# D206 Data CLEANING

**Research Question**

Our churn data set consist of communication customers information related to their demographic, customer satisfaction, services, and whether they are current customers or no longer customers. Today we’d like to know is there a relationship between customer demographics, monthly charges, tenure, usage, and customer churn? This question’s goal is to investigate what factors such as customer age, location, tenure, gender, contract type, usage, and customer satisfaction levels are correlated with customer churn. By analyzing and considering all variables we can explore potential patterns and identify factors that might contribute to customers leaving our company. To promote data consistency and usability all of variables, variable data types, and variable descriptions are provided below in the data dictionary. It will serve as a reference document for the dataset.

**Required Variables**

Our dataset is comprised of 10,000 rows and 52 columns. I’ll give a description of each variable, it’s data type and an example from the dataset to provide insight. The column names and datatypes represent the cleaned dataset.

|  |  |  |
| --- | --- | --- |
| **Column** | **Datatype** | **Example** |
| Unnamed: 0 | integer | Column Index Number: 0,1,2,3 |
| CaseOrder | string | Order of each case: 1,2,3 |
| Customer\_id | string | Customers Unique ID: D90850, K191035 |
| Interaction | string | Unique Interaction ID: fb76459f-c047-4a9d-8af9-e0f7d4ac2524 |
| City | string | Customer’s city: Del Mar, Yamhill, Needville |
| State | string | Customer’s state: GA, OK, TN |
| County | string | Customer’s county: Peach, Scott, Oklahoma |
| Zip | string | Customer’s zip: 31030, 37847,73109 |
| Lat | float | Customer’s location in Latitude: 32.57032, 35.43313 |
| Lng | float | Customer’s location in Longitude: -133.37571, -83.8904 |
| Population | integer | Customer’s city population: 23144, 2535 |
| Area | string | Type of area: Rural, Urban, Suburban |
| Timezone | string | Timezone customer lives in: America/Chicago, America/New\_York |
| Job | string | Customer’s Job: Applications developer, Broadcast presenter |
| Children | float | Number of children: 0, 7, 2 |
| Age | float | Customer’s age: 50, 27, 83 |
| Education | string | Customer’s education level: Regular High School Diploma, Master’s Degree |
| Employment | string | Customer’s employment status: Retired, Student, Full Time |
| Income | float | Customer’s income: 21704.77, 28561.99 |
| Marital | string | Customer’s marital status: Widowed, Married, Never Married |
| Gender | string | Customer’s gender: Female, Male |
| Churn | string | Churn status: Yes, No |
| Outage\_sec\_perweek | float | Outages in seconds per week: 44.72520233, 8.15350077 |
| Email | integer | Number of emails: 20, 16, 10 |
| Contacts | integer | Number of contacts: 2,3,0 |
| Yearly\_equip\_failure | integer | Yearly equipment failures: 0, 1, 2 |
| Techie | string | Technical savvy status: NA, No, Yes |
| Contract | string | Contract type: Month-to-month, One year |
| Port\_modem | string | Port modem status: Yes, No |
| Tablet | string | Tablet owner: Yes, No |
| InternetService | string | Internet service type: Fiber optic, DSL, None |
| Phone | string | Phone service status: Yes, No |
| Multiple | string | Multiple lines status: Yes, No |
| OnlineSecurity | string | Online security status: No, Yes |
| OnlineBackup | string | Online backup status: Yes, No |
| DeviceProtection | string | Device protection status: Yes, No |
| TechSupport | string | Tech support status: Yes, No |
| StreamingTv | string | Streaming tv status: Yes, No |
| StreamingMovies | string | Streaming movies status: Yes, No |
| PaperlessBilling | string | Paperless billing status: Yes, No |
| PaymentMethod | string | Payment method: Electronic Check, Mailed Check, Credit Card(automatic) |
| Tenure | integer | Tenure in months: 9.563362997, 6.732948946 |
| MonthlyCharge | float | Monthly service charge: 142.3770231, 116.095233 |
| Bandwidth\_GB\_Year | float | Bandwidth GB usage yearly: 1324.330108 |
| Timely response | string | Timely response customer satisfaction ratings: 1-7 |
| Timely fixes | string | Timely fixes customer satisfaction ratings: 1-7 |
| Timely replacements | string | Timely replacements customer satisfaction ratings: 1-8 |
| Reliability | string | Reliability customer satisfaction ratings: 1-7 |
| Options | string | Options customer satisfaction ratings: 1-7 |
| Respectful response | string | Respectful response customer satisfaction ratings: 1-8 |
| Courteous exchange | string | Courteous exchange customer satisfaction ratings: 1-7 |
| Evidence of active listening | string | Evidence of active listening customer satisfaction ratings: 1-8 |

