D206 Performance Assessment

Data Cleaning

Instructor: Dr.Straw

Student Name: Nicole Haibach

Student ID: 001260374

Email: nhaibac@wgu.edu

# Research Question

## Description of Question

My research question for the assessment will be evaluating what factors led to people discontinuing service. What factors led to people discontinuing service? I will look at what people discontinued their service in the past month and compare the yes versus no people. Then look at the outages per week they experience, equipment failures, tenure, and monthly charges to see if there is a correlation between any of them and discontinuing service.

## Identifying Variables

| **Variable Name** | **Data Type** | **Description** | **Example** |
| --- | --- | --- | --- |
| CaseOrder | Qualitative | Numerical placeholder to keep data in order | 1 |
| Customer\_id | Qualitative | Unique number for each customer | K662701 |
| Interaction, UID | Qualitative | Unique transaction number for customers | 70ddaa89-b726-49dc-9022-2d655e4c7936 |
| City | Qualitative | Customer city based on billing statement | Saint Cloud |
| State | Qualitative | Customer state based on the billing statement | FL |
| County | Qualitative | Customer county based on the billing statement | Osceola |
| Zip | Qualitative | Zip code of billing statement | 34771 |
| Latitude | Qualitative | Latitude of residence | 28.27646 |
| Longitude | Qualitative | Longitude of residence | -81.16273 |
| Population | Quantitative | Total population within a mile radius | 17351 |
| Area | Qualitative | Type of area based on census | Suburban |
| TimeZone | Qualitative | The time zone the customer resides in | America/New\_York |
| Job | Qualitative | Job title of customer | Teaching laboratory technician |
| Children | Quantitative | The number of children the customer has | 2 |
| Age | Quantitative | Age of Customer | 49 |
| Education | Qualitative | Amount of education the customer has received | In some College, Less than 1 Year |
| Employment | Qualitative | Employment status of the customer | Full Time |
| Income | Quantitative | Total income of the customer | 58634.51 |
| Marital | Qualitative | Customers marital status | Separated |
| Gender | Qualitative | Customer Gender identified by the customer | Prefer not to answer |
| Churn | Qualitative | If customers discontinue services | No |
| Outage\_sec\_perweek | Quantitative | Average seconds of outages per week in the area | 6.637258801 |
| Email | Quantitative | Emails sent to customers last year | 20 |
| Contacts | Quantitative | Number of Technical support contacts | 2 |
| Yearly\_equip\_failure | Quantitative | Number of equipment failures within a year | 3 |
| Techie | Qualitative | If customers feel technically inclined | NA |
| Contract | Qualitative | Length of contract | Month-to-month |
| Port\_modem | Qualitative | If the customer has a portable modem | Yes |
| Tablet | Qualitative | If the customer has a tablet | No |
| InternetService | Qualitative | Internet provider | DSL |
| Phone | Qualitative | If the customer has phone services | Yes |
| Multiple | Qualitative | If the customer has multiple lines | No |
| OnlineSecurity | Qualitative | If the customer has security for their computer | Yes |
| Online backup | Qualitative | If the customer keeps a backup | Yes |
| DeviceProtection | Qualitative | If the customer has the protection add-on | No |
| TechSupport | Qualitative | If the customer has a technical support add-on | No |
| StreamingTV | Qualitative | If the customer has a streaming television | No |
| StreamingMovies | Qualitative | If the customer has streaming movies | No |
| PaperlessBilling | Qualitative | If the customer opted for paperless billing | Yes |
| payment method | Qualitative | The method of payment chosen by the customer | Bank Transfer(automatic) |
| Tenure | Quantitative | length of time the customer has stayed with the provider in months | 8.220686373 |
| MonthlyCharge | Quantitative | The average amount charged to customers per month | 118.3668439 |
| Bandwith\_GB\_Year | Quantitative | Average GBs of data used by customers per year | 1312.874964 |
| Item1 | Qualitative | Importance scale rating for timely responses | 5 |
| Item2 | Qualitative | Importance scale rating for timely fixes | 4 |
| Item3 | Qualitative | Importance scale rating for timely replacements | 4 |
| Item4 | Qualitative | Importance scale rating for reliability | 3 |
| Item5 | Qualitative | Importance scale rating for options | 4 |
| Item6 | Qualitative | Importance scale rating for Respectful responses | 3 |
| Item7 | Qualitative | Importance scale rating for courteous exchange | 4 |
| Item8 | Qualitative | Importance scale rating for evidence of active listening | 4 |

# Data Cleaning Plan

## Propose a Plan

Several functions were used to detect duplicates, missing values, outliers, and other data quality issues. The first was to identify any duplicates in the data, and the number of duplicates was summed. The next task would be to utilize visualization charts to view the amount of missing data to decide further how the missing data will be handled. The first step was to use column sums to sum the amount of missing data for each column, then use a separate line to create a graph to visualize the missing data. The process used to detect outliers was to create a boxplot for all the quantitative columns to visualize the outliers using length to detect the number of outliers.

