D206 Data Cleaning Performance Assessment

Western Governors University

Kai Dumlao

D206

April 4, 2024

A. Research Question

Which telecommunication service factors contribute to customers churning?

B. Describe All Variables in Dataset

Variable Name	Data Type	Definition	Example
CaseOrder	Categorical	Index of the row	1
Customer_id	Categorical	Unique ID specific to the customer.	K409198
Interaction	Categorical	ID given to each transaction with technical support.	aa90260b-4141-4a24-8e36-b04ce1f4f77b
City	Categorical	City of customer residence.	Point Baker
State	Categorical	Abbreviation of State of customer residence.	AK
County	Categorical	County of customer residence.	Prince of Wales-Hyder
Zip	Categorical	Zip code of customer residence.	99927
Lat	Numeric	Latitude GPS coordinates of customer residence.	56.251
Lng	Numeric	Longitude of customer residence.	-133.37571
Population	Numeric	Number of people living within a mile of customer residence.	38
Area	Categorical	Type of environment the customer lives in based off of the census data.	Urban
Timezone	Categorical	Time zone the customer resides in.	America/Sitka

Job	Categorical	Occupation the customer reported to have.	Environmental Health Practitioner
Children	Numeric	Number of children the customer reported to have.	1
Age	Numeric	Age the customer signed up with.	68
Education	Categorical	The highest level of education the customer reported at time of sign up.	Master's Degree
Employment	Categorical	Type of employment the customer reported to work.	Part Time
Income	Numeric	Income the customer reported to make in a year.	28561.99
Marital	Categorical	Marital status of the customer.	Widowed
Gender	Categorical	Customer's identification of being Male, Female, or nonbinary	Male
Churn	Categorical	Whether or not the customer discontinued their service in the last month.	No
Outage_sec_perweek	Numeric	How many seconds per week on average that the service was down in the customer's neighborhood.	6.97256609
Email	Numeric	Number of Emails sent to the customer within the last year.	10
Contacts	Numeric	Number of times the customer contacted service support	0
Yearly_equip_failure	Numeric	Number of times the customer's equipment had to be replaced or reset in the past year.	1
Techie	Categorical	Answers if the customer thinks they are good with technology.	No
Contract	Categorical	Length of the customer's contract term.	One Year
Port_modem	Categorical	Does the customer have a portable modem?	Yes
Tablet	Categorical	Does the customer own a tablet?	Yes
InternetService	Categorical	Displays the internet service provider of the customer.	Fiber Optic
Phone	Categorical	Does the customer have a phone service?	Yes

Multiple	Categorical	Does the customer have multiple lines of service?	No
OnlineSecurity	Categorical	Does the customer have an add-on for online security?	Yes
OnlineBackup	Categorical	Does the customer have an add-on for online backup?	Yes
DeviceProtection	Categorical	Does the customer have an add-on for device protection?	No
TechSupport	Categorical	Does the customer have an add-on for technical support help?	No
StreamingTV	Categorical	Does the customer have a streaming TV?	Yes
StreamingMovies	Categorical	Does the customer have streaming movies?	Yes
PaperlessBilling	Categorical	Is the customer's billing paperless?	Yes
PaymentMethod	Categorical	Describes the payment method the customer uses.	Credit Card
Tenure	Numeric	How many months the customer has been with their provider.	6.79551295
MonthlyCharge	Numeric	The amount the customer is charged each month.	171.449762
Bandwidth_GB_Year	Numeric	Average number of GB of data used in a year.	904.5361102
Item1	Categorical	Rating from one to eight on the importance of a timely response.	5
Item2	Categorical	Rating from one to eight on the importance of timely fixes.	5
Item3	Categorical	Rating from one to eight on the importance of timely replacements.	5
Item4	Categorical	Rating from one to eight on the importance of reliability.	4
Item5	Categorical	Rating from one to eight on the importance of options.	4
Item6	Categorical	Rating from one to eight on the importance of a respectful response.	4
Item7	Categorical	Rating from one to eight on the importance of a courteous exchange.	3
Item8	Categorical	Rating from one to eight on the importance of evidence of active listening.	4

