D206 Data Cleaning Performance Task

Sean Simmons

WDU Data Analytics

MSDA D206

January 2023

All code and outputs for this project are attached in the associated file “D206_PA Code”

Part I: Research Question

A. 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.

The question I plan to address using the Churn data set is what are the characteristics(variables) more important to whether a customer will churn? This question is vital to the organization because it identifies the characters of customers who have left the company within the last month. Identifying the important variables here inform the retention, marketing, and organizational needs of this telemarketing company.

B. 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.

My question will be answered with the churn dataset. The Churn dataset included 10,000 rows of customer information from a telemarketing company. In this research question, the main dependent variable is the churn, while the independent variables are the remainder of the dataset.

The entire dataset was cleaned in this project, whose 50 variables can be summarized as follows:

- The orange highlighted variables are defined here, but not cleaned, per regulations and the identity of the variable.
- The yellow highlighted churn variable indicates the dependent variable.

Variable Name	Data Type	Description /Definition
Customer Identification		
CaseOrder	Qualitative, string	Placeholder for the order of the raw data, unique to each customer
Customer_id	Qualitative, string	The distinct id of the individual customer, unique to each customer
Interaction, UID	Qualitative, string	More distinct id's related to processes surrounding the customer, unique to each customer
Customer Demographics		
City	Qualitative, Categorical, String	City listed as the residence of the customer
State	Qualitative, Categorical, String	State listed as the residence of the customer
County	Qualitative, Categorical, String	County listed as the residence of the customer
Zip	Quantitative, numerical, discrete	Zip-Code listed as the residence of the customer
Lat, Lng	Quantitative, numerical, continuous	Latitude and Longitude coordinates listed as the residence of the customer
Population	Quantitative, numerical, discrete	Census date showing the population in a 1-mile radius of the customer's location
Area	Qualitative, Categorical, String	Type of area surrounding the customer.
Timezone	Qualitative, Categorical, String	The timezone of the customer's location
Job	Qualitative, Categorical, String	Reported Job of customer

Children	Quantitative, numerical, discrete	How many children live in the customer's household
Age	Quantitative, numerical, discrete	Age of the customer
Education	Qualitative, Categorical, String	Highest education level of the customer
Employment	Qualitative, Categorical, String	Employment type/status of the customer
Income	Quantitative, numerical, continuous	Annual income of the customer
Marital	Qualitative, Categorical, String	Marital status of the customer
Gender	Qualitative, Categorical, String, “male”, “female”, “nonbinary”	Gender of the customer
Services Purchased		
DeviceProtection	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer purchased/has device protection
Phone	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer purchased/has a phone service
Multiple	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer purchased/has multiple lines
OnlineSecurity	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer purchased/has online security
OnlineBackup	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer purchased/has online backup
TechSupport	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer purchased/has technical support
StreamingTV	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer purchased/has TV streaming

StreamingMovies	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer purchased/has the movie streaming
Customer Information		
Churn	Qualitative, nominal, binary “Yes” and “No”	Whether the customer has “churned” or not. Yes = churn and that means the customer has left the company in the last month. No = they are still with the company.
Outage_Sec_perweek	Quantitative, numerical, continuous	Average seconds a week the system is nonfunctional for the customer
Email	Quantitative, numerical, discrete	Number of emails the company sent to the customer in the last calendar year
Contacts	Quantitative, numerical, discrete	Complete history of the amount of times the customer has contacted customer support
Yearly_equip_failure	Quantitative, numerical, discrete	The number of times the customers equipment had to be replaced due to failure in the last year
Techie	Qualitative, nominal, binary “Yes” and “No”	Self identified customer response as technically inclined or not.
Contract	Qualitative, nominal,	Contract length the customer has on file
Port_modem	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer has a portable modem for their system
Tablet	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer owns any type of tablet

InternetService	Qualitative, nominal, categorial	The customer's type of internet provider
Paperless Billing	Qualitative, nominal, binary “Yes” and “No”	Whether or not the customer is enrolled in paperless milling
Payment Method	Qualitative, nominal, categorial	The payment method on file for the customer
Tenure	Quantitative, numerical, continuous	The length of time the customer has been with the provider in months
MonthlyCharge	Quantitative, numerical, continuous	The average monthly charge to the customer
Bandwidth_GB_year	Quantitative, numerical, continuous	The average amount of data used by the customer, quantified in gigabytes per year
Item 1	Quantitative, integer, 1:8, discrete	Eight Question survey: Timely Response
Item 2	Quantitative, integer, 1:8, discrete	Eight Question survey: Timely Fixes
Item 3	Quantitative, integer, 1:8, discrete	Eight Question survey: Timely Replacements
Item 4	QQuantitative, integer, 1:8, discrete	Eight Question survey: Reliability
Item 5	Quantitative, integer, 1:8, discrete	Eight Question survey: Options
Item 6	Quantitative, integer, 1:8, discrete	Eight Question survey: Respectful Response
Item 7	Quantitative, integer, 1:8, discrete	Eight Question survey: Courteous Exchange
Item 8	Quantitative, integer, 1:8, discrete	Eight Question survey: Active Listening Evidence

Examples of the data are pulled from the programming environment using the code in the attached file “D206_PA Code.”

Output/examples:

```
   Unnamed: 0  CaseOrder Customer_id                           Interaction  \
0           1          1     K409198  aa90260b-4141-4a24-8e36-b04ce1f4f77b
1           2          2     S120509  fb76459f-c047-4a9d-8af9-e0f7d4ac2524

           City State                 County    Zip       Lat        Lng  \
0   Point Baker    AK  Prince of Wales-Hyder  99927  56.25100 -133.37571
1  West Branch    MI                Ogemaw  48661  44.32893  -84.24080

   Population   Area          Timezone                              Job  \
0          38  Urban     America/Sitka  Environmental health practitioner
1       10446  Urban   America/Detroit           Programmer, multimedia

   Children   Age                     Education Employment    Income  Marital  \
0       NaN  68.0               Master's Degree  Part Time  28561.99  Widowed
1       1.0  27.0  Regular High School Diploma    Retired  21704.77  Married

   Gender Churn  Outage_sec_perweek  Email  Contacts  Yearly_equip_failure  \
0    Male    No            6.972566     10         0                     1
1  Female   Yes           12.014541     12         0                     1

  Techie        Contract Port_modem Tablet InternetService Phone Multiple  \
0     No        One year        Yes    Yes     Fiber Optic   Yes       No
1    Yes  Month-to-month         No    Yes     Fiber Optic   Yes      Yes

  OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV  \
0            Yes          Yes               No          No          No
1            Yes           No               No          No         Yes

  StreamingMovies PaperlessBilling             PaymentMethod    Tenure  \
0             Yes              Yes   Credit Card (automatic)  6.795513
1             Yes              Yes  Bank Transfer(automatic)  1.156681

   MonthlyCharge  Bandwidth_GB_Year  item1  item2  item3  item4  item5  item6  \
0     171.449762         904.536110      5      5      5      3      4      4
1     242.948015         800.982766      3      4      3      3      4      3

   item7  item8
0      3      4
1      4      4
```

Part II: Data-Cleaning Plan

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.

