineBackup : chr "Yes" "No" "No" "No" ...  
## $ DeviceProtection : chr "No" "No" "No" "No" ...  
## $ TechSupport : chr "No" "No" "No" "No" ...  
## $ StreamingTV : chr "No" "Yes" "No" "Yes" ...  
## $ StreamingMovies : chr "Yes" "Yes" "Yes" "No" ...  
## $ PaperlessBilling : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ PaymentMethod : chr "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (automatic)" "Mailed Check" ...  
## $ Tenure : num 6.8 1.16 15.75 17.09 1.67 ...  
## $ MonthlyCharge : num 171 243 159 120 151 ...  
## $ Bandwidth\_GB\_Year : num 905 801 2055 2165 271 ...  
## $ item1 : int 5 3 4 4 4 3 6 2 5 2 ...  
## $ item2 : int 5 4 4 4 4 3 5 2 4 2 ...  
## $ item3 : int 5 3 2 4 4 3 6 2 4 2 ...  
## $ item4 : int 3 3 4 2 3 2 4 5 3 2 ...  
## $ item5 : int 4 4 4 5 4 4 1 2 4 5 ...  
## $ item6 : int 4 3 3 4 4 3 5 3 3 2 ...  
## $ item7 : int 3 4 3 3 4 3 5 4 4 3 ...  
## $ item8 : int 4 4 3 3 5 3 5 5 4 3 ...

## 

## Part II: Data-Cleaning Plan

**C1)** The plan for cleaning this data set includes:

* Determine uniqueness of specific columns to ensure removal of duplicates
* Finding NULL or missing values
* Impute NULL or missing values accordingly
* Removing redundant or irrelevant columns that do not relate to the research question
* Renaming columns that need more specification
* Checking for outliers and their relevancy in specific columns
* Altering negative values to positive where applicable

**C2)** The seven aspects of data quality I will assess are:

* Consistency: is the data contradictory to another source collecting the same information
* Accuracy: data relates to intended use and is not misleading
* Completeness: the data set offers a full picture into the information needed for the research question
* Availability: the data is accessible to perform analysis over time
* Validity: parameters for the measured values are appropriate and follow a set constraint
* Uniqueness: evaluating duplicate entries or redundant data
* Relevancy: the data pertains logically to the research question

The approach to assess the quality of the data will start with consistency. Due to this being a singular data set and no way to access another source I will assume the consistency is of quality. Next is accuracy, which will be assessed by finding NULL or missing values as well as imputation to create a balanced data set for analysis. The data set shows a complete amount of information in relation to answering the research question. Like the consistency aspect, the availability of the data is of quality. By verifying the structure of the data set I will determine if the validity is of quality. Searching for redundant or duplicate values will determine the uniqueness of the data. The data pertains logically to the research question though no period is given, which in normal circumstances would be an issue for further data collection and the quality of relevancy.

**C3)** I chose R for my programming language as it is highly effective at statistical analysis with downloadable packages containing a vast array of tools for data exploration and cleaning. The packages and libraries used account for data visualization, reading text data into R, formatting data, data manipulation, multivariate visualization, PCA, and mode imputation from categorical values. Each library below has an attached purpose.

Packages needed for data cleaning:

* install.packages(“tidyverse”)
* install.packages(“factoextra”)
* install.packages(“FactoMineR”)
* install.packages(“modeest”)

#Load libraries from packages - use case shown  
library(ggplot2) #Data visualizations   
library(readr) #Read text data into R  
library(tidyr) #Creates tidy data  
library(dplyr) #Data manipulation

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(factoextra) #Multivariate visualizations

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(FactoMineR) #Principal Component Analysis  
library(modeest) #Enables mode imputation for categorical values

