**D206: Data Cleaning**

**Performance Assessment – Task 1**

**Leng Yang**

**ID: 012298452**

**July 20, 2024**

**A1. Question or Decision**

The research question of focus is: what major patient factors influence hospital readmission? This question is essential to the organization because it facilitates an analysis of potential factors that may lead to patient readmission. Identified factors aid in delivering quality patient care and better patient outcomes.

**A2. Required Variables**

Below is the table containing all variables of the medical data set and each variable’s data type, a brief description of the variable, and an example taken from the first row.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Data Type | Description | Example |
| CaseOrder | Quantitative | An index to preserve the order of the data | 1 |
| Customer\_id | Qualitative | Unique patient ID | C412403 |
| Interaction | Qualitative | Unique ID relating to a patient’s transactions, procedures, and admissions | 8cd49b13-f45a-4b47-a2bd-173ffa932c2f |
| UID | Qualitative | Unique ID relating to a patient’s transactions, procedures, and admissions | 3a83ddb66e2ae73798bdf1d705dc0932 |
| City | Qualitative | Patient's city of residence | Eva |
| State | Qualitative | Patient's state of residence | AL |
| County | Qualitative | Patient’s county of residence | Morgan |
| Zip | Qualitative | Patient’s zip code of residence | 35621 |
| Lat | Quantitative | GPS coordinate of patient’s residence | 34.34960 |
| Lng | Quantitative | GPS coordinate of patient’s residence | -86.72508 |
| Population | Quantitative | Population of area where patient resides | 2951 |
| Area | Qualitative | The type of area the patient resides in | Suburban |
| TimeZone | Qualitative | The time zone of the patient’s residence | America/Chicago |
| Job | Qualitative | Job title of patient or primary insurance holder | Psychologist, sport and exercise |
| Children | Quantitative | Number of children in the patient’s household | 1.0 |
| Age | Quantitative | Patient’s age | 53.0 |
| Education | Qualitative | Patient’s highest education level | Some College, Less than 1 Year |
| Employment | Qualitative | Patient’s employment status | Full Time |
| Income | Quantitative | Patient’s annual income | 86575.93 |
| Marital | Qualitative | Patient or primary insurance holder’s marital status | Divorced |
| Gender | Qualitative | Patient’s gender | Male |
| ReAdmis | Qualitative | Y/N - Patient’s readmission status after a month of release | No |
| VitD\_levels | Quantitative | Patient’s Vitamin D levels (ng/mL) | 17.802330 |
| Doc\_visits | Quantitative | Number of times patient was visited by the primary doctor during initial stay | 6 |
| Full\_meals\_eaten | Quantitative | Numbers of full meals eaten while admitted | 0 |
| VitD\_supp | Quantitative | Number of Vitamin D supplements administered to the patient | 0 |
| Soft\_drink | Qualitative | Y/N - whether patient’s soft drink intake was more than 3 drinks in a day | NaN |
| Initial\_admin | Qualitative | Type of admission of the patient | Emergency Admission |
| HighBlood | Qualitative | Y/N - Patient’s high blood pressure status | Yes |
| Stroke | Qualitative | Y/N - whether the patient has had a stroke | No |
| Complication\_risk | Qualitative | Assessment of patient’s complication risk level | Medium |
| Overweight | Qualitative | Y/N - whether the patient was considered overweight based on weight, gender, and height | 0.0 |
| Arthritis | Qualitative | Y/N - Patient’s arthritis status | Yes |
| Diabetes | Qualitative | Y/N - Patient’s diabetes status | Yes |
| Hyperlipidemia | Qualitative | Y/N - Patient’s hyperlipidemia status | No |
| BackPain | Qualitative | Y/N - whether patient has chronic back pain | Yes |
| Anxiety | Qualitative | Y/N - Patient’s anxiety status | 1.0 |
| Allergic\_rhinitis | Qualitative | Y/N - Patient’s allergic rhinitis status | Yes |
| Reflux\_esophagitis | Qualitative | Y/N - Patient’s reflux esophagitis status | No |
| Asthma | Qualitative | Y/N - Patient’s asthma status | Yes |
| Services | Qualitative | Primary service the patient received during admission | Blood work |
| Initial\_days | Quantitative | Days the patient was admitted during initial visit | 10.585770 |
| TotalCharge | Quantitative | Daily amount charged to the patient during the hospitalization, calculated based on the total bill and the number of days stayed | 3191.048774 |
| Additional\_charges | Quantitative | Average amount charged for miscellaneous procedures | 17939.403420 |
| Item1 | Qualitative | Response rating regarding survey topic of the importance of timely admission | 3 |
| Item2 | Qualitative | Response rating regarding survey topic of the importance of timely treatment | 3 |
| Item3 | Qualitative | Response rating regarding survey topic of the importance of timely visits | 2 |
| Item4 | Qualitative | Response rating regarding survey topic of the importance of reliability | 2 |
| Item5 | Qualitative | Response rating regarding survey topic of the importance of options | 4 |
| Item6 | Qualitative | Response rating regarding survey topic of the importance of hours of treatment | 3 |
| Item7 | Qualitative | Response rating regarding survey topic of the importance of courteous staff | 3 |
| Item8 | Qualitative | Response rating regarding survey topic of the importance of evidence of active listening from the doctor | 4 |

