D212 Task 2 v1

# Part I: Research Question

Which customer characteristics are most prevalent as pertaining to customers who churn versus customers who do not churn?

The goal in answering this question is to decrease costs associated with portfolio maintenance. As the cost to on-board a new customer is already known to be 10 times that of retaining an existing customer, one way to cut costs is by raising retention rates among the existing customer base. Therefore, it would be a valuable undertaking for an organization to have the ability to identify shared characteristics in their customer data set, especially as pertaining to customers likely to churn. Armed with this knowledge, new strategic initiatives could be developed as needed to target the retention of highly valued existing customers.

The research question will be addressed using the principal component analysis (PCA)

# Part II: Method Justification

Principal component analysis is chosen as it functions to reduce dimensionality in data. In this case to answer the research question an analyst could try to plot all of the many variables from the customer data set to try to find characteristics which are shared among those customers likely to churn. However, because of the high number of variables in the data set, such a visualization would quickly become overcrowded and highly obscure. Reducing the dimensions that the analyst is working with may require plotting only two or three at a time, but this approach limits the ability of the analyst to effectively deal with all of the variation in the data that exists in the other variables. Instead, PCA is a method of dimensionality reduction which preserves the “big picture” of variation existing in multiple variables. PCA works to fit multiple highly correlated variables to a single linear component. To account for the full variation in the data, multiple fits are created resulted in multiple principal components but still fewer than the original variable numbers. The analyst can then choose to work with only a small number of principal components which “tell the story” of the original data as opposed to the full array of variables in the original data set.

The assumption of PCA is that the variables used in the analysis will be numeric. In the case of the churn research question, the goal of performing PCA on the data set will be to reduce the original numeric variables chosen for analysis down to the principal components in order to decipher which variables have the strongest loading on the churn and not churn scenario. In other words, which variables account for the greatest variation and correlation among customers who churn and among those customers who did not churn.

R will be used for this analysis. R is open source software that was specifically made for statistical analysis. Using R, we can ingest the data set, and leveraging an extensive library of data manipulation and visualization packages, perform the necessary PCA steps. More information can be found on the R project website (<https://www.r-project.org/>).

The dplyr package will be used for data preparation and manipulation within R. More information for the dplyr package can be found on the tidyverse website (<https://dplyr.tidyverse.org/>).

PCA function construction will be done using the prcomp function in base R. The prcomp function accepts the data frame and scaling parameters as input.

PCA visualization will be done using the factoextra package which is a visualization package in R. More information on the factoextra package can be found on the CRAN R Project website (<https://cran.r-project.org/web/packages/factoextra/index.html>).

# Part III: Data Preparation

To prepare the data, first a check is run for missing values in the data set using the sapply function. No action is needed as no missing values were detected in the provided data set. However, had missing values been detected more analysis would have been required for each instance to determine handling of the records containing nulls. Depending on the variable in view, the nulls could be filled with variable mean or median values, or even removed from the data set.

Second, scaling is considered during the data preparation phase. However, since the prcomp function in R has a scaling paramter, no scaling is needed during the data preparation phase.

Next, prior to creating the prcomp function, the data will be prepared by selecting only the numeric variables to be used in the analysis. Categorical variables are not appropriate for a principal component analysis. In variable selection it is necessary to investigate the data set to distinguish between numeric and categorical variables. This is done using R’s str function. From the output, along with leveraging the data dictionary provided, the relevant numeric variables can then be selected. The churn indicator is also selected.

Next, any variable names which might be obscure will need to be renamed for clarity. This data set contains a series of survey questions which will be renamed using the data dictionary provided.