## Justify approach

The planned approach shows visually and numerically what is available in the data. When you can visualize your data, it helps to decide the best way to tackle the data cleaning. The command viss\_miss was used to visualize the missing data and give missingness percentages. This helps you to understand the best way to clean it up since you would not want to just delete all the rows of missing data after a certain percentage. If you just deleted all the missing data, you would ruin the large sample size. Boxplots are a great way to visualize the data along with the outliers. “A boxplot graphically represents the distribution of a quantitative variable by visually displaying five common location summary (minimum, median, first/third quartiles and maximum) and any observation that was classified as a suspected outlier using the interquartile range criterion” (Soetewey, 2020). The package diplyr was used for the recoding of values to apply the mode imputation to Phone, Techie, and TechSupport columns. Recoding the values makes mode imputation possible due to R being able to read the values as numbers to find the value that is used most often a lot easier. The packages plyr, tidyverse, and factoextra are all used to complete the principal component analysis.

## Justify Language

R is a programming language mainly used for statistics. “R has many functionalities for data analysis. R is great for statistical analysis, which are often cited in academic journals” (R or Python, 2023). This determines using R for data cleaning is reasonable as you use a lot of statistics and graphs to analyze your data to find the missing data, visualize where the data is, visualize the way each variable is distributed, and find outliers. The package used for this portion was visdat because it allowed the use of vis\_miss, which is a great way to visualize missing data values in a graph.

## Annotated Code

sum(duplicated(data))

colSums(is.na(data))

vis\_miss(data)

pop\_bp <- boxplot(dc$Population)$out

length(pop\_bp)

Children\_bp <- boxplot(dc$Children)$out

length(Children\_bp)

Age\_bp <- boxplot(dc$Age)$out

length(Age\_bp)

Income\_bp <- boxplot(dc$Income)$out

length(Income\_bp)

outage\_bp <- boxplot(dc$Outage\_sec\_perweek)$out

length(outage\_bp)

Email\_bp <- boxplot(dc$Email)$out

length(Email\_bp)

Contacts\_bp <- boxplot(dc$Contacts)$out

length(Contacts\_bp)

equip\_bp <- boxplot(dc$Yearly\_equip\_failure)$out

length(equip\_bp)

Tenure\_bp <- boxplot(dc$Tenure)$out

length(Tenure\_bp)

Charge\_bp <- boxplot(dc$MonthlyCharge)$out

length(Charge\_bp)

Bandwidth\_bp <- boxplot(dc$Bandwidth\_GB\_Year)$out

length(Bandwidth\_bp)

# Data Cleaning Process

## Findings

The churn data showed not to have any duplicate data. The data had eight columns of data with missing values. The columns with missing values were children, age income, techie, phone, tech support, tenure, and bandwidth. The columns were missing 2495, 2475, 2490, 2477, 1026, 991, 931, and 1021 values respectively. Outliers were detected in eight of eleven quantitative fields. The population was found to have 11 outliers with over 90,000. Children had 451 values that were over 6. Income had three outliers with values over 200,000. Outage seconds per week had over 500 outliers that reported over 30 seconds and 11 that reported less than zero. The email column had 38 outliers that were less than 4 or greater than 20. Contacts had eight outliers that were greater than five. Yearly equipment failure had 94 outliers that were greater than two. The monthly charge had five outliers that were greater than 300.

## Justify methods

The methods used to clean up the missing values in the various columns were based on the distribution that was present. Children and income both had skewed distributions, which allows for the median to be used during imputation. Age had a uniform distribution so the mean was used to impute the data of missing data. Tenure and bandwidth both had a bimodal distribution, which means mode or median could be used to impude the data for both columns median was used over mode. Techie, phone, and tech support are all non-numerical columns, so mode was used to impute the data of missing values. Several outliers were retained due to very plausible values while out of the normal are still plausible. The outliers that were retained were population, children, income, email, contacts, and monthly charges. The outliers in outages were changed to NA values and then imputed using mean due to uniform distribution. The outliers in yearly equipment failure were changed to NA values and then imputed using the median due to a skewed distribution.

## Summarize outcomes

The data had several imputations completed to fill in gaps of missing data for both numerical and logical data based on distribution. The data also had two sets of outliers cleaned up to align the data. The vis\_miss chart with the clean data will be provided.

