**Part I: Research Question**

1. Describe **one** question or decision that you will address using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation.

How can we reduce the number of customers leaving?

1. Describe the variables in the data set and indicate the specific type of data being described. Use examples from the data set that support your claims.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| CaseOrder | INT (CONTINOUS) | A placeholder variable to preserve the original order of the raw data file |
| Customer\_id | STRING | Unique customer ID |
| Interaction, UID | STRING | Unique IDs related to customer transactions, technical support, and sign-ups |
| City | STRING | Customer city of residence as listed on the billing statement |
| State | STRING | Customer state of residence as listed on the billing statement |
| County | STRING | Customer county of residence as listed on the billing statement |
| Zip | STRING | Customer zip code of residence as listed on the billing statement |
| Lat, Lng | INT | GPS coordinates of customer residence as listed on the billing statement |
| Population | INT | Population within a mile radius of customer, based on census data |
| Area | STRING | Area type (rural, urban, suburban), based on census data |
| TimeZone | STRING | Time zone of customer residence based on customer’s sign-up information |
| Job | STRING | Job of the customer (or invoiced person) as reported in sign-up information |
| Children | INT | Number of children in customer’s household as reported in sign-up information |
| Age | INT | Age of the customer as reported in sign-up information |
| Education | STRING | Highest degree earned by customer as reported in sign-up information |
| Employment | STRING | Employment status of customer as reported in sign-up information |
| Income | INT | Annual income of customer as reported at time of sign-up |
| Marital | STRING | Marital status of customer as reported in sign-up information |
| Gender | STRING | Customer self-identification as male, female, or nonbinary |
| Churn | STRING | Whether the customer discontinued service within the last month (yes,no) |
| Outage\_sec\_perweek | INT | Average number of seconds per week of system outages in the customer’s neighborhood |
| Email | INT | Number of emails sent to the customer in the last year (marketing or correspondence) |
| Contacts | INT | Number of times customer contacted technical support |
| Yearly\_equip\_failure | INT | The number of times customer’s equipment failed and had to be reset/replaced in the past year |
| Techie | STRING | Whether the customer considers themselves technically inclined (based on customer questionnaire when they signed up for services) (yes,no) |
| Contract | STRING | The contract term of the customer (month-to-month, one year, two year) |
| Port\_modem | STRING | Whether the customer has a portable modem (yes, no) |
| Tablet | STRING | Whether the customer owns a tablet (yes,no) |
| InternetService | STRING | Customer’s internet service provider (DSL, fiber optic, None) |
| Phone | STRING | Whether the customer has a phone service (yes, no) |
| Multiple | STRING | Whether the customer has multiple lines (yes, no) |
| OnlineSecurity | STRING | Whether the customer has an online security add-on (yes,no) |
| OnlineBackup | STRING | Whether the customer has an online backup add-on (yes,no) |
| DeviceProtection | STRING | Whether the customer has device protection add-on (yes,no) |
| TechSupport | STRING | Whether the customer has a technical support add-on (yes,no) |
| StreamingTV | STRING | Whether the customer has streaming TV (yes, no) |
| StreamingMovies | STRING | Whether the customer has streaming movies (yes,no) |
| PaperlessBilling | STRING | Whether the customer has paperless billing (yes, no) |
| PaymentMethod | STRING | The customer’s payment method (electronic check, mailed check, bank (automatic bank transfer), credit card (automatic)) |
| Tenure | INT | Number of months the customer has stayed with the provider |
| MonthlyCharge | INT | The amount charged to the customer monthly. This value reflects an average per customer |
| Bandwidth\_GB\_Year | INT | The average amount of data used, in GB, in a year by the customer |
| Item 1 | INT | Timely Response |
| Item 2 | INT | Timely fixes |
| Item 3 | INT | Timely replacements |
| Item 4 | INT | Reliability |
| Item 5 | INT | Options |
| Item 6 | INT | Respectful response |
| Item 7 | INT | Courteous exchange |
| Item 8 | INT | Evidence of active listening |