Finally, Once the variables are selected, the data frame is split into churn and not churn sets. The data sets are labeled accordingly and the categorical churn indicator is dropped. The data is now ready for PCA.

library(dplyr)

## Warning: replacing previous import 'vctrs::data\_frame' by 'tibble::data\_frame'  
## when loading 'dplyr'

##   
## 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

# Load Data Set  
df<-read.csv("c:/users/shua/documents/Data Mining II\_D212/churn\_clean.csv")  
  
#Scaling not needed during data prep, will be entered as a plotting parameter  
  
# Check for missing values  
sapply(df, function(x) sum(is.na(x)))

## CaseOrder Customer\_id Interaction   
## 0 0 0   
## UID City State   
## 0 0 0   
## County Zip Lat   
## 0 0 0   
## Lng Population Area   
## 0 0 0   
## TimeZone Job Children   
## 0 0 0   
## Age Income Marital   
## 0 0 0   
## Gender Churn Outage\_sec\_perweek   
## 0 0 0   
## Email Contacts Yearly\_equip\_failure   
## 0 0 0   
## Techie Contract Port\_modem   
## 0 0 0   
## Tablet InternetService Phone   
## 0 0 0   
## Multiple OnlineSecurity OnlineBackup   
## 0 0 0   
## DeviceProtection TechSupport StreamingTV   
## 0 0 0   
## StreamingMovies PaperlessBilling PaymentMethod   
## 0 0 0   
## Tenure MonthlyCharge Bandwidth\_GB\_Year   
## 0 0 0   
## Item1 Item2 Item3   
## 0 0 0   
## Item4 Item5 Item6   
## 0 0 0   
## Item7 Item8   
## 0 0

# Identify numeric columns for use  
str(df)

## 'data.frame': 10000 obs. of 50 variables:  
## $ 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" ...  
## $ UID : chr "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a2959829830fb8a" "f1784cfa9f6d92ae816197eb175d3c71" "dc8a365077241bb5cd5ccd305136b05e" ...  
## $ 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 0 1 4 1 0 3 0 2 2 1 ...  
## $ Age : int 68 27 50 48 83 83 79 30 49 86 ...  
## $ Income : num 28562 21705 9610 18925 40074 ...  
## $ Marital : chr "Widowed" "Married" "Widowed" "Married" ...  
## $ Gender : chr "Male" "Female" "Female" "Male" ...  
## $ Churn : chr "No" "Yes" "No" "No" ...  
## $ Outage\_sec\_perweek : num 7.98 11.7 10.75 14.91 8.15 ...  
## $ 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 172 243 160 120 150 ...  
## $ 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 ...

#Select columns  
df<-df%>%select(Children, Age, Income, Outage\_sec\_perweek, Tenure, MonthlyCharge, Bandwidth\_GB\_Year, Item1, Item2, Item3, Item4, Item5, Item6, Item7, Item8, Churn)  
  
#Transform Obscure Variable Names  
colnames(df)

## [1] "Children" "Age" "Income"   
## [4] "Outage\_sec\_perweek" "Tenure" "MonthlyCharge"   
## [7] "Bandwidth\_GB\_Year" "Item1" "Item2"   
## [10] "Item3" "Item4" "Item5"   
## [13] "Item6" "Item7" "Item8"   
## [16] "Churn"

colnames(df)[8:15]<-c("TimelyResponse", "TimelyFix", "TimelyReplacement","Reliability","Options","RespectfulResponse","CourteousExchange","ActiveListening")  
  
#Split data set  
unique(df$Churn)

## [1] "No" "Yes"

colnames(df)

## [1] "Children" "Age" "Income"   
## [4] "Outage\_sec\_perweek" "Tenure" "MonthlyCharge"   
## [7] "Bandwidth\_GB\_Year" "TimelyResponse" "TimelyFix"   
## [10] "TimelyReplacement" "Reliability" "Options"   
## [13] "RespectfulResponse" "CourteousExchange" "ActiveListening"   
## [16] "Churn"

df\_Churn<-df[df$Churn=="Yes",-16]  
nrow(df\_Churn)

## [1] 2650

df\_Not\_Churn<-df[df$Churn!="Yes",-16]  
nrow(df\_Not\_Churn)

## [1] 7350

colnames(df\_Not\_Churn)

## [1] "Children" "Age" "Income"   
## [4] "Outage\_sec\_perweek" "Tenure" "MonthlyCharge"   
## [7] "Bandwidth\_GB\_Year" "TimelyResponse" "TimelyFix"   
## [10] "TimelyReplacement" "Reliability" "Options"   
## [13] "RespectfulResponse" "CourteousExchange" "ActiveListening"

colnames(df\_Churn)

