**D206 Data Cleaning Performance Assessment**

**Crystal Ford, Student ID 011377930**

**A: Research Question**

My research question is “What customer factors influence monthly charge?” This question would give valuable insight into what specific factors have an influence on the number of services a customer will subscribe to. The question could be further broken down and explored by age, income, education, etc. This would be great for marketing to understand who to offer discounted packages to, and who to advertise specific services to.

**B: Description of Variables**

This dataset contains 51 columns. Each of the columns contains a separate variable relating to customers and service information for a cable company.

1. CaseOrder (quantitative) row 1 example: 1

This column is used as a placeholder variable to keep the data in the original order of raw data file.

1. Customer\_id (qualitative) row 2 example: K409198

This column is the unique customer id used to identify customers.

1. Interaction (qualitative) row 1 example: aa90260b-4141-4a24-8e36-b04ce1f4f77b

This column has a unique ID related to customer transactions.

1. City (qualitative) row 1 example: Point Baker
2. State (qualitative) row 1 example: AK
3. County (qualitative) row 1 example: Prince of Wales-Hyder
4. Zip (qualitative) row 1 example: 99927
5. Lat (quantitative) row 1 example: 56.251
6. Lng (quantitative) row 1 example: -133.37571

These columns identify the location data for the customer residence including city, state, county, zip as well as GPS coordinates latitude and longitude as listed on the billing statement.

1. Population (quantitative) row 1 example: 38

This column represents the population within a mile radius of customer residence, based on census data.

1. Area (qualitative) row 1 example: Urban

This column represents the type of area of customer residence, based on census data.

1. Timezone (qualitative) row 1 example: America/Sitka

This column represents timezone of customer residence, based on sign-up information.

1. Job (qualitative) row 1 example: Environmental health practitioner

This column represents job of the customer based on sign-up information.

1. Children (quantitative) row 1 example: NA

This column represents number of children in customer’s household, based on sign-up information.

1. Age (quantitative) row 1 example: 68

This column represents the age of customer based on sign-up information.

1. Education (qualitative) row 1 example: Master’s Degree

This column represents the highest earned degree of customer based on sign-up information.

1. Employment (qualitative) row 1 example: Part Time

This column represents employment status of customer based on sign-up information.

1. Income (quantitative) row 1 example: 28561.99

This column represents annual income of customer based on sign-up information.

1. Marital (qualitative) row 1 example: Widowed

This column represents marital status of customer based on sign-up information.

1. Gender (qualitative) row 1 example: Male

This column represents self-identification of customer gender as male, female, or nonbinary.

1. Churn (qualitative) row 1 example: No

This column represents whether or not the customer has discontinued service in the last month.

1. Outage\_sec\_perweek (quantitative) row 1 example: 6.972566093

This column represents the average number (in seconds per week) of system outages in customer’s neighborhood.

1. Email (quantitative) row 1 example: 10

This column represents number of emails sent to customer in last year.

1. Contacts (quantitative) row 1 example: 0

This column represents number of times customer contacted technical support. It is not indicated in dictionary what the time frame is for number of contacts, whether it’s an annual number, or lifetime.

1. Yearly\_equip\_failure (quantitative) row 1 example: 1

This column represents the number of times the customer’s equipment has failed and had to be either reset or replaced in the last year.

1. Techie (qualitative) row 1 example: No

This column represents a self indication on the customer’s part of whether they consider themselves to be technically inclined. This is based on a customer questionnaire from when they signed up for services.

1. Contract (quantitative) row 1 example: One year

This column represents the term of the contract for the customer.

1. Port\_modem (qualitative) row 1 example: Yes

This column represents whether the customer has a portable modem or not.

1. Tablet (qualitative) row 1 example: Yes

This column represents whether the customer has a tablet or not.

1. InternetService (qualitative) row 1 example: FiberOptic

This column represents the type of customer’s internet service provider. Options are DSL, Fiber Optic, or None.

1. Phone (qualitative) row 1 example: Yes

This column represents whether or not the customer has a phone service.

1. Multiple (qualitative) row 1 example: No

This column represents whether or not the customer has multiple lines.

