1. What factors influenced the different levels of the complication risk?
2. List Variables in the dataset

* CaseOrder – Quantitative, Example: 1

Index variable which is ordering the rows in this dataset

* Patient\_id – Qualitative, Example: C412403

Unique id that can identify each patient

* Interaction – Qualitative, Example: 8cd49b13-f45a-4b47-a2bd-173ffa932c2f

Another column of unique ID for transactions, procedures, and admissions for the patients

* UID – Qualitative, Example: 3a83ddb66e2ae73798bdf1d705dc0932

Like Interaction, UID is also an assigned to each patient for admissions, transactions, and procedures

* City – Qualitative, Example: Eva

Patient’s city of residence that is on the billing statement

* State – Qualitative, Example: AL

Patient’s state of residence that is on the billing statement

* County – Qualitative, Example: Morgan

Patient’s county of residence that is on the billing statement

* Zip – Qualitative, Example: 35621

Patient’s zip code of residence that is on the billing statement

* Lat – Quantitative, Example: 34.3496

Latitude from GPS coordination based on the patient’s residence on the billing statement

* Lng – Quantitative, Example: -86.72508

Longitude from GPS coordination based on the patient’s residence on the billing statement

* Population – Quantitative, Example: 2951

Population that is in one-mile radius of patient

* Area – Qualitative, Example: Suburban

Different types of areas that the patient is residing in

* Timezone – Qualitative, Example: America/Chicago

Time zone of the patient based on the patient’s residence during the sign-up

* Job – Qualitative, Example: Psychologist, sport and exercise

Title of a job title of the patient when patient reported during the sign-up

* Children – Quantitative, Example: 1

ow many children are in the patient’s household when patient reported during the sign-up

* Age – Quantitative, Example: 53

How old the patient is when patient reported during the time of sign-up

* Education – Qualitative, Example: Some College, Less than Year

Level of education the patient achieved when patient reported during the sign-up

* Employment – Qualitative, Example: Full Time

Employment status of the patient when patient reported during the sign-up

* Income – Quantitative, Example: 86575.93

Total annual income of the patient when patient reported during the sign-up

* Marital – Qualitative, Example: Divorced

Marital status of patient when patient reported during the sign-up

* Gender - Qualitative, Example: Male

Identifying the patient as male, female, or nonbinary

* ReAdmis – Qualitative, Example: No

Patient’s readmission within one month of release

* VitD\_levels – Quantitative, Example: 17.80233049

Vitamin D level of patient measured in nanogram per milliliter

* Doc\_visits – Quantitative, Example: 6

How many times the patient was visited by the primary physician during the initial hospitalization

* Full\_meals\_eaten – Quantitative, Example: 0

How many full meals the patient had eaten during the hospitalization

* VitD\_supp – Quantitative, Example: 0

How many times that the patient was administered the vitamin D supplements

* Soft\_drink – Qualitative, Example: NA

If the patient is drinking the soda minimum 3 or more times a day

* Initial\_admin - Qualitative, Example: Emergency Admission

How the patient was admitted to the hospital such as emergency, elective, or observation)

* HighBlood - Qualitative, Example: Yes

If the patient has a high blood pressure

* Stroke - Qualitative, Example: No
* If the patient has a stroke
* Complication\_risk - Qualitative, Example: Medium

The assessment of the patient by the primary doctor for the complication risk level

* Overweight – Qualitative, Example: 0

If the patient is overweight depending on patient’s gender, height, and age

* Arthritis - Qualitative, Example: Yes

If the patient has arthritis

* Diabetes - Qualitative, Example: Yes

If the patient has diabetes

* Hyperlipidemia - Qualitative, Example: No

If the patient has hyperlipidemia

* BackPain - Qualitative, Example: Yes

If the patient has a back pain

* Anxiety - Qualitative, Example: 1

If the patient has a disorder of anxiety

* Allergic\_rhinitis - Qualitative, Example: Yes

If the patient has an allergic rhinitis

* Reflux\_esophagitis - Qualitative, Example: No

If the patient has a reflux esophagitis

* Asthma – Qualitative, Example: Yes

If the patient has an asthma

* Services – Qualitative, Example: Blood Work

Primary service that was given to the patient during the hospitalization

* Initial\_days – Quantitative, Example: 10.58576971

How many days the patient was hospitalized during the initial visit

* TotalCharge – Quantitative, Example: 3191.048774

Average amount of daily charge for the patient. Total amount / number of days stayed at the hospital

* Additional\_charges - Quantitative, Example: 17939.40342

Any other miscellaneous treatments, procedures, anesthesiology, etc that was charged to the patient in daily average

* Item1 – Qualitative, Example: 5

A survey question of timely admission. This is from scale of 1 which is the most important to 8 with the least important.

* Item2 – Qualitative, Example: 5

A survey question of timely treatment. This is from scale of 1 which is the most important to 8 with the least important.

* Item3 – Qualitative, Example: 5

A survey question of timely visits. This is from scale of 1 which is the most important to 8 with the least important.

* Item4 – Qualitative, Example: 3

A survey question of reliability. This is from scale of 1 which is the most important to 8 with the least important.

* Item5 – Qualitative, Example: 4

A survey question of options. This is from scale of 1 which is the most important to 8 with the least important.