**B1. Plan to Assess Quality of Data**

Assessment of the dataset was performed using Python and its multiple libraries. The Pandas library was used to load the raw data file into a tabular format as a data frame. After the data was loaded, an .info() method was used on the data frame to give an overview of all the variables, the number of non-null records, and the initial data types the variables were loaded in.

A few Pandas methods were used to detect duplicates in the dataset. The .nunuique() method was used to count the total number of unique records for the variables that should only contain unique values: case order, customer ID, interaction, and UID. Additionally, as a double check of the entire dataset, the .duplicated() method, chained together with the .value\_counts() method, was used to count the total duplicate records.

A couple of Pandas methods were used to detect missing values in the dataset. The .isna() method, chained with the .sum() method, was used to count each variable's missing records. Additionally, the data dictionary was used to determine the legitimacy of the missing records, including how each variable was originally represented and how Pandas would initially load that data into a data frame.

Boxplots were used as the primary method to detect outliers. Additionally, interquartile ranges were calculated for each variable, with outliers visually depicted from the boxplots. These values were then used to calculate the outliers' counts and ranges for the respective variables.

To first detect how categorical variables were represented when the data was imported, a .info() method was used. Afterward, a .value\_counts() method was used to look at the different responses for each variable.

**B2. Justification of Approach**

The .nunique() method was used to detect duplicates because it would return the total number of unique records for each variable. Since the original data file contained 10,000 records, if the resulting value were less than that, it would indicate possible duplicate records. Additionally, the .duplicated() method was used because it would return true values for any records that were duplicates. Otherwise, the record would be labeled as false. The .value\_counts() method was chained to this to aggregate and summarize the results for more straightforward interpretability.

The .isna() and the .sum() methods were used for the detection of missing values because the first method would determine and label each record as missing or not for each variable. Then, the second method would summate the missingness for each variable as Python treats true and false values as 1’s and 0’s, respectively.

For the detection of outliers, the primary method was using boxplots because not all variables were normally or uniformly distributed. Additionally, the boxplots provided a quick and easy determination of outliers through its simple-to-interpret visualization.

To detect whether the re-expression of categorical variables was necessary, the .info() method was used because it would display what data type each variable was loaded in. For the categorical variables, if the data types didn’t match, this would indicate an inconsistency and facilitate the necessity of re-expression. Additionally, the .value\_counts() method was used because the result displays the number of different responses and the counts for each of the respective responses. The variable could be converted into a categorical type based on how many different responses there were to utilize memory allocation for a large dataset better.

**B3. Justification of Tools**

Python was the primary programming language for this assessment mainly because of prior academic experience and familiarity. Python’s simple syntax makes it easy to read and write code consistently. Additionally, as it is open-source, there are a lot of additional libraries and packages that can be imported for use specifically related to data analytics and data science.

