D206: Data Cleaning

Part I: Research Question

A. Research Question

The telecommunications industry is a highly competitive industry where customers have the option to have one service provider, multiple service providers that they can choose between, and customers can cancel at any point in time. The term customer "churn" can be described as the number, more specifically the percentage, of customers who cancel a service during a specific time frame. The service providers are now reaching out to analysts to determine if there is a root cause and my job, as an analyst, is to find a solution to prevent customer churn from increasing. Therefore, I must clean the raw data using the "churn raw data.csv" file to be able to identify and explore my research question:

Are there specific characteristics of the customers that are cancelling their service, affecting customer churn? If so, what are those characteristics?

B. Variables & their data type

The data set contains information about 10,000 customers and there are 50 columns/variables. As mentioned in part A, the dependent variable being analyzed is customer churn from the "churn_raw_data.csv" file. I have listed the following variables, their data types, and a description in the table below that are also from the churn data set.

Variable Name	Data Type	Description	Example
CaseOrder	Quantitative	Number assigned to keep the order of the original data file	1
Customer_id	Qualitative	Unique ID assigned to identify the customers	K409198
Interaction	Quantitative	Unique ID assigned to identify the customer's transactions, sign-ups, and support needed from tech	aa90260b-4141-4a24-8e36-b04ce1f4f77b
City	Qualitative	The city the customer resides in according to the billing statement	Point Baker
State	Qualitative	The state the customer resides in according to the billing statement	AK
County	Qualitative	The county the customer resides in according to the billing statement	Prince of Wales-Hyder
Zip	Qualitative	The zip code the customer resides in according to the billing statement	99927
Lat	Quantitative	The GPS coordinates the customer resides in according to the billing statement	56.251
Lng	Quantitative	The GPS coordinates the customer resides in according to the billing statement	-133.37571
Population	Quantitative	According to the census data of the customer, the population within a mile radius	38
Area	Qualitative	The type of area the customer resides in	Urban
TimeZone	Qualitative	Customer's time zone according to customer's location given on the sign-up information	America/Sitka
Job	Qualitative	Customer's job according to the	Environmental health

		sign-up information	practitioner
Children	Quantitative	Number of kids in the household of the customer according to the sign-up information	1
Age	Quantitative	Customer's age according to the sign-up information	68
Education	Qualitative	Customer's degree earned according to the sign-up information	Master's Degree
Employment	Qualitative	Customer's current employment status according to the sign-up information	Part Time
Income	Quantitative	Customer's annual amount of income according to the sign-up information	28561.99
Marital	Qualitative	Customer's marital status according to the sign-up information	Widowed
Gender	Qualitative	Customer's gender, how the customer self-identifies	Male
Churn	Qualitative	If the customer cancelled their service within the past month	No
Outage_sec_perweek	Quantitative	Amount, in seconds, per week of outages in the customer's neighborhood	6.972566093
Email	Quantitative	In the past year, amount of emails customer was sent	10
Contacts	Quantitative	Amount the customer contacted technical support	0
Yearly_equip_failure	Quantitative	Amount the customer's equipment failed and needed to be fixed in the past year	1
Techie	Qualitative	If customer is tech-savvy according to questionnaire	No
Contract	Qualitative	Customer's contract term	One year
Port_modem	Qualitative	If the customer has a portable modem or not	Yes
Tablet	Qualitative	If the customer owns some type of tablet or not	Yes
InternetService	Qualitative	Internet provider customer has	Fiber Optic
Phone	Qualitative	If customer has a phone service	Yes
Multiple	Qualitative	If the customer has multiple phone lines	No
OnlineSecurity	Qualitative	If the customer has an add-on for online security	Yes
OnlineBackup	Qualitative	If the customer has an add-on for online backup	Yes
DeviceProtection	Qualitative	If the customer has an add-on for device protection	No
TechSupport	Qualitative	If the customer has an add-on for technical support	No
StreamingTV	Qualitative	If the customer has streaming TV	No
StreamingMovies	Qualitative	If the customer has streaming movies	Yes
PaperlessBilling	Qualitative	If the customer has paperless billing	Yes
PaymentMethod	Qualitative	The type of payment method the customer uses	Credit card (automatic)
Tenure	Quantitative	Number of months the customer has been with the provider	6.795512947
MonthlyCharge	Quantitative	Monthly charged to the customer	171.4497621
Bandwidth_GB_Year	Quantitative	Amount of data customer used yearly, in GB	904.5361102
Item1	Qualitative	Timely response, customer ranked the importance of each factor 1-8, 1 being the most important and 8 being the least	5
Item2	Qualitative	Timely fixes, customer ranked the importance of each factor 1-8, 1	5