## [1] "Children" "Age" "Income"   
## [4] "Outage\_sec\_perweek" "Tenure" "MonthlyCharge"   
## [7] "Bandwidth\_GB\_Year" "TimelyResponse" "TimelyFix"   
## [10] "TimelyReplacement" "Reliability" "Options"   
## [13] "RespectfulResponse" "CourteousExchange" "ActiveListening"

write.csv(df, "c:/users/shua/documents/Data Mining II\_D212/processed\_dataset\_pca.csv")

# Part IV: Analysis

To begin the analysis phase the prcomp function is used on the not churned data frame. The data frame is passed to the function along with the scaling parameter and center parameter set to TRUE. A matrix of each principal component and its standard deviation, proportion of variance, and cumulative proportion can be inspected using the summary function. Of the 15 total principal components, we see that the first 8 account for over 75% of the cumulative proportion of variance. Building a scree plot based on the the proportion of variance for each of the principal components, we can clearly see an elbow after the 4th principal component where the proportion of variance for each principal component after the 4th is minimally increased. Again using the summary function, we can confirm that the first 4 principal components, after which we observed only minimal increase of accounting for variance, account for 51.98% of the total variance in the dataset.

When plotting the PCA results on a biplot the visualization is overcrowded due to the presence of individual observations. Instead, the factoextra package is used to produce a plot only of the variable loading in the first four principal components. The Timely Response and Timely Fix variables showed the strongest negative loadings for principal component 1 while the Options variable showed the strongest positive loading for principal component 1. For principal component 2 Tenure and Bandwidth usage demonstrated the strongest loadings. For principal component 3, Options and Reliability showed the strongest loadings. For principal component 4 Children (positive) and Age (negative) showed the strongest loadings. These results can be accessed tabularly by calling the rotation variable from the PCA object. Positive loadings for a variable indicate a positive correlation between the variable and the principal component. Negative loadings indicate a negative correlation between the variable and the principal component. The variables with the higher loading values indicate a stronger correlation. Therefore, the the PCA for the not churned data set is indicating the following variables demonstrate a large share of responsibility for variance in the data set and therefore should be examined for primary characteristics of a not churned customer:

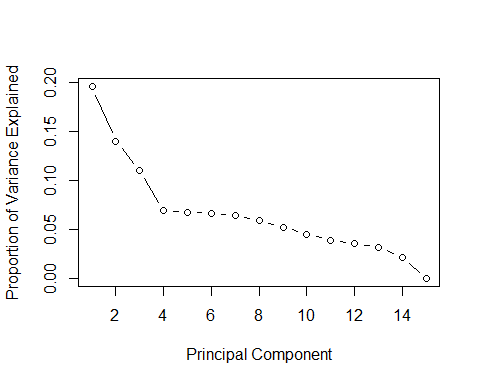
1. Timely Response
2. Timely Fix
3. Options
4. Reliability
5. Tenure
6. Bandwidth
7. Children
8. Age

Now the process is repeated with the churn dataset. In this case again based on the elbow in the scree plot the number of principal components for use is 4. Interestingly nearly all values are very similar pertaining to the top loading variables in the top 4 principal components between the churn and not churned data sets with one exception. Outage seconds per week stands out with a loading value of .60 in the 4th principal component of the churn data set. This suggests that the duration and frequency of service interruptions is carries more influnce as a characteristic of a churned customer than it does as a not churned customer.