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

- **characteristics of the data being assessed,**
- **the approach used to assess the quality.**

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

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

1. My plan to clean the data is to first, import the data into my python environment. I will use jupyter labs and several packages that are listed as required in step 3. Once the csv is loaded into my programming environment, I will evaluate the format of the data using read and print commands to verify the spellings of the headers and examples of the data. Next, I will evaluate the type of data, column names, and which columns need to be changed to a different data type and which columns are missing values. Finally, I will begin cleaning the data, looking for the following pieces of information:
   a. Duplicate values- Duplicates will be searched for with the distinct function using the unique customer Id and removed from the data set.
   b. Missing Values- I will search for null values in columns and then assign the median value or mode categorical string for the respective column to the missing value.

c. Duplicate columns- Redundant columns that hold the same or similar identifying information will be removed. I can print out the first 2 rows of my data and visually identify any columns that meet this criteria and then remove the columns from my data set.

d. Outliers- Outliers can greatly skew that data analysis process, so I will search for outliers using statistical methods including finding the relevant Z score and removing outliers that are 3 or more standard deviations away for all quantitative values.

2. My approach to clean this data started from the appearance of numerous missing values when I loaded it into jupyter. To begin working with the data in a future project or for this organization, those would have to be taken care of. Once the missing values were dealt with, the normal data cleaning techniques to do were to eliminate outliers and cut down on duplicates. Finally, I wanted to streamline the data set so cutting down on the redundant columns was my last step in identifying the approach I needed to take.

In my initial view of the data, I also discovered the eight items from the survey were not properly labeled according to the data dictionary, so they were relabeled to better reflect their purpose. Finally, the categorical state abbreviations needed to be changed to numeric values as described below. In summary, I plan to download the data and then drop redundant columns and change categorical columns to numeric. Then I will clean the missing values, duplicates, and outliers as described above. The final tasks will be to take the cleaned data and perform a PCA to help analyze my research question (Nguyen, 2022).

3. My selected programming language is Python. Python can handle larger data sets easier than R and I am much more familiar with how Python and its extended libraries work. I used jupyter labs as my coding environment since it provided a more user-friendly look and I could easily import the code into this paper. Finally, I imported the following packages:
    a. Pandas package to import and work with the raw churn data. It is the best to use here because it works very well with already labeled data, which is what we have.
    b. Numpy Package to perform the necessary mathematical operations on the arrays in the data set. It is used here because it is excellent in scientific computing.
    c. Scipy Package to provide the algorithms for numpy and help with the statistical computing needed in this project.
    d. Sklearn package to provide a method to visualize z scores, scree plot.
    e. Matplotlib package to create the visuals for this report (histograms). This package contains lots of helpful statistical analysis tools and visualization tools
    f. Seaborn package for boxplot visualization. Seaborn excels in easy to input and read boxplots.
4. Please see the attached file “D206_PA Code” for all executable code, which is labeled within.

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.

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

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

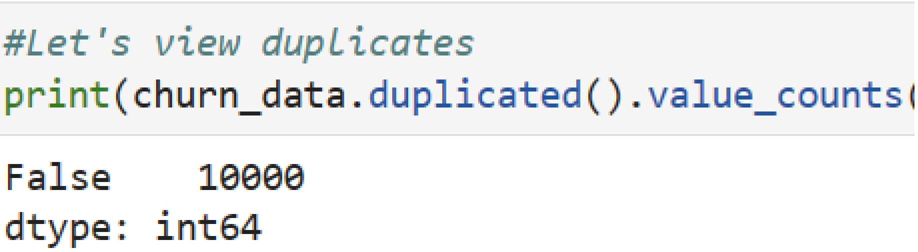
4. Provide the code used to mitigate anomalies.

5. Provide a copy of the cleaned data set.

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

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

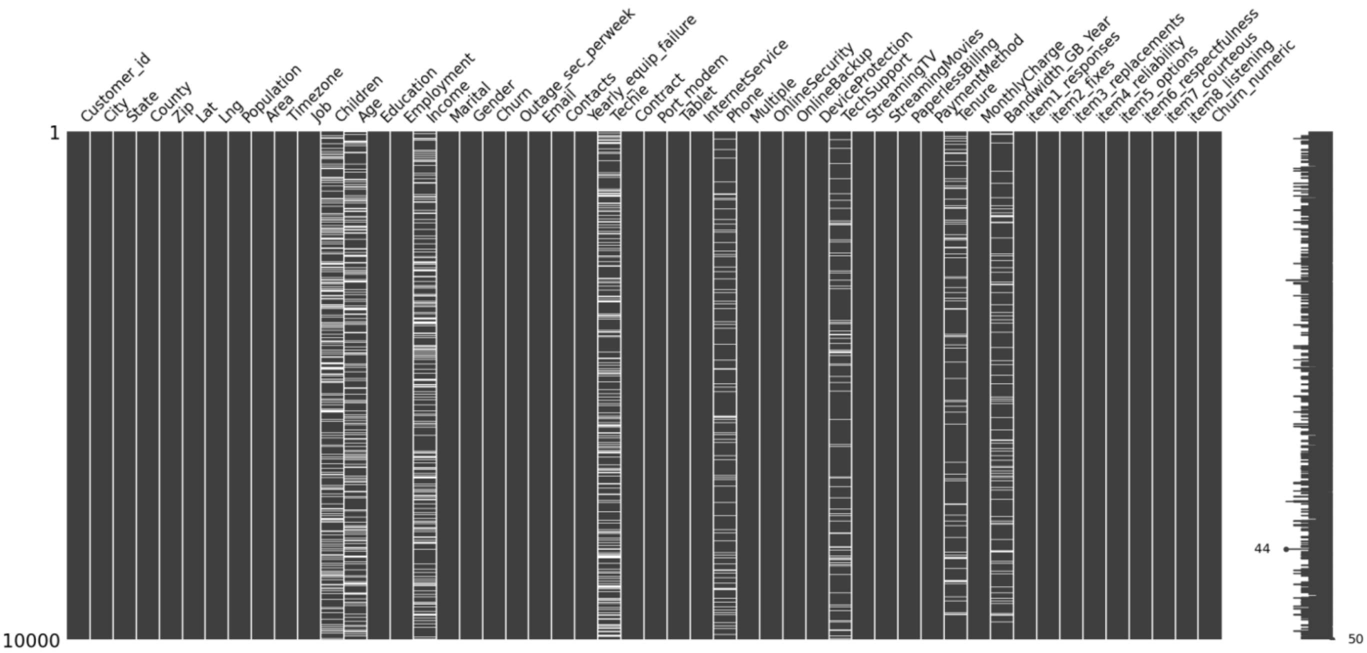
1. Through data cleaning, I found several of each anomaly that was tested for. The anomalies in the data set and were as follows:
    a. Duplicates- There were no duplicate rows found.



```
#Let's view duplicates
print(churn_data.duplicated().value_counts(
```

```
False    10000
dtype: int64
```

    b. Missing values- there were many missing values listed as NA in the children and age columns. These were found using a count null function. There were also missing values in several other columns such as tenure, bandwidth_gb_year, income, techie, techsupport, and phone. See below for all missing anomalies (Each white line in a column represents a missing or Null value).



c. Duplicate Columns- The first column without a header, caseorder, and interaction ID were removed as they serve as redundant unique keys similar to the customer ID using the drop column function.

d. Outliers- There were several outliers in the quantitative columns that include: 'Income', 'Tenure', 'Children', 'Population', 'Age', 'Contacts', 'Emails', 'Yearly_equip_failure', 'Outage_sec_perweek', 'MonthlyCharge', and 'Bandwidth_GB_Year’ and can be seen in more detail in the histograms below and in the output of the code attached to the file. These outliers were normalized and I found the z scores for each of them. All outliers seem to be attached to valid rows of data, giving the impression they should be kept in the dataset to show the true representation of the churn data set, further explained in the outcome section of this report.

e. Categorical Data- There were many columns with categorical data that were changed, including: Employment, Area, Education, Marital, Gender, Contract, PaymentMethod, InternetService, and then several columns that were Yes/No, including: Techie, Portmodem, Tablet, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling.