1. OnlineSecurity (qualitative) row 1 example: Yes

This column represents whether customer has an online backup security add-on.

1. OnlineBackup (qualitative) row 1 example: Yes

This column represents whether custoemr has an online backup add-on.

1. DeviceProtection (qualitative) row 1 example: No

This column represents whether customer has device protection add-on.

1. TechSupport (qualitative) row 1 example: No

This column represents whether customer has a technical support add-on.

1. StreamingTV (qualitative) row 1 example: No

This column represents whether customer has streaming Tv.

1. StreamingMovies (qualitative) row 1 example: Yes

This column represents whether customer has streaming movies.

1. PaperlessBilling (qualitative) row 1 example: Yes

This column represents whether the customer has paperless billing.

1. PaymentMethod (qualitative) row 1 example: Credit Card (automatic)

This column represents customer’s payment method. Options include electronic check, mailed check, bank (automatic bank transfer), and credit card (automatic)

1. Tenure (quantitative) row 1 example: 6.795512947

This column represents the number of months customer has been with the provider.

1. MonthlyCharge (quantitative) row 1 example: 171.4497621

This column represents the amount charged to customer monthly. This value reflects an average per customer.

1. Bandwidth\_GB\_Year (quantitative) row 1 example: 904.5361102

This column represents the average amount of data used (in GB) annually by customer.

1. Item1 (qualitative) row 1 example: 5

This column represents customer’s response to a survey question regarding the importance of timely response. The survey uses 1 to indicate that this is "most important" and an 8 to indicate that this is "least important".

1. Item2 (qualitative) row 1 example: 5

This column represents customer’s response to a survey question regarding the importance of timely fixes. The survey uses 1 to indicate that this is "most important" and an 8 to indicate that this is "least important".

1. Item3 (qualitative) row 1 example: 5

This column represents customer’s response to a survey question regarding the importance of timely replacements. The survey uses 1 to indicate that this is "most important" and an 8 to indicate that this is "least important".

1. Item4 (qualitative) row 1 example: 3

This column represents customer’s response to a survey question regarding the importance of reliability. The survey uses 1 to indicate that this is "most important" and an 8 to indicate that this is "least important".

1. Item5 (qualitative) row 1 example: 4

This column represents customer’s response to a survey question regarding the importance of options. The survey uses 1 to indicate that this is "most important" and an 8 to indicate that this is "least important".

1. Item6 (qualitative) row 1 example: 4

This column represents customer’s response to a survey question regarding the importance of respectful response. The survey uses 1 to indicate that this is "most important" and an 8 to indicate that this is "least important".

1. Item7 (qualitative) row 1 example: 3

This column represents customer’s response to a survey question regarding the importance of courteous exchange The survey uses 1 to indicate that this is "most important" and an 8 to indicate that this is "least important".

1. Item8 (qualitative) row 1 example: 4

This column represents customer’s response to a survey question regarding the importance of evidence of active listening. The survey uses 1 to indicate that this is "most important" and an 8 to indicate that this is "least important".

**C1: Plan to Assess Quality of Data**

My first step will be to import the CSV into R Studio. After that, I will look over the data and identify any immediate issues including NAs or missing values. I’ll verify that the datatypes for each column make sense for the data within, and then scan through the data in each column to ensure that it looks like it’s in the proper format. After looking at the data myself, I will use R Studio to detect duplicates and nulls. After detection of duplicates and nulls, I will determine the best method to rectify each instance. To detect duplicates, I will be using the duplicated() function.

After addressing all duplicates, I will look for missing values using the colSums(is.na) function to return the number of NAs in each column. I will further investigate this by using vismiss() function in the visdata package to see how high the percentage of missing information is to help determine how best to remedy the missing variables.

To determine the outliers in the dataset, I will be using the scale() function to determine the z-score for each quantitative variable. Z-scores are used to tell how many standard deviations a number is from the mean or average. If a dataset is more than 3 standard deviations from the mean, it is considered to be an outlier.   
  
After detecting outliers, I will determine whether any re-expression of categorical variables is necessary in order to make sense within the dataset, and to ensure that they represent the data within.

