Data Cleaning D206

Nicole Reiswig

College of Information Technology, Western Governors University

Keiona Middleton

August 23, 2023

Data Cleaning D206

Table of Contents

A) Research Question 3

B) Data Set Variables 3

C) Data Cleaning Plan (Detection) 5

D) Data Cleaning (Treatment) 12

E) Principal Component Analysis 25

F) Presentation 28

G) References 28

Research Question

A. The research question we will address is, “What factors affect the average length of customer tenure?”

B. The data set utilized to answer the research question is the churn data set. It is comprised of 50 columns and 10,000 customers. The variables contained in the data set are:

Variable	Data Type	Description	Example
Case order	Qualitative	Placeholder variable	1
Customer ID	Qualitative	Unique customer id	K409198
Interaction uid	Qualitative	Transaction id	Aa90260b-4141…
City	Qualitative	Customer residence	Point Baker
State	Qualitative	Customer residence	Arkansas
County	Qualitative	Customer residence	Cook
Zip	Qualitative	Customer residence	99362
Lat	Qualitative	GPS coordinate	56.251
Lng	Qualitative	GPS coordinate	-133.376
Population	Quantitative	Population within 1mile of customer	38
Area	Qualitative	Area type	Rural
Timezone	Qualitative	Customer timezone	America/Chicago
Job	Qualitative	Customer job	Engineer
Children	Quantitative	# of children in household	2
Age	Quantitative	Age of customer	30
Education	Qualitative	Customer education	Masters
Employment	Qualitative	Customer employment	Employed
Income	Quantitative	Income of customer	28561
Marital	Qualitative	Customer marital status	Single
Gender	Qualitative	Gender of customer	Female
Churn	Qualitative	If customer d/c service	Yes
Outage sec per week	Quantitative	Avg # of outage seconds a week	6.9
Email	Quantitative	# of emails sent to customer in week	5

Contacts	Quantitative	# of times customer contacted tech support	0
Yearly equip failure	Quantitative	# of customer equipment failures a year	0
Techie	Qualitative	Customer is technically inclined	Yes
Contract	Qualitative	Contract terms	One year
Port modem	Qualitative	Customer has port modem	Yes
Tablet	Qualitative	Customer has tablet	Yes
Internet service	Qualitative	Customers internet service provider	Fiber optics
Phone	Qualitative	Customer has phone service	Yes
Multiple	Qualitative	Customer has multiple lines	Yes
Online security	Qualitative	Customer has add on security	Yes
Online backup	Qualitative	Customer has add on backup	Yes
Device protection	Qualitative	Customer has device protection	yes
Tech support	Qualitative	Customer has add on tech support	Yes
Streaming tv	Qualitative	Customer has streaming tv	Yes
Streaming movies	Qualitative	Customer has streaming movies	Yes
Paperless billing	Qualitative	Customer has paperless billing	Yes
Payment method	Qualitative	Customers payment method	Check
Tenure	Quantitative	# of months customer has been with provider	12
Monthly charge	Quantitative	Amount of monthly charge	171.45
Bandwidth gb year	Quantitative	Amount of gb used per year	904
Timely response	Quantitative	Survey response rating importance of timely response	1

Timely fixes	Quantitative	Survey response rating importance of timely fixes	2
Timely replacements	Quantitative	Survey response rating importance of timely replacements	3
Reliability	Quantitative	Survey response rating importance of reliability	4
Options	Quantitative	Survey response rating importance of having options	5
Respectful response	Quantitative	Survey response rating importance of respectful response	6
Courteous exchange	Quantitative	Survey response rating importance of courteous exchange	7
Evidence of active listening	Quantitative	Survey response rating importance of active listening	8

Data Cleaning Plan

C.

1. The plan to clean the data set will be to utilize R Studio. Functions utilized to import file are df<-read_csv("C:/….), getwd(), setwd("C:/…..) and df<-read_csv("C:/…..). The function utilized for data profiling is str(df). Functions that will be utilized to detect duplicates include str(df), duplicated(df), sum(duplicated(df)), df <- distinct(df) sum(duplicated(df)). The function utilized to detect missing values is the function colSums(is.na(df)), sum(is.na(df$col1)), and which(is.na(df$col1)). The function utilized to visualize the data is, library(visdat) vis_miss(df). To detect outliers the function that will be utilized is geom_boxplot()

geom of ggplot 2, b <- boxplot(df$columnname). The function utilized to visualize the distribution of the data is hist(df$columnname). Outliers will be detected for all quantitative variables.

2. These functions were utilized in the detection of the data quality issues because they were the most efficient way to accomplish the detection of missing values, duplicates, and outliers utilizing the chosen programming language R with the libraries and packages dplyr, tidyr, stringr, and ggplot2. For example, when utilizing R studio to calculate Z scores, no additional download for libraries and packages is needed as it is included with the installation of R. However with Python to calculate Z scores we must also import numpy as np, import pandas as pd, from pandas import data frame, and import scipy.stats as stats.

