D206: Data Cleaning Task 1

Justin Huynh

Student ID: 012229514

M.S. Data Analytics

A. Research Question

My research question for this project is “Does a relationship exist between people who have a Master’s degree and people who have a longer contract?”

B: Description of Variables

Name	Data Type	Description	Example
CaseOrder	Qualitative	This acts as an index and is ordered numerically.	1
Customer_id	Qualitative	This is a unique identifier for all customers.	K409198
Interaction	Qualitative	This is the snippet for each interaction.	aa90260b-4141-4a24-8e36-b04ce1f4f77b
City	Qualitative	This is the city the customer resides in.	Point Baker
State	Qualitative	This is the state the customer resides in.	AK
County	Qualitative	This is the county the customer resides in.	Prince of Wales-Hyder
Zip	Qualitative	This is the zip code the customer resides in.	99927
Lat	Qualitative	The latitude of the customer's address.	56.251
Lng	Qualitative	The longitude of the customer's address.	-133.37571
Population	Quantitative	How large the amount of people are in that certain location.	38
Area	Qualitative	The customer's residence that are urban, rural, or suburban.	Urban
Timezone	Qualitative	This is the customer's timezone.	America/Sitka
Job	Qualitative	The customer's occupation.	Environmental health practitioner
Children	Quantitative	How many children do the customers have?	1
Age	Quantitative	How old is the customer?	68
Education	Qualitative	What is their highest level of education completed?	Master's Degree
Employment	Qualitative	The customer's current employment status.	Part Time
Income	Quantitative	How much does the customer make a year?	28561.99
Marital	Qualitative	This is the customer's marital status.	Widowed
Gender	Qualitative	This is the gender the customer self-identifies as.	Male
Churn	Qualitative	Did the customer churn?	No
Outage_sec_perweek	Quantitative	How long was the outage in seconds per week?	6.972566093
Email	Qualitative	This is the customer's email address.	10
Contacts	Qualitative	How many contacts do they have?	0
Yearly_equip_failure	Quantitative	How often does their equipment fail per year?	1
Techie	Qualitative	Are they considered a technical person?	No
Contract	Qualitative	How long is their contract on their service?	One year
Port_modem	Qualitative	This is whether the customer uses a port modem or not.	Yes
Tablet	Qualitative	This is whether the customer uses a tablet or not.	Yes
InternetService	Qualitative	This is the type of internet service the customer uses.	Fiber Optic
Phone	Qualitative	Does the customer have a phone number?	Yes
Multiple	Qualitative	This is if the customer has multiple lines of service.	No
OnlineSecurity	Qualitative	This is if the customer has security online.	Yes
OnlineBackup	Qualitative	This is if the customer has a back up service.	Yes
DeviceProtection	Qualitative	This is if the customer has protection for their device.	No
TechSupport	Qualitative	This is if the customer has technical support.	No
StreamingTV	Qualitative	This is if the customer has streaming television.	No
StreamingMovies	Qualitative	This is if the customer has streaming movie services.	Yes
PaperlessBilling	Qualitative	This is if the customer has paperless billing.	Yes
PaymentMethod	Qualitative	This is the type of payment method the customer uses.	Credit Card (automatic)
Tenure	Quantitative	The length of the service with a provider by months.	6.795512947

MonthlyCharge	Quantitative	How much is the customer getting charged per month?	171.4497621
Bandwidth_GB_Year	Quantitative	This is the bandwidth of the customer's gigabyte on average.	904.5361102
item1	Qualitative	The rating a customer gives for timely response with 1 being the most important and 8 being the least.	5
item2	Qualitative	The rating a customer gives for timely fixes with 1 being the most important and 8 being the least.	5
item3	Qualitative	The rating a customer gives for timely replacement with 1 being the most important and 8 being the least.	5
item4	Qualitative	The rating a customer gives for reliability with 1 being the most important and 8 being the least.	3
item5	Qualitative	The rating a customer gives for having different options with 1 being the most important and 8 being the least.	4
item6	Qualitative	The rating a customer gives for having respectful responses with 1 being the most important and 8 being the least.	4
item7	Qualitative	The rating a customer gives for courteous interactions with 1 being the most important and 8 being the least.	3
item8	Qualitative	The rating a customer gives for how well the helper listened to the customer's concerns with 1 being the most important and 8 being the least.	4

C1: Plan to Assess Data Quality in the Data Set

We're first going to detect any duplicate values by profiling our data set with .info() and then using the .duplicated() and .value_counts() functions to see whether we have any duplicate values we need to address. We'll then check for any missing values by using the isnull().sum() function to add up all the missing values per variable. We're then going to look at the histograms using matplotlib.pyplot to see what type of imputation we should use (mean, median, or mode functions). Lastly, we're using .boxplot() to find our outliers.