**C2: Justification of Approach**

By looking over the data first as a whole, I can ensure that the columns for the data are representative of the data contained within. I can visually spot where there is missing data and come up with a plan to impute relevant replacement values as needed in order to keep the data as close to original as possible.

Once I have an idea of how many duplicates there are, I will use the distinct() function to eliminate all duplicates and ensure that all entries are unique.

Next I will use the colSums(is.na) function to detect any NA values within the dataset. After I determine missing values, I will be using the hist() function before and after each imputation to visually verify that the imputations I use don’t change the overall skew of the data. The imputations for missing data will be determined on a case by case basis using either the median or mean for quantitative variables, with the mean being used for normally distributed data, and median being used for skewed data. For the imputation of missing data in categorical data, the mode will be used.

After using the scale() function to determine the outliers in the dataset via the z-scores, I will inspect the data to see if the outliers require treatment. I will count the number of outliers in each category using the which() function. After ascertaining the number and severity of outliers, I will treat each individually depending on what makes the most sense in each case.

**C3: Justification of Tools**

For this project, I chose to use R. I have no background in computer science or programming, and was introduced to R through a Google data analytics course this summer. It is easy for me to understand, and I enjoy using it. R is great for statistical models and because it is open source, many packages are created by people to do many different things. I am slowly working on learning Python, but I am nowhere near comfortable enough with it at this point to try to tackle a project using it.

I used the tidyverse library which includes lots of different packages. Readr was used to read the .csv I’m working with. Ggplot2 was used to make the histograms I used to compare before and after I imputed for missing NAs in each column. I also used the vis\_miss() function in the visdat package to visualize all the missing data in the dataset and compared that visualization to one taken after I imputed the missing variables to ensure that there were no more NAs in the dataset.

**C4: Code Used to Identify Anomalies**

#import dataset  
read.csv("churn\_raw\_data.csv")

#install packages tidyverse  
 library(tidyverse)

#import churn\_raw\_data dataset  
churn\_raw\_data <- read\_csv("churn\_raw\_data.csv")

#Look for duplicates  
View(churn\_raw\_data)  
sum(duplicated(churn\_raw\_data))

#Rename dataset for less typing  
churn<-churn\_raw\_data

#Look for NAs in dataset  
colSums(is.na(churn))

#Install visdat package and visualize missing data in dataset  
library(visdat)  
vis\_miss(churn)

#Determine datatype for each column  
str(churn)  
#Look for missing values  
colSums(is.na(churn))

#Verify distributions before and after imputations  
hist(churn$Children)  
hist(churn$Tenure)  
hist(churn$Income)  
hist(churn$Bandwidth\_GB\_Year)

#Determine the outliers for all quantitative variables   
churn$Children\_z<-scale(x=churn$Children)  
churn$Age\_z<-scale(x=churn$Age)  
churn$Income\_z<-scale(x=churn$Income)  
churn$Outage\_sec\_perweek\_z<-scale(x=churn$Outage\_sec\_perweek)  
churn$Contacts\_z<-scale(x=churn$Contacts)  
churn$Yearly\_equip\_failure\_z<-scale(x=churn$Yearly\_equip\_failure)  
churn$Tenure\_z<-scale(x=churn$Tenure)  
churn$MonthlyCharge\_z<-scale(x=churn$MonthlyCharge)  
churn$Bandwidth\_GB\_Year\_z<-scale(x=churn$Bandwidth\_GB\_Year)  
view(churn)

#Determine range of values of outliers

b\_Children<-boxplot(churn$Children)  
b\_income<-boxplot(churn$Income)  
b\_outage<-boxplot(churn$Outage\_sec\_perweek)  
b\_contacts<-boxplot(churn$Contacts)  
b\_yearly<-boxplot(churn$Yearly\_equip\_failure)

#Determine quantity of outliers

children\_query<-churn[which(churn$Children>6),]  
income\_query<-churn[which(churn$Income>75000),]  
outage\_query<-churn[which(churn$Outage\_sec\_perweek>20),]  
contacts\_query<-churn[which(churn$Contacts>5),]  
yearly\_query<-churn[which(churn$Yearly\_equip\_failure>2),]

#Examine all z-scores to determine if outliers need to be treated or are reasonable  
hist(churn$Age\_z)  
hist(churn$Income\_z)  
hist(churn$Outage\_z)  
hist(churn$Contacts\_z)  
hist(churn$Yearly\_equip\_failure\_z)  
hist(churn$Tenure\_z)  
hist(churn$MonthlyCharge\_z)  
hist(churn$Bandwidth\_GB\_Year\_z)

\*\*\*See code attached.