In this assessment, many libraries were imported to facilitate the data-cleaning process. Pandas was the main library to upload data into a tabular format, which could be further manipulated. NumPy was imported for use with numeric arrays and matrix calculations. Both Matplotlib and Seaborn were imported for visualization use. Finally, sklearn was imported for use with PCA.

**C1. Cleaning Findings**

The number of unique values was explicitly checked for four variables: case order, customer ID, interaction ID, and UID. Each variable returned a count of 10,000 unique values, thus indicating that there were no duplicates in the dataset. The .duplicated() method also returned all false values, suggesting no duplicates.

In detecting missing values, seven variables met this criterion: children, age, income, soft drink, overweight, anxiety, and initial days. The responses were checked against what the data dictionary had listed as possible responses to confirm the legitimacy of missing values for the categorical variables. Upon confirmation, the missing values were confirmed as legitimate. The variables and their missing values are listed below:

|  |  |
| --- | --- |
| **Variable** | **Count** |
| Children | 2588 |
| Age | 2414 |
| Income | 2464 |
| soft drink | 2467 |
| Overweight | 982 |
| Anxiety | 984 |
| Initial days | 1056 |

For the detection of outliers there were eight quantitative variables detected. The variables, counts of outliers, and ranges of outliers are listed below:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Count of Outliers** | **Range of Outliers** |
| Population | 855 | 33894 - 122814 |
| Children | 303 | 8 - 10 |
| Income | 252 | 106220 - 207250 |
| Vitamin D Levels | Lower: 26  Upper: 508 | Lower: 9 - 12  Upper: 24 - 54 |
| Full Meals Eaten | 8 | 6 - 7 |
| Vitamin D Supp | 70 | 3 - 5 |
| Total Charge | 466 | 14159 - 21525 |
| Additional Charge | 424 | 27088 - 30567 |

Upon checking how categorical variables were expressed, some inconsistencies were found. Instead of Yes/No as the expected response, there was a mix between 1/0 and Yes/No. Additionally, as qualitative variables were imported as an ‘object’ data type, these variables were checked for the feasibility of converting them to a ‘category’ data type existent within pandas for better memory utilization. The variables that matched this criterion were state, area, time zone, education, employment, marital, gender, initial admin, complication risk, and services. Lastly, as the zip code was imported, it was treated as an integer, so the leading zeroes were lost in translation, thus not meeting the 5-digit length of US zip codes.

**C2. Justification of Mitigation Techniques**

The dataset contained no duplicate records upon inspection, so there was no reason for the treatment. Most treatments focused on missing values, the reasoning for outliers, and the re-expression of categorical variables.

The missing values were treated utilizing univariate imputation to maintain statistical integrity. The missing values were treated this way because simply dropping the missing values would result in too much of a loss in data, resulting close to over 70% in a potential loss of data. Below are key points in each variable containing missing values that were treated:

* Children: treated using median as the histogram showed a right skew of data
* Age: treated using mean as the histogram showed a uniform distribution
* Income: treated using median as the histogram showed a right skew of data
* Initial Days: treated using median as the histogram showed a bimodal distribution
* Soft drink, overweight, and anxiety were all treated using the mode as they were categorical in nature

Although the data contained many outliers, none were removed; thus, no treatment was performed. Even if deemed outliers within the dataset, none are beyond a reasonable range that would warrant any treatment. Population numbers can vary widely per area, so its exclusion is unwarranted. The value of outliers for the ‘children’ variable does seem unusually high, but this only represents a small portion of the data. Additionally, these values may indicate that the patient’s household may comprise foster or children’s homes. Income ranges in the U.S. vary widely, so these outliers are reasonable to keep. Vitamin D levels and supplement values are reasonable because patients who need care are usually admitted to hospitals, so these values may indicate issues that need to be addressed regarding patient care. For the ‘full meals eaten’ variable, the range and count are barely beyond the normal and are kept, preserving data diversity. Both of the ‘charges’ variables are within a reasonable range as medical bills tend to vary widely depending upon the type and number of medical procedures performed.