		being the most important and 8 being the least	
Item3	Qualitative	Timely replacements, customer ranked the importance of each factor 1-8, 1 being the most important and 8 being the least	5
Item4	Qualitative	Reliability, customer ranked the importance of each factor 1-8, 1 being the most important and 8 being the least	3
Item5	Qualitative	Options, customer ranked the importance of each factor 1-8, 1 being the most important and 8 being the least	4
Item6	Qualitative	Respectful response, customer ranked the importance of each factor 1-8, 1 being the most important and 8 being the least	4
Item7	Qualitative	Courteous exchange, customer ranked the importance of each factor 1-8, 1 being the most important and 8 being the least	3
Item8	Qualitative	Evidence of active listening, customer ranked the importance of each factor 1-8, 1 being the most important and 8 being the least	4

Part II: Data Cleaning Plan

C1. Plan to Identify Anomalies

To identify the anomalies in the provided churn dataset, I will go by the following approach (Larose, 2019, p. 29-43):

1. Open the required packages and libraries using the import command.
2. Read the dataset into Python using the read_csv() command from *pandas* and the specified file path of the churn file.
3. Replicate the *churn* variable, naming it *churn_df* to prepare to replace the categorical values with numeric values and the data frame that has been sliced as *df*.
4. Explore the dataset in order to determine how to evaluate the input data by using the head() command to print the first 5 rows of the dataset.
5. Remove any duplicate columns that are clearly noticeable and less meaningful columns using the drop command.
6. Find any noticeable problems such as misspellings, any missing data, and variables/columns that need a more specific name. To rename columns, using the rename commands and then list out the original column name with the replacement columns name.
7. Identify if there are any missing values. To do so, use the following code to find the rows that contain missing values: df.isnull().any(axis=1) and to find the columns that contain missing values: df.isna().any().
8. If there is missing data, then it will say “True” beside the column name.
9. Verify the missing values by finding the number of records containing missing values in each column. I named this date the “Missing values distribution and used the following code: print(df.isnull().sum()).
10. Check the datatype in each column.
    * Numerical = int64 and float64; categorical = object

11. Inspect the categorical data for misspellings or outliers using the unique() command.
12. Double check to confirm all duplicates have been removed using the following code: df.loc[df.duplicated()] and print the columns.
13. To finally treat the missing values, use the imputation method by calculating the mean, median, and mode and replacing the missing values with one of the measures of central tendency calculated. To decide which central tendency value will replace the missing values, use histograms to determine its statistical distribution.
    *Asymmetrical and skewed = median; symmetrical = mean; categorical data = mode
14. Once the missing values have been replaced, find the count of missing values to double check there is not any. All values should be 0 if done correctly.
15. Identify any possible outliers that are outside of the reasonable range of values in the dataset using boxplots.
16. After isolating the outliers, determine how to treat the outlier; whether they should be kept, replaced, or removed from the dataset.
17. Extract the clean dataset.

C2. Approach for the Plan

As previously mentioned in section C1, I will be following the approach according to the book assigned called *Data Science Using Python and R* written by Chantal D. Larose and Daniel T. Larose. The definitions and concepts were much better explained in simpler terms in the videos made by Dr. Middleton. For example, there is missing data in some of the columns in the provided churn csv file. According to the "Getting Started with D206 Detecting and Treating Missing Values" video, missing data in Python will be shown as null, none, or NaN values. There are many methods that are used to treat the missing data, however the method I will be using for the performance assessment will be the univariate imputation method which involves calculating the central tendencies, creating a histogram, and analyzing the distribution in order to decide which central tendency to utilize. If normal distributions, use the mean; for skewed (right or left) distributions, use the median; for bimodal distributions, use the mode. This will be a good place to start in order to clean the data.