**D1: Cleaning Findings**

After importing the dataset, I used sum(duplicated) to determine that there were no duplicates within the dataset. Next, I used colSums(is.na) to determine that there was missing data in the following categories: (see table)

|  |  |
| --- | --- |
| **Categories** | **Fields Missing Data** |
| Age | 2475 |
| Techie | 2477 |
| Phone | 1026 |
| TechSupport | 991 |
| Tenure | 931 |
| Income | 2490 |
| Bandwidth\_GB\_Year | 1021 |
| Children | 2495 |

After determining the categories with missing data, I used the scale() function to determine the z-scores of all quantitative variables. The z-score is used to determine how many deviations a number is from the statistical mean. If a z-score is +/- 3, it is considered to be an outlier. The following categories had outliers: (see table)

|  |  |
| --- | --- |
| **Categories** | **Number of Outliers** |
| Children | 451 |
| Income | 858 |
| Outage\_sec\_perweek | 503 |
| Contacts | 8 |
| Yearly\_equip\_failure | 94 |

The Timezone category had inconsistencies, so I used the str\_replace() function to put all cities into standard US timezones.

**D2: Justification of Mitigation Methods**

To treat missing variables in the dataset, I first determined how many missing variables there were by using the colSums(is.na) function. After getting this information, I found out the datatypes for each variable with missing values using the str() funciton. For any missing values that were categorical, I imputed the mode for the missing values using the which.max() function.

For any missing values that were numeric, I looked at the distributions of the data and determined its structure. For any categories that were either skewed or bi-modal, I imputed the median for the missing values. For any categories that were normal in distribution, I imputed the mean.

After treating the duplicates and missing values, I next determined the outliers for all quantitative values in the dataset. I did this by using boxplots to visualize the outliers and determine their range. After determining the range of outliers, I was able to find the number of outliers in each category using the query() function.

For the Lat, Lng, and Population variables, I determined that outliers did not make sense because the data is representing specific locations and would be unique to each observation.

For the Children variable, I found a few outliers that were representative of families having more than 6 children and while this is less common, it is not unheard of in any way, therefore I left it as is.

For the Income variable I found some families reporting income over the mean annually, I do not consider this to be an outlier because income can vary so greatly from household to household.

For the variable of Outages measured in seconds per week, I found that there were 500 observations of over 30 seconds per week. Because this number is so large, I feel like this data is valuable to the company.

For the variable of Contacts, I left this number alone because there were only 8 outliers in the whole dataset, and the data represents the number of times a customer contacted the company, therefore the information seems valueable.

For the variable of Yearly Equipment Failure, I found outliers, however this number represents the number of times equipment failed in a year and this seems like valuable data for the company.

For the variables of Tenure, Monthly Charge, and Bandwidth there were no outliers detected.

**D3: Summary of Outcomes**

This data cleaning process included verifying there was no duplicate data, investigating and replacing all missing values with reasonable estimates, and investigating all outliers. These are all excellent steps on the way to cleaning data. No one wants incomplete or duplicated data clogging the system. There were lots of missing values, and it’s difficult to compare things with missing values.