C1. Data Cleaning Detection

In order to detect the abnormalities in the data, I first call str() on the churn dataframe to get familiar with the structure and all variables within the dataset. I will use a combination of sum() and duplicated() to detect duplicates within the churn data set as well as n_distinct on specific columns like Interaction in order to make sure each interaction is not listed more than once. In order to spot missing values, the function sum(is.na()) will allow me to reveal any missing values in the data frame. In addition to duplicates and missing values, it is necessary to check for outliers. The scale() function is crucial for scaling the numeric values into z-scores, allowing me to look at the data through a standardized lens. In addition to standardizing the data, ggplot() and boxplot() will also make spotting the outliers evident. For re-expressing categorical variables, I can use view() on the entire dataframe to check if there are any variables which can be re-expressed ordinally and check for consistency.

C2. Justify functions

Ultimately, the functions I chose help in the simplest way to detect duplicates, missing values, outliers, and variables which need to be re-expressed. The sum(duplicated()) function was chosen as it adds up the entire amount of duplicate numbers in the data frame. If the number is any higher than 0, I would then have knowledge there are duplicate rows which need to be fixed. Inclusion of the n_distinct() function allows me to check certain columns which are supposed to be unique. For checking missing values, the sum(is.na()) function is best in order to count the number of NA values in the dataframe. Like duplicates, if the output is larger than 0, the missing values should be cleaned. In order to break down which columns specifically include the NA values, I use colSums(is.na()) to view the individual columns with missing values. The path for detecting outliers is through the scale() function as it converts the numeric

column values into z-scores. This function standardizes the values to allow the outliers to stand out. Outliers in this case, according to WGU Course Ware Materials (n.d.) are when the z-score is less than -3 or greater than 3. I am also choosing to use ggplot() and boxplot() to plot outliers on an easy to read graph of the outlier range. By applying view() on the entire churn dataframe, I can check all categorical variables and their values. For example, the categorical variables with binary values are consistent and need no changes. However, Education is clearly ordinal, therefore I can re-express the Education variable as ordinals since they have specific rankings.

C3. Language and Packages

R is the programming language I choose to clean the dataframe with. Since I have already completed projects in Python, switching to R for this data cleaning can help me gain experience in both languages. One factor in choosing R over Python is because the ease of use with R in the early stages of data analysis whereas Python is useful for API production in later stages, according to WGU Information Technology (n.d.). One example is standardizing data in R with the simple scale() function, as opposed to the many steps for standardizing the data in Python manually by combining multiple formulas. Scale completes all steps of standardization in one.

For R, there are a few libraries I call on to allow for use of specific functions. First of all, I need the readr library in order to import the csv file format into R Studio. Visdat is a package which grants access to visualizing missing data. The dyplr library gives access data manipulation scores such as the scale() function to standardize data. This library is useful for checking integrity of data by transforming the data shape into one where outliers are easily spotted. Another way to detect outliers comes from the ggplot2 library. The functions accompanied with

ggplot2 allow for creating visualizations such as histograms and boxplots. In addition, the plyr package allows for the revalue() function which grants access to ordinal encoding. Finally, the factoextra package allows for easy PCA. Factoextra allows for normalization and creation of scree plots essential for the PCA process.

C4. Detection Input Code

```
#Detection for duplicates
duplicated(churn_raw_data)
n_distinct(churn_raw_data$CaseOrder)
n_distinct(churn_raw_data$Customer_id)
n_distinct(churn_raw_data$Interaction)
sum(duplicated(churn_raw_data))

#Detection for missing values
is.na(churn_raw_data)
sum(is.na(churn_raw_data))
colSums(is.na(churn_raw_data))

#Detection for outliers. Obtain z-values for variable Lat
churn_raw_data$latz <- scale(churn_raw_data$Lat, center = TRUE, scale = TRUE)
#Subset outliers into separate df
lat_outliers <- churn_raw_data[which(churn_raw_data$latz < -3 | churn_raw_data$latz > 3),]
str(lat_outliers)
```

```
#Obtain outlier range from boxplot and query the range
boxplot(churn_raw_data$Lat)
lat_query <- churn_raw_data[which(churn_raw_data$Lat < 25 | churn_raw_data$Lat > 50), ]
select(lat_query, Customer_id, Lat)

#Taking a look at all the classes the categorical variable Education contains
unique(churn_raw_data$Education)
```

For complete input code for each individual variable , see code attached.