**Plan to Assess Quality of Data**

My plan in assessing the data begins with data exploration. I’ll load the churn dataset into a Jupiter notebook using pandas ‘pd.read\_csv(‘churn\_raw\_data.csv’) to begin this phase then use ‘.head()’ to get a quick look at the first five rows of data, making sure it loaded in correctly. I’ll gather information on the variables, dataset size, and structure of the dataset. I’ll use ‘.info()’ to get an understanding of data types, and missing values in each column. I’ll use ‘.describe()’ to get statistical descriptions for the columns in the dataset and ‘.dtypes’ to see all columns and data types by themselves. The next step is to validate data types. I need to verify each variable is the correct data type and make changes if they are labeled incorrectly. To do this, I can use the output of ‘.head()’, and the output of ‘.info()’. I want to verify that categorical information isn’t being represented as numbers and numerical data is the correct data type. I then proceed to check consistency in my variables making changes were needed. Misspellings and column name updates will be addressed using ‘.rename(columns=rename\_columns, inplace=True)’ and numerical consistency will be addressed by using ‘.round()’ where needed. I’ll then move on to identifying missing values to verify if there is enough data to move forward with my assessments. I’ll use ‘.isnull().sum()’ to give me the sum of missing values in each column. In the event of missing data, I’ll then use ‘.isnull().sum() / len() \* 100’ to get the percentage of data missing from each column. I’ll have to decide on how to handle the missing values. If there are reasons why the data may be missing, I may impute the missing values or if there is too much missing data, I may delete the column all together since there’s so much data unavailable. In this case no column is missing over 25% of its values, so I decided to impute the missing values using ‘.mean()’ for the numerical columns and ‘.mode()’ for categorical columns. Once I’ve gotten this far in the data, I’ll move on to checking for duplicates using ‘.duplicated().sum()’. This will give me the number of duplicated rows in my dataset. Fortunately, we have no duplicates in our dataset, this is an import step because redundant data could result in inaccurate analysis. Identifying outliers in my variables will be my next step since outliers can influence the results of my analysis. I like to create boxplots to get a visual of how the outliers look in each column and from there I create a function that uses the quartile range to separate outliers and replace them with the median().

**Justification of Approach**

My approach on assessing the data allows thorough review of the dataset in a systematic way. In data exploration I need to get an understanding of the dataset. My first steps taken using ‘.info()’, ‘.describe()’, and ‘.head()’ provide a snapshot of the data types, dataset structure, and a quick look at any unexpected values in the data. I also get a look at the descriptive statistics of the dataset which allow me to identify basic patterns and if there may be a skewed distribution. Duplicates in data can cause misleading results so identifying and removing duplicates ensures I have a unique dataset that promotes data integrity. Missing values can significantly affect the quality of my analysis, so I decided to impute the values, so I have a complete dataset. Incorrect data types can cause errors, categorical data types being represented as numerical can affect your analysis, so for data accuracy it is best I verify data types are correct and make changes if not. I chose to replace outliers because they can skew results and lead to incorrect conclusions. Detecting and handling outliers in data help maintain data quality. It’s better to work with normal data rather than dealing with extreme outliers. My methods used to assess the data provide a more reliable dataset as a result.

**Justification of Tools**

Python will be the programming language I will use for my data analysis. Python offers me ease of use and the libraries needed to fully analyze the churn dataset. Pandas, NumPy, Matplotlib, and Seaborn to help me perform my analysis. Pandas has functions and methods for handling duplicates, outliers, missing values, and data types. NumPy helps with numerical data operations and has arrays that come in handy for data manipulation. Matplotlib and Seaborn will allow me to provide data visualization helping illustrate outliers and data distribution. Scikit-learn, I’ll use to standardize, scale the data, encode, transform, pipeline, and do the principal component analysis.