To identify outliers that are outside of the reasonable range of values in the dataset, I used boxplots. Boxplots highlight these extremities and make it visually easy to identify. After investigation, it will have to be determined if these outliers represent bad data that needs to be removed in order to move forward with a more accurate analysis.

There were also some obvious issues that were noticeable from simply looking at the first five rows. For example, there were duplicate columns that listed the number of rows. There were also columns that were not specific enough. Therefore, I had to rename those columns to describe what the column consisted of.

C3. Programming Language, Libraries, & Packages Used

I have chosen to use Python as my programming language of choice to complete the performance assessment. The reason I have decided to choose Python is because I have a little familiarity with it since I have been teaching myself the language for a little over a year. It is also one the most popular programming languages used in many data science-related jobs which is more like what I want to use in my future career fields, although I do plan on learning R later as well.

I downloaded Anaconda Navigator as the interactive development environment to use to launch Python programs. It provides Jupyter Notebook, an interactive web application for code, data, and notebooks, to annotate my code. There are so many libraries and packages that Python has to offer, but I will be using the following:

- Pandas = To load datasets
- NumPy = To work with arrays
- Sci-kit Learn = For machine learning
- SciPy = For mathematical problems
- Matplotlib = For basic plotting generally consisting of bars, lines, pies, scatter plots, and graphs
- Seaborn = For a variety of visualization patterns

C4. Annotated Code to Identify Anomalies

See attached "Data Cleaning PA.ipynb" - In [1] to out [30]

Part III: Data Cleaning

D1. Findings for the Data Quality Issues

- The column "Unnamed" was removed from the data frame because it is considered a duplicate column since the "CaseOrder" column has the exact same numbers to represent each customer. Additionally, I removed the "CaseOrder" column since it is irrelevant in identifying each customer since there is a unique customer ID that is assigned to each customer.
- The columns named "Item1" through "Item8" were renamed to better describe what the variables stand for.
- After doing the code to find the missing values in the dataset columns, the output shows that columns "Children", "Age", "Income", "Techie", "InternetService", "Phone", "TechSupport", "Tenure", and "Bandwidth_GB_Year" all contain Nas since missing values are shown in the output as "True".
- To be more specific, the following is how many records are missing in each column out of 10,000 records:
    - Children – 2,495
    - Age – 2,475
    - Income – 2,490
    - Techie – 2,477

    - InternetService – 2,129
    - Phone – 1,026
    - TechSupport – 991
    - Tenure – 931
    - Bandwidth_GB_Year – 1,021
- I created histograms for the quantitative variables with missing data to determine whether to replace the missing values with the mean or median depending on the symmetry of the graph. Therefore, “Children”, “Income”, “Tenure”, and “Bandwidth_GB_Year” had their missing data replaced with the median while the “Age” column was replaced with the mean.
- All the missing categorical data had their values replaced with mode. This included the “Techie”, “InternetService”, “Phone”, and “TechSupport” columns.
- After verifying there is no longer any missing values in the dataset, I created boxplots to check for outliers in the quantitative variables. There were outliers in the following columns: “Population”, “Children”, “Income”, “Outage_sec_perweek”, “Email”, “Contacts”, “Yearly_equip_failure”, “MonthlyCharge”, and “Bandwidth_GB_Year”. I came to the conclusion after looking at the statistics for those columns keep them in the dataset.

D2. Methods to Correct Findings

To treat the missing values, I used the imputation method to replace the values with either the mean, median, or mode. To determine which one to go with, I created histograms for only the quantitative variables and examine each of their statistical distributions. If the histogram is symmetrical, replace with the mean; if the histogram is asymmetrical/bimodal or skewed, use the median; and for categorical data, use the mode. To detect outliers, I used the boxplots because they give a clear visualization of the values that fall outside of the range. You can use matplotlib or seaborn for boxplots.

