#visdat

local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)

+ if(nchar(pkg)) library(pkg, character.only=TRUE)})

#read exl

local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)

+ if(nchar(pkg)) library(pkg, character.only=TRUE)})

#plyr

local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)+ if(nchar(pkg)) library(pkg, character.only=TRUE)})

#tidyverse

local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)+if(nchar(pkg)) library(pkg, character.only=TRUE)})

#factoextra

local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)+if(nchar(pkg)) library(pkg, character.only=TRUE)})

data <- read.csv("C:/Users/nshai/OneDrive/Desktop/churn.csv")

print(summary(data))

#detecting duplicates or missing data

sum(duplicated(data))

colSums(is.na(data))

vis\_miss(data)

#Clean data name change

dc <- data

#Cleaning All Missing Values

#got rid of children na (Skewed distribution)

dc$Children[is.na(dc$Children)]<-median(dc$Children, na.rm = TRUE)

#Age NA (uniform Distribution)

dc$Age[is.na(dc$Age)]<-mean(dc$Age, na.rm = TRUE)

#Income NA (Skewed distribution)

dc$Income[is.na(dc$Income)]<-median(dc$Income, na.rm = TRUE)

#Tenure(bimodal distribution)

dc$Tenure[is.na(dc$Tenure)]<-median(dc$Tenure, na.rm = TRUE)

#Bandwith (bimodal distribution)

dc$Bandwidth\_GB\_Year[is.na(dc$Bandwidth\_GB\_Year)]<-median(dc$Bandwidth\_GB\_Year, na.rm = TRUE)

#Techie(used Mode due to logical data type)

dc$Techie[is.na(dc$Techie)]<-mode(dc$Techie)

#Phone(used Mode due to logical data type)

dc$Phone[is.na(dc$Phone)]<-mode(dc$Phone)

#TechSupport(used Mode due to logical data type)

dc$TechSupport[is.na(dc$TechSupport)]<-mode(dc$TechSupport)

#Showing no more missing values

colSums(is.na(dc))

summary(dc)

vis\_miss(dc)

#Detection of Outliers

#has 11 outliers over 90000

pop\_bp <- hist(dc$Population, xlab ="Population")

length(pop\_bp)

pop\_out <- str(dc[which(dc$Population>90000),])

#has 451 outliers greater than 6

Children\_bp <- boxplot(dc$Children, xlab ="Children")

length(Children\_bp)

Children\_out <- str(dc[which(dc$Children>6),])

#has no outliers

Age\_bp <- boxplot(dc$Age, xlab ="Age")$out

length(Age\_bp)

#has 3 outliers over 200,000

Income\_bp <- boxplot(dc$Income, xlab ="Income")$out

Income\_hist <- hist(dc$Income, xlab ="Income")

Income\_out <- str(dc[which(dc$Income>200000),])

length(Income\_bp)

#has 500 over 30 sec and 11 under 0 sec

outage\_bp <- boxplot(dc$Outage\_sec\_perweek, xlab ="Outage")$out

outage\_out <- str(dc[which(dc$Outage\_sec\_perweek>30),])

outagelow\_out <- str(dc[which(dc$Outage\_sec\_perweek<0),])

length(outage\_bp)

#has 38 outliers that are less than 4 or greater than 20

Email\_bp <- boxplot(dc$Email, xlab ="Email")$out

length(Email\_bp)

#has 8 outliers that are greater than 5

Contacts\_bp <- boxplot(dc$Contacts, xlab ="Contacts")$out

length(Contacts\_bp)

#has 94 outliers that are greater than 2

equip\_bp <- boxplot(dc$Yearly\_equip\_failure, xlab ="Equip failure")$out

length(equip\_bp)

#has no Outliers

Tenure\_bp <- boxplot(dc$Tenure, xlab ="Tenure")$out

length(Tenure\_bp)

#has 5 outliers that are greater than 300

Charge\_bp <- boxplot(dc$MonthlyCharge, xlab ="MonthlyCharge")$out

length(Charge\_bp)

#has no Outliers

Bandwidth\_bp <- boxplot(dc$Bandwidth\_GB\_Year, xlab ="Bandwidth")$out

length(Bandwidth\_bp)

#outages

dc$Outage\_sec\_perweek[dc$Outage\_sec\_perweek > 13] <- NA

dc$Outage\_sec\_perweek[dc$Outage\_sec\_perweek < 2] <- NA

colSums(is.na(dc))

#Outages NA (uniform Distribution)

dc$Outage\_sec\_perweek[is.na(dc$Outage\_sec\_perweek)]<-mean(dc$Outage\_sec\_perweek, na.rm = TRUE)

outage\_bp <- boxplot(dc$Outage\_sec\_perweek, xlab ="Outage")$out

#equipment failure

dc$Yearly\_equip\_failure[dc$Yearly\_equip\_failure > 2] <- NA

#got rid of Yearly\_equip\_failure na (Skewed distribution)

dc$Yearly\_equip\_failure[is.na(dc$Yearly\_equip\_failure)]<-median(dc$Yearly\_equip\_failure, na.rm = TRUE)

colSums(is.na(dc))

equip\_bp <- boxplot(dc$Yearly\_equip\_failure, xlab ="Equip failure")$out

#Download clean Data

write.csv(dc, "C:/Users/nshai/OneDrive/Desktop/clean\_churn.csv")

#creating subset for pca"Outage\_sec\_perweek",

pcaValues <- dc[,c( "Email","Yearly\_equip\_failure", "Contacts", "Tenure", "Age")]

view(pcaValues)

mtdc.pca <- prcomp(pcaValues[,],center = TRUE, scale= TRUE)

mtdc.pca

fviz\_eig(mtdc.pca, choice = "eigenvalue", addlabels = TRUE)