D206: Data Acquisition

Performance Assessment

Western Governor’s University

Mariel Vicente

May 23, 2024

**Performance Assessment**

1. **Research Question**

My research question for this performance assessment is, “What patient characteristics influence high soft drink consumption?” It is known that high soft drink consumption can lead to negative health outcomes. Medical providers may be interested in this information to appropriately offer preventative advice to patients and to prescribe care to those suffering from the negative affects of high soft drink consumption. Providers will want to know which subpopulations of their patients are more susceptible to this behavior, such as whether they are younger children or senior patients. The subpopulation will determine how they administer care.

1. **Required Variables**

For this performance assessment, I will be utilizing the medical data set provided. The data set contains 50 variables, listed below:

* + CaseOrder (quantitative): This is a placeholder variable to maintain the order of the raw data. Example: 1
  + Customer\_id (qualitative): This is the unique patient ID. Example: C412403
  + Interaction, UID (both qualitative): These are different unique IDs related to patient transactions, procedures, and admissions. Examples: 8cd49b13-f45a-4b47-a2bd-173ffa932c2f, 3a83ddb66e2ae73798bdf1d705dc0932
  + City (qualitative): This is the city where the patient lives. Example: Eva
  + State (qualitative): This is the state where the patient lives. Example: AL
  + County (qualitative): This is the county where the patient lives. Example: Morgan
  + Zip (qualitative): This is the patient’s zip code. Example: 35621
  + Lat, Lng (both quantitative): These are the GPS coordinates of where the patient lives. Example: 34.3496, -86.7251
  + Population (quantitative): This is the population within a mile radius of the patient. Example: 2951
  + Area (qualitative): This is the classification of what type of area the patient lives in. Example: Suburban
  + TimeZone (qualitative): This is the time zone the patient lives in. Example: America/Chicago
  + Job (qualitative): This is the reported job of the patient. Example: Psychologist, sport and exercise
  + Children (quantitative): This is the reported number of children in the patient’s household. Example: 1
  + Age (quantitative): This is the patient’s age. Example: 53
  + Education (qualitative): This is the highest level of education completed by the patient. Example: Some College, Less than 1 Year
  + Employment (qualitative): This is the employment status of the patient. Example: Full Time
  + Income (quantitative): This is the reported annual income of the patient, or primary insurance holder. Example: 86575.93
  + Marital (qualitative): This is the reported marital status of the patient, or primary insurance holder. Example: Divorced
  + Gender (qualitative): This is the reported gender of the patient. Example: Male
  + ReAdmis (qualitative): This is whether the patient was readmitted within a month of release. Example: No
  + VitD\_levels (quantitative): This is the patient’s vitamin D level. Example: 17.802330
  + Doc\_visits (quantitative): This the reported number of times the primary physician visited the patient. Example: 6
  + Full\_meals\_eaten (quantitative): This is the number of full meals the patient ate while hospitalized. Example: 0
  + VitD\_supp (quantitative): This is the number of times that vitamin D supplements were given to the patient. Example: 0
  + Soft\_drink (qualitative): This is whether the patient habitually drinks three or more sodas in a day. Example: Yes
  + Initial\_admin (qualitative): This is how the patient was admitted to the hospital initially. Example: Emergency Admission
  + HighBlood (qualitative): This is whether the patient has high blood pressure. Example: Yes
  + Stroke (qualitative): This is whether the patient has had a stroke. Example: No
  + Complication\_risk (qualitative): This is the assessed level of complication risk for the patient. Example: Medium
  + Overweight (qualitative): This is whether the patient is considered overweight. Example: No
  + Arthritis (qualitative): This is whether the patient has arthritis. Example: Yes
  + Diabetes (qualitative): This is whether the patient has diabetes. Example: Yes
  + Hyperlipidemia (qualitative): This is whether the patient has hyperlipidemia. Example: No
  + BackPain (qualitative): This is whether the patient has chronic back pain. Example: Yes
  + Anxiety (qualitative): This is whether the patient has an anxiety disorder. Example: No
  + Allergic\_rhinitis (qualitative): This is whether the patient has allergic rhinitis. Example: Yes
  + Reflux\_esophagitis (qualitative): This is whether the patient has reflux esophagitis. Example: No
  + Asthma (qualitative): This is whether the patient has asthma. Example: Yes
  + Services (qualitative): This includes any service the patient received while hospitalized. Example: Blood Work
  + Initial\_days (quantitative): This is the number of days the patient was initially hospitalized. Example: 10.585770
  + TotalCharge (quantitative): This is the mean average amount charged to the patient daily. Example: 3191.048774
  + Additional\_charges (quantitative): This is the average amount charged to the patient for additional services. Example: 17939.403420