**Part II: Data-Cleaning Plan**

*Note: You may use Python, R, or any other programming language for implementing your coding solutions, manipulating the data, and creating visual representations.*

C.  Explain the plan for cleaning the data by doing the following:

1. Propose a plan that includes the relevant techniques and specific steps needed to identify anomalies in the data set.

To detect outliers for all variables I will be using creating z-score columns for all necessary columns and will then extract the data to a csv file for further observation. Before creating the z-scores, I will recategorize all categorical data to numeric. Duplicated data will also be checked. I will also create a dataframe for each of the z-score columns created to quickly check to see if there are any outliers before viewing the csv file. Missing data will be detected by first counting the amount of missing values using the is.na.sum() method in Python/Pandas and then creating histograms for each of the columns to determine how I will lately mitigate the missing values.

1. Justify your approach for assessing the quality of the data, include:

Outliers can be detected using z-scores after data transformation is completed, according to Chantal D. Larose, & Daniel T. Larose. (2019). Any value that is not between the range of -3 and 3 is an outlier. I chose this method instead of a histogram or boxplot, so that I can visually see the values and compare it to the rest of the data.

1. Justify your selected programming language and any libraries and packages that will support the data-cleaning process.

According to Lesson 2: Introduction to Programming with Python and R , both programming languages have their pros and cons (UCERTIFY).My choice of language, Python, is known for its ease of integration and also its data analysis tools/libraries.

1. Provide the code you will use to identify the anomalies in the data.

import pandas as pd

import matplotlib.pyplot as plt

from scipy import stats

import numpy as np

churndata = pd.read\_csv(r'C:\Users\Aaliyah Smith\Desktop\Data Cleaning\churn\_raw\_data.csv')

#Count of Missing Data for each column

churndata.isna().sum()

#Check Missing Column Distribution to determine type of imputation

churndata.hist(['Age']) #uniform

churndata.hist(['Income'])#skewed

churndata.hist(['Children'])#skewed

churndata.hist(['Tenure'])#bimodial

churndata.hist(['Bandwidth\_GB\_Year'])#bimodal

#Check for duplicates

churndata.duplicated().sum()

#Recategorize Categorical data to numeric

#Area

churndata['Area'] = churndata['Area'].replace(['Urban'],0)

churndata['Area'] = churndata['Area'].replace(['Suburban'],1)

churndata['Area'] = churndata['Area'].replace(['Rural'],2)

#Employment

churndata['Employment'] = churndata['Employment'].replace(['Unemployed'],0)

churndata['Employment'] = churndata['Employment'].replace(['Full Time'],1)

churndata['Employment'] = churndata['Employment'].replace(['Part Time'],2)

churndata['Employment'] = churndata['Employment'].replace(['Retired'],3)

churndata['Employment'] = churndata['Employment'].replace(['Student'],4)

#Marital

churndata['Marital'] = churndata['Marital'].replace(['Never Married'],0)

churndata['Marital'] = churndata['Marital'].replace(['Married'],1)

churndata['Marital'] = churndata['Marital'].replace(['Divorced'],2)

churndata['Marital'] = churndata['Marital'].replace(['Separated'],3)

churndata['Marital'] = churndata['Marital'].replace(['Widowed'],4)

#Gender

churndata['Gender'] = churndata['Gender'].replace(['Male'],0)

churndata['Gender'] = churndata['Gender'].replace(['Female'],1)

churndata['Gender'] = churndata['Gender'].replace(['Prefer not to answer'],2)

#Churn

churndata['Churn'] = churndata['Churn'].replace(['No'],0)

churndata['Churn'] = churndata['Churn'].replace(['Yes'],1)

#Techie

churndata['Techie'] = churndata['Techie'].replace(['No'],0)

churndata['Techie'] = churndata['Techie'].replace(['Yes'],1)

#Contract

churndata['Contract'] = churndata['Contract'].replace(['Month-to-month'],0)

churndata['Contract'] = churndata['Contract'].replace(['One year'],1)

churndata['Contract'] = churndata['Contract'].replace(['Two Year'],2)