For categorical variables with relatively few numbers of different responses, these were retyped as the Pandas ‘category’ data type. This treatment was done to enhance memory utilization and faster processing. These variables included state, area, time zone, education, employment, marital, gender, initial admin, complication risk, and services. Furthermore, the ‘zip’ variable was retyped to a string and then filled with zeroes to match the 5-digit length of a U.S. postal code.

Categorical variables with Yes/No responses had all their responses transformed using ordinal encoding for consistency. This included these variables: readmission, soft drink, high blood pressure, stroke, arthritis, diabetes, hyperlipidemia, back pain, allergic rhinitis, reflux esophagitis, and asthma.

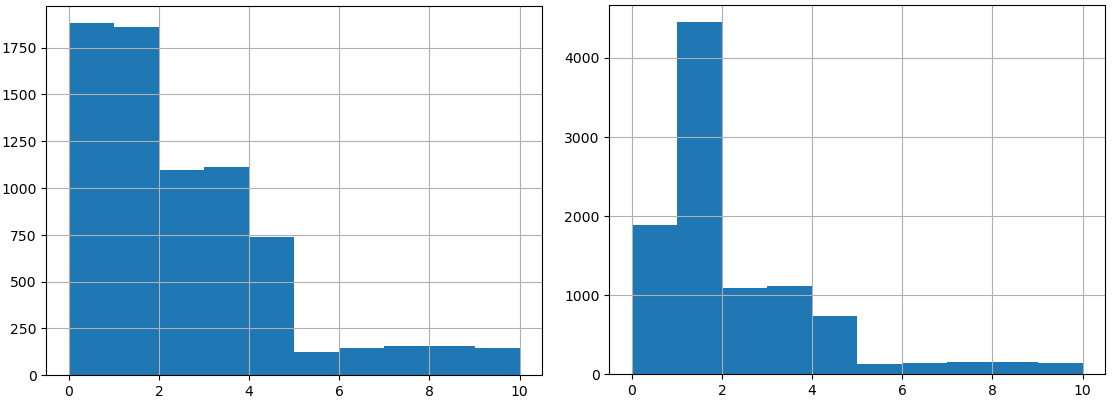
The survey responses were transformed into an ordered ‘category’ data type using Pandas’ CategoricalDtype class. The order was preserved using 1 (most important) > 8 (least important), as defined by the data dictionary. This was necessary as one would usually be less than eight if sorted typically.

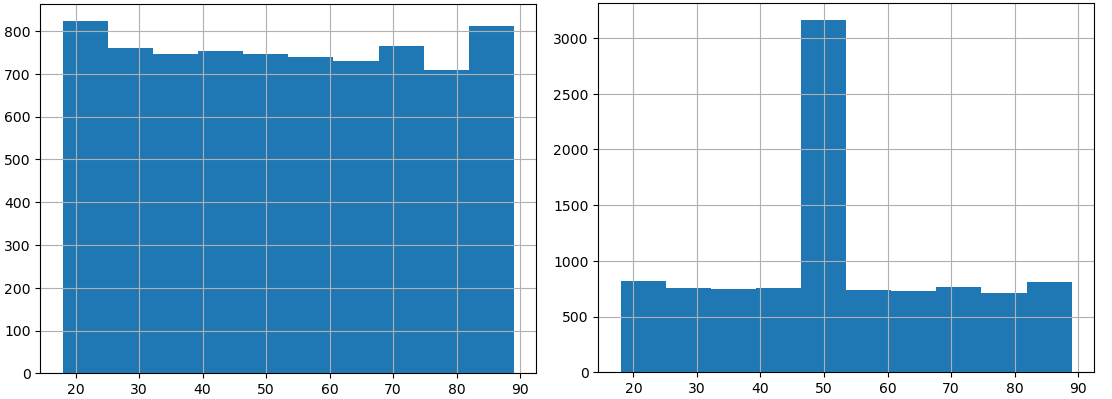
Other data transformations included re-casting children, age, and initial days as integers, as these variables are considered discrete in their perception. Lastly, the Vitamin D levels, total charge, and additional charge variables were rounded to two decimal places to conform to how they are presented typically.