2. My methods and reasons for mitigating are as follows (Nguyen, 2022):
   a. Duplicates- If there were any duplicates, they would be found using the unique customer id column and searching for duplicates with the function of the same name. There were no duplicates found here.
   b. Missing values- All missing values were replaced with their respective median values for quantitative data and the mode for qualitative data using the replace function. As seen below in part 3, there are no longer missing values in the data set.
   c. Duplicate Columns were removed from the data set using the drop column function. These columns served as unique identifiers for the customers, which resulted in 3 total rows providing unique id’s and did not provide any data analytics purpose. As a result, I dropped two of them because doing so would not alter any further analytics that would need to be done on the dataset and still

provide a unique key. I also dropped the unnamed column for similar reasons and it seemed to be an error.

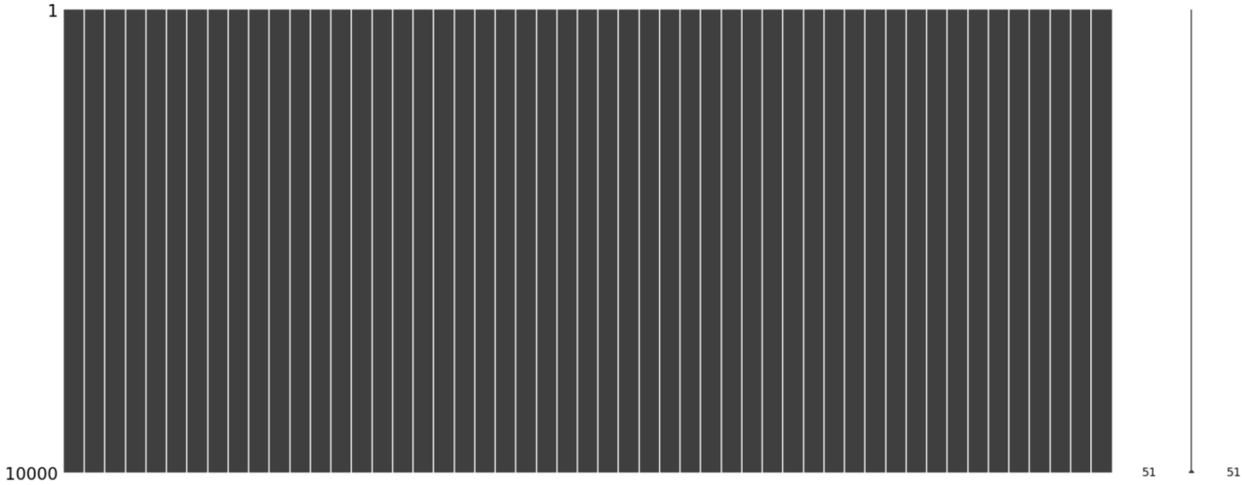
d. Outliers were imputed using a histogram for identification and Z score statistical method as shown in the plots below. The Z score cutoff was -3 and 3. The histograms and Z scores were reviewed and it was determined the entries that were identified as outliers were viable to be retained in the data. The reason I looked for outliers was to try and find any errors in the dataset. Histograms provide a visually friendly way to identify any values that were out of the Z score range. Z scores were used to identify any outliers due to their effectiveness in outlier identification. Through this method, I found no egregious values that seemed to be true outliers that needed to be removed. Everything was within my reasonable expectations for what a customer database should look like.

e. All categorical yes/no columns were replaced with quantitative measures on a scale from 0 counting up. The Yes/No columns were replaced where Yes = 0 and No = 1. I wanted to provide a data environment where all of the variables could be analyzed in the same way, if needed to in the future. As a result, I added many columns to the dataset that were the same variable with the tag "_numeric" added to the end. This was done to preserve the integrity of the original dataset and to provide any future projects on this dataset greater analytical possibilities for this organization.

3. The outcome of the data cleaning resulted in a more efficient data set:

   a. Duplicates- There were no duplicate rows found, so none were removed.

b. Missing values- All missing values were replaced with their respective median values from the variable's dataset to provide a more accurate picture. This chart below shows empty space as missing values in columns, and the solid black bars represent no missing values.



c. Duplicate Columns were removed from the data set to reduce redundancy and trim the data set to only what is important. There is still a unique customer_id to act as a key, but the dataset has been trimmed of excess that would not help any analytics and may even provide confusion in calling the right key throughout the project.

d. Outliers were retained due to qualitative investigation. The major points for retaining them were that this data is a record of customer information with presumably a wide variety of people. An example was a value that came back as an outlier was an income of $115,114.57 at above a 3 Z score. This may statistically be an outlier, but for all accounts it is a valid data entry as this database lists customer’s attributes and characteristics. Removing these outliers or replacing them with the median will be taking the true evaluation of the customer, and changing it, which is not what we want in this data cleaning process. The outliers have been identified as follows and are retained in the dataset to preserve that original integrity of the data and key information that may need to be considered for this organization. There were also several variables that fit the ability to be tested that had no outliers, also seen below with the histograms.

Children had several values with a > 3 Z score, but those might be families with large numbers of children and should not be removed or attributed to error.

Population had several values with a >3 Z score, but these may be customers that live in large cities whose identification could be helpful to the organization, so they should not be removed or attributed to error.

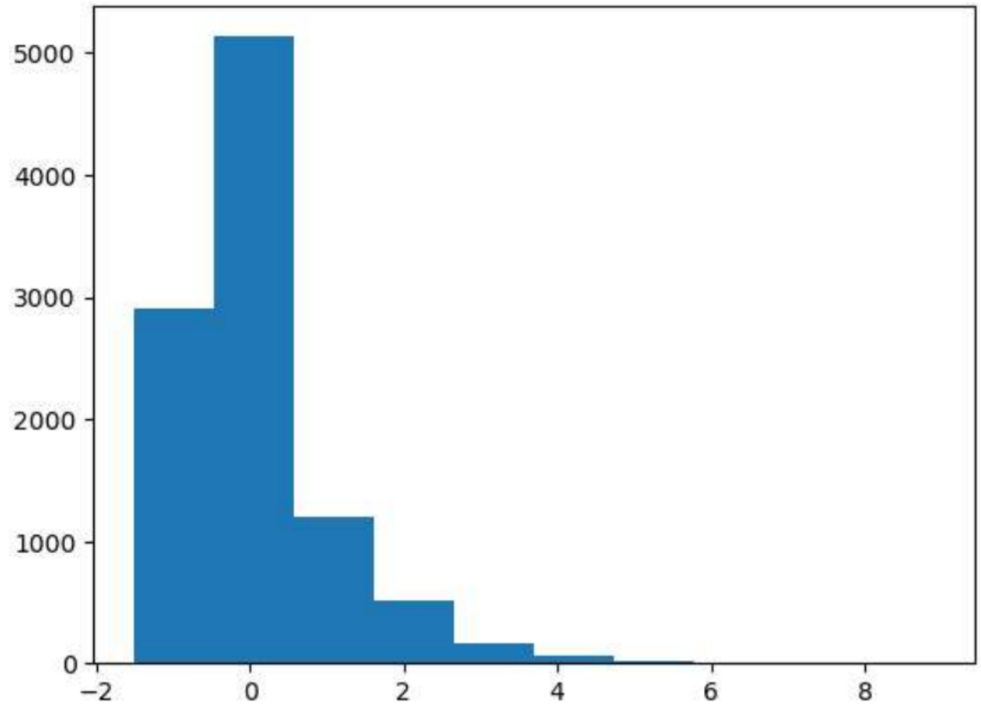
There were only a couple of values with a >3 Z score for Contacts and Email, but knowing which customers get contacted more than the standard distribution could be important to the organization and need to be kept into the dataset.