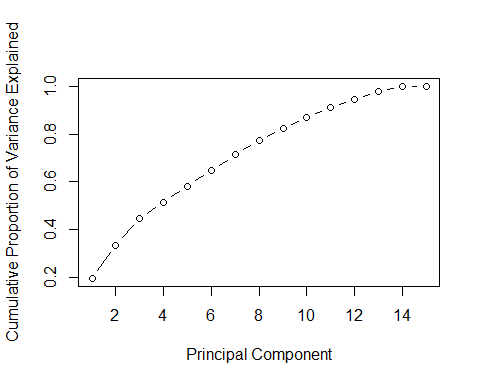
# Not Churned PCA  
pr.telcom.notchurned<-prcomp(x=df\_Not\_Churn, scale=TRUE, center=TRUE)  
summary(pr.telcom.notchurned)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.7135 1.4473 1.2860 1.01747 1.00558 1.00010 0.98496  
## Proportion of Variance 0.1957 0.1396 0.1103 0.06902 0.06741 0.06668 0.06468  
## Cumulative Proportion 0.1957 0.3354 0.4457 0.51467 0.58208 0.64877 0.71344  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 0.94382 0.88678 0.82587 0.76697 0.73211 0.69460 0.5718  
## Proportion of Variance 0.05939 0.05243 0.04547 0.03922 0.03573 0.03216 0.0218  
## Cumulative Proportion 0.77283 0.82525 0.87072 0.90994 0.94567 0.97784 0.9996  
## PC15  
## Standard deviation 0.07400  
## Proportion of Variance 0.00037  
## Cumulative Proportion 1.00000

pr.var<-pr.telcom.notchurned$sdev^2  
pve<-pr.var / sum(pr.var)  
  
plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained", type="b")



plot(cumsum(pve), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained", type="b")



summary(pr.telcom.notchurned)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.7135 1.4473 1.2860 1.01747 1.00558 1.00010 0.98496  
## Proportion of Variance 0.1957 0.1396 0.1103 0.06902 0.06741 0.06668 0.06468  
## Cumulative Proportion 0.1957 0.3354 0.4457 0.51467 0.58208 0.64877 0.71344  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 0.94382 0.88678 0.82587 0.76697 0.73211 0.69460 0.5718  
## Proportion of Variance 0.05939 0.05243 0.04547 0.03922 0.03573 0.03216 0.0218  
## Cumulative Proportion 0.77283 0.82525 0.87072 0.90994 0.94567 0.97784 0.9996  
## PC15  
## Standard deviation 0.07400  
## Proportion of Variance 0.00037  
## Cumulative Proportion 1.00000

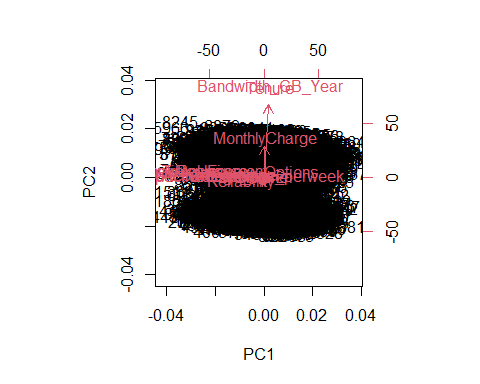
biplot(pr.telcom.notchurned)  
  
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.0.5

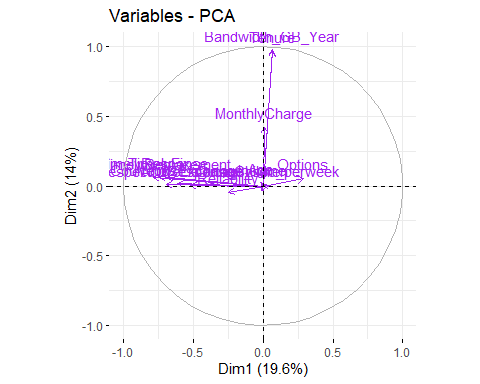
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.5

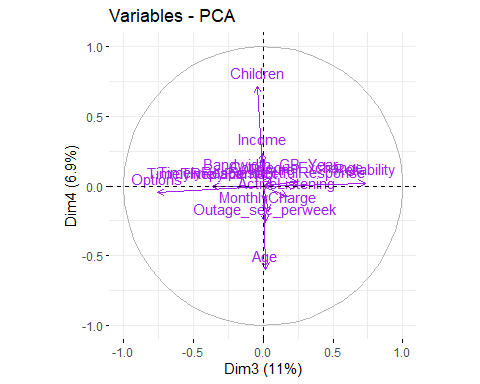
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa



fviz\_pca\_var(pr.telcom.notchurned, col.var="purple", fill.var="white", axes=c(1,2))