#Port\_modem

churndata['Port\_modem'] = churndata['Port\_modem'].replace(['No'],0)

churndata['Port\_modem'] = churndata['Port\_modem'].replace(['Yes'],1)

#Tablet

churndata['Tablet'] = churndata['Tablet'].replace(['No'],0)

churndata['Tablet'] = churndata['Tablet'].replace(['Yes'],1)

#InternetService

churndata['InternetService'] = churndata['InternetService'].replace(['None'],0)

churndata['InternetService'] = churndata['InternetService'].replace(['Fiber Optic'],1)

churndata['InternetService'] = churndata['InternetService'].replace(['DSL'],2)

#Phone

churndata['Phone'] = churndata['Phone'].replace(['No'],0)

churndata['Phone'] = churndata['Phone'].replace(['Yes'],1)

#Multiple

churndata['Multiple'] = churndata['Multiple'].replace(['No'],0)

churndata['Multiple'] = churndata['Multiple'].replace(['Yes'],1)

#Online Security

churndata['OnlineSecurity'] = churndata['OnlineSecurity'].replace(['No'],0)

churndata['OnlineSecurity'] = churndata['OnlineSecurity'].replace(['Yes'],1)

#Online Backup

churndata['OnlineBackup'] = churndata['OnlineBackup'].replace(['No'],0)

churndata['OnlineBackup'] = churndata['OnlineBackup'].replace(['Yes'],1)

#DeviceProtection

churndata['DeviceProtection'] = churndata['DeviceProtection'].replace(['No'],0)

churndata['DeviceProtection'] = churndata['DeviceProtection'].replace(['Yes'],1)

#TechSupport

churndata['TechSupport'] = churndata['TechSupport'].replace(['No'],0)

churndata['TechSupport'] = churndata['TechSupport'].replace(['Yes'],1)

#StreamingTV

churndata['StreamingTV'] = churndata['StreamingTV'].replace(['No'],0)

churndata['StreamingTV'] = churndata['StreamingTV'].replace(['Yes'],1)

#StreamingMovies

churndata['StreamingMovies'] = churndata['StreamingMovies'].replace(['No'],0)

churndata['StreamingMovies'] = churndata['StreamingMovies'].replace(['Yes'],1)

#PaperlessBilling

churndata['PaperlessBilling'] = churndata['PaperlessBilling'].replace(['No'],0)

churndata['PaperlessBilling'] = churndata['PaperlessBilling'].replace(['Yes'],1)

#PaymentMethod

churndata['PaymentMethod'] = churndata['PaymentMethod'].replace(['Credit Card (automatic)'],0)

churndata['PaymentMethod'] = churndata['PaymentMethod'].replace(['Bank Transfer(automatic)'],1)

churndata['PaymentMethod'] = churndata['PaymentMethod'].replace(['Mailed Check'],2)

churndata['PaymentMethod'] = churndata['PaymentMethod'].replace(['Electronic Check'],3)

#Education

churndata['Education'] = churndata['Education'].replace(['9th Grade to 12th Grade, No Diploma'],2)

churndata['Education'] = churndata['Education'].replace(["Associate's Degree"],7)

churndata['Education'] = churndata['Education'].replace(["Bachelor's Degree"],8)

churndata['Education'] = churndata['Education'].replace(["Doctorate Degree"],11)

churndata['Education'] = churndata['Education'].replace(['GED or Alternative Credential'],4)

churndata['Education'] = churndata['Education'].replace(["Master's Degree"],9)

churndata['Education'] = churndata['Education'].replace(["No Schooling Completed"],0)

churndata['Education'] = churndata['Education'].replace(["Nursery School to 8th Grade"],1)

churndata['Education'] = churndata['Education'].replace(["Professional School Degree"],10)

churndata['Education'] = churndata['Education'].replace(["Regular High School Diploma"],3)

churndata['Education'] = churndata['Education'].replace(["Some College, 1 or More Years, No Degree"],6)

churndata['Education'] = churndata['Education'].replace(["Some College, Less than 1 Year"],5)