File name: Detection_D206.R

D1. Findings

Duplicates:

While running the functions to check for duplicates and ensuring the columns CaseOrder, Customer_id, and Interaction contained unique values, I found 0 duplicates, and the columns contained unique values. Therefore, there are no changes needed to be made in dealing with duplicate data.

Missing Values:

While searching for missing values, 13906 NA values were spotted in the entire dataframe. By narrowing the NA values by column, the variables containing NA values include: Children, Age, Income, Phone, Techie, TechSupport, Tenure, and Bandwidth_GB_Year.

Outiers:

When checking for outliers , each variable is individually observed as they each have their own criteria of outliers. Because outliers are sometimes natural, they will not all need to be imputed or taken away. The following table describes the variable, number of outliers detected, and the range the outliers fall into.

Variable Name	**Outlier Count**	**Outlier Range**
Lat	151	> 50 | < 25
Lng	102	> -65 | < -125
Population	219	> 40000 | < 0
Children	144	> 7 | < 0
Age	0	> 90 | <0
Income	110	>120000 | < 0
Outage_sec_perweek	491	> 20 | < 1
Email	12	> 21 | < 3
Contacts	165	>5 | < 0
Yearly_equip_failure	94	> 2 | < 0
Tenure	0	> 73 | < 0
MonthlyCharge	3	> 300 | < 60
Bandwidth_GB_Year	0	> 7200 | < 0

Re-expressing categorical data:

While viewing categorical data, all columns with binary values were consistent with "Yes/No" values. Because of this, the categorical data falling into these binary values do not need alterations. However, one column Education has clear ranking values. To make the rankings more easily visible in the data, I will change the strings to numeric values with ordinal encoding.

D2. Data Cleaning Treatment

Duplicates:

To begin with, while checking for duplicates, the output returned 0. As well as zero duplicates, I checked for unique values in the CaseOrder, Customer_id, and Interaction. All of

these values are unique. Since the variables checked are unique, and there are no duplicates, there is no need to fix any duplicates as there are none.

Missing Data:

Next, the missing values (NA) in the Children, Age, Income, Phone, Techie, TechSupport, Tenure, and Bandwidth_GB_Year need to be fixed. I will fix each variable by univariate statistical imputation after researching the best type of imputation for each column. I choose to use imputation over deletion as it will preserve as much data from the original dataset as possible.

Variable Name	**Data Distribution**	**Imputation Method**
Children	Positive Skew	Median
Age	Uniform	Mean
Income	Positively Skewed	Median
Phone	Categorical Variable	Mode
Techie	Categorical Variable	Mode
TechSupport	Categorical Variable	Mode
Tenure	Bimodal	Mode
Bandwidth_GB_Year	Bimodal	Mode

After treating a variable for missing values, I then run colSums(is.na()) to verify there are no more columns with missing data after all imputation has occurred. In addition, I check the histogram distributions of each individual variable after imputation, to verify the distribution has remained consistent from before and after the imputation. Since the Age column contained values with many decimal places, I rounded the column values to the nearest whole value in order to make logical sense with the previous data.

Outlier Treatment:

For the variables with outliers, the values will be converted to NA and then imputed with the median. Since we are converting existing values to NA, all missing data must be treated before this step. Outliers heavily influence the mean, so the median is the most practical option of imputation to keep the distribution consistent and maintain as much data as possible.

Re-expressing Categorical Data:

For the variable Education, the strings will be converted into ordinals through ordinal encoding. With the revalue() function, the levels of education will be ordered with the lowest level of education (No Schooling Completed) starting at 0 and the highest (Doctorate Degree) ending at 11. These rankings are also converted to doubles with the as.numeric() function.

D3. Data Cleaning Summary

To begin with, detecting and treating duplicates was the first priority in the data cleaning process. Through summing all duplicates and checking specific columns for unique values, no duplicates were found. In order to check for missing values, I summed all NA values by column to get a full picture on the variables needing to be adjusted. The variables with missing values were visualized with histograms to check their distribution. Depending on distribution, the variables were imputed with the mean or median. For categorical variables, the mode was utilized for imputation. After missing values were mitigated, I checked for outliers with two methods. For numeric variables, new columns containing their corresponding z-scores were created and filtered to check for z-scores greater than three or less than negative three. Boxplots of the variables assisted with finding the range of outliers. Then, the outliers were imputed by converting to NAs and imputed with the median. The last step of data cleaning involved re-expressing the Education variable with ordinal encoding. The strings were converted to ranking numeric values in correspondence with their obvious ranks.