The few values with a >3 Z score for Outage_sec_perweek and Yearly_equip_failure provide key metrics on customers and should be retained as they were very close to 3 and whose existence and identification may be beneficial to the organization in a future project.

All necessary histograms are on the following pages.

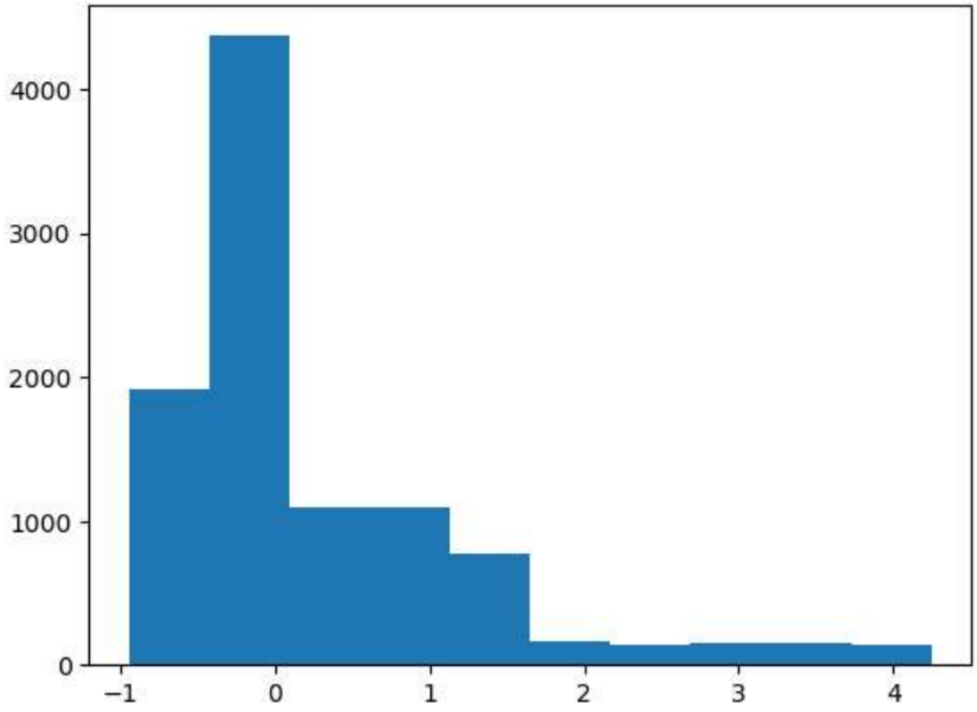
```
[10000 rows x 2 columns]>
```

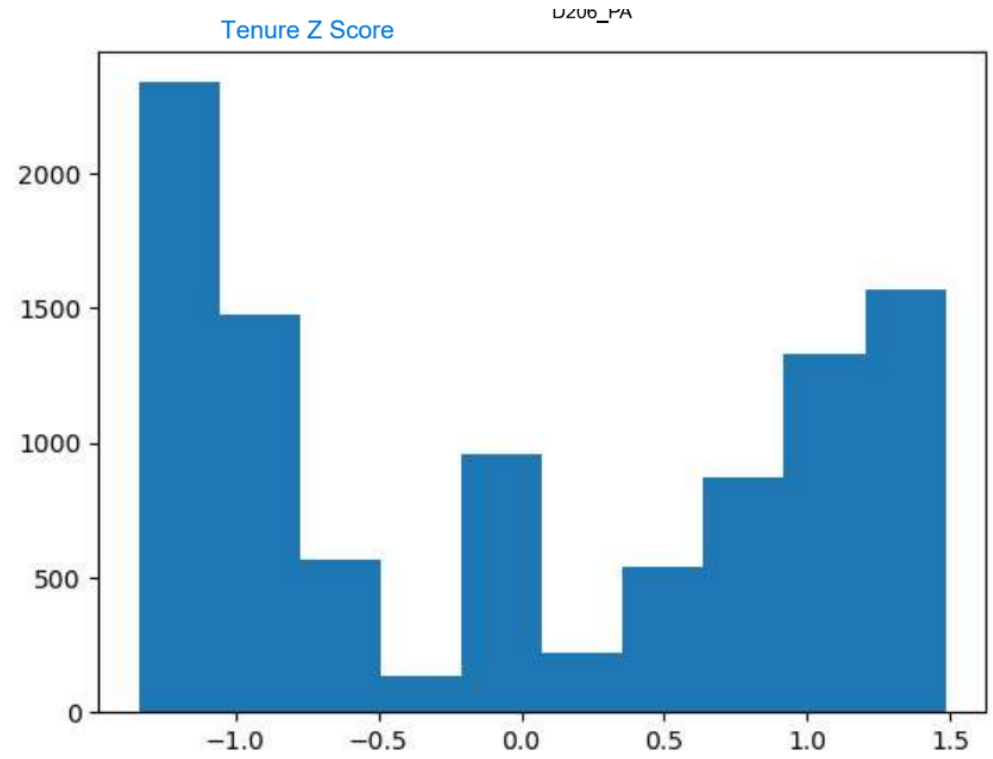
Income Z Scores

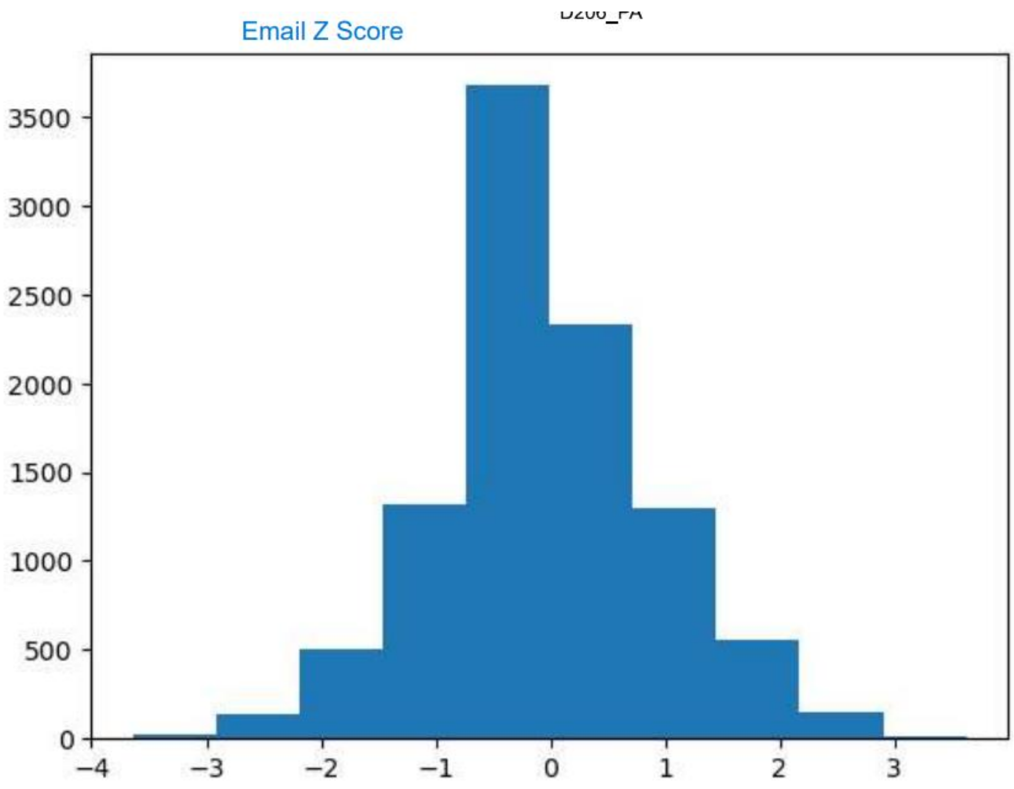


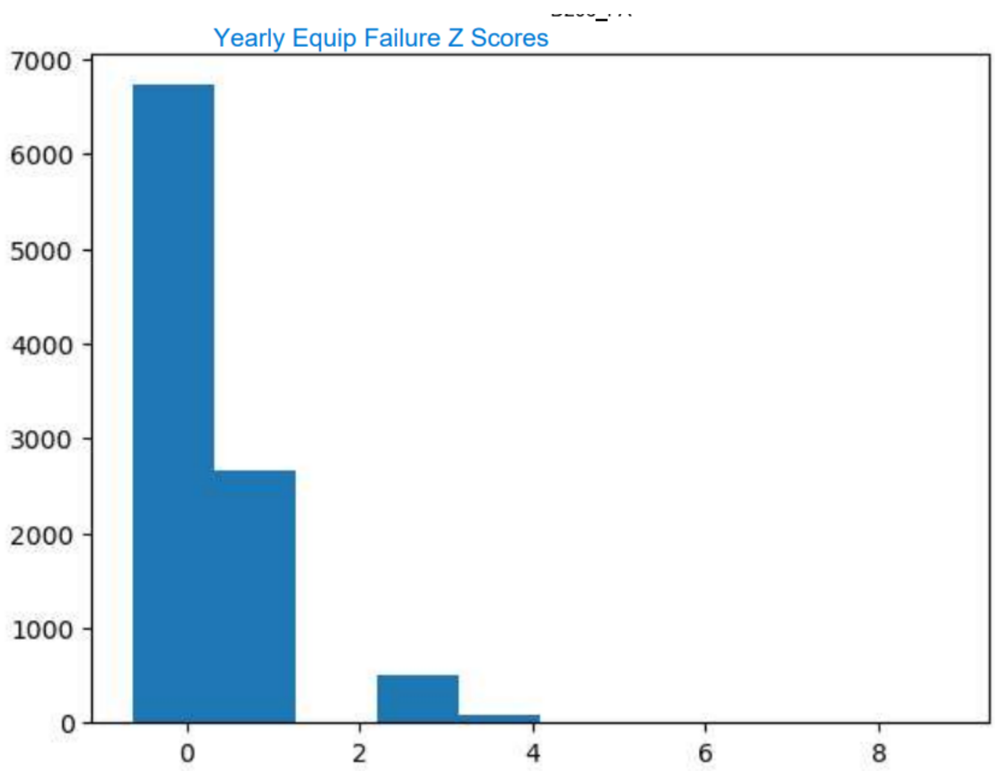
D206_PA

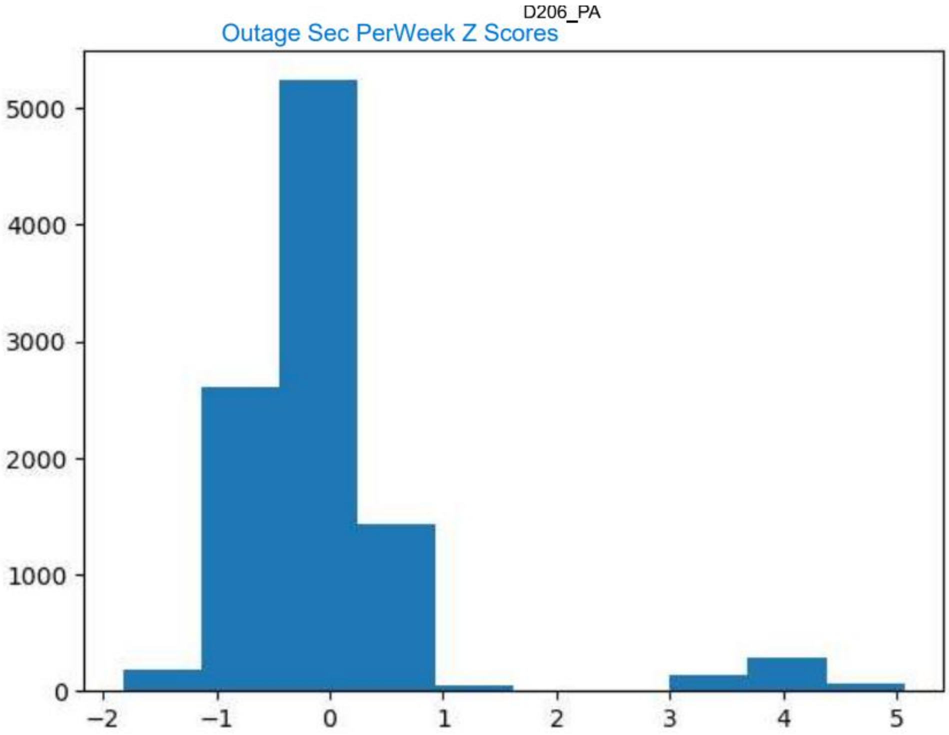
Children Z Scores

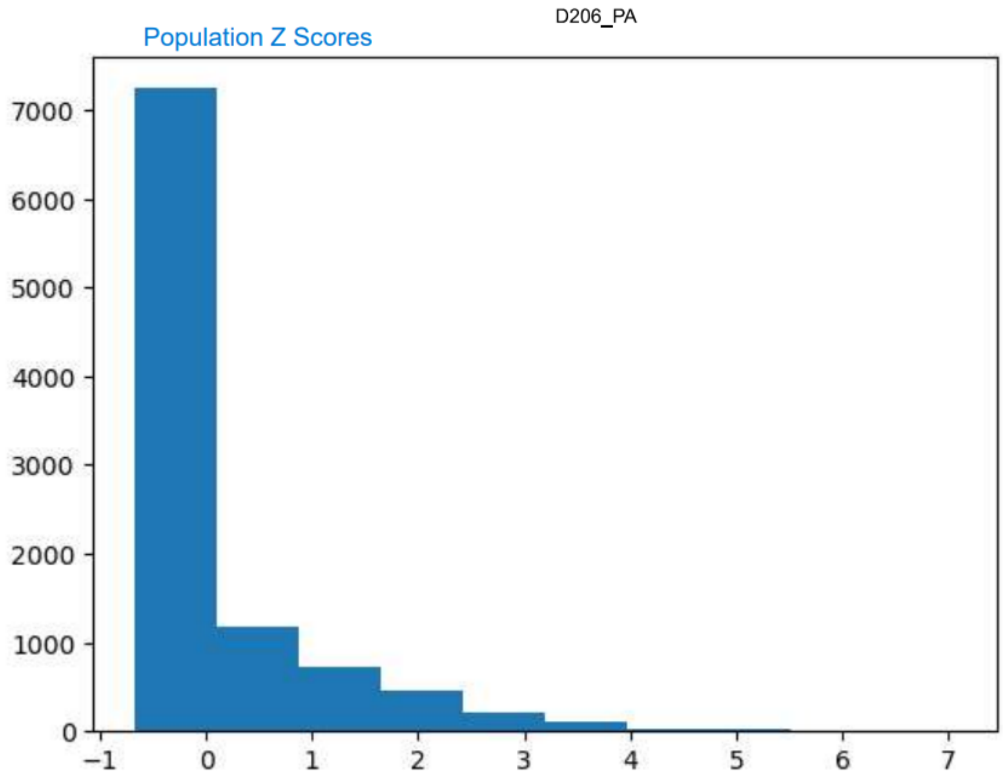


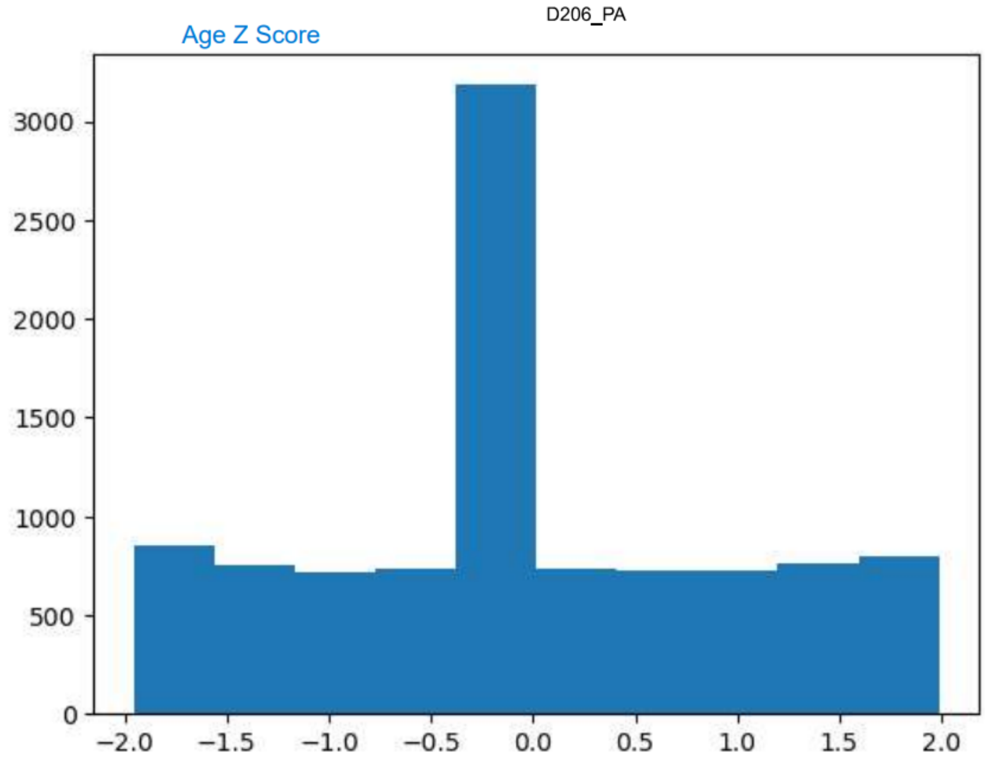

D206_PA