## Annotated code

#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)

#outages

dc$Outage\_sec\_perweek[dc$Outage\_sec\_perweek > 30] <- NA

dc$Outage\_sec\_perweek[dc$Outage\_sec\_perweek < 0] <- 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)

#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))

## Copy of cleaned data set

This will be attached when turned in.

## Summarize limits

The disadvantage of using the mean or mode method of imputation is that it tends to skew your statistics. The mode method also tends to leave a large peak in your histogram, which does look a bit off as far as pleasing bell curve graphs go. Imputing outliers can change your summary statistics, especially with taking out all of them. These are all disadvantages because they move the data set and can change the meaning behind some of the numbers. The best way to prevent any misinterpretations is to make sure you keep your values as close to the original as possible and only make the necessary changes.

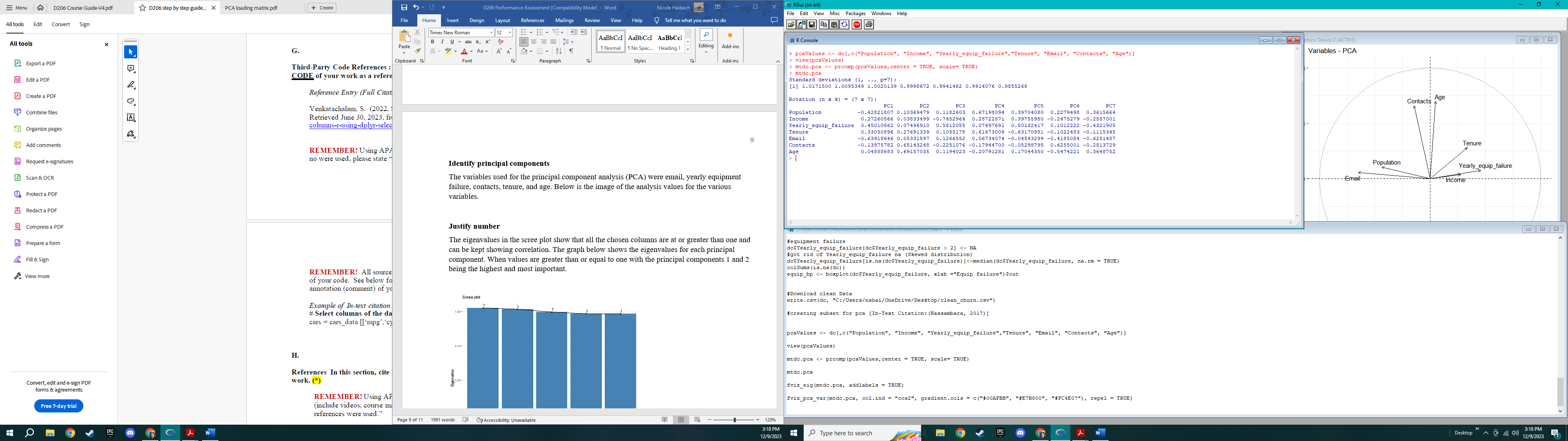
## Discuss limits

The limits of using the data set that was cleaned are that cleaning the data may skew the viewpoint of the data slightly. Especially in regards to answering my research question, which is about understanding why people stopped using the company for the internet. It can skew the answer when you take out the outliers since there are no longer values that could be the real reason for them leaving. An example would be if you took out all the high numbers in the contacts column since they were deemed outliers then you are taking away the real reason they left the service, which is probably due to poor service.

# Principal Component Analysis

## Identify principal components

The variables used for the principal component analysis (PCA) were email, yearly equipment failure, contacts, tenure, and age. Below is the image of the analysis values for the various variables.



## Justify number

The eigenvalues in the scree plot show that all the chosen columns are at or greater than one and can be kept showing correlation. The graph below shows the eigenvalues for each principal component. When values are greater than or equal to one with the principal components 1 and 2 being the highest and most important.



## Describe the benefits

The organization can benefit from the PCA because it shows that their values do correlate with one another and can be used to conclude. The data can be used to conclude how tenure or the length of time someone stays with the company correlates with age and the number of emails sent according to PC2. This justifies being able to answer my research question using the variables.

**Code References**

Kassambara. (2017, October 7). *Principal component analysis in R: Prcomp VS princomp*. STHDA. http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

**References**

*R or Python*. (2023, July 7). Western Governors University. https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html

Soetewey, A. (2020, January 22). *Descriptive statistics in R*. Stats and R. https://statsandr.com/blog/descriptive-statistics-in-r/#boxplot