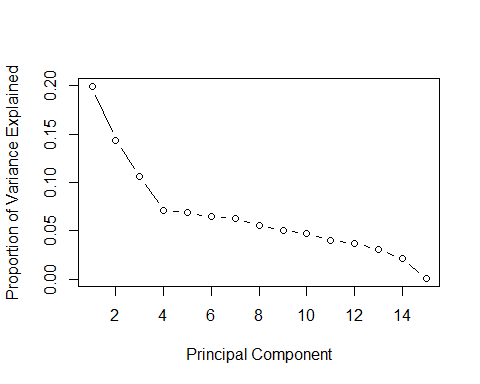
fviz\_pca\_var(pr.telcom.notchurned, col.var="purple", fill.var="white", axes=c(3,4))



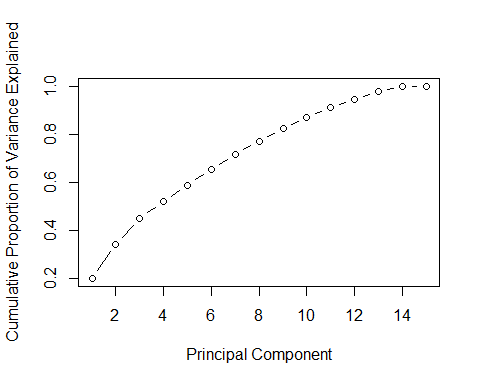
pr.telcom.notchurned$rotation

## PC1 PC2 PC3 PC4  
## Children -0.006119598 0.007811615 -0.033995713 0.699767625  
## Age -0.008717707 0.014101131 0.015248828 -0.588139113  
## Income -0.002749007 0.008197636 -0.006747587 0.231963350  
## Outage\_sec\_perweek 0.019925463 0.004152763 0.011979042 -0.256418434  
## Tenure 0.038270544 0.669998801 0.047378861 0.025486417  
## MonthlyCharge 0.006817632 0.291498419 0.028757879 -0.178176860  
## Bandwidth\_GB\_Year 0.038395316 0.676392358 0.047801675 0.058527541  
## TimelyResponse -0.459222735 0.044336506 -0.277772660 -0.001522628  
## TimelyFix -0.432756562 0.044988109 -0.281598094 -0.005415748  
## TimelyReplacement -0.403347281 0.041484118 -0.281840514 -0.005523545  
## Reliability -0.143701080 -0.029776004 0.568078414 0.019168213  
## Options 0.170013503 0.038611983 -0.587003182 -0.045252417  
## RespectfulResponse -0.405409361 0.005255097 0.190411688 0.004422074  
## CourteousExchange -0.359079319 0.012041158 0.188771063 0.034561705  
## ActiveListening -0.304622005 0.004319423 0.134163364 -0.073121783  
## PC5 PC6 PC7 PC8  
## Children 0.0174325845 8.889226e-02 0.691620330 -0.130748268  
## Age 0.3199120389 -3.663305e-01 0.642562884 0.017692885  
## Income -0.3568397316 -8.903682e-01 -0.119996105 -0.053369605  
## Outage\_sec\_perweek -0.8535697997 2.232839e-01 0.302651742 0.226218632  
## Tenure 0.0502173312 -1.787262e-02 -0.008201169 0.224673139  
## MonthlyCharge -0.1742874213 7.702962e-02 -0.007899857 -0.915723701  
## Bandwidth\_GB\_Year 0.0295057417 7.758689e-05 -0.008259476 0.171142445  
## TimelyResponse -0.0003661952 3.004363e-02 -0.012661987 -0.030742627  
## TimelyFix -0.0233111099 1.672103e-02 0.010073728 -0.015498796  
## TimelyReplacement 0.0316323623 3.154495e-02 -0.019572313 0.010305228  
## Reliability 0.0316929408 4.246961e-02 -0.015575620 -0.011045927  
## Options 0.0007575741 -2.332071e-02 -0.019926455 0.037763653  
## RespectfulResponse -0.0093489634 -1.879311e-02 