**C4)** Code used for identifying anomalies:

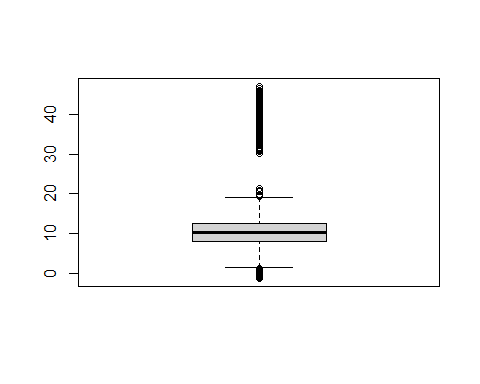
#Determine unique values for Customer\_id  
require(dplyr)  
n\_distinct(Churn\_Raw\_Data$Customer\_id)

## [1] 10000

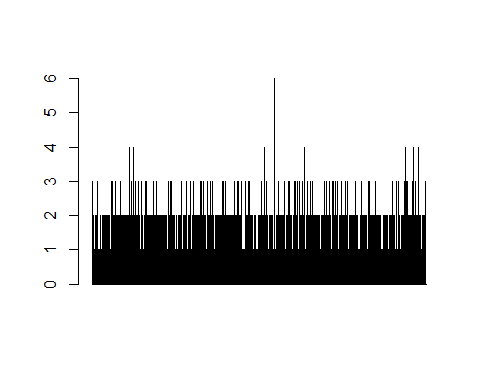
#Find columns with NULL values  
colSums(is.na(Churn\_Raw\_Data))

## X CaseOrder Customer\_id   
## 0 0 0   
## Interaction City State   
## 0 0 0   
## County Zip Lat   
## 0 0 0   
## Lng Population Area   
## 0 0 0   
## Timezone Job Children   
## 0 0 2495   
## Age Education Employment   
## 2475 0 0   
## Income Marital Gender   
## 2490 0 0   
## Churn Outage\_sec\_perweek Email   
## 0 0 0   
## Contacts Yearly\_equip\_failure Techie   
## 0 0 2477   
## Contract Port\_modem Tablet   
## 0 0 0   
## InternetService Phone Multiple   
## 0 1026 0   
## OnlineSecurity OnlineBackup DeviceProtection   
## 0 0 0   
## TechSupport StreamingTV StreamingMovies   
## 991 0 0   
## PaperlessBilling PaymentMethod Tenure   
## 0 0 931   
## MonthlyCharge Bandwidth\_GB\_Year item1   
## 0 1021 0   
## item2 item3 item4   
## 0 0 0   
## item5 item6 item7   
## 0 0 0   
## item8   
## 0

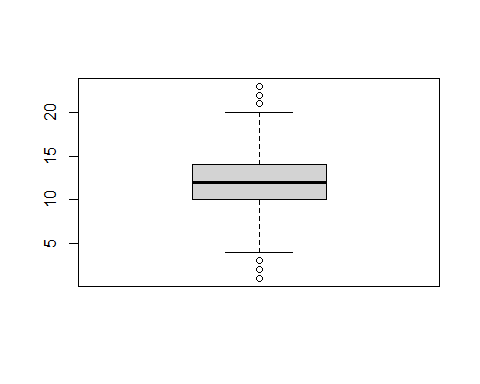
#Visualize for outliers in Outage\_sec\_perweek, Yearly\_equip\_failure, and Email columns as they are viewed as most likely contributors to churn  
boxplot.default(Churn\_Raw\_Data$Outage\_sec\_perweek)



barplot.default(Churn\_Raw\_Data$Yearly\_equip\_failure)



boxplot.default(Churn\_Raw\_Data$Email)



## Part 3: Data Cleaning

**D1)** I determined that each Customer\_id was unique, and all 10,000 rows were valid for further analysis. My next step was to find NULL or missing values, which produced eight columns out of the original 52. I found two columns (Lng and Lat) that due to their redundancy and having columns that already described the location of each customer by City, State, and Zip code I will remove from the data set. By manually scanning the data I also realized the last eight columns (Item1-8) needed to be labeled more specifically with appropriate titles and the Outage\_sec\_perweek column has negative values that needed conversion to positive. Lastly, I checked for outliers with three columns that I believe pertain most to churn rate. The Email column was evenly distributed. The Outage\_sec\_perweek column had a large portion of positive outliers. The Yearly\_equip\_failure column had positive outliers as well.