C2: Justification of Plan

We're using the given functions to look at all the information of our data set with .info() so we can use .duplicated() and .value_counts() to find the duplicates so we can clean them. Any missing values will be summed up with isnull().sum() so we can address them so we can properly plot the variables to identify any outliers using .boxplot(). This is all to get a more narrow perspective of our variables so we can evaluate our columns for our organizational objectives. The data we're assessing either has missing values, duplicates, or outliers once we've parsed them. We will only be

addressing the quantitative variables as they are more objective for comparisons and can be used in statistical modeling for the future.

C3: Justification of Programming Language, Libraries and Packages

I'm using Python as the preferred programming language due to the versatility and capabilities to acquire, clean and analyze data through various libraries and packages that make the analyses more efficient. I'm using the pandas library because it provides the data frame structure needed to store the CSV files and transform the data as needed. I'm also using matplotlib.pyplot to display the histograms to find any skews in the variables and seaborn display a boxplot to identify any outliers.

C4: Annotated Code to Assess Data Quality

```
In [67]: import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sb

         # importing the churn_raw_data.csv through the file path
         df = pd.read_csv('/Users/justinhuynh/Desktop/churn_raw_data.csv')
         # check all information about this file
         df.info()
```

```
In [63]: df.duplicated()
```

```
In [64]: print(df.duplicated().value_counts())
```

```
In [45]: df.isnull().sum()
```

```
In [7]: plt.hist(df['Children'])
        plt.show()
```

```
In [51]: boxplot=sb.boxplot(x='Children',data=df)
```

```
In [69]: # put outliers into new dataframe called children_query
         children_query = df.query('Children > 7')
         # look at number of outliers
         children_query.info()
```

See code attached in WGU_D206_Task_1.ipynb.

D1: Cleaning Findings

After cleaning the data, this is what we've found for data quality issues:

- Children had NA values
- Age had NA values
- Income had NA values
- Tenure had NA values
- Bandwidth_GB_Year had NA values
- Techie had NA values
- Phone had NA values
- TechSupport had NA values

D2: Justification of Mitigation Methods

For the missing NA values, we would use the .fillna() method to fill all the missing values by using either the .mean() method to replace it with the mean value or the .median() method to replace the missing values with the median value of that variable.

As for any outliers, we would use the .boxplot() to check for any outliers and how many are outside the range. For those that are outside the range, we'd use the df.query() method to filter the dataframe based on a condition to put the outliers into a new dataframe.

D3: Summary of Outcomes

The operations we performed will fill in all missing NA values as well as identify outliers and put them into a separate data frame, making it simpler to understand which variables had outliers and by how many. This makes the analyses of our data set a lot more efficient and able to identify the necessary values more quickly as opposed to manually looking for the missing values or outliers.

D4: Mitigation Code

```
In [7]: plt.hist(df['Children'])
        plt.show()
```

```
In [17]: df['Children'].fillna(df['Children'].median(), inplace=True)
```

```
In [50]: boxplot=sb.boxplot(x='Population',data=df)
```

```
In [68]: # put outliers into new dataframe called population_query
         population_query = df.query('Population > 3000')
         # look at number of outliers
         population_query.info()
```

See code attached in WGU_D206_Task_1.ipynb.

D5: Clean Data

A copy of the cleaned data will be submitted as ‘cleaned_data.csv’.

D6: Limitations

One of the limitations encountered during this process was subjectivity in cleaning decisions. An example is determining what to do with the outliers. I decided to retain the outliers and not discard them because I’d like to preserve the sample size and diversity of the dataset, which could be an opposing opinion for another analyst. Another limitation was data integrity and governance due to tracking changes and wanting the original data set to stay intact. This would potentially be a bigger problem with larger datasets. The risk of over cleaning was also a factor due to not knowing how far to take the data cleaning process and risking removing data.