**D4: Mitigation Code**

#Impute median for missing values in Children, Tenure, Income, Bandwidth\_GB\_Year due to bi-modal distribution

churn$Children[is.na(churn$Children)]<-median(churn$Children, na.rm=TRUE)  
churn$Tenure[is.na(churn$Tenure)]<-median(churn$Tenure, na.rm=TRUE)  
churn$Income[is.na(churn$Income)]<-median(churn$Income, na.rm=TRUE)  
churn$Bandwidth\_GB\_Year[is.na(churn$Bandwidth\_GB\_Year)]<-median(churn$Bandwidth\_GB\_Year, na.rm=TRUE)

#Impute mode for missing values in Techie and Phone due to categorical variables  
churn$Techie[is.na(churn$Techie)]<-names(which.max(table(churn$Techie)))  
churn$Phone[is.na(churn$Phone)]<-names(which.max(table(churn$Phone)))

#Consolidate Timezones into standard US Timezones  
library(stringr)

#US/Eastern timezones  
churn$Timezone<-str\_replace(churn$Timezone, "America/Indiana/Marengo", "US/Eastern")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Indiana/Vincennes", "US/Eastern")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Detroit", "US/Eastern")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Kentucky/Louisville", "US/Eastern")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Indiana/Indianapolis", "US/Eastern")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Indiana/Petersburg", "US/Eastern")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Indiana/Winamac", "US/Eastern")  
churn$Timezone<-str\_replace(churn$Timezone, "America/New\_York", "US/Eastern")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Toronto", "US/Eastern")

#US/Mountain timezones  
churn$Timezone<-str\_replace(churn$Timezone, "America/Boise", "US/Mountain")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Ojinaga", "US/Mountain")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Denver", "US/Mountain")

#US/Central timezones  
churn$Timezone<-str\_replace(churn$Timezone, "America/Indiana/Knox", "US/Central")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Indiana/Tell\_City", "US/Central")  
churn$Timezone<-str\_replace(churn$Timezone, "America/North\_Dakota/New\_Salem", "US/Central")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Menominee", "US/Central")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Chicago", "US/Central")

#US/Pacific timezones  
churn$Timezone<-str\_replace(churn$Timezone, "America/Los\_Angeles", "US/Pacific")

#US/Alaska timezones  
churn$Timezone<-str\_replace(churn$Timezone, "America/Sitka", "US/Alaska")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Nome", "US/Alaska")  
churn$Timezone<-str\_replace(churn$Timezone, "America/Anchorage", "US/Alaska")

#US/Arizona timezones  
churn$Timezone<-str\_replace(churn$Timezone, "America/Phoenix", "US/Arizona")

#US/Hawaii timezones  
churn$Timezone<-str\_replace(churn$Timezone, "Pacific/Honolulu", "US/Hawaii")

#US/Puerto\_Rico timezones  
churn$Timezone<-str\_replace(churn$Timezone, "America/Puerto\_Rico", "US/Puerto\_Rico")

\*\*\*See code attached.

**D5: Clean Data**

A copy of the cleaned dataset is attached to this as cfor224\_churn\_clean\_data.

**D6: Limitations**

Using the mean or median for missing values can be misleading because it’s making assumptions about the dataset. In a perfect world, there would be no missing values, and the best treatment must be chosen based on information available. For this example, there might be one missing value out of 52 possible categories. It would reduce the sample size if you threw out all the data for each category with one missing value. An informed estimation is a good choice, but it’s still an estimation and not as accurate as actual data would be.

**D7: Impact of Limitations**

The impact of using my cleaned dataset for analysis could be skewed as far as quantitative values are concerned because I did use imputation to fill in missing blanks. While I tried to keep the data as close to the original when possible, it’s difficult to assume a value such as Income for someone based on the mean of all the other incomes in the dataset. Income varies widely and is based on so many different things that this dataset isn’t covering.

**E1: Principal Components**

The variables used in this Principal Component Analysis (PCA) were Lat (latitude), Lng (longitude), Population, Income, Outage\_sec\_perweek, Tenure, MonthlyCharge, and Bandwidth\_GB\_Year. To perform a PCA, you need quanitifiable numberic data that is continuous. This PCA includes all of the variables that fit that description in this dataset.