#Create Z-Score for every numeric column

cl = churndata[list(churndata.loc[:,'Population':'Area']) + list(churndata.loc[:,'Children':'item8'])]

for column in cl:

col\_zscore = column + '\_zscore'

churndatanumeric[col\_zscore] = (churndatanumeric[column] - churndatanumeric[column].mean())/churndatanumeric[column].std(ddof=0)

#Create Outlier Dataframes

PopulationOutlier = churndata.query('Population\_zscore > 3 | Population\_zscore < -3')

AreaOutlier = churndata.query('Area\_zscore > 3 | Area\_zscore < -3')

ChildrenOutlier = churndata.query('Children\_zscore > 3 | Children\_zscore < -3')

AgeOutlier = churndata.query('Age\_zscore > 3 | Age\_zscore < -3')

EducationOutlier = churndata.query('Education\_zscore > 3 | Education\_zscore < -3')

EmploymentOutlier = churndata.query('Employment\_zscore > 3 | Employment\_zscore < -3')

IncomeOutlier = churndata.query('Income\_zscore > 3 | Income\_zscore < -3')

MaritalOutlier = churndata.query('Marital\_zscore > 3 | Marital\_zscore < -3')

GenderOutlier = churndata.query('Gender\_zscore > 3 | Gender\_zscore < -3')

OutageOutlier = churndata.query('Outage\_sec\_perweek\_zscore > 3 | Outage\_sec\_perweek\_zscore < -3')

EmailOutlier = churndata.query('Email\_zscore > 3 | Email\_zscore < -3')

ContactsOutlier = churndata.query('Contacts\_zscore > 3 | Contacts\_zscore < -3')

YearlyEquipFailureOutlier = churndata.query('Yearly\_equip\_failure\_zscore > 3 | Yearly\_equip\_failure\_zscore < -3')

TechieOutlier = churndata.query('Techie\_zscore > 3 | Techie\_zscore < -3')

ContractOutlier = churndata.query('Contract\_zscore > 3 | Contract\_zscore < -3')

PortModemOutlier = churndata.query('Port\_modem\_zscore > 3 | Port\_modem\_zscore < -3')

TabletOutlier = churndata.query('Tablet\_zscore > 3 | Tablet\_zscore < -3')

InternetServiceOutlier = churndata.query('InternetService\_zscore > 3 | InternetService\_zscore < -3')

PhoneOutlier = churndata.query('Phone\_zscore > 3 | Phone\_zscore < -3')

MultipleOutlier = churndata.query('Multiple\_zscore > 3 | Multiple\_zscore < -3')

OnlineSecurityOutlier = churndata.query('OnlineSecurity\_zscore > 3 | OnlineSecurity\_zscore < -3')

OnlineBackupOutlier = churndata.query('OnlineBackup\_zscore > 3 | OnlineBackup\_zscore < -3')

DeviceProtectionOutlier = churndata.query('DeviceProtection\_zscore > 3 | DeviceProtection\_zscore < -3')

TechSupportOutlier = churndata.query('TechSupport\_zscore > 3 | TechSupport\_zscore < -3')

StreamingTVOutlier = churndata.query('StreamingTV\_zscore > 3 | StreamingTV\_zscore < -3')

StreamingMoviesOutlier = churndata.query('StreamingMovies\_zscore > 3 | StreamingMovies\_zscore < -3')

PaperlessBillingOutlier = churndata.query('PaperlessBilling\_zscore > 3 | PaperlessBilling\_zscore < -3')

PaymentMethodOutlier = churndata.query('PaymentMethod\_zscore > 3 | PaymentMethod\_zscore < -3')

TenureOutlier = churndata.query('Tenure\_zscore > 3 | Tenure\_zscore < -3')