Tenure Z Score
0
500
1000
1500
2000
−1.0
−0.5
0.0
0.5
1.0
1.5

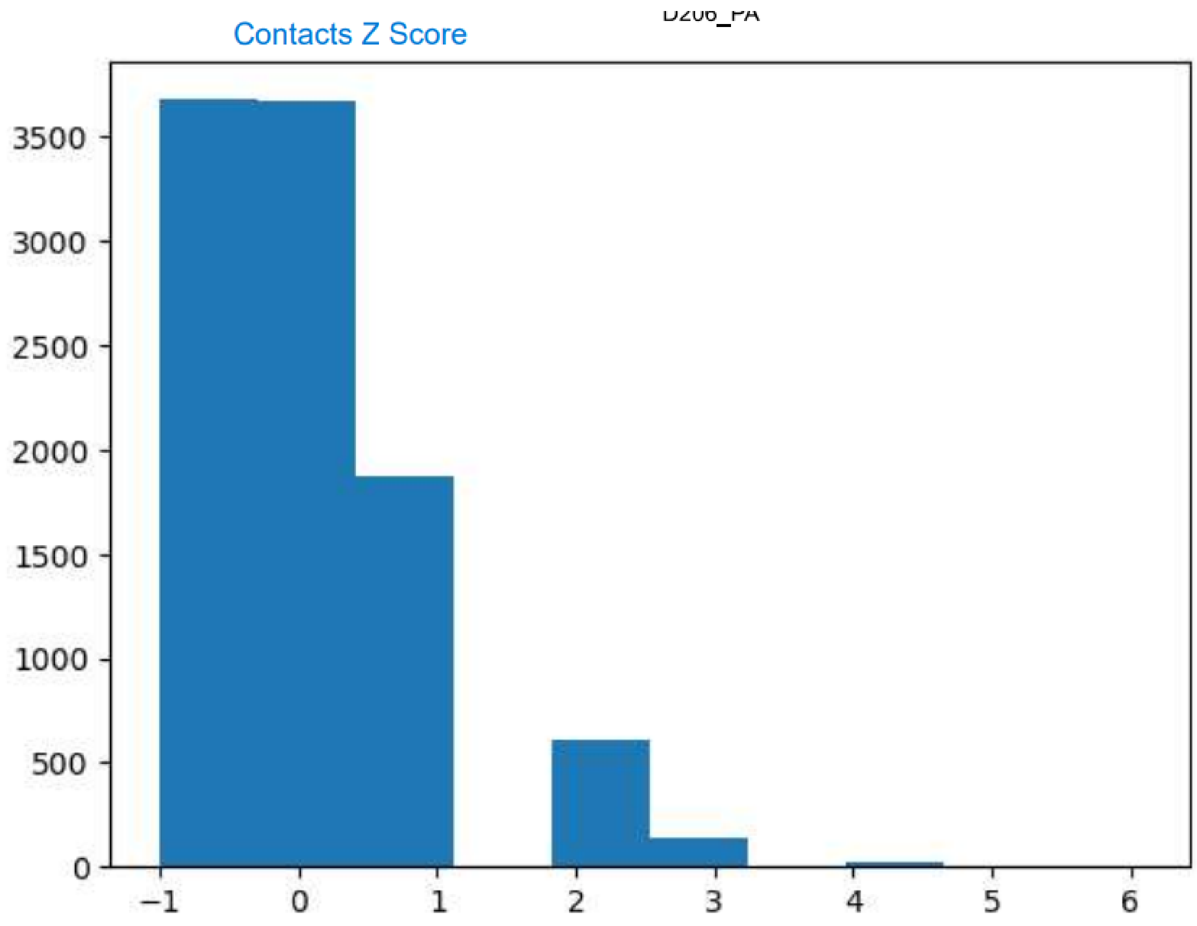

D206_PA
Email Z Score
0
500
1000
1500
2000
2500
3000
3500
−4
−3
−2
−1
0
1
2
3

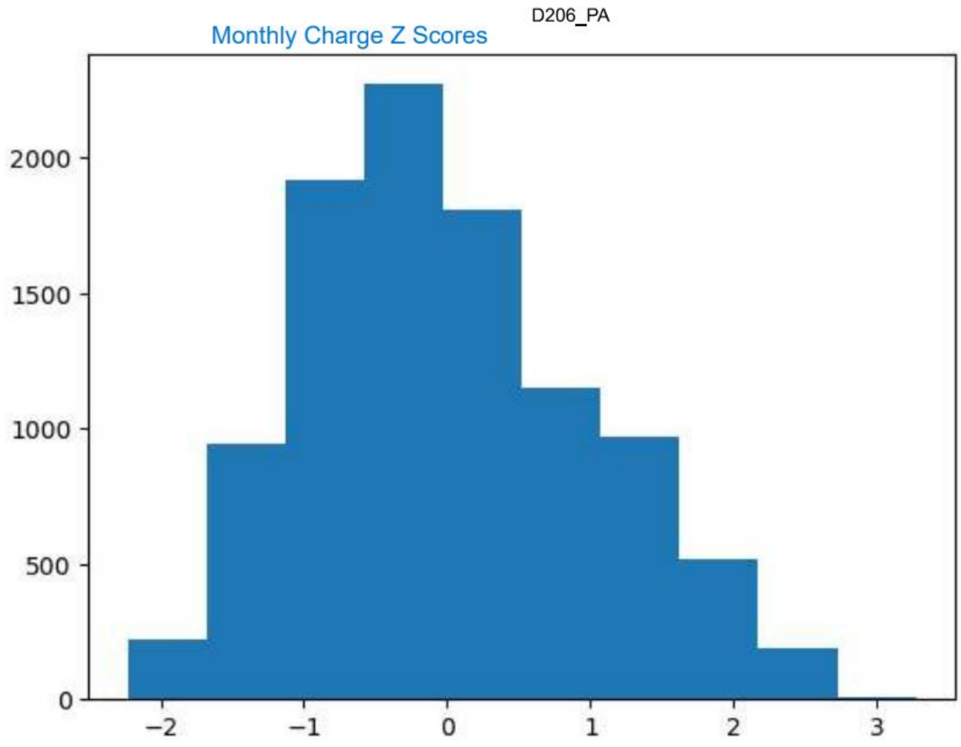

D206_PA
Yearly Equip Failure Z Scores
7000
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8

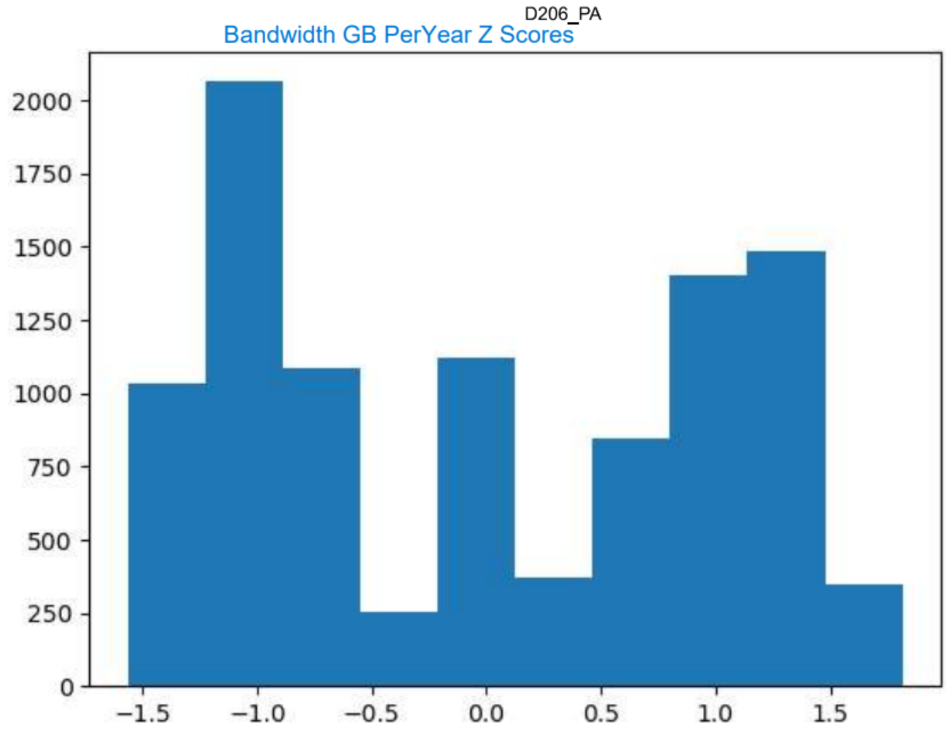

D206_PA
Outage Sec PerWeek Z Scores
5000
4000
3000
2000
1000
0
−2
−1
0
1
2
3
4
5


D206_PA
Population Z Scores
7000
6000
5000
4000
3000
2000
1000
0
−1
0
1
2
3
4
5
6
7


D206_PA
Age Z Score
3000
2500
2000
1500
1000
500
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


D206_PA
Contacts Z Score
3500
3000
2500
2000
1500
1000
500
0
−1
0
1
2
3
4
5
6


D206_PA
Monthly Charge Z Scores
2000
1500
1000
500
0
−2
−1
0
1
2
3



e. Categorical columns were replaced with their numeric counters so we can better analyze the data and have a cohesive data set. All Yes/No columns were replaced with 0 for Yes, and 1 for No. The rest of the categorical values were replaced with a 0… and counting scale to represent their variables, which can be seen below in the code executing the change in part 4.

5. Please see the attached file "D206_PA Code" for all executable code, which is labeled within.

4. The cleaned data set is attached as the csv file "churn_data_cleaned.csv"

5. The limitations of the data-cleaning process stem from this being an external data set. The variables are in many different formats, there is no explanation for missing or "NA" values, and there is no note as to why there are redundant keys in the dataset. Removing duplicate columns may provide issues to the organization unknown to me at this time. Outliers that are identified may or may not be viable to the organization and although through my investigation, they should be retained, this might not be the case and it may be human error that is not known to me. The Z score and histogram method may not provide the clearest picture of which outliers should be removed due to the non-explanatory nature of them, but I believe they were a valid choice with limitations.