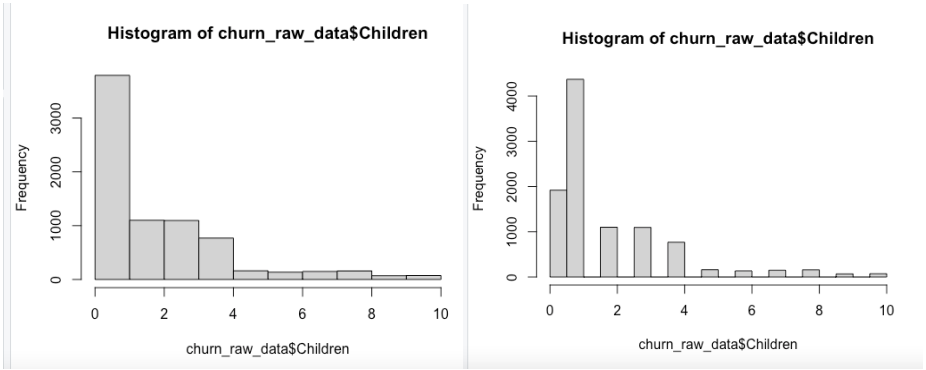
Zero duplicates:

```
> sum(duplicated(churn_raw_data))
[1] 0
>
```

The dataset is free from missing values:

```
> colSums(is.na(churn_clean))
                ...1             CaseOrder           Customer_id
                   0                     0                     0
         Interaction                  City                 State
                   0                     0                     0
              County                   Zip                   Lat
                   0                     0                     0
                 Lng            Population                  Area
                   0                     0                     0
            Timezone                   Job              Children
                   0                     0                     0
                 Age             Education            Employment
                   0                     0                     0
              Income               Marital                Gender
                   0                     0                     0
               Churn    Outage_sec_perweek                 Email
                   0                     0                     0
            Contacts    Yearly_equip_failure                Techie
                   0                     0                     0
            Contract           Port_modem                Tablet
                   0                     0                     0
     InternetService                 Phone              Multiple
                   0                     0                     0
      OnlineSecurity          OnlineBackup      DeviceProtection
                   0                     0                     0
         TechSupport           StreamingTV          StreamingMovies
                   0                     0                     0
    PaperlessBilling         PaymentMethod                Tenure
                   0                     0                     0
      MonthlyCharge     Bandwidth_GB_Year                 item1
                   0                     0                     0
               item2                 item3                 item4
                   0                     0                     0
               item5                 item6                 item7
                   0                     0                     0
               item8     Education_numeric
                   0                     0
> #PCA
```

Variable Children distribution remains the same before (left) and after (right) imputation:



Education Ordinal Encoding:

	Education	Education_numeric
	<chr>	<dbl>
1	Master's Degree	9
2	Regular High School Diploma	4
3	Regular High School Diploma	4
4	Doctorate Degree	11
5	Master's Degree	9
6	No Schooling Completed	0
7	Associate's Degree	7
8	Bachelor's Degree	8
9	Some College, Less than 1 Year	5
10	GED or Alternative Credential	3