**Cleaning Findings**

Once I read the churn data file into Python and began the cleaning process, I found there are 2495 missing values in the ‘Children’ column, 2475 missing values in the ‘Age’ column, 2490 missing values in the ‘Income’ column, 2477 missing values in the Techie column, 1026 missing values in the ‘Phone’ column, 991 missing values in the ‘TechSupport’ column, 931 missing values in the ‘Tenure’ column, and 1021 missing values in the ‘Bandwidth\_GB\_Year’ column. I moved on to checking for duplicates using duplicated().sum(). There were no duplicates found in the dataset. There are inconsistencies in the column naming regarding the survey responses listed as item1 – item8. I decided to change the names to reflect the description of the survey question. 'item1' is now 'Timely response', 'item2': 'Timely fixes’ 'item3': 'Timely replacements', 'item4': 'Reliability', 'item5': 'Options', 'item6': 'Respectful response', 'item7': 'Courteous exchange', 'item8': 'Evidence of active listening'. The data types for each survey questions were numerical. I decided to change each to categorical due to the context of how it’s being used in the dataset. CaseOrder and Zip were numerical data types, but needed to be changed to categorical, we don’t want to use those for computation. In the categorical columns there were inconsistencies in capitalization. Example, the ‘Job’ column would capitalize ‘Chief Executive Office’ but wouldn’t for a job like ‘Social research officer’. I decided to make all categorical values lowercase to promote consistency. I also found outliers in the dataset for the numerical columns. ‘Population’, ‘Children’, ‘Income’, ‘Outage\_sec\_perweek’, ‘Email’, ‘Contacts’, ‘Yearly\_equip\_failure’, and ’MonthlyCharge’ all had outliers that could possible skew analysis. The column ‘Unnamed: 0’ also had no relevance to the data besides indexing.

**Justification of Mitigation Methods**

When dealing with missing values in the dataset I decided to use mean imputation, because it uses a central value and minimizes bias. When dealing with my outlier columns I went with median imputation because it is less affected by outliers and skewed distributions. Mode imputation, I used for my categorical columns with missing values. I used mode because it uses the most common characteristic in the data and keeps the distribution. The choice on how to deal with outliers came down to the IQR method and the Z-Score method. I chose to use Interquartile Range(IQR) because it’s a versatile tool for detecting outliers in various distributions. I converted certain column data types to improve data integrity and prevent mathematical calculations that are not meaningful for those columns. I decided to change column names for the ‘item’ column series and lowercase the values of the dataset to promote readability and consistency amongst the data. No duplicates were found in the dataset but if there were any duplicates they should be removed because they can make analysis biased and removing them gives a unique value for each instance in the dataset. The ‘Unnamed: 0’ isn’t meaningful in the dataset, so to focus on only the meaningful data I decided to remove it.

**Summary of The Outcomes**

Transforming the ‘item’ columns, ‘CaseOrder’ and ‘Zip’ to str/objects in Python prevented unintended numerical operations and ensured the columns are treated as categorical. The mean imputation for missing values ensures I have complete data without introducing extreme values. Median imputation for outliers provided a central value so the data is less affected by outliers and skewed distribution. Mode imputation in my categorical columns gives my missing values the most frequent values and my outliers the most frequent value to maintain consistency. Using the IQR method for outliers ensured the dataset still has variability without being skewed by extreme values. Checking for and removing duplicates makes sure each observation in the data is unique and maintains data integrity. Making each columns data lowercase created uniformity in the values and improved data consistency and checking for column relevance allowed me to decide to remove the ‘Unnamed: 0’ column because it has no value in analysis. These cleaning steps create a quality dataset ready for analysis.

