D206 PA

Jon Isak

D206: Data Cleaning

Eric Straw

Contents

[D206 PA 3](#_Toc149690486)

[Part I: Research Question 3](#_Toc149690487)

[Part II: Data-Cleaning Plan 7](#_Toc149690488)

[Part III: Data Cleaning 8](#_Toc149690489)

[Part IV. Supporting Documents 16](#_Toc149690490)

D206 PA

A. What customer satisfaction factors relate to customer churn rate? This question allows the company to identify which factors are most important to customer satisfaction and lead to customer churn.

B.  The data comes from a telecommunications company and the data set contains 10k customers and 50 columns. The variables are:

* CaseOrder: Placeholder variable (Quantitative) (Example: 1,2)
* Customer\_id: Identifier for each customer (Qualitative) (Example: K409198).
* Interaction, UID: Unique identifiers associated with customer transactions, technical support, and sign-ups (Qualitative) (Example: aa90260b-4141-4a24-8e36-b04ce1f4f77b)
* City: Customer's city of residence as listed on the billing statement (Qualitative) (Example: Point Baker)
* State: Customer's state of residence as listed on the billing statement (Qualitative) (Example: AK)
* County: Customer's county of residence as listed on the billing statement (Qualitative) (Example: Prince of Wales-Hyder)
* Zip: Customer's zip code of residence as listed on the billing statement (Qualitative) (Example: 99927)
* Lat, Lng: Geographic coordinates (latitude and longitude) of customer residence as listed on the billing statement (Quantitative) (Example: 56.251, -133.37571)
* Population: Population within a one-mile radius of the customer, based on census data (Quantitative) (Example: 38)
* Area: Type of area based on census data (Qualitative) (Example: rural, urban, suburban)
* TimeZone: Time zone of customer residence based on sign-up information (Qualitative) (Example: America/Sitka)
* Job: Occupation of the customer (or invoiced person) as reported in sign-up information (Qualitative) (Example: Environmental health practitioner)
* Children: Number of children in the customer's household as reported in sign-up information (Quantitative) (Example: 1)
* Age: Customer's age as reported in sign-up information (Quantitative) (Example: 68)
* Education: Highest educational degree earned by the customer as reported in sign-up information (Qualitative) (Example: Master's Degree)
* Employment: Employment status of the customer as reported in sign-up information (Qualitative) (Example: Part Time)
* Income: Annual income of the customer as reported at the time of sign-up (Quantitative) (Example: 28561.99)
* Marital: Marital status of the customer as reported in sign-up information (Qualitative) (Example: Widowed)
* Gender: Customer's self-identified gender (Qualitative) (Example: male, female, nonbinary)
* Churn: Whether the customer discontinued service within the last month (Qualitative) (Example: Yes, No)
* Outage\_sec\_perweek: Average weekly duration of system outages in the customer's neighborhood (Quantitative) (Example: 6.972566093)
* Email: Number of emails sent to the customer in the past year (marketing or correspondence) (Quantitative) (Example: 10)
* Contacts: Number of times the customer contacted technical support (Quantitative) (Example: 1)
* Yearly\_equip\_failure: Number of times the customer's equipment failed and required resetting or replacement in the past year (Quantitative) (Example: 1)
* Techie: Whether the customer considers themselves technically inclined (based on customer questionnaire when signing up for services) (Qualitative) (Example: yes, no)
* Contract: Customer's contract term (Qualitative) (Example: month-to-month, one year, two years)
* Port\_modem: Whether the customer has a portable modem (Qualitative) (Example: yes, no)
* Tablet: Whether the customer owns a tablet such as an iPad, Surface, etc. (Example: yes, no) (Qualitative)
* InternetService: Customer's internet service provider (Qualitative) (Example: DSL, fiber optic, None)
* Phone: Whether the customer has phone service (Qualitative) (Example: yes, no)
* Multiple: Whether the customer has multiple lines (Qualitative) (Example: yes, no)
* OnlineSecurity: Whether the customer has an online security add-on (Qualitative) (Example: yes, no)
* OnlineBackup: Whether the customer has an online backup add-on (Qualitative) (Example: yes, no)
* DeviceProtection: Whether the customer has device protection add-on (Qualitative) (Example: yes, no)
* TechSupport: Whether the customer has a technical support add-on (Qualitative) (Example: yes, no)
* StreamingTV: Whether the customer has streaming TV (Qualitative) (Example: yes, no)
* StreamingMovies: Whether the customer has streaming movies (Qualitative) (Example: yes, no)
* PaperlessBilling: Whether the customer has paperless billing (Qualitative) (Example: yes, no)
* PaymentMethod: The customer's payment method (Qualitative) (Example: electronic check, mailed check, bank (automatic bank transfer), credit card (automatic))
* Tenure: Number of months the customer has been with the provider (Quantitative) (Example: 6.795512947)
* MonthlyCharge: The monthly charge amount for the customer (Quantitative) (Example: 171.4497621)
* Bandwidth\_GB\_Year: The average data usage in gigabytes (GB) per year by the customer (Quantitative) (Example: 904.5361102)
* Item1: Evaluation of timely response (Qualitative) (Example: 1)
* Item2: Evaluation of timely fixes (Qualitative) (Example: 1)
* Item3: Evaluation of timely replacements (Qualitative) (Example: 1)
* Item4: Evaluation of service reliability (Qualitative) (Example: 1)
* Item5: Evaluation of available options (Qualitative) (Example: 1)
* Item6: Evaluation of respectful response (Qualitative) (Example: 1)
* Item7: Evaluation of courteous exchange (Qualitative) (Example: 1)
* Item8: Evaluation of evidence of active listening (Qualitative) (Example: 1)

C1. To assess the quality of the data, I will first read it into a Pandas DataFrame using pd.read\_csv(). Then, I will use the describe() method to generate a summary of the data, including the mean, count, standard deviation, minimum, and maximum values of each column. I will also use info() to verify data types. I will also use duplicated() to check for duplicates within the data. Outliers will be detected using quartile() and boxplot() will be used to visualize.

C2.  The describe() technique will allow me to understand the structure, skew of the data, and identify potential outliers. Info() will allow me to compare the data types to what’s expected by the dictionary and identify missing values. Duplicated() will make sure data integrity is maintained by checking if data is duplicated. These methods will allow me to gain insight into the data, assess the data quality and what needs to be done to clean it. The quartile technique will allow me to find data that’s out of the boundary, and the box plot will be to see how the outliers compare to other data.

C3.   Python will be used to assess and clean the data. Python was selected due to its versatility, ability to handle large data sets, and many useful packages like Pandas. Pandas will be used to read the CSV file, for manipulation, and analysis. Matplotlib will be used to create visualizations. Sklearn will be utilized to perform the PCA. Numpy will be used for array operations.

C4.  The file is attached.

D1.  During the data quality review, multiple issues were found. There are multiple columns with null values. The children column has 2495 null values, age has 2475, income has 2490, techie has 2477, phone has 1026, tenure has 931, and bandwidth\_GB\_year has 1021. There were also outliers. Population has 214 outliers, children has 77 outliers, income has 67 outliers, Outage\_sec\_week has 123, email has 9, Contacts has 3, and Yearly\_equip\_failure has 23. Customer\_id and interaction had no duplicates, indicating there were no duplicate customers or interactions. There were rows where children exceeded the population, which is an error because they are part of the household according to the dictionary.

D2. The strategy of replacing null values with the column's average is employed to maintain the representative nature of the imputed values, ensuring their alignment with the central tendency of the column's distribution within the dataset. The strategy of replacing null values with the column's mode is used to ensure that the imputed values are aligned with the most frequently occurring categories within their respective columns. The 'fillna()' method will be used to impute null values. The 2495 null values in the 'children' column are filled with 0, under the assumption that the customers who did not provide this information might not have any children. As such, imputing 0 in place of the missing values aligns with the assumption that a null value signifies an absence of reported children. The 2475 null values in the 'age' column will be replaced by the average age of the dataset. The 'income' column contains, 2490 missing values, which will be replaced with the average income. The 2477 missing values in the 'Techie' column will be replaced with the mode, ensuring that the imputed values align with the most frequently occurring category within the dataset. The 2490 missing values in the 'income' column will be replaced with the average income. The 931 null values in the 'Tenure' column will be replaced with the average tenure. The 1021 null values in the 'Bandwidth\_GB\_Year' column will be replaced with the average bandwidth usage per year. The 1026 null values in the 'Phone' column and the 991 null values in the 'TechSupport' column will be replaced with the mode value. Outliers outside the lower bound (q1 - 1.5 \* IQR) and upper bound (q3 + 1.5 \* IQR) will be dropped. The 'children' column exhibits 77 outliers in the range of 8 to 10. These outliers will be retained in the analysis based on domain knowledge indicating that it is possible for individuals to have 8 to 10 children. 'Population' has 214 outliers ranging from, 31371 to 94512. This data will be retained based on domain knowledge that it was acquired from census data, which is obtained from reputable sources like the government. 'Income' has 67 outliers, ranging from 105193 to 172884. This data will be kept based on domain knowledge of average salaries for different jobs. Retaining these values ensures that the analysis accounts for the possible variation in reported income levels. The 'outage\_sec\_perweek' column has 123 outliers ranging from -0.78 to 47. The outliers will be dropped because they are outside the bounds, and a customer can’t have negative downtime. 'Email' has 9 outliers, ranging from 1 to 21. These outliers will be retained based on domain knowledge that customers may receive marketing emails. 'Contacts' has 3 outliers ranging from 6 to 7. These outliers will be retained based on the domain knowledge that customers send multiple emails to address an issue or get clarification. 'Yearly\_equip\_failure' has 23 outliers ranging from 3 to 4. These values will be retained based on domain knowledge because equipment failure is possible and can occur for many reasons without a limit. Rows where the reported population of the customer city is less than the count of children in the customer's household will be dropped. This decision is based on the assumption that the reported population should inherently include the individuals within the household, including the children. Therefore, instances where the reported population is lower than the number of children in the household are considered inconsistent.

D3.  To handle the missing values in the dataset, different imputation strategies were applied based on the characteristics of each column. For the numerical columns 'age', 'Income', and 'Tenure', the mean values were used to replace the null values, preserving the central tendencies of the distributions. For the categorical columns 'Techie', 'Phone', and 'TechSupport', the mode values were used to replace the null values, reflecting the most common categories in each column. For the 'children' column, the null values were assumed to indicate no children and were replaced with 0. To deal with the outliers in the dataset, some columns were trimmed and some were kept based on domain knowledge and relevance. The 'outage\_sec\_perweek' column was trimmed by removing the outliers beyond the lower and upper bounds, ensuring the validity of the downtime data. The outliers in the 'children' column were kept based on the assumption that some customers may have large families with 8 to 10 children. The outliers in the 'Population' and 'income' columns were also kept based on the census data and average salaries for different regions. The outliers in the 'Email', 'Contacts', and 'Yearly\_equip\_failure' columns were retained as well, as there were no specific limits for these variables in terms of customer interactions and equipment failure scenarios. Finally, some rows were dropped from the dataset due to data inconsistency and inaccuracy, such as those where the reported population within a one-mile radius of the customer was less than the number of children in the customer's household. These data-cleaning steps aimed to ensure the integrity of the dataset by aligning the imputed values with the underlying distributions and retaining or excluding data points based on their relevance and consistency with domain knowledge.

D4.  The code is attached.

D5.  A copy of the cleaned data set is included.

D6.  Data cleaning is a vital process that improves the quality and reliability of the dataset, but it also has some limitations that need to be acknowledged. Imputing missing values with the mean or mode may introduce bias and change the original distribution, affecting the accuracy of further analyses. Excluding outliers based only on statistical criteria may result in losing valid data points, possibly missing important insights and trends in the dataset. Imputing null values with the mode may oversimplify the data, especially when the mode is an extreme or uncommon category, leading to a distorted representation of the true underlying patterns. The data cleaning process depends on the initial data quality, and any inherent errors or inconsistencies in the raw data may persist or even worsen through subsequent cleaning steps. Moreover, the data includes third-party data from census, which is subject to data decay and may have different quality standards and formats (Clearbit, n.d.). Cleaning data involves making assumptions and decisions that could influence the interpretation of results. Cleaning data often requires a strong understanding of the specific domain and context. Without adequate domain knowledge, there is a risk of misinterpreting data patterns and making erroneous cleaning decisions.

D7.  The quality of the data cleaning process has a significant impact on the analysis of customer satisfaction factors related to the customer churn rate. One of the limitations is the imputation of missing customer satisfaction data with mean or mode values. This could skew the analysis by distorting the actual distribution of satisfaction levels among customers. For example, if a large proportion of customers have low or high satisfaction levels, but their data is missing and replaced with average values, the analysis may fail to capture the true relationship between customer satisfaction factors and churn rate. Therefore, imputing missing values with mean or mode values may lead to biased conclusions. Another limitation is the exclusion of outliers without careful consideration. They could represent important trends or anomalies that are critical for understanding the factors contributing to customer churn. For instance, high or low satisfaction levels indicated by outliers could reflect customer behavior patterns that are not captured by the majority of the data. If these outliers are excluded without proper justification, the analysis may overlook crucial insights related to customer satisfaction and churn rate. Another limitation is the oversimplification of customer satisfaction factors by imputing them with mode values. Mode values are the most frequently occurring values in a dataset. However, they may not reflect the diversity of customer experiences and preferences. Customer satisfaction factors may vary depending on various factors such as demographics, services, downtime, equipment failure. If these factors are imputed with mode values, the analysis may neglect the nuances that lead to customer churn. An overemphasis on the most frequently occurring satisfaction factors might fail to capture the complexity of customer behavior. Another limitation is the lack of domain knowledge and contextual understanding of customer behavior and satisfaction dynamics within the telecommunications industry. Without this knowledge, it may be difficult to identify and address relevant issues and challenges related to data quality and analysis. For example, without knowing what factors affect customer satisfaction and churn rate in a particular industry, it may be hard to determine what data points are outliers or what values are appropriate for imputation. These limitations could affect the analysis of customer satisfaction factors related to customer churn rate by introducing bias, error, oversimplification, incompleteness, or misinterpretation into the results.

E1.  The total number of principal components identified through the PCA analysis was 11. The loading matrix for the principal components is attached below.

|  | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** | **PC8** | **PC9** | **PC10** | **PC11** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Population** | -0.001873 | -0.063108 | -0.327003 | -0.368601 | -0.018352 | 0.583280 | 0.537289 | -0.134856 | -0.325246 | -0.003657 | -0.000591 |
| **Children** | -0.001779 | 0.052564 | 0.510184 | -0.212489 | 0.047294 | 0.378901 | -0.451198 | 0.324787 | -0.485771 | 0.026442 | -0.015926 |
| **Age** | -0.012146 | -0.054361 | -0.288785 | 0.571200 | -0.200840 | 0.284340 | -0.396083 | -0.459291 | -0.290341 | 0.116010 | 0.021430 |
| **Income** | 0.004868 | -0.009208 | 0.264166 | 0.198538 | 0.742023 | 0.382231 | 0.040894 | -0.272158 | 0.335989 | -0.073209 | 0.001252 |
| **Outage\_sec\_perweek** | 0.022694 | 0.704925 | 0.004582 | -0.010470 | 0.015532 | 0.039457 | 0.078973 | -0.018236 | 0.070389 | 0.699366 | 0.000343 |
| **Email** | -0.020746 | 0.053761 | -0.413460 | -0.442598 | -0.065338 | 0.278131 | -0.540620 | 0.039248 | 0.500766 | -0.060006 | 0.005851 |
| **Contacts** | 0.005236 | -0.007697 | -0.408699 | 0.426828 | 0.230113 | 0.161355 | 0.055830 | 0.753687 | -0.015431 | 0.017345 | -0.003086 |
| **Yearly\_equip\_failure** | 0.016198 | 0.055675 | 0.363690 | 0.267482 | -0.590116 | 0.427247 | 0.216926 | 0.130520 | 0.425513 | -0.129938 | -0.002376 |
| **Tenure** | 0.704990 | -0.057934 | -0.017037 | -0.003080 | -0.004040 | 0.003294 | -0.018877 | -0.014589 | 0.009672 | 0.036612 | -0.705196 |
| **MonthlyCharge** | 0.043799 | 0.695569 | -0.095047 | 0.057572 | 0.027681 | -0.042054 | -0.004277 | -0.064077 | -0.154588 | -0.684883 | -0.047905 |
| **Bandwidth\_GB\_Year** | 0.706864 | -0.008496 | -0.001072 | -0.015505 | 0.003225 | -0.000017 | -0.011494 | 0.006127 | -0.006654 | -0.012889 | 0.706851 |

E2.  The number of principal components was reduced to 8, adhering to the Kaiser rule, where components with an eigenvalue greater than or equal to 1 are retained. The components that were retained include PC1, PC2, PC3, PC4, PC5, PC6, PC6,PC7, PC8, PC9. This decision was supported by the screen plot, which demonstrated a drop in the eigenvalue below 1 at the 8th component.

A blue line graph with numbers

Description automatically generated

The plot shows the number of components and eigenvalue. The plot shows the eigenvalue dropping below 1 at component 8.

E3 In the telecommunications industry, there are a lot of things that can impact the churn rate of a customer. One of these is customer satisfaction factors. To help identify what these factors are and to streamline data analysis, organizations can benefit from PCA. Through this method, organizations should be able to better understand what satisfaction metrics are and why customers leave or stay. Reduction of data dimensions helps uncover some of the most influential factors in customer satisfaction by capturing the variations across the data set in fewer dimensions. Having these insights can help create ways to tackle them head on before they lead to churn. Furthermore, PCA enables machine learning, which can further identify novel correlations. By utilizing PCA, companies may be able to make more informed decisions based off correlations and dependencies among satisfaction factors that lead to retention and satisfaction of a customer (Joshi, 2020).

F. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b22dffc5-f418-4e2d-acb9-b0ad003f299c

G.

Chaudhary, S. (2020, August 26). Why “1.5” in IQR Method of Outlier Detection? Medium. https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097

Nkmk. (n.d.). pandas: Find and remove duplicate rows of DataFrame, Series | note.nkmk.me. Retrieved October 31, 2023, from https://note.nkmk.me/en/python-pandas-duplicated-drop-duplicates/

Western Governors University. (n.d.). Data Cleaning—D206. Retrieved October 31, 2023, from https://my.wgu.edu/courses/course/29780019

H.

Chaudhary, S. (2020, August 26). *Why “1.5” in IQR Method of Outlier Detection?* Medium. https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097

Clearbit. (n.d.). *What is first-party vs third-party data?* Retrieved October 31, 2023, from https://clearbit.com/resources/books/b2b-data/what-is-data

Joshi, Y. (2020, March 26). *Applications of Principal Component Analysis (PCA)*. OpenGenus IQ: Computing Expertise & Legacy. https://iq.opengenus.org/applications-of-pca/