**D2)** Mitigation Methods:

* Renamed the columns (Item1-8) to their appropriate names using a rename code and creating my clean churn data set.
* Removed the redundant columns (Lng, Lat) and (X) from the data set using a code to retain only the columns I desired to have in my clean churn data set.
* Imputed data for all columns discovered to have NULL or missing values using the code is.na (Zach, 2020) with either their mean, median or mode. I used median for the Children and Age columns as they involved whole numbers. The mean function used for Income, Tenure, and Bandwidth\_GB\_Year as they involved numeric or decimals. The mode function was used for Techie, Phone, and TechSupport as they involved a non-continuous set of numbers.
* Converted negative values to positive in the Outage\_sec\_perweek column using the abs() code.
* No removal of outliers occurred. The Outage\_sec\_perweek column had such a high level of outliers that removal would have disrupted the data. The Email column was evenly distribute with outliers in both negative and positive and had no illogical data points among those outliers. The Yearly\_equip\_failure column had a few positive outliers, but the logical nature of assessing the column suggests these data points are plausible.

**D3)** Looking at the customer\_id column to identify the unique values there were 10,000 distinct values, which suggests no duplicate entries in the data. Then the NULL values were imputed with their appropriate central tendency of either mean, median or mode. I then removed two columns that were redundant and added no value to the data set which was longitude and latitude. The outlier discovery for the selected columns proved to not an issue for the integrity of the data after logically viewing the values and determining any removal would disrupt the data set too much certain values though high were plausible. Altering the names of the final 8 columns to more detailed titles was done to offer a better understanding of the values represent. Finally, I changed the negative values of the Outage\_sec\_perweek column since a negative value is an not logically feasible.

**D4)** Code for Mitigating Anomalies

#Rename columns 'item1' through 'item8' to detailed name  
churn\_data\_clean <- Churn\_Raw\_Data %>% rename(Timely\_response = item1, Timely\_fixes = item2, Timely\_replacements = item3, Reliability = item4, Options = item5, Respectful\_response = item6, Courteous\_exchange = item7, Active\_listening = item8)  
  
#Delete columns that are redundant or irrelevant to analysis  
churn\_data\_clean <- churn\_data\_clean[,c(2:8,11:52)]  
  
#Impute missing or NULL values using mean, median or mode  
churn\_data\_clean$Children[is.na(churn\_data\_clean$Children)] <- median(churn\_data\_clean$Children, na.rm=TRUE)  
churn\_data\_clean$Age[is.na(churn\_data\_clean$Age)] <- median(churn\_data\_clean$Age, na.rm=TRUE)  
churn\_data\_clean$Income[is.na(churn\_data\_clean$Income)] <- mean(churn\_data\_clean$Income, na.rm=TRUE)  
churn\_data\_clean$Tenure[is.na(churn\_data\_clean$Tenure)] <- mean(churn\_data\_clean$Tenure, na.rm=TRUE)  
churn\_data\_clean$Bandwidth\_GB\_Year[is.na(churn\_data\_clean$Bandwidth\_GB\_Year)] <- mean(churn\_data\_clean$Bandwidth\_GB\_Year, na.rm=TRUE)  
churn\_data\_clean$Techie[is.na(churn\_data\_clean$Techie)] <- mfv(churn\_data\_clean$Techie, na.rm=TRUE)

## argument 'na.rm' is soft-deprecated, please start using 'na\_rm' instead

churn\_data\_clean$Phone[is.na(churn\_data\_clean$Phone)] <- mfv(churn\_data\_clean$Phone, na.rm=TRUE)