**Limitations**

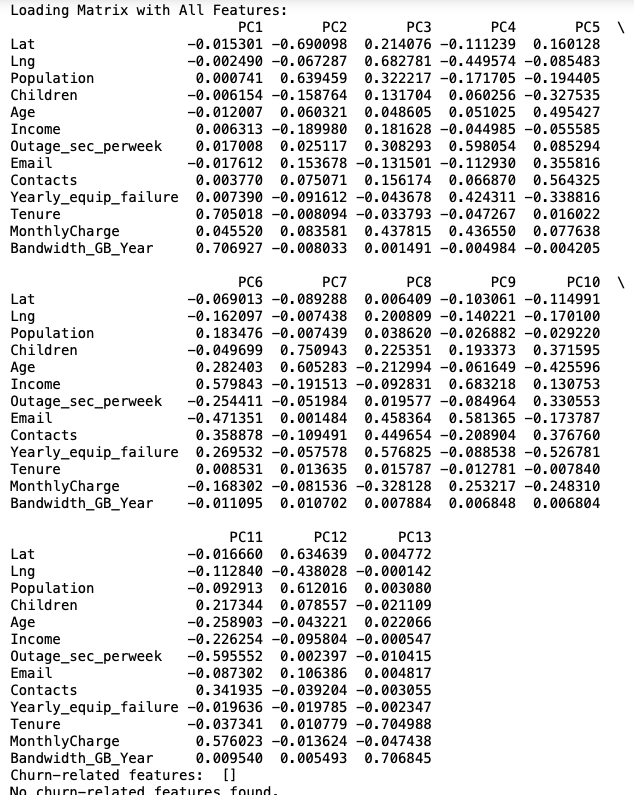
When dealing with missing values mean imputation can introduce bias in a non-normally distributed dataset. It doesn’t account for relationships between variables. Median imputation is good for outliers, but it still can miss the relationships between variables and introduce bias. Mode imputation is more suitable for categorical variables, but it can lead to over representation of the most frequent value which could distort the data. Changing the data type of variables from numerical to categorical could result in the loss of valuable information so careful consideration is needed when making this decision because it has an impact on analysis. Using the IQR method for outliers runs the risk of removing extreme values in the data that are valid and aren’t one off situations. The limitations of removing duplicates stand out if repeat occurrences are valid. Data consistency is important but in some instances distinctions in spelling or punctuation may indicate different context and that’s something you must be aware of when trying to ensure data consistency. The limitations of removing columns comes down to domain knowledge. The overall biggest limitation in dealing with this dataset is you must be aware of context, imputation assumptions, and the potential for bias.

**Impact of Limitations**

The impact of these limitations is reduced accuracy, bias, and loss of variability. Imputing data can lead to biased estimates, loss of information, distorted distribution, and it doesn’t account for the relationship of variables. Incorrectly labeling data types and column names results in inappropriate analysis and incorrect conclusions. The IQR method could classify too many data points as outliers that could result in substantial data loss. When dealing with duplicates, the incorrect removal could lead to inconsistencies and data loss. Removing columns for relevance or over simplifying data categories for consistency could result in the loss of insight that is crucial to analysis.

**Principal Components**

I have provided a screenshot of the loading matrix for the principal components and scree plot below. The loading matrix consist of eleven features specifically related to churn. We started with fifty-two columns but with the use of PCA, the scree plot, and loading matrix we can reduce the dimensionality of the dataset without losing important information. The first eleven or 12 principal components provide most of the variance in the data. The first eleven components explain about 92.81% of total variance and the first twelve components explain about 99.21% of the total variance.



**A graph with a line

Description automatically generated**

A graph of a graph with a line

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated

**Benefits**

Principal Component Analysis (PCA) benefits the organization greatly when analyzing churn. It simplifies the dataset by reducing dimensionality while maintaining variance. This provides a more manageable and easier dataset to work with that saves time and resources. The key components are identified so insights and patterns are easier to find and this helps stakeholders in important decision making. The removal of unimportant features allows quicker analysis and visualizations which allows the company to make changes in key areas to reduce churn. In summary PCA gives the organization a streamlined analysis process, improved accuracy and efficiency during analysis, and better insight when deciding retention strategies.

**References:**

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA) accessed 09 June 2024, https://apps.cgp-oex.wgu.edu.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

Scikit-learn, accessed 09 June 2024, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

W3schools.com accessed, 09 June 2024, https://www.w3schools.com/python/numpy/numpy\_array\_iterating.asp.

W3schools.com accessed, 09 June 2024, https://www.w3schools.com/python/pandas/default.asp.