0.019578729 0.008970887  
## CourteousExchange 0.0038480608 -6.299384e-02 -0.025090908 0.030264339  
## ActiveListening -0.0708275770 -1.377154e-02 0.016787097 0.080093591  
## PC9 PC10 PC11 PC12  
## Children 0.055436372 0.006109413 -0.027158841 0.0235171648  
## Age -0.048137323 -0.010684417 -0.020187528 -0.0134179046  
## Income -0.019767413 -0.087693334 -0.014364293 -0.0123425428  
## Outage\_sec\_perweek -0.105920455 0.007643153 -0.019422144 -0.0002046076  
## Tenure -0.011366672 -0.011955510 0.009256821 -0.0088602541  
## MonthlyCharge 0.046306565 0.040237168 -0.014330334 0.0201429828  
## Bandwidth\_GB\_Year -0.002235094 -0.010069414 0.007964243 -0.0067094728  
## TimelyResponse -0.070416122 -0.121299642 -0.038564541 -0.0154883467  
## TimelyFix -0.103174244 -0.178980444 -0.066697149 -0.0145056646  
## TimelyReplacement -0.174784332 -0.231686326 -0.160775354 -0.2874513979  
## Reliability -0.185409065 -0.461736218 -0.459669228 0.4384996142  
## Options 0.115273579 0.082160195 -0.212610645 0.7290235363  
## RespectfulResponse -0.064354971 0.048767896 0.754925382 0.4334269099  
## CourteousExchange -0.176740555 0.816738847 -0.353068848 0.0582239162  
## ActiveListening 0.922781120 -0.020473935 -0.122732860 -0.0223721599  
## PC13 PC14 PC15  
## Children 0.022704570 -0.0004579323 2.268675e-02  
## Age -0.006095510 0.0143055457 -2.413404e-02  
## Income 0.011551697 0.0126844246 1.138052e-03  
## Outage\_sec\_perweek 0.023192097 0.0170015752 1.349617e-04  
## Tenure -0.009752506 0.0077094952 7.019724e-01  
## MonthlyCharge 0.037374276 -0.0205011663 3.598125e-02  
## Bandwidth\_GB\_Year -0.005810523 0.0021869434 -7.105116e-01  
## TimelyResponse -0.223358952 0.7979252107 -2.754450e-03  
## TimelyFix -0.607236788 -0.5604742882 7.253100e-04  
## TimelyReplacement 0.723059818 -0.1952423246 5.372571e-04  
## Reliability 0.024672939 0.0211374211 2.245719e-04  
## Options 0.149788593 -0.0382478224 -5.514418e-05  
## RespectfulResponse 0.170325104 -0.0716473613 7.956627e-05  
## CourteousExchange 0.027668055 -0.0380329613 -5.294746e-04  
## ActiveListening 0.054658446 -0.0378880547 2.467051e-03