The variables below correspond to survey questions about the patient’s experience during care:

* Item1 (qualitative): Question about timely admission. Example: 3
* Item2 (qualitative): Question about timely treatment. Example: 3
* Item3 (qualitative): Question about timely visits. Example: 2
* Item4 (qualitative): Question about reliability. Example: 2
* Item5 (qualitative): Question about options. Example: 4
* Item6 (qualitative): Question about hours of treatment. Example: 3
* Item7 (qualitative): Question about courteous staff. Example: 3
* Item8 (qualitative): Question about evidence of active listening from doctor. Example: 4

**C1. Plan to Assess Quality of Data**

To detect data quality issues within the data set, I have used Python within the Jupyter application. I used the .info() function to assess the data type of each variable to ensure each data type aligns with the data dictionary and check if re-expression is necessary. I used the .duplicated() and .value\_counts() functions to assess for duplicates. The .isnull().sum() function was used to count the number of missing values for each variable. The matplotlib and seaborn packages were imported to examine the distribution of the variables and to check for outliers.

**C2. Justification of Approach**

The functions listed above serve to clean the data set. The data set must be cleaned so that we may draw accurate conclusions from any analysis done with the data set. Duplicates, missing values, and outliers can all skew the results of an analysis. The .info() function provides an overview of the data which helps me decide what variables to focus on for cleaning. The duplicate functions allow me to find the duplicate and remove them from the data set. The null functions locate the nulls for me, which I may then fill however is most appropriate. The visualizations allow me to locate outliers and help me decide if and how to remove them.

**C3. Justification of Tools**

For this performance assessment, I chose to utilize Python to clean my data. Data cleaning can be accomplished using either Python or R, however I have chosen to begin the program using Python and will switch to R at a later time. Python includes a number of useful packages such as NumPy, panda, matplotlib, and seaborn. NumPy and panda can be used to detect data quality issues. Matplotlib and seaborn are great for visualizations, which can be used to view the distribution of different variables.

**C4. Code Used to Detect Data Quality Issues**

See code in file attached (D206 PA – Mariel Vicente.ipynb).

*Duplicates:*

A close-up of a computer code

Description automatically generated

*Missing Values:*

*A close up of text

Description automatically generated*

*Outliers:*

*A screenshot of a computer code

Description automatically generated*

*Re-expression:*

A close-up of a text

Description automatically generated

**D1. Cleaning Findings**

No duplicates were found in the data after assessment. I could see from the summary that not all variables had 10,000 non-nulls, which indicated that there were missing values. After assessing for missing values, the output found that 7 variables had null values. The variable children had 2,588 nulls; age had 2,414 nulls; income had 2,464 nulls; soft drink had 2,467 nulls; overweight had 982 nulls; anxiety had 984 nulls; and initial days had 1,056 nulls. The summary also showed that 2 variables that were supposed to be qualitative per the data dictionary were expressed as numerals and would need re-expression. After creating box plots for all quantitative variables using the seaborn package, I could also see that 4 variables contained outliers. The variable population contained 805 outliers from 3,500 to 121,000. The variable income contained 714 outliers from 80,000 to 210,000. The variable total charge contained 472 outliers from 140,000 to 220,000. The variable additional charges contained 432 outliers from 27,000 to 31,000.

