**Data Cleaning**

Ryan S. Russon

ID: 010337779

11/11/2023

*Western Governors University*

**Part I: Research Question**

*A. Research Question/Decision*

What factors contribute most to customer monthly charge in our telecom service? Understanding these factors and how they contribute to monthly charges for customers can help in developing strategies to improve business revenue.

*B. Description of Variables*

The dataset contains the following variables:

CaseOrder (Integer): Unique order number. 3.

Customer\_id (String): Unique identifier for each customer. K409198.

Interaction (String): Identifier for customer interactions. Aa90260b-4141-4a24-8e36-b04ce1f4f77b.

City (String): Geographic location details. Point Baker.

State (String): Geographic location details. AK.

County (String): Geographic location details. Ogemaw.

Zip (Integer): Postal code. 99927.

Lat (Float): Latitude coordinates. 56.251.

Lng (Float): Longitude coordinates. -133.37571.

Population (Integer): Population in the customer's area. 10446.

Area (String): Type of area (Urban/Rural). Urban.

Timezone (String): Time zone of the customer's location. America/Sitka.

Job (String): Customer's occupation. Surveyor.

Children (Integer): Demographic details. 4.

Age (Integer): Demographic details. 68.

Education (String): Highest education achieved. Master’s Degree.

Income (Float): Demographic details. 28561.99.

Marital (String): Marital status. Widowed.

Gender (String): Gender. Male.

Churn (String): Whether the customer churned or not. No.

Outage\_sec\_perweek (Float): Average weekly outage duration in seconds. 10.2179.

Email (Integer): Contact’s email count. 10.

Contacts (Integer): Contact count. 3.

Yearly\_equip\_failure (Integer): Equipment failure details. 1.

Techie (String): Service related data. Yes.

Contract (String): Contract length. One year.

Port\_modem (String): Port option. Yes.

Tablet (String): Tablet use. Yes.

InternetService (String): Specific internet used. DSL.

Phone (String): Yes/No phone available. Yes.

Multiple (String): Yes/No. No.

OnlineSecurity (String): Service-related features. Yes.

OnlineBackup (String): Back-up availability. No.

DeviceProtection (String): Protection available. Yes.

TechSupport (String): Tech support available. No.

StreamingTV (String): Streaming option. Yes.

StreamingMovies (String): Streaming (movie) option. No.

PaperlessBilling (String): Paperless billing. Yes.

PaymentMethod (String): Payment option. Credit Card.

Tenure (Float): Tenure or time as member. 1.156681.

MonthlyCharge (Float): Monthly charge. 242.948.

Bandwidth\_GB\_Year (Float): Data used yearly. 1948.694.

item1 (Integer): Uknown item. 3.

item2 (Integer): Uknown item. 5.

item3 (Integer): Uknown item. 6.

item4 (Integer): Uknown item. 4.

item5 (Integer): Uknown item. 3.

item6 (Integer): Uknown item. 5.

item7 (Integer): Uknown item. 4.

item8 (Integer): Uknown item. 2.

**Part II: Data-Cleaning Plan**

*C1. Methods for Detecting Data Quality Issues*

Duplicates: Use pandas.DataFrame.duplicated() to find duplicate rows. This is crucial to ensure each data point is unique and avoid skewing the analysis.

Missing Values: Employ pandas.DataFrame.isnull() or .isna() methods to identify missing values in the dataset. This is essential for understanding the extent of the missing data.

Outliers: For numerical data, use statistical methods like Z-score (scipy.stats.zscore) and IQR (Interquartile Range). Graphical methods like boxplots (using Matplotlib/Seaborn) can also be helpful.

Re-expression of Categorical Variables: Examine categorical variables for inconsistencies using pandas.Series.value\_counts(). For re-encoding medical condition variables, pandas.DataFrame.replace() or pandas.get\_dummies() can be used for consistent encoding.

Data Type Validation: Use pandas.DataFrame.info() or .dtypes to ensure correct data types.

Range Validation: For numerical data, simple conditional checks can highlight values outside expected ranges.

*C2. Justification for Methods Used*

Duplicates: Detecting duplicates is crucial for data integrity. Duplicate entries can lead to erroneous analysis outcomes.

Missing Values: Understanding the pattern of missing data helps in choosing appropriate imputation techniques.

Outliers: Outliers can significantly impact statistical analyses and model performance. Identifying them is crucial for robust data analysis.

Re-expression of Categorical Variables: Ensuring consistent encoding of categorical variables is important for accurate classification and analysis, especially in medical datasets.

Data Type and Range Validation: Ensures data is in a correct format and within normal bounds, which is critical for accurate calculations and analyses.

*C3. Programming Language, Libraries, and Packages*

Python: Chosen for its simplicity, readability, and extensive ecosystem of data science libraries. It's particularly well-suited for data manipulation and analysis.

Libraries and Packages:

Pandas: For data manipulation and cleaning.

NumPy: For numerical computations and specific data structuring.

Matplotlib and Seaborn: For data visualization, crucial for outlier detection.

Scipy: For statistical functions, such as computing Z-scores.

Python's popularity and extensive community support make it a good choice for data cleaning tasks.

*C4. Detection Code*

*See attached Jupyter Notebook*

**Part III: Data Cleaning**

*D. Data-Cleaning Process*

Findings from Data Quality Assessment:

Potential missing values.

Possible outliers in numerical columns like 'Income', 'Outage\_sec\_perweek'.

Inconsistencies in categorical data (e.g., 'Area', 'Job').

Methods for Mitigating Issues:

Imputation for missing values (mean/median/mode or KNN imputation).

Outlier treatment using IQR or Z-score methods.

Standardization of categorical variables.

Outcome Summary:

Reduced missing data impacting analysis.

Controlled impact of outliers on statistical analysis.

Consistent and clean categorical data.

Annotated Code for Data Cleaning:

Includes imputation, outlier treatment, and data standardization steps.

Cleaned Data Set:

The CSV file will contain the cleaned dataset.

Limitations:

Imputation might introduce bias.

Outlier removal might lead to loss of valuable information.

Effect of Limitations on Analysis:

Biased imputation may skew analysis results.

Removing outliers could oversimplify the model, missing key insights.

E. Principal Component Analysis (PCA)

Total Number of Principal Components and Loading Matrix:

Code to perform PCA and display the loading matrix.

Justification for Reduced Components and Scree Plot:

Scree plot to determine the number of components capturing significant variance.

Benefits of PCA:

Reduces dimensionality, highlighting key features.

Helps in understanding the underlying structure of the data.