D3. Results of Cleaning the Data

The original churn csv file no longer has missing values in the dataset, irrelevant data, misspellings, and has more specific column names where necessary. The NAs in the columns "Children", "Age", "Income", "Techie", "InternetService", "Phone", "TechSupport", "Tenure", and "Bandwidth_GB_Year" were dealt with. The outliers in the quantitative variables were also analyzed and kept in the dataset. Now, we have a clean data file that we could use to answer the research question in part A.

D4. Annotated Code

See attached “Data Cleaning PA.ipynb”

D5. Copy of Cleaned Dataset

A copy of the cleaned dataset has been attached as "D206_churn_clean.csv" file. However, a preview can be found below. As a reminder, I removed the duplicate columns that listed the customers 1 to 10,000 and renamed the columns "item1" to "item8" to be more specific to what their columns consisted of.

	CaseOrder	Customer_id	Interaction	City	State	County	Zip	Lat	Lng	Population	...	MonthlyCharge	Bandwidth_GB_Year	Responses	Fixes	Replacements	Reliability	Options	Respectfulness	Courteous	Listening
0	1	K409198	aa90260b-4141-4a24-8e36-b04ce1f4f77b	Point Baker	AK	Prince of Wales-Hyder	99927	56.25100	-133.37571	38	...	171.449762	904.536110	5	5	5	3	4	4	3	4
1	2	S120509	fb76459f-c047-4a9d-8af9-e0f7d4ac2524	West Branch	MI	Ogemaw	48661	44.32893	-84.24080	10446	...	242.948015	800.982766	3	4	3	3	4	3	4	4
2	3	K191035	344d114c-3736-4be5-98f7-c72c281e2d35	Yamhill	OR	Yamhill	97148	45.35589	-123.24657	3735	...	159.440398	2054.706961	4	4	2	4	4	3	3	3
3	4	D90850	abfa2b40-2d43-4994-b15a-989b8c79e311	Del Mar	CA	San Diego	92014	32.96687	-117.24798	13863	...	120.249493	2164.579412	4	4	4	2	5	4	3	3
4	5	K662701	68a861fd-0d20-4e51-a587-8a90407ee574	Needville	TX	Fort Bend	77461	29.38012	-95.80673	11352	...	150.761216	271.493436	4	4	4	3	4	4	4	5

D6. Limitations

There were quite a few cons I noticed while treating the data. I have listed them below:

- The first con in this situation would be the inconvenience of not being able to contact the telecommunications company about the missing data in the file that they provided. In a real-life scenario, I would be able to contact them to ask why it wasn't provided and if it is relevant to keep regarding the question they are wanting me to answer.
- If I were to use the imputation method, the limitation would be that it could possibly manipulate the data and give an inaccurate distribution of the data.
- According to the provided scenario, there are many columns based on the customer's sign-up information such as the job of the customer, number of the children in the customer's household, etc. That means another limitation could be not having updated customer information which could also give inaccurate results.

D7. How Limitations Affect the Analysis of the Question

As mentioned in part D6, the limitations ultimately affect the analysis of the question by not being provided with accurate, updated information and missing data from the telecommunications company. As a result, this could affect the results that I come up with to answer their question.