MonthlyChargeOutlier = churndata.query('MonthlyCharge\_zscore > 3 | MonthlyCharge\_zscore < -3')

BandwidthOutlier = churndata.query('Bandwidth\_GB\_Year\_zscore > 3 | Bandwidth\_GB\_Year\_zscore < -3')

Item1Outlier = churndata.query('item1\_zscore > 3 | item1\_zscore < -3')

Item2Outlier = churndata.query('item2\_zscore > 3 | item2\_zscore < -3')

Item3Outlier = churndata.query('item3\_zscore > 3 | item3\_zscore < -3')

Item4Outlier = churndata.query('item4\_zscore > 3 | item4\_zscore < -3')

Item5Outlier = churndata.query('item5\_zscore > 3 | item5\_zscore < -3')

Item6Outlier = churndata.query('item6\_zscore > 3 | item6\_zscore < -3')

Item7Outlier = churndata.query('item7\_zscore > 3 | item7\_zscore < -3')

Item8Outlier = churndata.query('item8\_zscore > 3 | item8\_zscore < -3')

**Part III: Data Cleaning**

D.  Summarize the data-cleaning process by doing the following:

1.  Describe the findings, including all anomalies, from the implementation of the data-cleaning plan from part C.

Missing Data

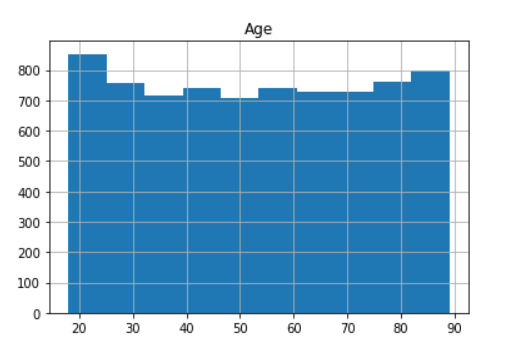
* Children has 2,495 missing values in the churn dataset. However, this data doesn’t have an associated pattern with them when compared to the other variables which according to the textbook considers those values to be Missing Completely at Random (MCAR).
* Age has 2,475 missing values in the churn dataset. However, this data doesn’t have an associated pattern with them when compared to the other variables which according to the textbook considers those values to be Missing Completely at Random (MCAR).
* Income has 2,490 missing values in the churn dataset. However, this data doesn’t have an associated pattern with them when compared to the other variables which according to the textbook considers those values to be Missing Completely at Random (MCAR).
* Techie has 2,477 missing values in the churn dataset. However, this data doesn’t have an associated pattern with them when compared to the other variables which according to the textbook considers those values to be Missing Completely at Random (MCAR).
* Phone has 1,026 missing values in the churn dataset. However, this data doesn’t have an associated pattern with them when compared to the other variables which according to the textbook considers those values to be Missing Completely at Random (MCAR).
* TechSupport has 991 missing values in the churn dataset. However, this data doesn’t have an associated pattern with them when compared to the other variables which according to the textbook considers those values to be Missing Completely at Random (MCAR).
* Tenure has 931 missing values in the churn dataset. However, this data doesn’t have an associated pattern with them when compared to the other variables which according to the textbook considers those values to be Missing Completely at Random (MCAR).
* Bandwidth\_GB\_Year has 1021 missing values in the churn dataset. However, this data doesn’t have an associated pattern with them when compared to the other variables which according to the textbook considers those values to be Missing Completely at Random (MCAR).