   These limitations of my methods provide a clear picture of why communication with the organization is important. Outliers and missing information skew data, not knowing whether outliers are human or experimental error also provides a limitation, because the missing values could even be important to the organization. Finally, the lack of naming for the survey item columns and the variety in data type provide multiple limitations in trying to observe, clean, or work with the data. The categorical data was limited to what could be done, analytically speaking, so columns were added to provide a numeric value to them.

6. The previously mentioned limitations affect my data question of " what are the characteristics(variables) more important to whether a customer will churn?" in several ways. Firstly, the lack of redundant columns may cause confusion when looking at the original data set in comparison to the cleaned one when searching for a key. Secondly, the missing, duplicate, and outlier values may be reduced if the data acquisition phase could be informed. If the outliers were actually errors and not key pieces of customer

information, then their retainment would be problematic. The variety in data type provided an initial limitation in what could be observed for PCA. All of these limitations affect the original research question posed in this report by altering which columns have the greatest impact on the churn column.

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.**
2. **Describe how you identified the principal components of the data set.**
3. **Describe how the organization can benefit from the results of the PCA**

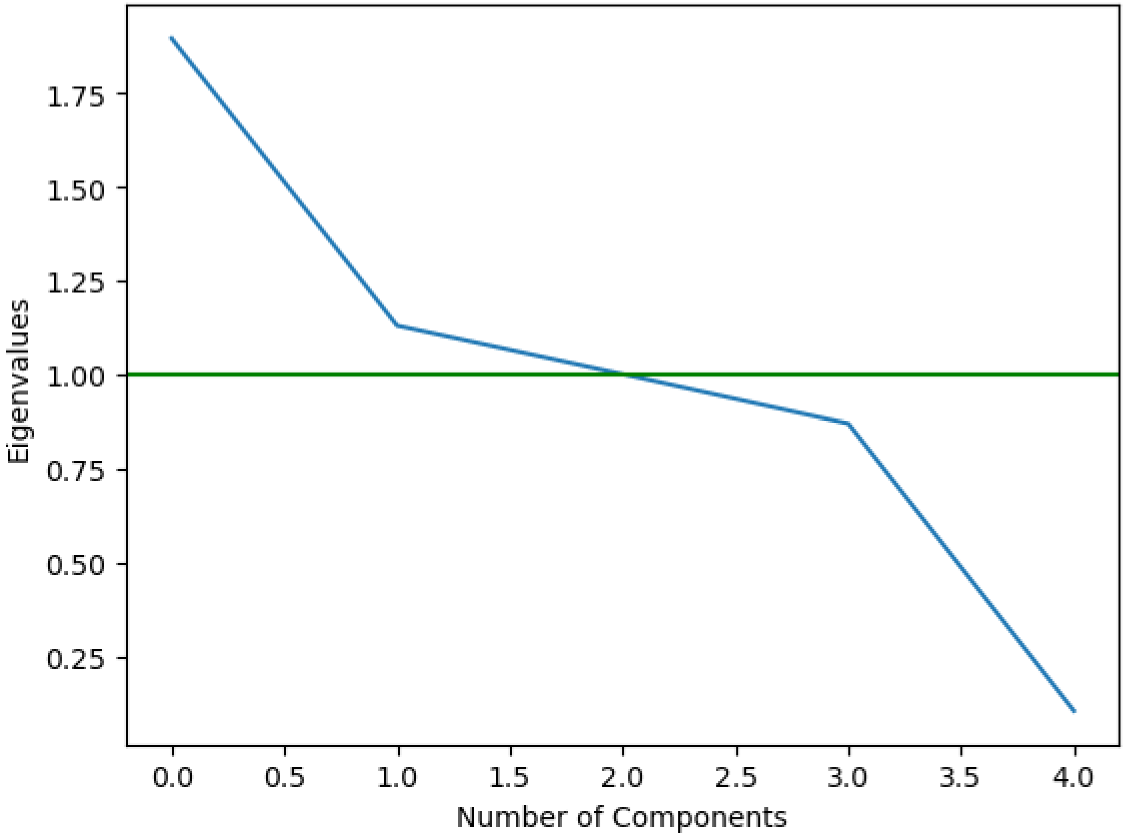
    1. The principal component variables in this data set are Income, Outage_sec_perweek, Tenure, MonthlyCharge, Bandwidth_GB_Year.
        a. The variable that makeup the matrix are:

<li>Income</li>
<li>Outage_sec_perweek</li>
<li>Tenure</li>
<li>MonthlyCharge</li>
<li>Bandwidth_GB_Year</li>

    2. I identified the principal components initially by limiting my search to all continuous variables due to the nature of PCA. I also wanted to focus the PCA on variables related to the financial aspect of this organization and service performance to better inform that area of business. Next, I found the following matrix and then performed a scree plot evaluation (Brems 2017).

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Income	0.005975	-0.001356	0.997951	-0.063676	0.000843
Tenure	0.705360	-0.057918	-0.001425	0.035745	-0.705573
Outage_sec_perweek	0.022378	0.705639	0.045918	0.706728	0.000158
MonthlyCharge	0.045469	0.706129	-0.044166	-0.703600	-0.048063
Bandwidth_GB_Year	0.707010	-0.009953	-0.005625	-0.012242	0.707005

3. The organization can benefit from these results by refining the metrics they use to retain customers. By completing the PCA, the organization can reduce the number of variables they are looking at when trying to reduce churn, helping reduce resources and time spent on this process. Specifically, this can aid the company in price setting, length of customer's contract, how much data usage a customer has, and how often their internet is out to better inform decisions about retention and pricing. From the scree plot below, the components to the left of the intersection point should be retained as important variables, the first two components.



Part IV. Supporting Documents

F. Provide a Panopto recording that demonstrates the warning- and error-free functionality of the code used to support the discovery of anomalies and the data cleaning process and summarizes the programming environment.

Panopto video link:

This link is also provided in the task submission.

G. Reference the web sources used to acquire segments of third-party code to support the application. Be sure the web sources are reliable.

For help with the outlier mitigation with Z score, the code used was informed by:

Zach, Z. (2020, August 6). *How to remove outliers in Python*. Statology. Retrieved January 18, 2023, from https://www.statology.org/remove-outliers-python/#:~:text=How%20to%20Remove%20Outliers%20in%20Python%201%20import,%20%2899%2C3%29%204%20Interquartile%20range%20method%3A%20More%20items

For help with the PCA, the code used was informed by:

Principal Components Analysis(PCA) in Python - Step by Step - Kindson The Genius. (n.d.). https://www.kindsonthegenius.com/principal-components-analysispca-in-python-step-by-step/#:~:text=These%20are%20the%20following%20eight%20steps%20to%20performing

No other third-party code was used in this assignment. The code not mentioned above was informed by the learning done in this course and previous experience working with datasets in Python. The original raw churn data was provided by WGU in the performance assessment assignment page.

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

All of the bolded language that includes header and descriptions of the various elements of this assignment are taken directly from the rubric outlined in the task overview to provide clarity to my evaluator.

Brems, M. (2017, April 17). *A One-Stop Shop for Principal Component Analysis*. Medium; Towards Data Science. https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Nguyen, M. (2022, November 30). *All You Need To Know About Different Types Of Missing Data Values And How To Handle It*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

I. Demonstrate professional communication in the content and presentation of your submission.

This section cannot be summarized and I hope it has shown throughout this document.