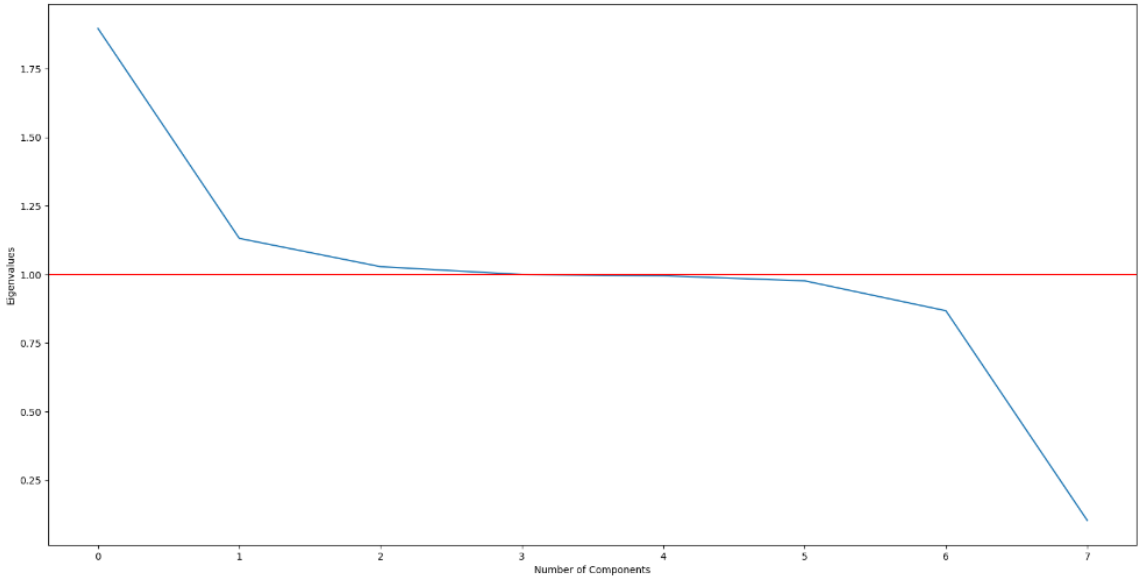
E1. Total Number of Principal Components & Output

I used 8 quantitative, continuous variables. The variables I used for the Principal Component Analysis were 'Population', 'Outage_sec_perweek', 'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure', 'MonthlyCharge', and 'Bandwidth_GB_Year'. The output can be found below:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Population	-0.000437	-0.053644	0.515655	-0.425926	0.515193	0.533274	-0.001708	-0.000784
Outage_sec_perweek	0.022311	0.705254	-0.011774	-0.042591	0.042696	0.009342	0.705873	0.000145
Email	-0.021184	0.058663	0.637311	-0.248775	-0.093277	-0.719096	-0.047164	0.005641
Contacts	0.004731	0.001993	0.257825	0.800978	0.526818	-0.118255	0.020190	-0.002188
Yearly_equip_failure	0.015966	0.057290	-0.509243	-0.331953	0.665998	-0.410746	-0.121116	-0.002565
Tenure	0.705063	-0.057060	0.014936	-0.007196	-0.010416	-0.016213	0.035530	-0.705494
MonthlyCharge	0.045318	0.699757	0.040315	0.055431	-0.052714	0.124348	-0.695006	-0.048099
Bandwidth_GB_Year	0.706829	-0.009790	0.012085	-0.000614	-0.008625	-0.003247	-0.011976	0.707051

E2. Reduced Number of Principal Components

I identified the Principal Components by using visualization and creating a scree plot to identify which PCs to keep. To determine which components to keep, you would examine the graph and only keep the values greater than 1. According to the graph, only PC1 through PC3 would be kept. The other values would be discarded since those values are less than 1. I also printed the eigenvalues from the graph for a more accurate interpretation since the exact values that are plotted are given. A screenshot of scree plot graph and eigenvalues can be found below:



E3. Benefits of using PCA

Principal Component Analysis only uses quantitative variables. In this scenario, this is already beneficial considering there are very few quantitative variables versus the many qualitative variables provided in the given csv file. PCA reduces the dimensionality of a larger data set while keeping the important data. It also helps to identify if there are any patterns present in the data, which are then grouped based on their relations to one another. The telecom business can use PCA to determine what quantitative variables are affecting churn. Regarding the research question based on our PCA results, the relationship between the tenure and the amount of data customer used yearly in GB, for example, seems to be a positive one. The analysis shows that customers who use more GB yearly are customers who have stuck around the business the longest to give just one example.

Part IV: Supporting Documents

F. Panopto Video

G. Web Sources

N/A

H. Sources

Chantal D. Larose, Daniel T. Larose. Data Science Using Python and R. Wiley; 2019. Accessed December 21, 2023. https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2091371&site=eds-live&scope=site

Middleton, Keiona. Panopto. Retrieved December 24, 2023, from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=767749d2-ba19-4f94-bec8-b058017b2f5e.

I. Professional Communication

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