**D2. Justification of Mitigation Methods**

Since no duplicates were found, no treatment was necessary for duplicates. After finding nulls in the data, I assessed the distribution of each of these variables with a histogram using the .hist() function. I performed univariate imputation on all variables, imputing the quantitative variables (children, age, income, initial days) with the median and the qualitative variables (soft drink, overweight, anxiety). with the mode as recommended in Dr. Middleton’s lectures (Middleton, 2024). After discovering the variables with outliers (population, income, total charge, additional charges), I replaced the outliers with nulls and imputed the nulls with the median in order to preserve the sample size. After treatment of the outliers, I did find new ranges of outliers after altering the data. However, due to the sensitive nature of the health data, I chose not to do another round of imputation. I re-expressed the inconsistent variables (overweight & anxiety) as categorical values by replicating them and replacing the values using a dictionary.

**D3. Summary of Outcomes**

After assessing for duplicates and finding none, the data can be considered reliable as duplicates can skew the data and affect conclusions drawn from the data. Imputing the nulls of the data ensures that all rows in the data are included in any analysis conducted. Removing the outliers from the quantitative variables improves the ability to use the data for more advanced analytics, such as predictive modeling. Re-expression of the variables ensures the data matches what is outlined in the data dictionary as well as uniformity in the data. Below you will find outputs confirming the data was cleaned.

*Duplicates:*

A close up of a text

Description automatically generated

*Missing Values:*

A screenshot of a computer program

Description automatically generated

*Outliers:*

A screen shot of a graph

Description automatically generated A screen shot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

*Re-expression:*

*A screenshot of a computer code

Description automatically generated*

**D4. Mitigation Code**

See code in file attached (D206 PA – Mariel Vicente.ipynb).

*Duplicates:*

No treatment necessary.

*Missing Values:*

A screenshot of a computer code

Description automatically generated

*Outliers:*

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

*Re-expression:*

A screenshot of a computer

Description automatically generated

**D5. Clean Data**

See output file attached (medical\_raw\_data\_clean.csv).

**D6. Limitations**

No treatment was necessary for duplicates, however this type of treatment can cause an unintended loss of data. Imputing nulls with an average, while better for statistical analysis, can misrepresent the data. Imputing outliers with an average can also skew the data and cause bias.

**D7. Impact of Limitations**

Since the data represent actual patients, there are some ethical concerns with altering the data, particularly when it comes to health data. For the research question, the variable soft drink is essential for the research and included many nulls. Treatment methods such as imputing do not reflect what the reality is for these patients. Imputing the nulls for this variable limits the conclusions we can draw. It may be better to remove rows with these data quality issues completely to ensure all the data accurately reflects the patients.

**E1. Principal Components**

For the PCA, I chose to only use the quantitative, continuous variables in the data. I excluded categorical and discrete quantitative values from the analysis. The PCA variables were lat, lng, income, vitamin D levels, initial days, total charge, and additional charges.

A table with numbers and symbols

Description automatically generated**E2. Criteria Used**

After running the PCA and plotting the scree plot, I would recommend retaining PC1, PC2, PC3, and PC4. This decision is based on the Kaiser Rule, which recommends using PCs with eigenvalues greater than 1 (Middleton, 2024). This is because the high eigenvalues represent a strong correlation of variables in the PC.

**E3. Benefits**

An organization can benefit from the results of the PCA because it provides them with a high-level look at the correlation between variables in a simple way that doesn’t require building a model. It quickly reduces the dimension of the data and can inform an analyst of whether more advanced analysis can be done. It is also useful for visualization of the data, as the scree plot can clearly demonstrate the strength of the variable relationships in the data.

**F. Video**

See Panopto recording here: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7de98bad-9a8a-442b-b0ba-b17a002fba66>

**G. Sources of Third-Party Code**

Middleton, Keiona. *D206 Data Cleaning Lecture Videos.* Retrieved May 23, 2024.

**H. Sources**

Bigabid. *What is Principal Component Analysis (PCA) & How to use it?* Bigabid. Retrieved May 23, 2024 from <https://www.bigabid.com/what-is-pca-and-how-can-i-use-it/>.

Data Camp. *D206 - Data Cleaning Organization Track*. Data Camp. Retrieved May 23, 2024 from <https://app.datacamp.com/learn/custom-tracks/custom-d206-data-cleaning>.

Middleton, Keiona. *D206 Data Cleaning Lecture Videos.* Retrieved May 23, 2024.