3. The programming language of choice for this project is R Studio because according to Western Governors University (2023), "R is an open-source programming language and environment for statistical computing and graphics. R provides a wide variety of statistical(linear and nonlinear modeling, classical statistical tests, time series analysis, classification, and clustering) and graphical techniques, and is highly extensible… Statistical models can be written with only a few lines of code- very efficient!" For example, when utilizing R studio to calculate Z scores, no additional download for libraries and packages is needed as it is included with the installation of R. However with Python to calculate Z scores we must also import numpy as np, import pandas as pd, from pandas import data frame, and import scipy.stats as stats.

4. Detection code is as follows:

```
#checking working directory
```

```
getwd()

#data profiling

str("~/MSDA/churn_raw_data.csv")

str("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_raw_data.csv")

#detect duplicates [in-text citation: (Middleton, n.d.)]

duplicated("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_raw_data.csv")

#sum of duplicated rows [in-text citation: (Middleton, n.d.)]

sum(duplicated("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_raw_data.csv"))

#detect missing values [in-text citation: (Middleton, n.d.)]

colSums(is.na(churn_raw_data))

#visualize missing data [in-text citation: (Middleton, n.d.)]

library(visdat)

vis_miss(churn_raw_data)

#histogram of columns with missing data [in-text citation: (Middleton, n.d.)]

hist(churn_raw_data$Children)

hist(churn_raw_data$Age)

hist(churn_raw_data$Income)

hist(churn_raw_data$Bandwidth_GB_Year)

hist(churn_raw_data$Tenure)

#View data

churn_raw_data

#detect outliers [in-text citation: (Middleton, n.d.)]

b <- boxplot(churn_raw_data$CaseOrder)

b <- boxplot(churn_raw_data$Zip)
```

```
b <- boxplot(churn_raw_data$Lat)

#count and range of Lat [in-text citation: (Middleton, n.d.)]

lat_query <- churn_raw_data[which(churn_raw_data$Lat < 25), ]

str(lat_query)

lat_query2 <- churn_raw_data[which(churn_raw_data$Lat > 50), ]

str(lat_query2)

#detect outliers [in-text citation: (Middleton, n.d.)]

b <- boxplot(churn_raw_data$Lng)

#count and range of Lng [in-text citation: (Middleton, n.d.)]

lng_query <- churn_raw_data[which(churn_raw_data$Lng < -120), ]

str(lng_query)

#detect outliers [in-text citation: (Middleton, n.d.)]

b <- boxplot(churn_raw_data$Population)

#count and range of Population [in-text citation: (Middleton, n.d.)]

pop_query <- churn_raw_data[which(churn_raw_data$Population > 2e+04), ]

str(pop_query)

#detect outliers [in-text citation: (Middleton, n.d.)]

b <- boxplot(churn_raw_data$Children)

#count and range of Children [in-text citation: (Middleton, n.d.)]

children_query <- churn_raw_data[which(churn_raw_data$Children > 7), ]

str(children_query)

#detect outliers [in-text citation: (Middleton, n.d.)]

b <- boxplot(churn_raw_data$Age)

b <- boxplot(churn_raw_data$Income)
```

```
#count and range of Income [in-text citation: (Middleton, n.d.)]
income_query <- churn_raw_data[which(churn_raw_data$Income > 100000), ]
str(income_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$Outage_sec_perweek)
#count and range of outage_sec_week [in-text citation: (Middleton, n.d.)]
osw_query <- churn_raw_data[which(churn_raw_data$Outage_sec_perweek > 20), ]
str(osw_query)
osw2_query <- churn_raw_data[which(churn_raw_data$Outage_sec_perweek < 0), ]
str(osw2_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$Email)
#count and range of email [in-text citation: (Middleton, n.d.)]
email_query <- churn_raw_data[which(churn_raw_data$Email > 20), ]
str(contacts_query)
email2_query <- churn_raw_data[which(churn_raw_data$Email < 4), ]
str(contacts_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$Contacts)
#count and range of contacts [in-text citation: (Middleton, n.d.)]
contacts_query <- churn_raw_data[which(churn_raw_data$Contacts > 5), ]
str(contacts_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$Yearly_equip_failure)
```

```
#count and range of yearly_equip_failures [in-text citation: (Middleton, n.d.)]
yef_query <- churn_raw_data[which(churn_raw_data$Yearly_equip_failure > 2), ]
str(yef_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$Tenure)
b <- boxplot(churn_raw_data$MonthlyCharge)
#count and range of monthly charge [in-text citation: (Middleton, n.d.)]
mc_query <- churn_raw_data[which(churn_raw_data$MonthlyCharge > 300), ]
str(mc_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$Bandwidth_GB_Year)
b <- boxplot(churn_raw_data$item1)
#count and range of item1 [in-text citation: (Middleton, n.d.)]
item1_query <- churn_raw_data[which(churn_raw_data$item1 > 5), ]
str(item1_query)
item1.1_query <- churn_raw_data[which(churn_raw_data$item1 < 2), ]
str(item1.1_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$item2)
#count and range of item2 [in-text citation: (Middleton, n.d.)]
item2_query <- churn_raw_data[which(churn_raw_data$item2 > 5), ]
str(item1_query)
item2.1_query <- churn_raw_data[which(churn_raw_data$item2 < 2), ]
str(item2.1_query)
```

```
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$item3)
#count and range of item3 [in-text citation: (Middleton, n.d.)]
item3_query <- churn_raw_data[which(churn_raw_data$item3 > 5), ]
str(item3_query)
item3.1_query <- churn_raw_data[which(churn_raw_data$item3 < 2), ]
str(item3.1_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$item4)
#count and range of item4 [in-text citation: (Middleton, n.d.)]
item4_query <- churn_raw_data[which(churn_raw_data$item4 > 5), ]
str(item4_query)
item4.1_query <- churn_raw_data[which(churn_raw_data$item4 < 2), ]
str(item4.1_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$item5)
#count and range of item5 [in-text citation: (Middleton, n.d.)]
item5_query <- churn_raw_data[which(churn_raw_data$item5 > 5), ]
str(item5_query)
item5.1_query <- churn_raw_data[which(churn_raw_data$item5 < 2), ]
str(item5.1_query)
#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$item6)
#count and range of item6 [in-text citation: (Middleton, n.d.)]
```

```
item6_query <- churn_raw_data[which(churn_raw_data$item6 > 5), ]

str(item6_query)

item6.1_query <- churn_raw_data[which(churn_raw_data$item6 < 2), ]

str(item6.1_query)

#detect outliers [in-text citation: (Middleton, n.d.)]

b <- boxplot(churn_raw_data$item7)

#count and range of item7 [in-text citation: (Middleton, n.d.)]

item7_query <- churn_raw_data[which(churn_raw_data$item7 > 5), ]

str(item7_query)

item7.1_query <- churn_raw_data[which(churn_raw_data$item7 < 2), ]

str(item7.1_query)

#detect outliers [in-text citation: (Middleton, n.d.)]

b <- boxplot(churn_raw_data$item8)

#count and range of item8 [in-text citation: (Middleton, n.d.)]

item8_query <- churn_raw_data[which(churn_raw_data$item8 > 5), ]

str(item8_query)

item8.1_query <- churn_raw_data[which(churn_raw_data$item8 < 2), ]

str(item8.1_query)
```