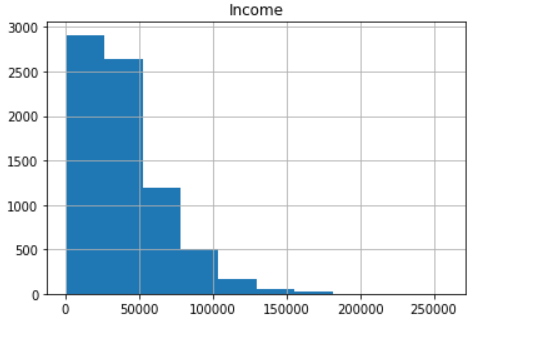
Outliers

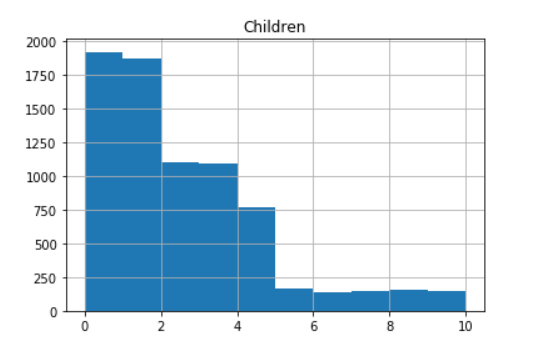
* Population has 219 outliers.
* Income has 193 outliers.
* Outage has 491 outliers.
* Email has 12 outliers.
* Contacts has 165 outliers.
* YearlyEquipFailure has 94 outliers.
* Phone has 846 outliers.
* MonthlyCharge has 3 outliers.
* Item 1 has 19 outliers.
* Item 2 has 13 outliers.
* Item 3 has 13 outliers.
* Item 4 has 9 outliers.
* Item 5 has 12 outliers.
* Item 6 has 13 outliers.
* Item 7 has 11 outliers.
* Item 8 has 15 outliers.
* Children has 302 outliers.

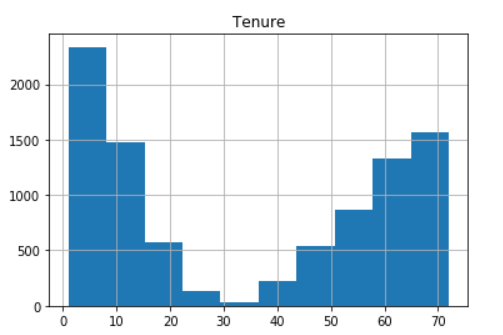
2.  Justify your methods for mitigating each type of discovered anomaly in the data set.

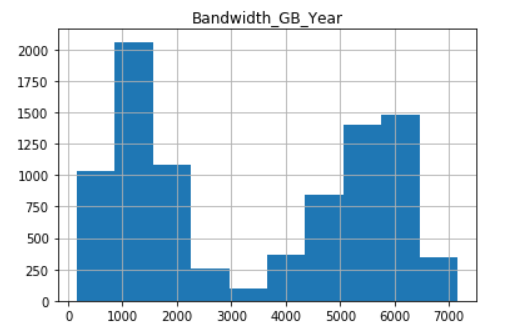
* Missing Data
  + Children, Age, Income, Techie, Phone, TechSupport, Tenure, and Bandwidth\_GB\_Year had missing values with characteristics of those values being Missing Completely at Random (MCAR). Because of that reason, I was able to use either the mean, median, or mode to replace the missing values. Before I could use one of those methods, I created a histogram for the abovementioned columns to observe the distribution. If there was a normal distribution, I was free to use the mean to fill the missing values. For Skewed distributions I could use the median and for categorical, mode.











* Outliers
  + All population outliers will be retained due to this data coming directly from census data and from larger cities.
  + Income outliers will be retained because these are legitimate entries from customer’s who have six-figure jobs.
  + All email outliers will be retained but noted.
  + All contact outliers will be retained. The minimum amount is 4 and the max is 7 which I believe is a legitimate occurrence.
  + All phone outliers will be retained. All of the outliers are customers who do not have phone service.
  + MonthlyCharge has 3 outliers which I believe to be all legitimate. All of customer’s has multiple add-ons associated with their account which I assume contributes to the MonthlyCharge.
  + Item1 – Item 8 all have outliers which I believe to be legitimate due to the “outliers” all being within the range of the survey scale (1-8).
  + All children outliers will be retained. All of the outliers are still possible outcomes even if the customer is 19. There could have been adoption, naturally birth, children through marriage, etc.