## argument 'na.rm' is soft-deprecated, please start using 'na\_rm' instead

churn\_data\_clean$TechSupport[is.na(churn\_data\_clean$TechSupport)] <- mfv(churn\_data\_clean$TechSupport, na.rm=TRUE)

## argument 'na.rm' is soft-deprecated, please start using 'na\_rm' instead

#Change negative values in Outage\_sec\_perweek column to positive  
churn\_data\_clean$Outage\_sec\_perweek <- abs(churn\_data\_clean$Outage\_sec\_perweek)

**D5)** Exported clean data sheet to CSV

library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

fwrite(churn\_data\_clean, "C:\\Users\\eyerm\\OneDrive\\Documents\\churn\_data\_clean.csv")

**D6)** The data cleaning process is limited due to the lack of access to the source of the data or individuals creating this data set. Insight into how data was collected and why certain values were illogical could help understand and enhance the cleaning process of the data set. For example, the population column has real values, but logically viewing the values shows an unlikely scenario where locations have either 0,1,2,3, etc. for their total population. There is no way to alter this data to a mean or median or mode because the numbers are technically valid and not outliers that disrupt the data since the distribution seems even from high to low values. Communicating with the source would clarify why these data values exist and if an error or scaling issue occurred upon collection. The major limiting factor is the lack of connection to the data collection process, which hinders the data cleaning process.

**D7)** The limitations create a problem for the research question strictly because to accurately reach a conclusion you would need to be able to continuously update the data and ensure accuracy within each variable collected. To determine the churn rate, you need to identify variables that are potentially contributing to the issue and address those specific variables in an actionable step to alter the rate. The inability to communicate with the data collection team or have access to the process creates a problem with the integrity of the conclusions and the replication moving forward as new data flows in. I suppose my answer leans more into the data collection or acquisition aspect, but since that step is the precursor to data cleaning it is the largest limiting factor.

**E1)** Principal components:

* Outage\_sec\_perweek
* Email
* Contact
* Yearly\_equip\_failure
* Tenure
* MonthlyCharge
* Timely\_response
* Timely\_fixes
* Timely\_replacements
* Reliability
* Options
* Respectful\_response
* Courteous\_exchange
* Active\_listening

**E2)** For my PCA I utilized 14 numeric variables in the data set that pertained to customer survey responses and customer interaction with the service company and devices utilized. The prcomp function was used for this analysis (*Prcomp Function | R Documentation*, n.d.).

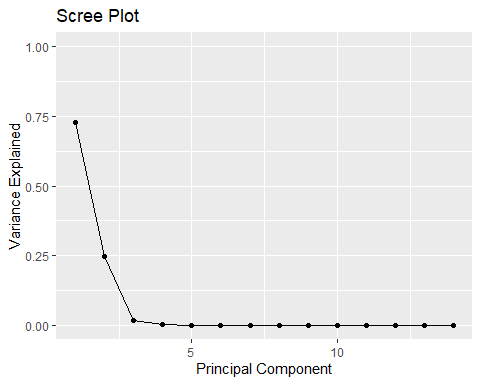
#PCA of selected numeric variables  
churn.pca <- prcomp(churn\_data\_clean[,c(20:23,39:40,42:49)], TRUE, TRUE)  
  
summary(churn.pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 43.3456 25.1778 6.96210 3.02574 1.77167 1.31466 0.98912  
## Proportion of Variance 0.7282 0.2457 0.01879 0.00355 0.00122 0.00067 0.00038  
## Cumulative Proportion 0.7282 0.9738 0.99263 0.99618 0.99739 0.99806 0.99844  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 0.90963 0.85557 0.79373 0.75320 0.7160 0.63677 0.58970  
## Proportion of Variance 0.00032 0.00028 0.00024 0.00022 0.0002 0.00016 0.00013  
## Cumulative Proportion 0.99876 0.99905 0.99929 0.99951 0.9997 0.99987 1.00000