(see code attached).

Data Cleaning

D.

1. In detection zero duplicates were found.

In detection, it was found that there were 2.7% overall missing data values in the data set. 97.3% of data was present in the churn data set. The variables with missing data were children,

age, income, techie, phone, tech support, bandwidth, and tenure. The percentage of missing values was between 9% and 25% per variable. Tenure had 9% missing values. Children, age, income, and techie had 25% missing values. Phone, tech support, and bandwidth had 10% missing values.

In detection, it was found that there were outliers in several variables which included, children, income, outage_sec_perweek, contacts, email, yearly_equip_failures, monthly charge, bandwidth, timely response, timely fixes, timely replacements, reliability, options, respectful response, courteous exchange and evidence of active listening. Quantitative variables that had zero outliers were tenure, bandwidth, and age. The monthly charge variable had three outliers with values of 306, 308, and 316. The children variable had 302 observations with the values between 8 and 10. The income variable had 299 observations with values from 100029- 258900. The outage_sec_week had 503 with values from 20- 47 and 11 with values from -.14 to -1.3. The contacts variable had 8 observations with values of 6 and 7. The email variable had 4 observations with a value of 21 and 1 observation of 1. The variable yearly_equip_failures had 94 observations with values from 3-6. The timely response variable had 218 observations with ranges from 6-7 and 224 observations with a value of 1. The timely fixes variable had 228 observations with ranges from 6-7 and 217 observations with a value of 1. The timely replacement variable had 216 observations with ranges from 6-7 and 202 observations with a value of 1. The reliability variable had 212 observations with ranges from 6-7 and 221 observations with a value of 1. The options variable had 216 observations with ranges from 6-7 and 206 observations with a value of 1. The respectful response variable had 223 observations with ranges from 6-7 and 190 observations with a value of 1. The courteous exchange variable had 235 observations with ranges from 6-7 and 219 observations with a value of 1. The evidence

of active listening variable had 220 observations with ranges from 6-7 and 206 observations with a value of 1.