D7: Impact of Limitations

In regards to my research question “Does a relationship exist between people who have a Master’s degree and people who have a longer contract?” and the impact of the limitations, the subjectivity of the data cleaning could change the outcome of the question if one decides to preserve the data vs removing outliers. Another impact would be if data integrity is affected and tracking changes were missed, it could also change the outcome drastically. Lastly, over cleaning would definitely impact the results of the research if we’ve removed too many necessary values due to not knowing our own limitations on when the data cleaning process is enough.

E1: Principal Components

```
In [90]: # store all quantitative variables in a new data frame for PCA
df_pca = df[['Children', 'Age', 'Income', 'Tenure', 'Population', 'Outage_sec_perweek', 'Yearly_equip_failure', 'Mon
# fill missing values with mean of each column
df_pca_filled = df_pca.fillna(df_pca.mean())
# normalize columns by subtracting the mean from the value and then dividing by the standard deviation
df_pca_normalized = (df_pca_filled-df_pca_filled.mean())/df_pca_filled.std()
# determine size of the PCA
pca = PCA(n_components=df_pca_filled.shape[1])
# put normalized data into PCA
pca.fit(df_pca_normalized)
# print data with PCA
pca_print = pd.DataFrame(pca.transform(df_pca_normalized), columns=["PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7",
# create dataframe for each component
pca_loadings = pd.DataFrame(pca.components_.T, columns = ["PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7", "PC8", "P
# print the component loadings
print(pca_loadings)
```