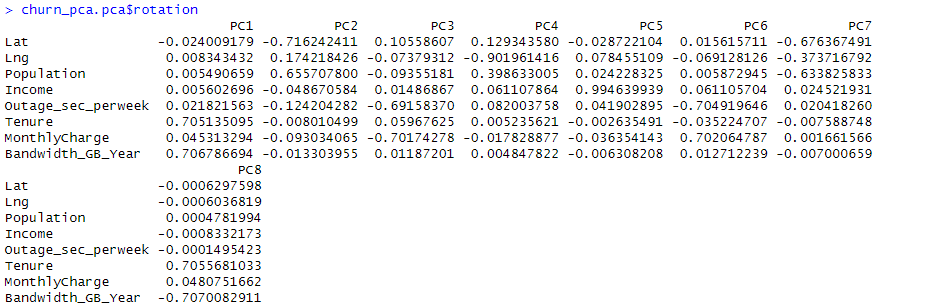
Here is the code I used to perform the PCA on the churn dataset.

#Import packages  
library(plyr)  
library(factoextra)

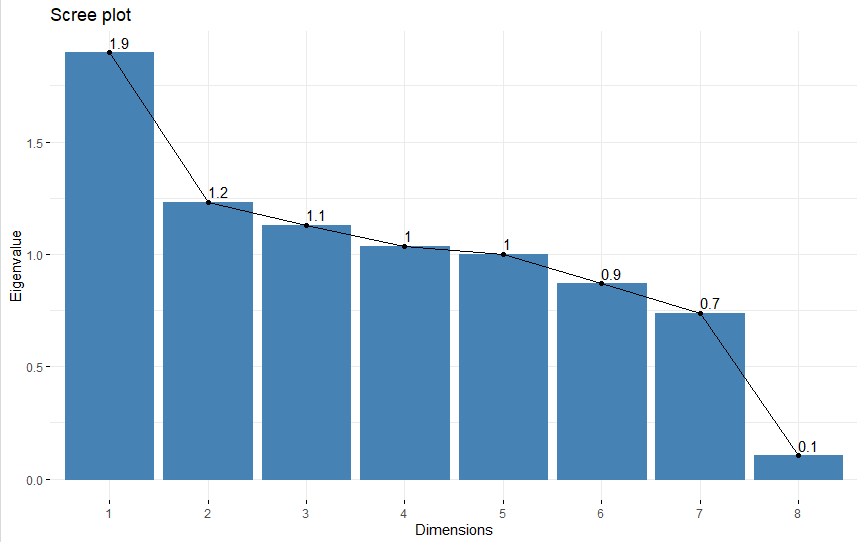
#Select columns of the dataframe and perform PCA  
churn\_pca.pca<-prcomp(churn[,c(9,10,11,19,23,42,43,44)], center=TRUE, scale=TRUE)  
churn\_pca.pca$rotation

#Create scree plot using eigen values  
fviz\_eig(churn\_pca.pca, choice="eigenvalue", addlabels=TRUE)

The Loading Matrix is shown here.



**E2: Criteria Used**

After determining the loading matrix, I created a scree plot using the fviz\_eig() function. Using the scree plot, we can use the Kaiser Rule to determine which PCs should be retained in our analysis. Any PCs with an Eigen Value of greater than one should be retained. For this dataset, there are five: PC1, PC2, PC3, PC4, and PC5.

**E3: Benefits of PCA**

Using PCA on this dataset helps us to break down the data into fewer dimensions and look at the factors that are similar to infer connections. For example, PC1 is showing a strong likeness between Tenure and Bandwidth\_GB\_Year. This is intuitive as it is telling us that when a customer has been using the service longer, they are using more of the service. However, PC3 is implying a strong likeness between Outage\_sec\_perweek and MonthlyCharge. This is something I would want to look into further. Is there a price difference between differing locations? Are we overcharging in an area where the equipment is less reliable? Or is it a case of a customer purchasing more services and having a higher number of outages simply because they have more services to be reported on?

Because we have these analyses, we can look deeper into our data and find out where adjustments need to be made. Maybe the equipment in certain areas is overdue for maintenance.

**F: Panopto Recording**  
See attached video.

**G: Code References**  
All code in this project was taken from the Lecture Series “Getting Started with D206” by Dr. Middleton on WGU’s website.

**H: Source References**

No references were used.