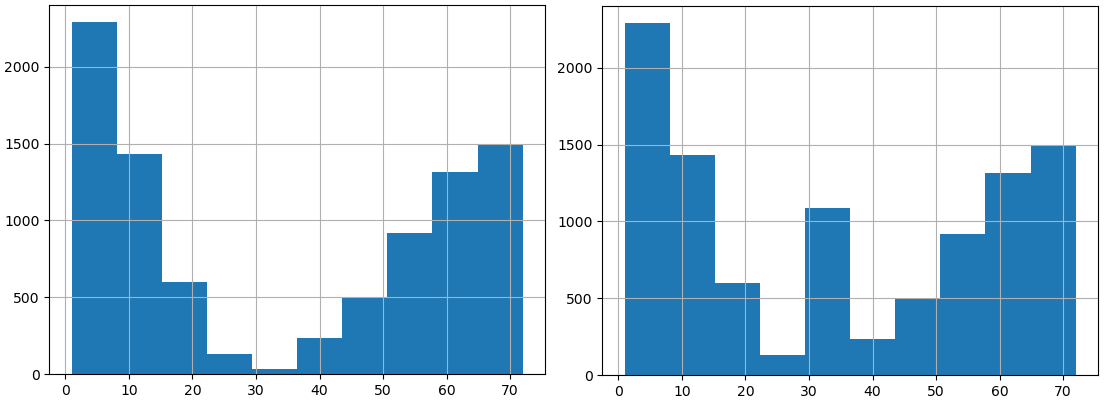
**C3. Summary of the Outcomes**

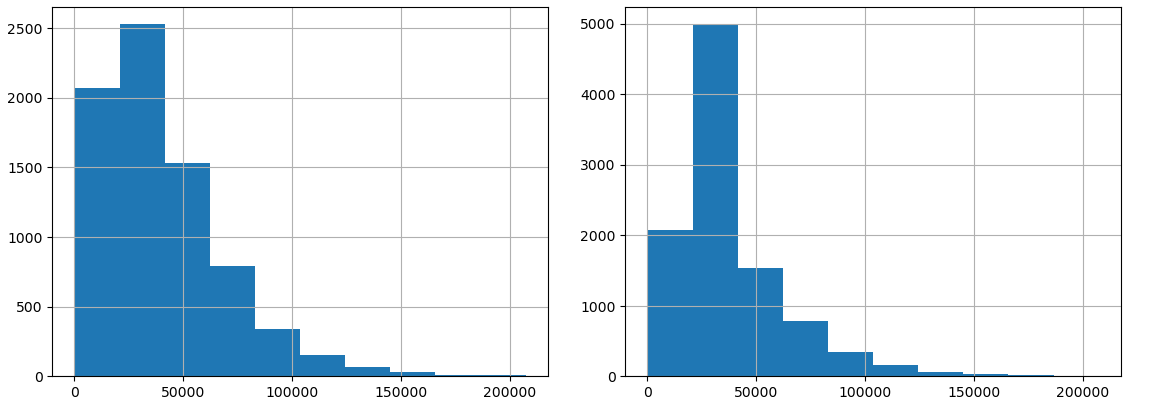
Outliers were kept in the data because the range of values was within a reasonable range. All missing values were treated utilizing univariate imputation methods. Categorical variables were re-expressed for consistency and more efficient memory utilization.