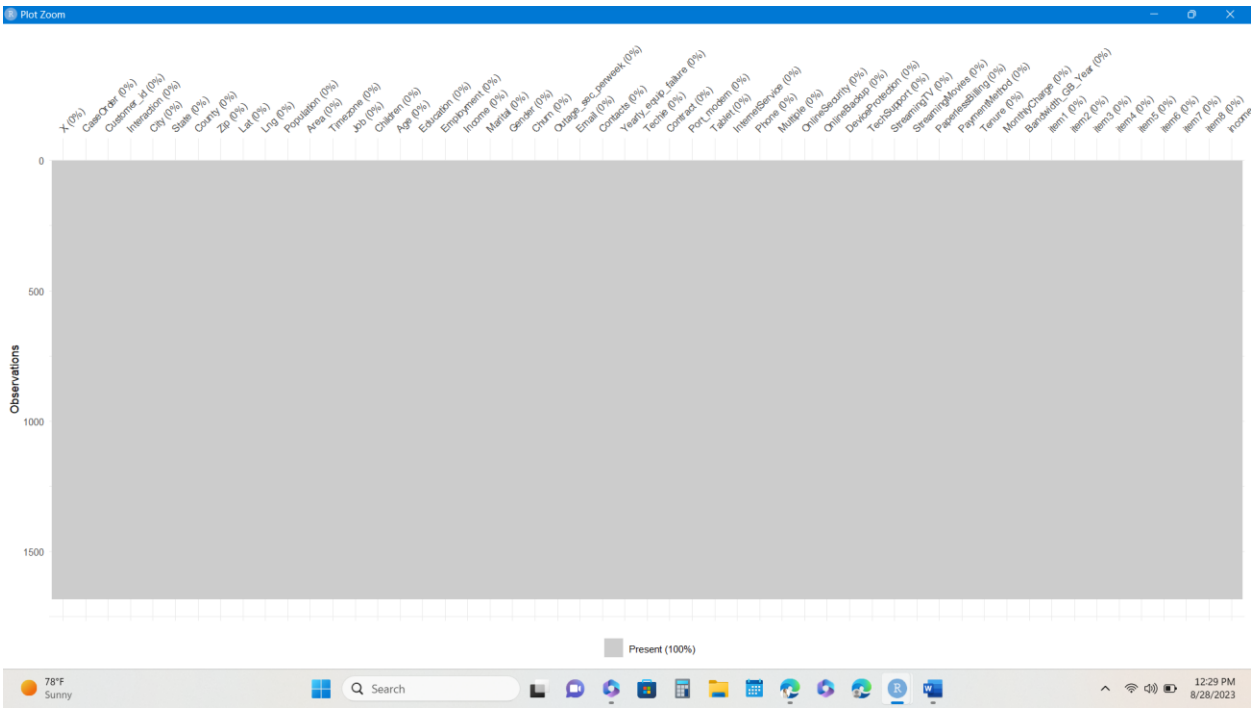
2. No action was taken to treat duplicates due to there being no duplicates detected.

To treat the missing values in the data set imputation was utilized. Imputation was utilized because no variable had missing values over 60% or under 5% to justify deletion. The children's variable was imputed with the median due to the distribution being positively skewed to the right. The age and income variables were imputed utilizing the mean due to the uniform distribution. The phone and tech support variables were imputed utilizing the mode due to being non-numeric. The tenure and bandwidth variables were imputed utilizing the median due to having bi-modal distributions.

To treat the outliers in the data set I utilized retain, impute, and exclude. Lat and Lng were retained as they are not quantitative data and the values are to be expected. Population and zip code were retained to preserve the sample size. Income was imputed with the median. Outage_sec_perweek was both imputed and excluded. There were values less than 0 which were imputed with the median because it is not possible to have less than 0 outage seconds per week. The values that were larger than 20 seconds were excluded because outages greater than 20 seconds we would want to investigate the cause to correct separately. The contacts variable that had greater than 5 contacts was excluded for further analysis and the outliers were minimal. Yearly_equip_failures were imputed with the median. The monthly charge was excluded for further analysis. The 8 survey question variables were imputed with the mean. Age and income were imputed with the mean due to the histogram results of distribution. Children, tenure, and bandwidth were imputed with the median due to the distribution of the histogram results. Phone, techie, and tech support were imputed with the mode due to being non-numerical values. Once

imputation, retention, and exclusion had been run on all quantitative variables detection utilizing histogram and boxplots were utilized to check for remaining outliers or distribution issues. Population, children, and income had additional outliers. Income was imputed again with the mean due to the distribution. Population, children, and income are then retained to preserve the sample size and quality of data.

3. To summarize all the work that has been performed, the data set has been cleaned. I utilized R Studio to detect and treat missing data, duplicates, and outliers. There are no missing values. There are no duplicates. All outliers have been treated utilizing retention, imputation, exclusion, or removal as described above. See the attached screenshots of the missing data, outliers, and duplicates being successfully treated.




```
> vis_miss(churn_raw_data)
> #re-expression of categorical variables not necessary due to consistent yes/no
> #detect duplicates
> duplicated("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_raw_data.csv")
[1] FALSE
> #sum of duplicated rows
> sum(duplicated("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_raw_data.csv"))
[1] 0
> 8
```