ℹ 9,990 more rows
ℹ Use `print(n = ...)` to see more rows

D4. Treatment Code

```
#Treating Missing Data in Children by observing distribution and imputing accordingly

hist(churn_raw_data$Children)

churn_raw_data$Children[is.na(churn_raw_data$Children)] <-

median(churn_raw_data$Children, na.rm = TRUE)

colSums(is.na(churn_raw_data))



#Treating Missing Categorical Data in Phone by mode imputation

churn_raw_data$Phone[is.na(churn_raw_data$Phone)] <-

(names(which.max(table(churn_raw_data$Phone))))
```

```
#Treating outliers in Population by converting outliers to NA and imputing with median

churn_raw_data$Population[churn_raw_data$populationz <= -3 | churn_raw_data$populationz
>= 3] <- NA

colSums(is.na(churn_raw_data))

sum(is.na((churn_raw_data$populationz)))

churn_raw_data$Population[is.na(churn_raw_data$Population)] <-
median(churn_raw_data$Population, na.rm = TRUE)

colSums(is.na(churn_raw_data))



#Re-expressing categorical variable Education with ordinal encoding

edu.num <- revalue(x = churn_raw_data$Education, replace = c("No Schooling Completed" =
0,"Nursery School to 8th Grade" = 1, "9th Grade to 12th Grade, No Diploma" = 2, "GED or
Alternative Credential" = 3, "Regular High School Diploma" = 4, "Some College, Less than 1
Year" = 5, "Some College, 1 or More Years, No Degree" = 6,"Associate's Degree" = 7,
"Bachelor's Degree" = 8, "Master's Degree" = 9, "Professional School Degree" = 10, "Doctorate
Degree" = 11  ))

churn_raw_data$Education_numeric <- as.numeric(edu.num)
```

For complete treatment code on all variables, see code attached.

File name: Fixing_D206.R

D5. Cleaned Data CSV

See file attached: churn_clean.csv

D6. Disadvantages of Chosen Methods

Although data cleaning is a necessary part of managing data sets, it is also immensely time consuming. Many data analysts refer to data cleaning as the most time consuming secta of data analysis. As well as the immense amount of time it takes to clean the data, no cleaning method is perfect. As a result of imputation for missing values and outliers, the data does not reflect the customer's observations to one hundred percent accuracy. Instead, imputation assumes the new values were close to the average and middle range of the rest of the dataset. As a result, imputation is only an estimate of the true data, and can miss variability which naturally can occur between customers. According to Middleton (n.d.), Imputation possibly can distort the data distribution. Distorting the distribution is a shortcoming as it can provide an answer later which does not accurately reflect the true data. However, to mitigate these problems, the distributions were reviewed before and after imputation to reveal the new data remained similar in distribution to the unaltered data.

D7. Challenge for a Data Analyst

If a data anlayst were to take the now-cleaned churn dataset for analysis, it could be a challenge depending on their view of cleaning data. Since the steps are not written in stone and they are up to the interpretation of the user cleaning the data, inconsistencies can arise. One solution to the challenge of cleaning data in the way the analyst would want would be to go over the parameters with the analyst ahead of time. If the parameters are pre-set, the analyst have knowledge of

exactly what to expect when utilizing the cleaned dataset. For example, I chose to leave categorical variables with the values "Yes/No" alone. However, when analyzing the data, if no prior communication took place, the analyst might be expecting values such as "1/0" instead. The choice to leave the variables alone may throw off the analyst when working with the data, and they would have to revalue the variables on their own. In addition to categorical variables, retaining the outliers and imputing them may throw off their analysis as it may not be accurate. Instead, the analyst may have favored deletion of the outliers, which would lessen the sample size. All in all, these problems can be resolved with effective communication with the analyst prior to starting the data cleaning process.

E1. Perform PCA

In order to perform PCA, I choose to select as many continuous quantitative variables as possible. The variables fitting the criteria include: Lat, Lng, Income, Outage_sec_perweek, Tenure, MonthlyCharge, and Bandwidth_GB_Year.