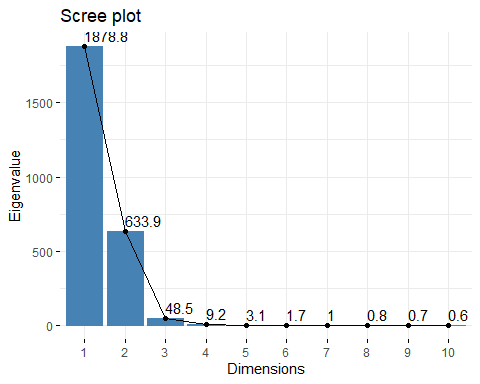
churn.pca$rotation

## PC1 PC2 PC3 PC4  
## Outage\_sec\_perweek 2.173487e-02 -1.820538e-03 -0.9997276137 0.0047839282  
## Email 2.297421e-04 1.200130e-03 -0.0048333864 -0.9999210004  
## Contacts 6.571279e-05 -6.406905e-05 0.0003109294 -0.0011084318  
## Yearly\_equip\_failure -6.597086e-05 -2.117118e-04 -0.0016317942 0.0036273355  
## Tenure -2.965732e-03 -9.999926e-01 0.0017477789 -0.0012128060  
## MonthlyCharge 9.997592e-01 -2.927450e-03 0.0217416597 0.0001216468  
## Timely\_response 2.148287e-04 2.603996e-04 0.0024399976 -0.0025437161  
## Timely\_fixes 2.419071e-05 -2.926107e-04 0.0028391424 -0.0016135711  
## Timely\_replacements -2.823071e-04 -2.493521e-04 0.0044096704 -0.0058124329  
## Reliability 3.786796e-05 2.185519e-04 -0.0014628459 -0.0007547233  
## Options -1.544580e-04 -8.361408e-04 0.0010621146 0.0033410966  
## Respectful\_response 9.217424e-05 2.953864e-04 0.0002124524 -0.0063517007  
## Courteous\_exchange -1.699970e-04 6.747929e-05 0.0019856853 -0.0046863071  
## Active\_listening 7.961573e-05 5.836355e-04 -0.0014516053 0.0015292050  
## PC5 PC6 PC7 PC8  
## Outage\_sec\_perweek 0.0041337077 -3.960060e-03 4.800012e-04 -2.551030e-03  
## Email -0.0086406212 -1.513643e-03 -1.379146e-03 4.517408e-03  
## Contacts -0.0079619748 8.042837e-03 9.969934e-01 4.970439e-02  
## Yearly\_equip\_failure -0.0032015632 -3.434367e-03 -6.272532e-03 -6.144210e-03  
## Tenure 0.0003738821 8.506198e-04 -9.468178e-05 4.239599e-04  
## MonthlyCharge -0.0001169931 -3.233533e-05 -7.390996e-05 -5.019995e-05  
## Timely\_response 0.4634230700 -2.829255e-01 1.020993e-02 -6.968651e-02  
## Timely\_fixes 0.4370842742 -2.845277e-01 1.922262e-02 -1.061969e-01  
## Timely\_replacements 0.3994713694 -2.790453e-01 -2.509601e-02 -1.727266e-01  
## Reliability 0.1422569046 5.690233e-01 -6.036293e-03 -1.737273e-01  
## Options -0.1711759220 -5.897969e-01 1.401892e-02 1.373959e-01  
## Respectful\_response 0.4052667143 1.903589e-01 1.758849e-02 -6.312569e-02  
## Courteous\_exchange 0.3556018183 1.860616e-01 4.440473e-02 -1.792731e-01  
## Active\_listening 0.3069551921 1.361337e-01 -4.842476e-02 9.307597e-01  
## PC9 PC10 PC11 PC12  
## Outage\_sec\_perweek 5.372169e-04 -0.0010049790 1.070845e-03 0.0010804465  
## Email -1.564435e-03 -0.0024017660 -1.357418e-03 -0.0033201121  
## Contacts -4.317264e-02 -0.0047406454 3.083990e-02 0.0227129353  
## Yearly\_equip\_failure -3.023261e-03 0.0177241106 -1.532643e-03 -0.0246643322  
## Tenure -1.245419e-05 0.0002319860 3.117623e-04 -0.0001716311  
## MonthlyCharge 1.123580e-04 -0.0001440333 1.929283e-05 0.0002557881  
## Timely\_response -1.206175e-01 -0.0490129642 -2.702898e-02 -0.2357662547  
## Timely\_fixes -1.723476e-01 -0.0708346509 -7.939263e-02 -0.5867055791  
## Timely\_replacements -2.469587e-01 -0.1568208959 3.920068e-01 0.6781837950  
## Reliability -4.818874e-01 -0.4258219662 -4.499318e-01 0.0966890597  
## Options 6.479221e-02 -0.1721846694 -7.028084e-01 0.2722569175  
## Respectful\_response 6.314981e-02 0.7774332847 -3.655745e-01 0.2214259661  
## Courteous\_exchange 8.076760e-01 -0.3732825558 -8.238272e-02 0.0672828259  
## Active\_listening -1.189024e-02 -0.1137085368 3.912186e-02 0.0457388695  
## PC13 PC14  
## Outage\_sec\_perweek -1.691253e-03 -0.0003353135  
## Email 3.568355e-03 -0.0002507833  
## Contacts 7.041333e-03 -0.0011243856  
## Yearly\_equip\_failure 9.860615e-01 -0.1631983559  
## Tenure -1.360533e-04 0.0004428541  
## MonthlyCharge 7.466385e-05 -0.0002127596  
## Timely\_response 1.241618e-01 0.7820076189  
## Timely\_fixes -1.077046e-01 -0.5652390354  
## Timely\_replacements -1.118990e-02 -0.1809128848  
## Reliability 1.200668e-02 0.0168438997  
## Options 2.321769e-04 -0.0429206821  
## Respectful\_response -1.732233e-02 -0.0618308433  
## Courteous\_exchange 4.756723e-03 -0.0421296136  
## Active\_listening 2.990222e-03 -0.0433612867