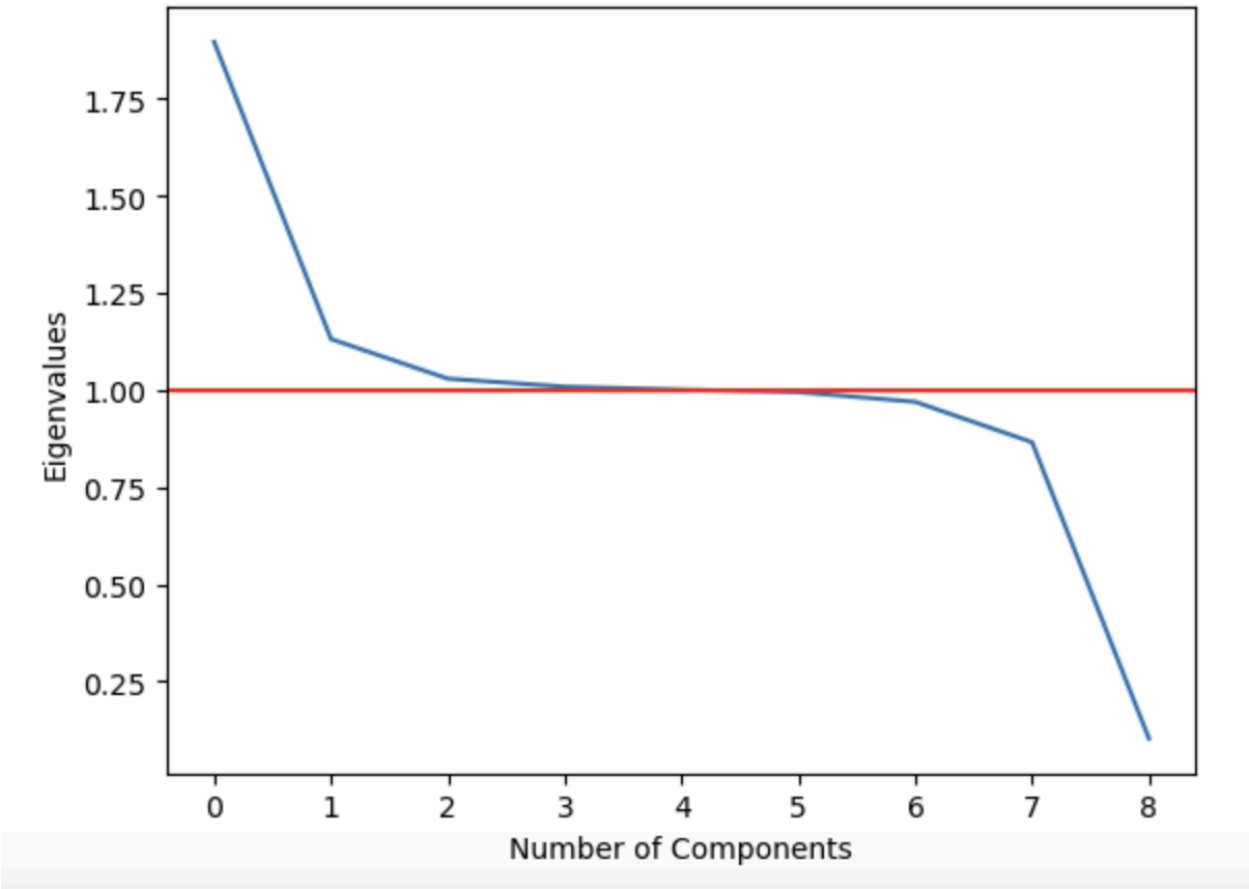
```
                           PC1       PC2       PC3       PC4       PC5  \
Children              0.000556 -0.005230  0.714106  0.001481 -0.082516
Age                  -0.012202 -0.043172 -0.506465  0.615788  0.217762
Income                0.005529 -0.000121  0.293582  0.116783  0.801022
Tenure                0.705104 -0.059268 -0.019660  0.009079  0.007342
Population            0.000099 -0.060970 -0.279857 -0.452772 -0.209469
Outage_sec_perweek    0.022728  0.706582  0.039716 -0.024357 -0.010723
Yearly_equip_failure  0.015615  0.062500  0.232900  0.632695 -0.507374
MonthlyCharge         0.045550  0.698283 -0.112914 -0.032801  0.051558
Bandwidth_GB_Year     0.706975 -0.010706  0.008902 -0.010356 -0.001505

                           PC6       PC7       PC8       PC9
Children              0.043079  0.693493 -0.001822 -0.020530
Age                   0.081950  0.541661  0.119533  0.021568
Income                0.447088 -0.235142 -0.057659  0.001180
Tenure               -0.004626 -0.000398  0.039331 -0.705141
Population            0.789262  0.214690 -0.007502 -0.000703
Outage_sec_perweek    0.069185 -0.039251  0.701151  0.000482
Yearly_equip_failure  0.403380 -0.326533 -0.120544 -0.002455
MonthlyCharge        -0.032227  0.126471 -0.688831 -0.048163
Bandwidth_GB_Year    -0.006669  0.011334 -0.012208  0.706797
```

See code attached in WGU_D206_Task_1.ipynb.

The principal components are PC1, PC2, PC3, PC4 and PC5.

E2: Criteria Used

Using the scree plot visualization, we've identified that PC1-PC5 are at or above the horizontal red line, which justifies the reduced number of PCAs.



See code attached in WGU_D206_Task_1.ipynb.

E3: Benefits

An organization would benefit the use of PCA by improving several factors of data analysis. One being the use of visualization such as a scree plot to make understanding trends and patterns a lot easier to interpret. Another benefit would be reducing the noise that comes from complex datasets. By eliminating redundancy and irrelevant values, cleaner datasets could be produced for more efficient analyses. Lastly, because of the reduction of irrelevant data, better decision making as well as improved strategic planning on the most impactful areas would be a great benefit to an organization.

F: Panopto Video

The URL link for the Panopto video can be found here. It will also be submitted in the Performance Assessment task submission and has been uploaded.

G: Sources of Third Party Code

1. "Export Jupyter Notebook to CSV." Retrieved from https://sourcetable.com/export-csv/jupyter-notebook#:~:text=The%20correct%20command%20is%20df,directory%20of%20your%20Jupyter%20Notebook.
2. WGU Courseware. Retrieved from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=19c24c56-0f37-408e-bb1f-b059002a77ac

H: Web Sources

1. "What is Principal Component Analysis (PCA)?" Retrieved from https://www.ibm.com/topics/principal-component-analysis
2. WGU Courseware. Retrieved from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95