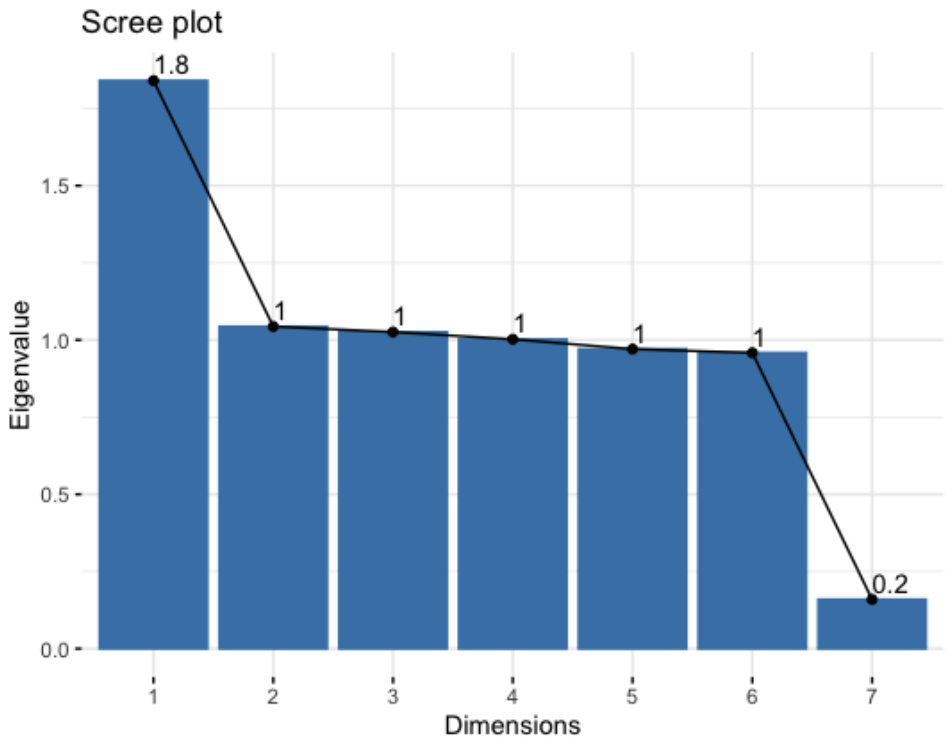
PCA Loadings Matrix:

```
> #Loading matrix
> churn.pca$rotation
                           PC1          PC2          PC3          PC4
Lat                 0.012147357 -0.154670451 -0.713776113  0.077905232
Lng                -0.004354312  0.115506262 -0.415029528 -0.804555050
Income              0.005275995 -0.388701274 -0.489010397  0.412697504
Outage_sec_perweek -0.021043926  0.606116923 -0.173681142  0.419496001
Tenure             -0.705294772 -0.060218310  0.008158608 -0.007415083
MonthlyCharge      -0.044215711  0.663784335 -0.221144870  0.016127998
Bandwidth_GB_Year  -0.707083484 -0.005749943 -0.002495539  0.003275260
                           PC5         PC6           PC7
Lat                -0.6366569025  0.23460479 -0.0035367851
Lng                 0.2093981693 -0.35099974 -0.0057095604
Income              0.6544831643 -0.10528396 -0.0004014361
Outage_sec_perweek -0.1368041343 -0.63818890  0.0060789116
Tenure             -0.0235806825 -0.02925714  0.7052655277
MonthlyCharge       0.3212487157  0.63430300  0.0522410879
Bandwidth_GB_Year   0.0001605947  0.01391838 -0.7069576946
```

E2. Select PCs

In total, there are seven PCs. In order to determine which PCs are the most important, I will utilize the Kaiser rule. The Kaiser rule includes all PCs with an eigenvalue which is greater or equal to one. With the help of a scree plot to depict the eigenvalue of each PC, it is evident PC1, PC2, PC3, PC4, PC5, and PC6 are the most significant principal components. These six PCs are observed on the scree plot to have an eigenvalue of one or greater. PC7 has an eigenvalue of .2, so PC7 will be removed.

Scree Plot:



E3. Benefits of PCA

Ultimately, PCA can help an organization by reducing the dimensionality of the data. Since an algorithm depends on the dimensions of the data, it is most efficient to use the PCs which hold the most weight. By dropping the unnecessary PCs, the algorithms can run faster and more efficiently. PCA can improve algorithm model performance while costing the model accuracy to decrease slightly (Bigabid, 2023). This statement reveals that the model accuracy may be reduced, however, the tradeoff is an increase in performance efficiency which cuts down the run time. As well as increased performance, PCA contributes to feature selection and reducing the noise of a dataset while increasing variance. Increased variance is important since more variance leads to a larger portion of information being gathered as a result of PCA.

F. References

Bigabid. (2023). *What Is Principal Component Analysis (PCA) & How to Use It?* www.bigabid.com/what-is-pca-and-how-can-i-use-it/

Middleton, K (n.d.). *Getting Started with D206 | Missing Values.*

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=767749d2-ba19-4f94-bec8-b058017b2f5e

WGU Course Ware Materials (n.d.). *Data Cleaning Lesson 6: Z-Scores.*

WGU Information Technology (n.d.). *R or Python.*

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