4. See the code below utilized to treat the data quality issues.

```
#Retain Lat, Lng outliers expected
#Retain population to preserve sample size
#Impute outliers in income [in-text citation: (Middleton, n.d.)]
churn_raw_data$Income[churn_raw_data$Income > 100000] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$income[is.na(churn_raw_data$Income)] <- median(churn_raw_data$income,na.rm
=TRUE)
colSums(is.na(churn_raw_data))
#Impute outliers in outage_sec_perweek < 0 [in-text citation: (Middleton, n.d.)]
churn_raw_data$Outage_sec_perweek[churn_raw_data$Outage_sec_perweek < 0] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$Outage_sec_perweek[is.na(churn_raw_data$Outage_sec_perweek)]<-
median(churn_raw_data$Outage_sec_perweek,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Exclude outliers in outage_sec_perweek > 20 [in-text citation: (Middleton, n.d.)]
Outliers_outage_sec_perweek<-churn_raw_data[which(churn_raw_data$Outage_sec_perweek > 20),]
str(Outliers_outage_sec_perweek)
churn_raw_data<-churn_raw_data[!(churn_raw_data$Outage_sec_perweek>20),]
str(churn_raw_data)
#Exclude contacts > 5 [in-text citation: (Middleton, n.d.)]
Outliers_contacts<-churn_raw_data[which(churn_raw_data$Contacts > 5),]
str(Outliers_contacts)
churn_raw_data<-churn_raw_data[!(churn_raw_data$Contacts > 5),]
str(churn_raw_data)
#Impute yearly_equip_failures [in-text citation: (Middleton, n.d.)]
churn_raw_data$Yearly_equip_failure[churn_raw_data$Yearly_equip_failure > 2] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$Yearly_equip_failure[is.na(churn_raw_data$Yearly_equip_failure)]<-
median(churn_raw_data$Yearly_equip_failure,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Exclude monthly charge > 300 [in-text citation: (Middleton, n.d.)]
Outliers_monthlycharge<-churn_raw_data[which(churn_raw_data$MonthlyCharge > 300),]
str(Outliers_monthlycharge)
churn_raw_data<-churn_raw_data[!(churn_raw_data$MonthlyCharge > 300),]
str(churn_raw_data)
#histogram of columns with outliers [in-text citation: (Middleton, n.d.)]
hist(churn_raw_data$item1)
hist(churn_raw_data$item2)
hist(churn_raw_data$item3)
hist(churn_raw_data$item4)
hist(churn_raw_data$item5)
hist(churn_raw_data$item6)
hist(churn_raw_data$item7)
hist(churn_raw_data$item8)
#Impute item 1 using mean due to normal distribution [in-text citation: (Middleton, n.d.)]
churn_raw_data$item1[churn_raw_data$item1 > 5] <- NA
colSums(is.na(churn_raw_data))
```

```
churn_raw_data$item1[is.na(churn_raw_data$item1)]<-mean(churn_raw_data$item1,na.rm =TRUE)
colSums(is.na(churn_raw_data))
churn_raw_data$item1[churn_raw_data$item1 < 2] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item1[is.na(churn_raw_data$item1)]<-mean(churn_raw_data$item1,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Impute item 2 using mean due to normal distribution [in-text citation: (Middleton, n.d.)]
churn_raw_data$item2[churn_raw_data$item2 > 5] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item2[is.na(churn_raw_data$item2)]<-mean(churn_raw_data$item2,na.rm =TRUE)
colSums(is.na(churn_raw_data))
churn_raw_data$item2[churn_raw_data$item2 < 2] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item2[is.na(churn_raw_data$item2)]<-mean(churn_raw_data$item2,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Impute item 3 using mean due to normal distribution [in-text citation: (Middleton, n.d.)]
churn_raw_data$item3[churn_raw_data$item3 > 5] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item3[is.na(churn_raw_data$item3)]<-mean(churn_raw_data$item3,na.rm =TRUE)
colSums(is.na(churn_raw_data))
churn_raw_data$item3[churn_raw_data$item3 < 2] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item3[is.na(churn_raw_data$item3)]<-mean(churn_raw_data$item3,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Impute item 4 using mean due to normal distribution [in-text citation: (Middleton, n.d.)]
churn_raw_data$item4[churn_raw_data$item4 > 5] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item4[is.na(churn_raw_data$item4)]<-mean(churn_raw_data$item4,na.rm =TRUE)
colSums(is.na(churn_raw_data))
churn_raw_data$item4[churn_raw_data$item4 < 2] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item4[is.na(churn_raw_data$item4)]<-mean(churn_raw_data$item4,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Impute item 5 using mean due to normal distribution [in-text citation: (Middleton, n.d.)]
churn_raw_data$item5[churn_raw_data$item5 > 5] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item5[is.na(churn_raw_data$item5)]<-mean(churn_raw_data$item5,na.rm =TRUE)
colSums(is.na(churn_raw_data))
churn_raw_data$item5[churn_raw_data$item5 < 2] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item5[is.na(churn_raw_data$item5)]<-mean(churn_raw_data$item5,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Impute item 6 using mean due to normal distribution [in-text citation: (Middleton, n.d.)]
churn_raw_data$item6[churn_raw_data$item6 > 5] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item6[is.na(churn_raw_data$item6)]<-mean(churn_raw_data$item6,na.rm =TRUE)
colSums(is.na(churn_raw_data))
churn_raw_data$item6[churn_raw_data$item6 < 2] <- NA
```

```
colSums(is.na(churn_raw_data))
churn_raw_data$item6[is.na(churn_raw_data$item6)]<-mean(churn_raw_data$item6,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Impute item 7 using mean due to normal distribution [in-text citation: (Middleton, n.d.)]
churn_raw_data$item7[churn_raw_data$item7 > 5] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item7[is.na(churn_raw_data$item7)]<-mean(churn_raw_data$item7,na.rm =TRUE)
colSums(is.na(churn_raw_data))
churn_raw_data$item7[churn_raw_data$item7 < 2] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item7[is.na(churn_raw_data$item7)]<-mean(churn_raw_data$item7,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Impute item 8 using mean due to normal distribution [in-text citation: (Middleton, n.d.)]
churn_raw_data$item8[churn_raw_data$item8 > 5] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item8[is.na(churn_raw_data$item8)]<-mean(churn_raw_data$item8,na.rm =TRUE)
colSums(is.na(churn_raw_data))
churn_raw_data$item8[churn_raw_data$item8 < 2] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$item8[is.na(churn_raw_data$item8)]<-mean(churn_raw_data$item8,na.rm =TRUE)
colSums(is.na(churn_raw_data))
#Find number of NAs in a single column [in-text citation: (Middleton, n.d.)]
sum(is.na(churn_raw_data$Children))
#Find rows that have NA [in-text citation: (Middleton, n.d.)]
which(is.na(churn_raw_data$Children))
#Find how many NAs are in each column [in-text citation: (Middleton, n.d.)]
colSums(is.na(churn_raw_data))
#histogram for missing data  [in-text citation: (Middleton, n.d.)]
hist(churn_raw_data$Children)
hist(churn_raw_data$Age)
hist(churn_raw_data$Income)
hist(churn_raw_data$Tenure)
hist(churn_raw_data$Bandwidth_GB_Year)
#impute age with mean [in-text citation: (Middleton, n.d.)]
churn_raw_data$Age[is.na(churn_raw_data$Age)] <- mean(churn_raw_data$Age,na.rm = TRUE)
colSums(is.na(churn_raw_data))
#impute income with mean [in-text citation: (Middleton, n.d.)]
churn_raw_data$Income[is.na(churn_raw_data$Income)] <- mean(churn_raw_data$Income,na.rm =
TRUE)
colSums(is.na(churn_raw_data))
#impute children with median [in-text citation: (Middleton, n.d.)]
churn_raw_data$Children[is.na(churn_raw_data$Children)] <- median(churn_raw_data$Children,na.rm
= TRUE)
colSums(is.na(churn_raw_data))
#impute tenure with median [in-text citation: (Middleton, n.d.)]
churn_raw_data$Tenure[is.na(churn_raw_data$Tenure)] <- median(churn_raw_data$Tenure,na.rm =
TRUE)
colSums(is.na(churn_raw_data))
```

```
#impute bandwidth with median [in-text citation: (Middleton, n.d.)]
churn_raw_data$Bandwidth_GB_Year[is.na(churn_raw_data$Bandwidth_GB_Year)] <-
median(churn_raw_data$Bandwidth_GB_Year,na.rm = TRUE)
colSums(is.na(churn_raw_data))
#impute phone with mode [in-text citation: (Middleton, n.d.)]
churn_raw_data$Phone[is.na(churn_raw_data$Phone)] <-
(names(which.max(table(churn_raw_data$Phone))))
colSums(is.na(churn_raw_data))
#impute techie with mode [in-text citation: (Middleton, n.d.)]
churn_raw_data$Techie[is.na(churn_raw_data$Techie)] <-
(names(which.max(table(churn_raw_data$Techie))))
colSums(is.na(churn_raw_data))
#impute tech support with mode [in-text citation: (Middleton, n.d.)]
churn_raw_data$TechSupport[is.na(churn_raw_data$TechSupport)] <-
(names(which.max(table(churn_raw_data$TechSupport))))
colSums(is.na(churn_raw_data))
#visualize missing data  [in-text citation: (Middleton, n.d.)]

library(visdat)

vis_miss(churn_raw_data)

#detect outliers [in-text citation: (Middleton, n.d.)]
b <- boxplot(churn_raw_data$Population)
b <- boxplot(churn_raw_data$Children)
b <- boxplot(churn_raw_data$Income)
#histogram for outliers  [in-text citation: (Middleton, n.d.)]
hist(churn_raw_data$Population)
hist(churn_raw_data$Children)
hist(churn_raw_data$Income)
#impute income with mean [in-text citation: (Middleton, n.d.)]
churn_raw_data$Income[churn_raw_data$Income > 100000] <- NA
colSums(is.na(churn_raw_data))
churn_raw_data$Income[is.na(churn_raw_data$Income)] <- mean(churn_raw_data$Income,na.rm =
TRUE)
colSums(is.na(churn_raw_data))
#round age to integer [in-text citation: (Middleton, n.d.)]
churn_raw_data$Age <- round(churn_raw_data$Age)
#round item 1-8 to integer [in-text citation: (Middleton, n.d.)]
churn_raw_data$item1 <- round(churn_raw_data$item1)
churn_raw_data$item2 <- round(churn_raw_data$item2)
churn_raw_data$item3 <- round(churn_raw_data$item3)
churn_raw_data$item4 <- round(churn_raw_data$item4)
churn_raw_data$item5 <- round(churn_raw_data$item5)
churn_raw_data$item6 <- round(churn_raw_data$item6)
churn_raw_data$item7 <- round(churn_raw_data$item7)
churn_raw_data$item8 <- round(churn_raw_data$item8)
#retain zip to preserve sample size  [in-text citation: (Middleton, n.d.)]
```

```
#re-expression of categorical variables not necessary due to consistent yes/no
#detect duplicates [in-text citation: (Middleton, n.d.)]
duplicated("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_raw_data.csv")
#sum of duplicated rows [in-text citation: (Middleton, n.d.)]
sum(duplicated("C:/Users/ntrei/OneDrive/Documents/MSDA/churn_raw_data.csv"))
#export cleaned data to csv [in-text citation: (Middleton, n.d.)]
write.csv(churn_raw_data, "C:/Users/ntrei/OneDrive/Documents/MSDA/cleaned.csv")
#export excluded data [in-text citation: (Middleton, n.d.)]
write.csv(Outliers_outage_sec_perweek,
"C:/Users/ntrei/OneDrive/Documents/MSDA/excluded_outage.csv")
write.csv(Outliers_contacts, "C:/Users/ntrei/OneDrive/Documents/MSDA/excluded_contacts.csv")
write.csv(Outliers_monthlycharge,
"C:/Users/ntrei/OneDrive/Documents/MSDA/excluded_charge.csv")
```

See the code attached.

5. See the attached CSV files that include the cleaned data and the excluded data in separate CSV files.

6. Disadvantages of the methods used to detect and treat missing values, duplicates, and outliers are as follows, excluding or removing some of the outlier data points reduces the sample size, when imputing missing integers with the median the result may not be a whole number and a second treatment will be required to have a clean data set. A disadvantage to utilizing boxplots and histograms in detecting data quality issues some issues may not be seen in qualitative data such as a zip code with 4 digits rather than 5.

7. Challenges a data analyst may encounter if they were to utilize this data for analysis is that the sample size has been reduced due to imputing some variables more than once. There were variables that had been imputed more than once and if I were to clean the data again, now knowing the result of treatments, I would have stopped after the first treatment to preserve the sample size and quality of the data. The data cleaning limitations may impact the pursuit of an answer to the research question because treating some variables more than once may have caused imperfections in the quality of the data. This means if this data were to be utilized to answer the research question it may be a biased result.

PCA

E.

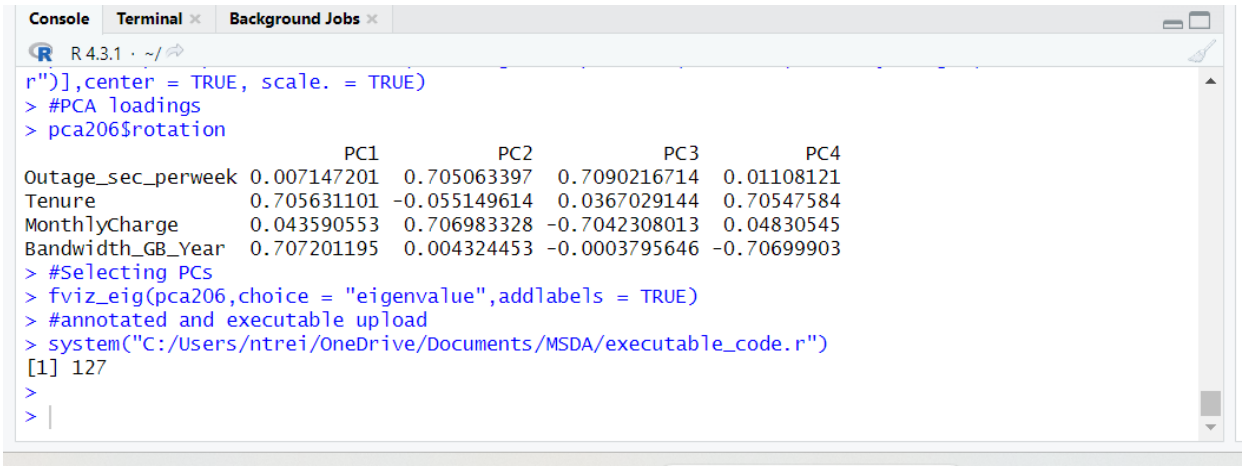
1. To perform a PCA on this data set I have chosen to utilize all quantitative and continuous data available. The variables that fit this criteria are outage_sec_perweek, tenure, monthlycharge, and bandwidth_gb_year. See the code utilized to perform the PCA, the output and scree plot results below.

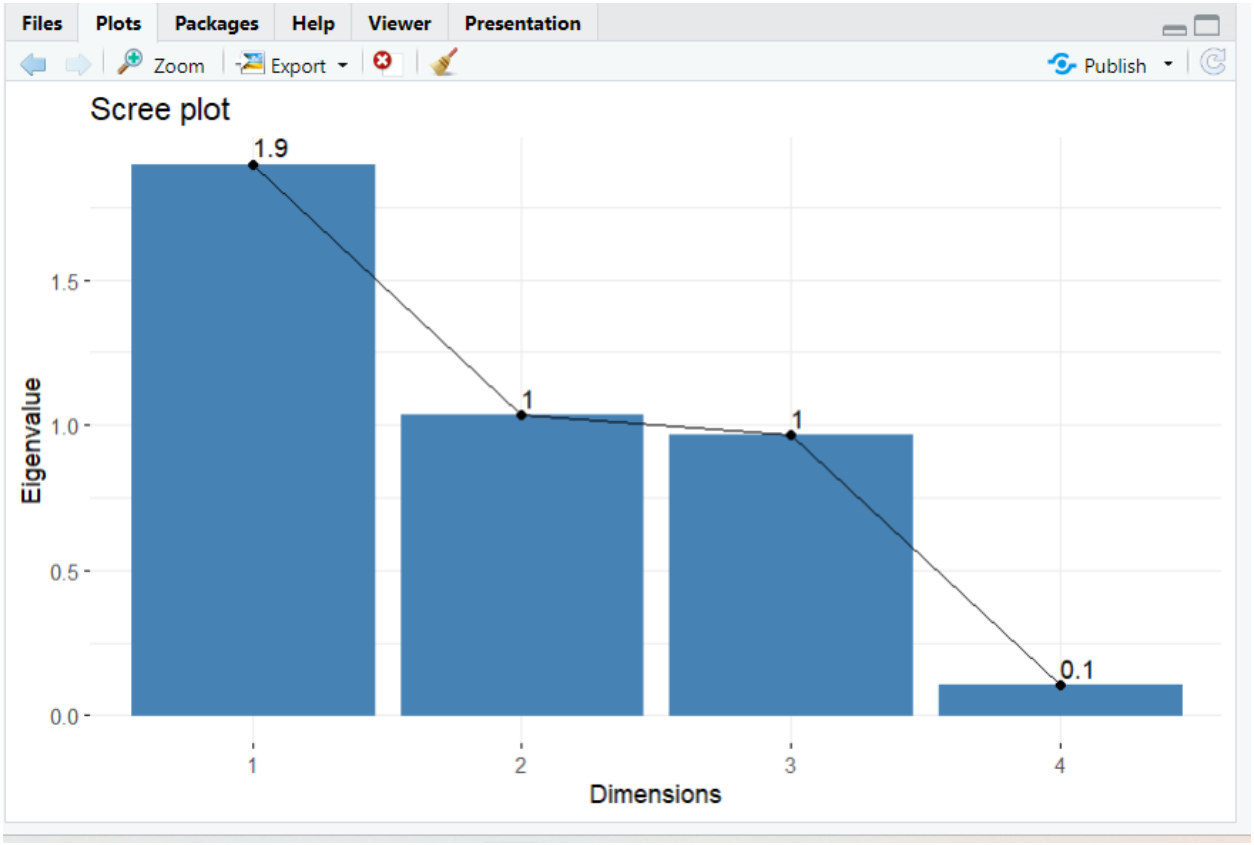
```
#install library for PCA factoextra
install.packages("factoextra")
library("factoextra")

#import cleaned.csv for pca

#new df for PCA
pca206<-select(cleaned, outage_sec_perweek,tenure, monthlycharge, bandwidth_gb_year)
pca206<-
prcomp(cleaned[,c("Outage_sec_perweek","Tenure","MonthlyCharge","Bandwidth_GB_Year")],center =
TRUE, scale. = TRUE)
#PCA loadings
pca206$rotation
#Selecting PCs
fviz_eig(pca206,choice = "eigenvalue",addlabels = TRUE)

#annotated and executable upload
system("C:/Users/ntrei/OneDrive/Documents/MSDA/executable_code.r")
```





2. The PCs that should be retained as a result of the PCA are PC1, PC2 and PC3 of the scree plot above because they are more than or equal to 1 eigenvalue. According to the Kaiser rule if a principal component is greater than or equal to 1 it needs to be retained. Therefore PC1, PC2 and PC3 are the most important.

3. According to Bigabid (nd), “PCA is a dimensionality reduction framework in machine learning.” The benefit of utilizing it is improved performance and it allows us to produce independent uncorrelated features of the data (Bigabid, nd). An organization can benefit from utilizing PCA in analysis because data can be analyzed in ulterior views not otherwise capable of. For example in the scree plot above PC1 includes data from all four variables that were input and there’s a correlation according to the Kaiser rule that demonstrates significance.

Supporting Documents

F. See attached Panopto recording.

G.

Code citation

Middleton, K. (n.d.). *Detecting and Treating Duplicates.* https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6eedfad4-240e-4c5c-8eab-b058003d3e6b

Middleton, K. (n.d.). *Detecting and Treating Missing Values.* https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=dc5f6cf5-a0cf-4e7d-b5ab-b053003703b2

Middleton, K. (n.d.). *Detecting and Treating Outliers.* https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=19c24c56-0f37-408e-bb1f-b059002a77ac

Middleton, K. (n.d.). *Reexpression of categorical variables.* D206 - Getting Started With D206 | Re-expression of Categorical Variables (panopto.com)

Zach. (2023). *Statology.*

R: How to Use drop_na to Drop Rows with Missing Values - Statology

H.

Content Citation

Bigabid. (n.d.). *What is PCA and how can I use it?*

What is Principal Component Analysis (PCA) & How to Use It? | Bigabid

Western Governors University. (2023). *R or Python?*

https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html

Western Governors University. (2020). *Welcome to data cleaning.*

https://cgp-oex.wgu.edu/courses/course-v1:WGUx+OEX0026+v02/courseware/1f468770545f494fa657b4dc0ed3762f/f8572a9b3754417b85680e23bb7de7e6/1?activate_block_id=block-v1%3AWGUx%2BOEX0026%2Bv02%2Btype%40vertical%2Bblock%407e75a31fbe604a2cb339b5d5ba9cebbe

Western Governors University. (n.d.). *Data files and associated dictionary files.*

https://lrps.wgu.edu/provision/354726028