#Calculate total variance with each principal component  
var\_by\_component <- churn.pca$sdev^2 / sum(churn.pca$sdev^2)  
  
#Display Scree Plot for PCA visualization  
require(ggplot2)  
qplot(c(1:14), var\_by\_component) +   
 geom\_line() +   
 xlab("Principal Component") +   
 ylab("Variance Explained") +  
 ggtitle("Scree Plot") +  
 ylim(0, 1)



#Display Eigenvalue visualization  
require(factoextra)  
fviz\_eig(churn.pca, choice = "eigenvalue", addlabels=TRUE, scale. = FALSE)



#Display total variance with each principal component  
churn.pca$sdev^2 / sum(churn.pca$sdev^2)

## [1] 0.7281598261 0.2456823008 0.0187852671 0.0035481300 0.0012164775  
## [6] 0.0006698325 0.0003791690 0.0003206793 0.0002836937 0.0002441670  
## [11] 0.0002198636 0.0001986790 0.0001571445 0.0001347700

**E3)** The PCA displayed that PC1 and PC2 accounted for over 97% with PC3 bringing the total cumulative variance to over 99%. The organization can utilize PC1 and PC2 to further explore the data set to determine if any applicable variance can predict or inform on the churn rate. Monthly charge and tenure may be the most related to the churn rate according to the PCA with Outage\_sec\_perweek being PC3 and accounting for around 2% of variation as a slight addition to that assumption, but all aspects would need further analysis.

**F)** [Wed Jun 29 2022 11:35:25 AM (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5643c820-2d4f-4ae6-bc86-aec30100ebce)

**G) Web Sources**

* Zach. (2020, October 12). *How to Impute Missing Values in R (With Examples)*. Statology. <https://www.statology.org/impute-missing-values-in-r/>
* *prcomp function | R Documentation*. (n.d.). Www.rdocumentation.org. <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp>