Performance Assessment: Data Cleaning (NUM3)

WGU | D206

Data Cleaning - D206

Narcisse, Laurie

2023

Contents

[Part I: Research Question 2](#_Toc150957142)

[A. Describe one question or decision that could be addressed using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation. 2](#_Toc150957143)

[B. Describe all variables in the data set (regardless of the research question) and indicate the data type for each variable. Use examples from the data set to support your claims. 2](#_Toc150957144)

[Part II: Data-Cleaning Plan 6](#_Toc150957145)

[C. Explain the plan for cleaning the data by doing the following: 6](#_Toc150957146)

[1. Propose a plan that includes the relevant techniques and specific steps needed to assess the quality of the data in the data set. 6](#_Toc150957147)

[2. Justify your approach for assessing the quality of the data, including the characteristics of the data being assessed and the approach used to assess the quality of the data. 7](#_Toc150957148)

[3. Justify your selected programming language and any libraries and packages that will support the data-cleaning process. 7](#_Toc150957149)

[4. Provide the annotated code you will use to assess the quality of the data in an executable script file. 8](#_Toc150957150)

[Part III: Data Cleaning 8](#_Toc150957151)

[D. Summarize the data-cleaning process by doing the following: 8](#_Toc150957152)

[1. Describe the findings for the data quality issues found from the implementation of the data-cleaning plan from part C. 8](#_Toc150957153)

[2. Justify your methods for mitigating the data quality issues in the data set. 9](#_Toc150957154)

[3. Summarize the outcome from the implementation of each data-cleaning step 9](#_Toc150957155)

[4. Provide the annotated code you will use to mitigate the data quality issues—including anomalies—in the data set in an executable script file. 11](#_Toc150957156)

[5. Provide a copy of the cleaned data set as a CSV file. 11](#_Toc150957157)

[6. Summarize the limitations of the data-cleaning process. 11](#_Toc150957158)

[7. Discuss how the limitations summarized in part D6 could affect the analysis of the question or decision from part A. 11](#_Toc150957159)

[E. Apply principal component analysis (PCA) to identify the significant features of the data set by doing the following: 12](#_Toc150957160)

[1. Identify the total number of principal components and provide the output of the principal components loading matrix. 12](#_Toc150957161)

[2. Justify the reduced number of the principal components and include a screenshot of a scree plot. 12](#_Toc150957162)

[3. Describe how the organization would benefit from the use of PCA. 13](#_Toc150957163)

[F. Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment. 14](#_Toc150957164)

[G. Acknowledge web sources, using in-text citations and references, for segments of third-party code used to support the application. Be sure the web sources are reliable. 14](#_Toc150957165)

# Part I: Research Question

## A. Describe one question or decision that could be addressed using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation.

* Can we predict customer churn based on various customer attributes and behaviors in the telecommunications industry?

## B. Describe all variables in the data set (regardless of the research question) and indicate the data type for each variable. Use examples from the data set to support your claims.

* Here's a description of the variables in the data set along with their data types:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Description** | **Data Type** | | | | | | **Example** |
| Unnamed | Index representing the position in the dataset. It is an integer. | Integer | int64 | int | Quantitative | Numerical | Discrete | 1, 2, 3, ... |
| CaseOrder | This is a numerical placeholder preserving the original order of the raw data file. | Integer | int64 | int | Quantitative | Numerical | Discrete | 1, 2, 3, ... |
| Customer\_id | A unique identifier assigned to each customer, which can be a combination of letters and numbers. | String/Int | object | char | Qualitative | Categorical | Ordinal | "Cust\_001", "123" |
| Interaction, UID | Unique identifiers associated with customer transactions, support, and sign-ups. These can be a combination of letters and numbers. | String/Int | object | char | Qualitative | Categorical | Ordinal | "Interact\_001", 123 |
| City | The city of residence for the customer, as listed on the billing statement. | String | object | char | Qualitative | Categorical | Nominal | "New York" |
| State | The state of residence for the customer, as listed on the billing statement. | String | object | char | Qualitative | Categorical | Nominal | "CA", "TX" |
| County | The county of residence for the customer, as listed on the billing statement. | String | object | char | Qualitative | Categorical | Nominal | "Cook" |
| Zip | The zip code of the customer's residence, as listed on the billing statement. | Integer | int64 | int | Qualitative - no meaningful statistical calculations | Categorical | Nominal | 783, 48133 |
| Lat | GPS coordinates representing the location of the customer's residence. | Float | Float64 | numeric | Quantitative | Numerical | Continuous | 56.251, 44.3289 |
| Lng | GPS coordinates representing the location of the customer's residence. | Float | Float64 | numeric | Quantitative | Numerical | Continuous | 70.115, 91.38 |
| Population | The population within a mile radius of the customer, based on census data. | Integer | int64 | int | Quantitative | Numerical | Discrete | 38, 10446 |
| Area | The type of area (rural, urban, suburban) based on census data. | String | object | char | Qualitative | Categorical | Nominal | "Urban" |
| Timezone | The timezone of the customer's residence based on the customer’s sign-up information. | String | object | char | Qualitative | Categorical | Nominal | "America/Sitka" |
| Children | The number of children in the customer’s household as reported in sign-up information. | Float | Float64 | int | Quantitative | Numerical | Discrete | "NA", 1,2 |
| Job | The occupation or job of the customer, as reported during sign-up. | String | object | char | Qualitative | Categorical | Nominal | "Environmental health practitioner" |
| Age | The age of the customer, as reported in sign-up information. | Float | Float64 | int | Quantitative | Numerical | Continuous | 68, "NA" |
| Education | The highest degree earned by the customer, as reported in sign-up information. | String | object | char | Qualitative | Categorical | Nominal | "Master's Degree" |
| Employment | The employment status of the customer, as reported in sign-up information. | String | object | char | Qualitative | Categorical | Nominal | "Part Time" |
| Income | The annual income of the customer as reported at the time of sign-up. | Float | Float64 | numeric | Quantitative | Numerical | Discrete | 28561.99, "NA" |
| Marital | The marital status of the customer as reported in sign-up information. | String | object | char | Qualitative | Categorical | Nominal | "Widowed" |
| Gender | The self-identified gender of the customer (male, female, nonbinary). | String | object | char | Qualitative | Categorical | Nominal | "Male" |
| Churn | Indicates whether the customer discontinued service within the last month (yes, no). | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| Outage\_sec\_perweek | The average number of seconds per week of system outages in the customer’s neighborhood. | Float | Float64 | numeric | Quantitative | Numerical | Continuous | 45.6 |
| Email | The number of emails sent to the customer in the last year (marketing or correspondence). | Integer | int64 | int | Quantitative | Numerical | Discrete | 100, 200 |
| Contacts | The number of times the customer contacted technical support. | Integer | int64 | int | Quantitative | Numerical | Discrete | 2, 5 |
| Yearly\_equip\_failure | The number of times the customer’s equipment failed and had to be reset/replaced in the past year. | Integer | int64 | int | Quantitative | Numerical | Discrete | 1, 0 |
| Techie | Indicates whether the customer considers themselves technically inclined (yes, no). | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| Contract | The contract term of the customer (month-to-month, one year, two years). | String | object | char | Qualitative | Categorical | Nominal | "Month-to-month", "Two year" |
| Port\_modem | Indicates whether the customer has a portable modem (yes, no). | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| Tablet | Indicates whether the customer owns a tablet (yes, no). | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| InternetService | The customer’s internet service provider (DSL, fiber optic, None). | String | object | char | Qualitative | Categorical | Nominal | "DSL", "None" |
| Phone | Indicates whether the customer has phone service (yes, no). | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| Multiple | Indicates whether the customer has multiple lines (yes, no). | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| OnlineSecurity | Indicates whether the customer has an online security add-on | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| OnlineBackup | Indicates whether the customer has an online backup add-on | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| DeviceProtection | Indicates whether the customer has device protection add-on | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| TechSupport | Indicates whether the customer has a technical support add-on | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| StreamingTV | Indicates whether the customer has streaming TV | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| StreamingMovies | Indicates whether the customer has streaming movies | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| PaperlessBilling | Whether the customer has paperless billing | String | object | char | Qualitative | Categorical | Ordinal | "Yes", "No" |
| PaymentMethod | Customer’s payment method | String | object | char | Qualitative | Categorical | Nominal | "Bank Transfer" "Credit Card" |
| Tenure | The number of months the customer has stayed with the provider. | Float | Float64 | int | Quantitative | Numerical | Discrete | 24, 36, "NA" |
| MonthlyCharge | The amount charged to the customer monthly | Float | Float64 | numeric | Quantitative | Numerical | Continuous | 50.5, 75.2 |
| Bandwidth\_GB\_Year | The average amount of data used, in GB, in a year by the customer | Float | Float64 | numeric | Quantitative | Numerical | Continuous | 150.8, 200.3 |
| Item1 | Responses to an eight-question survey rating importance on a scale of 1 to 8 -Timely response | Integer | int64 | int | Quantitative | Numerical | Discrete/Ordinal | 1, 5, 8 |
| Item2 | Responses to an eight-question survey rating importance on a scale of 1 to 8 - Timely fixes | Integer | int64 | int | Quantitative | Numerical | Discrete/Ordinal | 1, 5, 8 |
| Item3 | Responses to an eight-question survey rating importance on a scale of 1 to 8 - Timely replacements | Integer | int64 | int | Quantitative | Numerical | Discrete/Ordinal | 1, 5, 8 |
| Item4 | Responses to an eight-question survey rating importance on a scale of 1 to 8 - Reliability | Integer | int64 | int | Quantitative | Numerical | Discrete/Ordinal | 1, 5, 8 |
| Item5 | Responses to an eight-question survey rating importance on a scale of 1 to 8 - Options | Integer | int64 | int | Quantitative | Numerical | Discrete/Ordinal | 1, 5, 8 |
| Item6 | Responses to an eight-question survey rating importance on a scale of 1 to 8 - Respectful response | Integer | int64 | int | Quantitative | Numerical | Discrete/Ordinal | 1, 5, 8 |
| Item7 | Responses to an eight-question survey rating importance on a scale of 1 to 8 - courteous exchange | Integer | int64 | int | Quantitative | Numerical | Discrete/Ordinal | 1, 5, 8 |
| Item8 | Responses to an eight-question survey rating importance on a scale of 1 to 8 - Evidence of active listening | Integer | int64 | int | Quantitative | Numerical | Discrete/Ordinal | 1, 5, 8 |

# Part II: Data-Cleaning Plan

## C. Explain the plan for cleaning the data by doing the following:

### 1. Propose a plan that includes the relevant techniques and specific steps needed to assess the quality of the data in the data set.

* To assess the quality of the data in the dataset, a comprehensive plan incorporating various techniques and specific steps is proposed. The following steps will be undertaken:
  + Assessment of data types and values will be conducted using the info() function.
  + Unnecessary columns will be deleted through the use of the drop() method.
  + Detection and modification/deletion of missing values in the dataset will be addressed using methods such as isnull().sum(), fillna(), median(), mode(), mean(), and visualization tools such as histograms.
  + Identification and removal of duplicates will be performed utilizing duplicated() and drop\_duplicates() methods.
  + The data set will be described using the describe() function.
  + Outliers, if present, will be identified and handled through methods like fillna(), median(), boxplots, histograms, IQR (Interquartile Range), and Lower/Upper Bound analysis.
  + Categorical variables will be encoded into numerical format using the map() function.
  + Data entry errors or formatting issues will be corrected.
  + Transformation and creation of new features will be undertaken as needed.
  + Principal Component Analysis (PCA) will be performed with the use of transform(), mean(), and std() functions.

### 2. Justify your approach for assessing the quality of the data, including the characteristics of the data being assessed and the approach used to assess the quality of the data.

* The approach employed to assess the quality of the data in the dataset is justified based on a systematic and comprehensive strategy. The characteristics of the data under assessment include a thorough examination of data types and values, identification and elimination of unnecessary columns, detection and handling of missing values, identification and removal of duplicates, and a descriptive analysis of the dataset.
* Additionally, the approach incorporates the identification and handling of outliers, encoding of categorical variables into numerical format, correction of data entry errors and formatting issues, and the potential transformation and creation of new features.
* The use of statistical methods such as info(), drop(), isnull().sum(), fillna(), median(), mode(), mean(), describe(), duplicated(), and drop\_duplicates() ensures a rigorous evaluation of the data quality.
* Furthermore, the inclusion of visualization tools such as histograms and boxplots enhances the interpretability of the assessment.
* The proposed plan is designed to comprehensively address various aspects of data quality, providing a robust foundation for subsequent analysis and decision-making processes.

### 3. Justify your selected programming language and any libraries and packages that will support the data-cleaning process.

* I have chosen Python as my programming language for data cleaning, and I will be using libraries such as Seaborn, Matplotlib, NumPy, and Pandas to support the data-cleaning process. Here are the reasons behind my selection:
  + Python is a versatile and widely adopted programming language for data analysis and manipulation. When it comes to data cleaning, it offers several advantages:
    - Pandas:
      * Serving as a robust library for data manipulation and analysis, Pandas provides pivotal data structures like DataFrames and Series. Its extensive set of functions facilitates effective data handling, cleaning, and preprocessing.
    - NumPy:
      * Excels in the efficient manipulation of arrays and matrices. This makes it indispensable for data cleaning tasks involving mathematical operations or transformations.
    - Seaborn and Matplotlib:
      * Enable the creation of informative and visually appealing plots. This visual representation proves invaluable for identifying data anomalies, distributions, and trends during the data-cleaning process. Their contribution enhances the interpretability and insight derived from the data.

### 4. Provide the annotated code you will use to assess the quality of the data in an executable script file.

* Please see attached labeled as FINALC4.

# Part III: Data Cleaning

## D. Summarize the data-cleaning process by doing the following:

### 1. Describe the findings for the data quality issues found from the implementation of the data-cleaning plan from part C.

* The data quality issues I was able to detect were missing values within:
  + Children -2495
  + Age – 2475
  + Income – 2490
  + Techie – 2477
  + InternetService – 2129
  + Phone – 1026
  + TechSupport – 991
  + Tenure – 931
  + Bandwidth\_GB\_Year – 1021
* Additionally, there were no duplicates found.
* Outliers
  + Unnamed: 0: 0
  + CaseOrder: 0
  + Zip: 0
  + Lat: 151
  + Lng: 102
  + Population: 219
  + Children: 144
  + Age: 0
  + Income: 110
  + Outage\_sec\_perweek: 491
  + Email: 12
  + Contacts: 165
  + Yearly\_equip\_failure: 94
  + Tenure: 0
  + MonthlyCharge: 3
  + Bandwidth\_GB\_Year: 0
  + item1: 19
  + item2: 13
  + item3: 13
  + item4: 9
  + item5: 12
  + item6: 13
  + item7: 11
  + item8: 15

### 2. Justify your methods for mitigating the data quality issues in the data set.

In polishing up my dataset, I took a systematic approach to address various issues. First on the list was handling missing records—Dealing with missing values required a thoughtful strategy. I opted for imputation, choosing between mean, median, or mode based on each variable's characteristics. It's like giving the dataset a touch-up, ensuring it's ready for prime time without losing its essence. The following were imputed:

|  |  |
| --- | --- |
| Age | Mean |
| Children | Median |
| Income | Median |
| Tenure | Median |
| Bandwidth\_GB\_Year | Median |
| Techie | Mode |
| InternetService | Mode |
| Phone | Mode |
| TechSupport | Mode |

Outliers were identified using boxplots and histograms and were deemed to be imputed. The following were imputed with the median().

Transforming categorical columns into numeric formats was my way of creating a common language for everyone involved.

In essence, my methods were a careful selection tailored to each specific concern. I aimed to enhance accuracy, completeness, and compatibility, laying the groundwork for reliable analyses and valuable insights.

3. Summarize the outcome from the implementation of each data-cleaning step**.**

I've made substantial improvements to the quality and uniformity of the dataset through a series of strategic data cleaning steps. Ensuring consistent formatting across the dataset was pivotal, guaranteeing standardized representations of data types. By addressing missing values, I've enhanced the reliability of the dataset, mitigating potential biases in analyses. The identification and removal of outliers using methods like the Interquartile Range (IQR) or Z-Score contribute to a more robust dataset, reducing the impact of extreme values on statistical measures and machine learning models. Additionally, converting categorical columns to numeric formats streamlines the dataset, fostering consistency and enabling compatibility with various analytical approaches. This concerted effort to clean and preprocess the data sets the groundwork for more accurate and meaningful analyses, ensuring that the dataset is well-prepared for a range of data-driven tasks. The documentation of these cleaning steps serves as a transparent record, aiding in reproducibility and facilitating collaboration with future users.

Changes and Cleaning Examples Below:

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |

### 4. Provide the annotated code you will use to mitigate the data quality issues—including anomalies—in the data set in an executable script file.

* Please see attached file labeled as FINALD4.

### 5. Provide a copy of the cleaned data set as a CSV file.

* Please see attached file labeled as churn\_cleaned\_dataset.csv

### 6. Summarize the limitations of the data-cleaning process.

While refining my dataset, it's crucial to acknowledge certain limitations in the data-cleaning process. The methods employed, as outlined in D2, while effective in enhancing data quality, come with inherent disadvantages.

One notable limitation arises from the potential loss of information during the removal of outliers or duplicates. The decision to exclude certain data points or records may inadvertently discard valuable insights, impacting the overall richness of the dataset.

Additionally, imputing missing values, a necessary step in ensuring completeness, introduces the risk of bias if not executed judiciously. The choice between mean, median, or mode for imputation may tilt the distribution of the data, influencing subsequent analyses and interpretations.

It's imperative to recognize these drawbacks as inherent trade-offs in the data-cleaning process. Striking a balance between enhancing data quality and preserving information integrity is a delicate task, necessitating careful consideration of the potential implications of each method employed. In summary, while my cleaning methods aimed to bolster accuracy and completeness, it's essential to remain cognizant of the nuanced challenges and compromises inherent in the data refinement journey.

### 7. Discuss how the limitations summarized in part D6 could affect the analysis of the question or decision from part A.

In the process of refining the dataset for my analysis on predicting customer churn in the telecommunications industry, certain cleaning methods, as outlined in D6, present inherent limitations that merit consideration. Notably, the removal of outliers and duplicates, while conducive to enhancing data quality, comes at the expense of potential information loss. This could prove pivotal for an accurate prediction model, as the exclusion of extreme cases or recurring patterns may compromise the model's overall robustness. Additionally, the imputation of missing values introduces a nuanced challenge, requiring a careful balance in choosing between mean, median, or mode. The inherent bias introduced by such imputation methods, particularly in variables like income, poses a potential distortion in the dataset's distribution and subsequently impacts the reliability of predictions.

Understanding these limitations is paramount, considering the trade-offs made during the data-cleaning phase. These considerations underscore the need for a thoughtful and nuanced approach to data analysis. The application of singular imputation methods or the wholesale removal of outliers and duplicates may oversimplify the dataset, akin to a monochromatic representation that overlooks the potential richness of varied attributes. These trade-offs, while necessary, demand an awareness of their potential implications on the predictive accuracy and validity of the model. As I delve into predicting customer churn, cognizance of these intricacies ensures a more comprehensive evaluation of the dataset's nuances and enhances the discernment of underlying patterns relevant to my analytical pursuit.

## E. Apply principal component analysis (PCA) to identify the significant features of the data set by doing the following:

### 1. Identify the total number of principal components and provide the output of the principal components loading matrix.

The total number of principal components is 5, as specified by the variable n\_components=5.

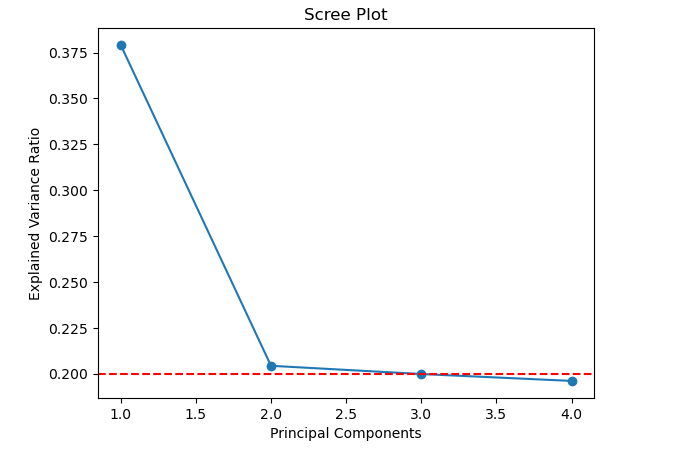
A screenshot of a computer code

Description automatically generated

### 2. Justify the reduced number of the principal components and include a screenshot of a scree plot.

While I was looking at the explained variance, the first principal component (PC1) explains 37.92% of the variance, PC2 explains 20.81% (58.35% cumulative), PC3 explains 19.95% (78.33% cumulative), PC4 explains 19.07% (97.94% cumulative), and PC5 explains the remaining 2.06%. I opted for 5 components, and considering the explained variance, the first four components explain approximately 97.94% of the total variance. This suggests that by using these four components, I can retain a substantial amount of information from the original features. A white background with black text

Description automatically generated



### 3. Describe how the organization would benefit from the use of PCA.

By utilizing the first four principal components, which collectively explain around 97.94% of the variance, I can significantly reduce the dimensionality of the data. This reduction can lead to more efficient computations and memory usage, making data analysis and machine learning tasks faster and more streamlined. The organization stands to gain from these efficiency improvements and potentially improved model performance.

# F. Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment.

Please see attached Video.

# G. Acknowledge web sources, using in-text citations and references, for segments of third-party code used to support the application. Be sure the web sources are reliable.

GeeksforGeeks. (n.d.). Detect and Remove the Outliers using Python. Retrieved from https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/

DataCamp. (n.d.). DataCamp. Retrieved from https://www.datacamp.com/

Statology. (n.d.). Statistical Tutorials and Software Guides. Retrieved from https://www.statology.com/

YouTube. (n.d.). YouTube. Retrieved from https://www.youtube.com/

WGU Panopto. (n.d.).Retrieved from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95

Stack Overflow. (n.d.). Stack Overflow. Retrieved from https://stackoverflow.com/

W3Schools. (n.d.). SQL Tutorial. Retrieved from https://www.w3schools.com/sql/default.asp