# Churned PCA  
pr.telcom.churned<-prcomp(x=df\_Churn, scale=TRUE, center=TRUE)  
  
pr.var<-pr.telcom.churned$sdev^2  
pve<-pr.var / sum(pr.var)  
  
  
plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained", type="b")



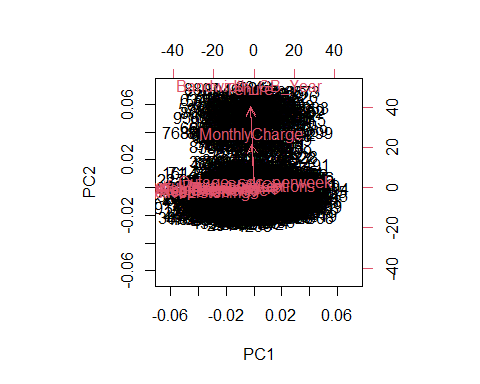
plot(cumsum(pve), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained", type="b")



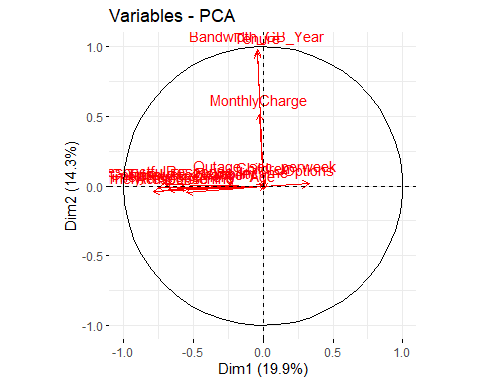
summary(pr.telcom.churned)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.7289 1.4651 1.2637 1.03143 1.01575 0.98578 0.96953  
## Proportion of Variance 0.1993 0.1431 0.1065 0.07092 0.06878 0.06478 0.06267  
## Cumulative Proportion 0.1993 0.3424 0.4489 0.51977 0.58855 0.65334 0.71600  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 0.91498 0.87190 0.84429 0.77530 0.74249 0.68257 0.56354  
## Proportion of Variance 0.05581 0.05068 0.04752 0.04007 0.03675 0.03106 0.02117  
## Cumulative Proportion 0.77181 0.82249 0.87002 0.91009 0.94684 0.97790 0.99907  
## PC15  
## Standard deviation 0.11795  
## Proportion of Variance 0.00093  
## Cumulative Proportion 1.00000

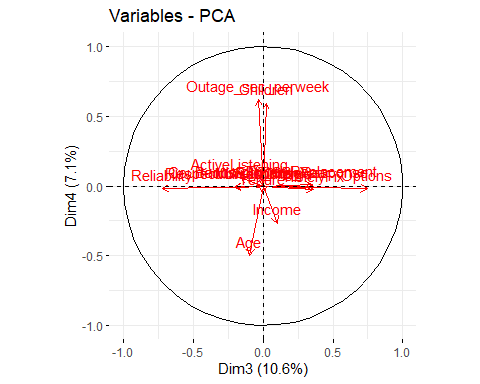
biplot(pr.telcom.churned)



library(factoextra)  
fviz\_pca\_var(pr.telcom.churned, col.var="red", fill.var="white", axes=c(1,2), col.circle="black")



fviz\_pca\_var(pr.telcom.churned, col.var="red", fill.var="white", axes=c(3,4), col.circle="black")