3.  Summarize the outcome from the implementation of *each* data-cleaning step.

* A missing data count was the first step completed. After that step, for all columns that had missing values, a histogram was created to classify the distribution step which according to Data Cleaning “Getting Started with D206” Powerpoint slide “Considerations and Limitations” column is key for determining the imputation method (Middleton,2021). After filling in the missing values with the correct measure, I recategorize the categorical data to numeric. Z-scores was created for all numeric columns by creating a calculation and applying it to all relevant columns. Fifteen columns had outliers with none of the columns outliers being replaced/deleted. This decision was made by looking thoroughly at the data and referencing the ChurnData dictionary file.

4.  Provide the code used to mitigate anomalies.

#Replace Missing Values -Numeric

churndata['Income'].fillna(churndata['Income'].median(),inplace = True)

churndata['Children'].fillna(churndata['Children'].median(),inplace = True)

churndata['Tenure'].fillna(churndata['Tenure'].mean(),inplace = True)

churndata['Bandwidth\_GB\_Year'].fillna(churndata['Bandwidth\_GB\_Year'].mean(),inplace = True)

churndata['Age'].fillna(churndata['Age'].mean(),inplace = True)

#Replace Missing Values -Categorical

TechieMode = max(list(churndata['Techie']), key = list(churndata['Techie']).count)

PhoneMode = max(list(churndata['Phone']), key = list(churndata['Phone']).count)

TechSupportMode = max(list(churndata['TechSupport']), key = list(churndata['TechSupport']).count)

churndata['Techie'].fillna(TechieMode,inplace = True)

churndata['Phone'].fillna(PhoneMode,inplace = True)

churndata['TechSupport'].fillna(TechSupportMode,inplace = True)

5.  Provide a copy of the cleaned data set.

A copy of the cleaned data set is attached to the submission.

6.  Summarize the limitations of the data-cleaning process.

* When handling the missing data, I referenced the Methods for Missing Data slide when determining the imputation method. I use either the mean, median, and mode fill in method for all missing values. As stated previously, that fill in method was determine by the data distribution and if it was categorical data. The limitation of using this method is that “it could possibly distort data/distribution of the data” (Middleton, 2021).
* Outliers can distort/distribution of the data as well. (Middleton,2021)

7.  Discuss how the limitations in part D6 affect the analysis of the question or decision from part A.

The customer data could potentially be affected in future analysis. I believe I lessened the amount of distortion by replacing the missing values with measures that has a lesser effect. Also, for all of the outliers that was retained was only kept because the values are legitimate entries that have a higher/lesser value than the majority.

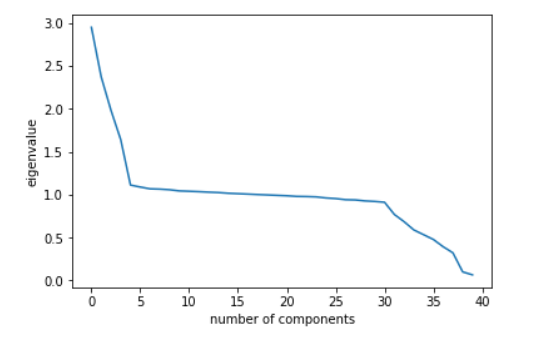
E.  Apply principal component analysis (PCA) to identify the significant features of the data set by doing the following:

1. List the principal components in the data set.

PC1 – PC30 are the principal components.

1. Describe how you identified the principal components of the data set.

Identified the principal components by observing the eigenvalue. All significant principal components have an eigenvalue greater than 1.



1. Describe how the organization can benefit from the results of the PCA  
    With the results of the PCA, we now know how to limit the amount of variables for future analysis and also which variables are the most significant.

Reference Page

Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley.

Middleton, K. (2021). *Data Cleaning: Getting Started with D206* [PowerPoint slides]. Western Governors University . Classroom Lecture

UCERTIFY. D206-Date Cleaning (2021). Retrieved from [https://wgu.ucertify.com/wgu](https://wgu.ucertify.com/)