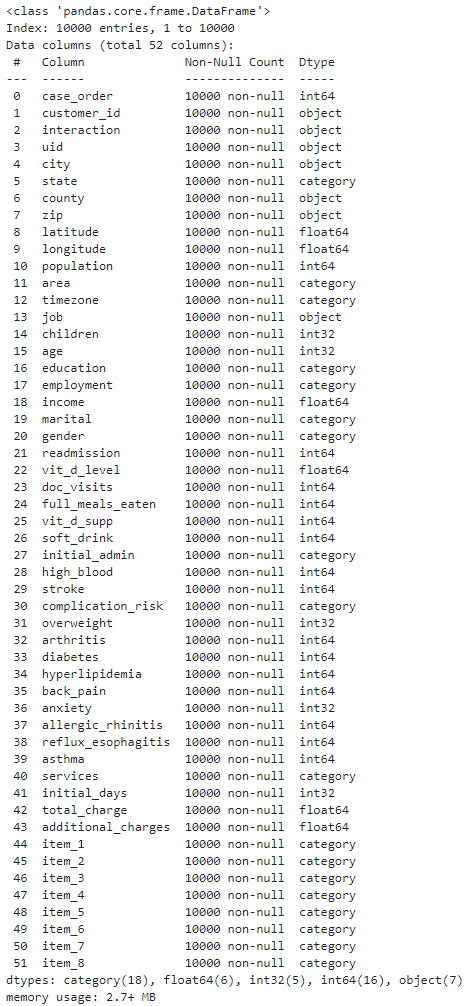
The figure displaying the data frame information shows 10,000 non-null values, thus indicating all missing values were dealt with. After the imputation of missing values for the quantitative variables, the histogram distribution still seems similar, albeit with spikes where the data was imputed. Additionally, many initial variables’ data types were cast to a more suitable one.

Figure 1. Children – Before & After

Figure 2. Age – Before & After

Figure 3. Initial Days – Before & After

Figure 4. Income – Before & After



**C4. Limitations**

One of the significant limitations while cleaning the data was the lack of domain knowledge. This lack of understanding limits the ability to tackle the data-cleaning process efficiently and optimally. This issue was especially prevalent when dealing with missing medical conditions data. It was challenging to decide whether to impute the mode or a specific value as this could impact patient care outcomes during analysis.

Another limitation was how missing data was treated. Since the method involved simple univariate imputation, this distorts how the data is distributed. This limitation can be seen with the new spikes in the histograms post-treatment.

Lastly, as outliers were kept in the dataset, this could severely affect the normality of statistical tests. These outliers affect the summary statistics of the data as well. Again, lacking domain knowledge makes it challenging to justify reasonable ranges at which outliers may exist entirely.

**C5. Impact of Limitations**

A significant impact of utilizing the cleaned dataset for analysis is that it may result in inaccurate analysis. The lack of domain knowledge during the cleaning step produces a dataset that may still contain a lot of underlying anomalies. Another issue is that data is cleaned for specific purposes, so this dataset may not be suitable for another type of analysis with a different focus.

How the data was cleaned could drastically affect the analysis results presented by the research question. The volume of imputations performed and the presence of outliers could drastically skew results. An example is the imputation of the ‘overweight’ and ‘age’ variables. It is intuitive to think that having an overweight condition and being older generally leads to a decline in health, so these imputed values would impact how results are generated. This analysis would lead to false conclusions about the major contributing factors.

**D1. Mitigation Code**

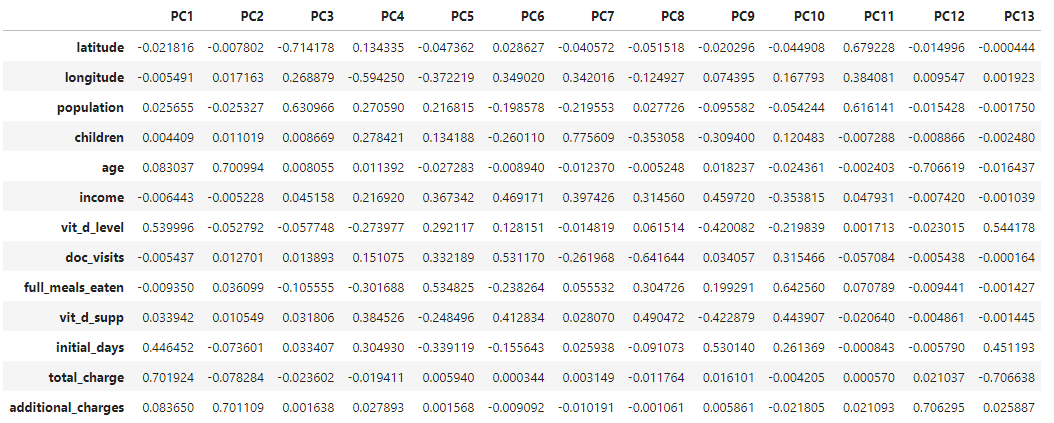
See code attached as D206\_PA\_Leng\_Yang\_code.ipynb.