pr.telcom.churned$rotation

## PC1 PC2 PC3 PC4  
## Children 1.417477e-03 0.019071014 0.016841900 0.5751169067  
## Age -8.261794e-05 -0.017591265 -0.076635230 -0.4834710386  
## Income 4.068090e-03 0.001725326 0.082064237 -0.2589973357  
## Outage\_sec\_perweek 1.003214e-02 0.030882524 -0.025645452 0.6010723467  
## Tenure -2.474600e-02 0.654038771 -0.004467659 -0.0501218007  
## MonthlyCharge -1.517680e-02 0.352304004 0.004485760 -0.0052135163  
## Bandwidth\_GB\_Year -2.349287e-02 0.666441362 0.003926485 -0.0017562741  
## TimelyResponse -4.550955e-01 -0.027546107 0.287475024 -0.0009494989  
## TimelyFix -4.352532e-01 -0.008619007 0.285279669 -0.0244038958  
## TimelyReplacement -3.910758e-01 -0.017385350 0.277043430 0.0097002670  
## Reliability -1.514712e-01 -0.003634756 -0.571406884 -0.0180805392  
## Options 1.908357e-01 0.010214898 0.591698707 -0.0174955585  
## RespectfulResponse -4.036343e-01 0.005934399 -0.163956517 -0.0088529886  
## CourteousExchange -3.545702e-01 -0.011206902 -0.158306457 -0.0014125318  
## ActiveListening -3.177831e-01 -0.030455382 -0.130615321 0.0564121836  
## PC5 PC6 PC7 PC8  
## Children -0.425007375 0.453536143 -0.5046402979 0.135142354  
## Age 0.357697699 0.773576401 -0.1368434323 0.086450300  
## Income -0.720184552 0.292693212 0.5467072111 -0.139699895  
## Outage\_sec\_perweek 0.389069453 0.307164556 0.6116834920 -0.090349584  
## Tenure 0.009109207 -0.007959251 0.0639077817 0.274879056  
## MonthlyCharge 0.041617906 0.044909135 -0.2026754829 -0.894660100  
## Bandwidth\_GB\_Year -0.024692223 -0.019091821 0.0260510775 0.201705921  
## TimelyResponse 0.047596063 0.011256248 -0.0123181448 0.011631794  
## TimelyFix 0.052911998 0.034885319 -0.0001947498 0.029802176  
## TimelyReplacement 0.040702217 -0.009123391 0.0446910669 0.033119459  
## Reliability -0.010447668 -0.017121816 0.0115492768 -0.015192399  
## Options 0.063029076 0.021581812 -0.0670953534 -0.044799765  
## RespectfulResponse -0.030055668 -0.068185734 -0.0143418789 0.007881936  
## CourteousExchange -0.083569725 -0.009325830 -0.0121162671 0.028327676  
## ActiveListening -0.016270504 0.080612275 -0.0242822856 -0.157164043  
## PC9 PC10 PC11 PC12  
## Children 0.060572586 -0.0635094219 -0.02025853 0.001660571  
## Age 0.034371998 0.0322205737 -0.04488296 -0.020513682  
## Income 0.007960123 -0.0421848942 -0.01311723 0.034793710  
## Outage\_sec\_perweek 0.054423129 0.0440279449 -0.01922574 0.057044011  
## Tenure -0.063361978 0.0045135068 0.01654716 -0.004380511  
## MonthlyCharge 0.158573796 -0.0104765291 -0.01417351 -0.039476337  
## Bandwidth\_GB\_Year -0.052323267 0.0007784373 0.01493936 0.001543437  
## TimelyResponse 0.044938635 -0.1103818197 0.06745393 0.119640425  
## TimelyFix 0.102193095 -0.1620004167 0.07853555 0.233449406  
## TimelyReplacement 0.135233085 -0.3179686405 0.09935833 -0.586609865  
## Reliability 0.069483931 -0.5267031117 0.41981867 0.384223485  
## Options -0.169814846 0.0313638588 0.20605941 0.555211254  
## RespectfulResponse 0.065329985 0.0737709813 -0.75720418 0.336839866  
## CourteousExchange 0.282672347 0.7442048914 0.41933859 0.047202708  
## ActiveListening -0.900909827 0.1177469842 0.08851202 -0.095036067  
## PC13 PC14 PC15  
## Children -0.0086153410 -0.0025557744 0.0338110448  
## Age -0.0401202001 0.0131664752 -0.0326739667  
## Income -0.0048793264 0.0132889543 -0.0006141607  
## Outage\_sec\_perweek -0.0204696718 0.0026897427 -0.0022131082  
## Tenure -0.0012593990 0.0146601888 0.6962435798  
## MonthlyCharge 0.0191573082 -0.0004450851 0.0545843171  
## Bandwidth\_GB\_Year 0.0004939025 0.0049024664 -0.7141142620  
## TimelyResponse 0.2565420002 0.7796832769 -0.0044763041  
## TimelyFix 0.5117901225 -0.6035679171 0.0042460410  
## TimelyReplacement -0.5236385147 -0.1322728742 -0.0027527103  
## Reliability -0.2073822351 0.0177949582 -0.0006306535  
## Options -0.4684650947 -0.0456051350 0.0060435508  
## RespectfulResponse -0.3251246734 -0.0449828643 0.0021497377  
## CourteousExchange -0.1651506225 -0.0444844547 -0.0004080010  
## ActiveListening -0.0107962227 -0.0574626641 0.0016873938

# Part V. Conclusion

Using Principal Component Analysis on the telecommunciations data set we were able to reduce the dimensionality of the data in such a way as to account for the variation in the data set while making it more consumable and readily understandable. Comparing and contrasting the PCA results between the churned and not churned data sets we were able to glean that lengthier or more frequent outages via service interruptions are a stronger shared characteristic among customers who churn rather than customers who do not churn. Armed with this knowledge, the organization can begin root cause analysis on the source of the service interruptions as well as develop customer service initiatives to attempt to retain those customers experiencing outages and thereby cut company costs associated with defending their portfolio of customers.

A limitation of Principal Component Analysis is that it sacrifices some accuracy in order to reduce dimensionality. However, this can be offset by using validation models and monitoring of information gleaned from the PCA that otherwise would not have been observable by the audience pre dimension reduction. In this case, it is recommended that specific KPIs also be built around service outages and monitored in conjunction with retention rates to assure ongoing benefit from the PCA performed.