**D2. Clean Data**

See cleaned data as a csv file attached as medical\_data\_clean.csv.

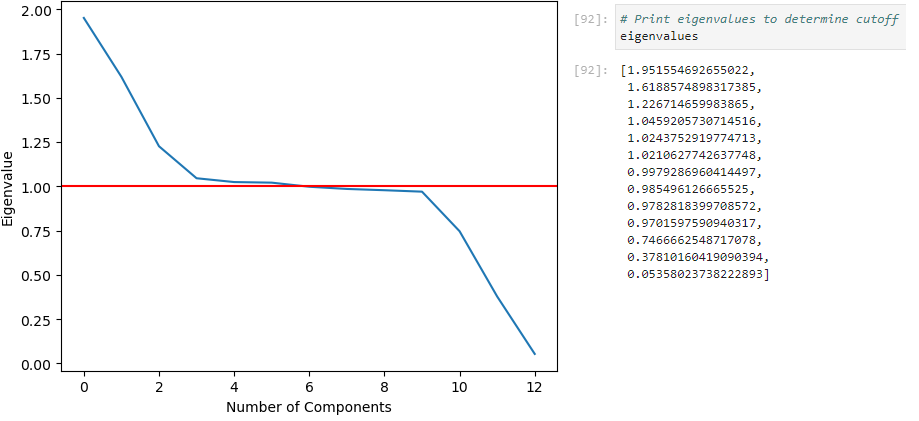
**E1. Principal Components**

A subset of the data was created using only quantitative variables to perform PCA. These variables included latitude, longitude, population, children, age, income, Vitamin D level, doctor visits, full meals eaten, Vitamin D supplement, initial days, total charge, and additional charges. Below is the loadings matrix of the PCA result.



**E2. Criteria Used**

The principal components (PCs) were selected using the Kaiser rule. This rule selects PCs with eigenvalues greater than 1. A scree plot was used to represent the eigenvalues to aid in this selection visually. Additionally, as it was hard to visually determine where the cutoff was, the eigenvalues were also looked at. Based on these results PC1, PC2, PC3, PC4, PC5, and PC6 should be retained.



**E3. Benefits**

The main benefit of performing PCA is that it reduces dimensionality. This reduction allows machine learning models to run more efficiently at the slight cost of model accuracy. In contrast, when running models without PCA, high-dimensional datasets would perform slowly or even require heavy resources due to the “curse of dimensionality” (Bigabid, 2023).

The organization can benefit from PCA because it would help reduce the dimensionality of their data. Given that the dataset contains many variables, machine learning models may run poorly without PCA, so PCA and PC selection would reduce dimensionality and complexity, thus drastically improving runtimes.

**F. Video**

A recording of the code being ran was also submitted and can be found here:

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4e57909e-909a-41d5-ad42-b1b30166e8fd>

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

*Categorical Data*. Categorical data - pandas 2.2.2 documentation. (n.d.). <https://pandas.pydata.org/pandas-docs/stable/user_guide/categorical.html>

Middleton, Keiona. (n.d.). *D206 – Getting Started with D206 | PCA*. Western Governors University. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95>

**H. Sources**

*What is Principal Component Analysis (PCA) & How to use it?*. Bigabid. (2023, February 8). <https://www.bigabid.com/what-is-pca-and-how-can-i-use-it/>