* Item6 – Qualitative, Example: 4

A survey question of hours of treatment. This is from scale of 1 which is the most important to 8 with the least important.

* Item7 – Qualitative, Example: 3

A survey question of courteous staff. This is from scale of 1 which is the most important to 8 with the least important.

* Item8 – Qualitative, Example: 4

A survey question of evidence of active listening from doctor. This is from scale of 1 which is the most important to 8 with the least important.

C1. Discuss the methods used to detect the data quality issues

1. Duplicates  
   Import the packages.  
   Read the dataset.  
   Use info() function to see a general detail of the data and duplicated() to find any duplicates.  
   Values\_counts().count() function will be conducted on CaseOrder, Customer\_id, Interaction, and UID columns to check any other duplicates on those columns only.
2. Missing Values  
   Use isnull().sum() to see the missing values in number.  
   Visaualize the data using missingno the data set and histograms for each column.
3. Outliers  
   Use seaborn to visualize the quantitative columns in boxplot.
4. Re-expression of categorical variables and other quality issues  
   Compare the actual values in the data, the given consideration and dictionary pdf for medical data, and the result of info() in the beginning for datatypes.  
   Use unique() to check the categories of qualitative columns and created another column for re-express in numeric.

C2. Why did I used the methods discussed in C1

1. Duplicates  
   To identify anomalies within the dataset, I commenced by importing the necessary libraries, loading the dataset, and scrutinizing the data types and presence of null values in each column using the info() function. Subsequently, I employed the duplicated() function to ascertain the existence of duplicate rows, and removing them if found. Further validation was performed utilizing the value\_contents().count() functions, specifically focusing on the CaseOrder, Customer\_id, Interactions, and UID, ensuring their uniqueness as per the requisite dataset constraints.
2. Missing Values  
   I employed the isnull().sum() function to assess the quantity of missing values, providing a numerical overview in contrast to the info() function. Additionally, I utilized Missingno to visualize the distribution of missing values across the entire dataset, complemented by histogram charts specifically tailored for quantitative columns exhibiting null values.
3. Outliers  
   I involved seaborn to generate boxplots for the quantitative columns, facilitating the identification of outliers. To refine the analysis, I employed the query() method to filter out values beyond the specified minimum and maximum thresholds.
4. Re-expression and others  
   Building upon the insights obtained from the info() function, I conducted an examination to ensure the alignment of column data types with the prescribed standards outlined in both the provided dataset and accompanying medical data guidelines. This involved cross-referencing actual values within the CSV file against the specified criteria and reference documentation. Furthermore, I employed the unique() function to pertinent columns slated for conversion to categorical types, meticulously scrutinizing for discrepancies such as capitalization, misspellings, leading spaces, and other potential anomalies.

C3. Which programming language used and Packages

Python was selected as the primary programming language for this project due to its intuitive nature and accessibility, making it conducive for rapid development and ease of comprehension.   
The utilized packages encompassed pandas, numpy, matplotlib.pyplot, Missingno, seaborn, statistics, and PCA from SKlearn. Pandas played a pivotal role in handling various tasks, including CSV file parsing, data manipulation, transformation, and storage. Numpy facilitated the conversion of outliers to nan values, ensuring data integrity. Matplotlib.pyplot and Missingno were leveraged for data visualization, aiding in the identification of anomalies more effectively than conventional tabular representations. Seaborn, in conjunction with boxplots, provided a concise means to detect outliers with quantitative columns. Statistics was instrumental in computing mode values for columns, enabling the imputation of missing values within categorical variables. Additionally, PCA will be also used to perform the PCA analysis.

C4.See code attached (Saemi Ramirez D206 Data Cleaning Detection.ipynb).

D1. What I found for treatment

1. Duplicates  
   The data did not have any duplicates.
2. Missing Values  
   Children column was missing 2588 values.  
   Age column was missing 2414 values.  
   Income column was missing 2464 values.  
   Soft\_drink column was missing 2467 values.  
   Overweight column was missing 982 values.  
   Anxiety column was missing 984 values.  
   Initial\_days column was missing 1056 values.
3. Outliers  
   No outliers were found in CaseOrder, Age, VitD\_levels, Initial\_days column.  
   Lat column had total 150 outliers when less than 25 or greater than 50.  
   Lng column had 628 outliers when less than -120.  
   Population had 805 outliers when greater than 35000.  
   Children had 457 outliers when greater than 6.  
   Income had 870 outliers when greater than 75000.  
   VitD\_levels had 549 outliers when less than 12 (42) or greater than 25 (507).  
   Full\_meals\_eaten had 8 outliers when greater than 5.  
   VitD\_supp had 70 outliers when greater than 2.  
   TotalCharge had 488 outliers when greater than 13750.  
   Additional\_charges had 358 outliers when greater than 27500.

D2. How I treated the anomalies

* 1. Duplicates  
     The info() function served the dual purpose of examining column data types and ascertaining the total count of values within each column. To identify duplicate entries within the dataset, I utilized the duplicated() function. Despite its comprehensive nature, duplicated() checks for duplicates across entire rows, prompting a supplementary verification step. Thus, I executed value\_counts().count() individually on the Case Order, Customer\_id, Interaction, and UID columns. By specifically scrutinizing these 4 columns, mandated to possess unique values, I ensured the absence of any duplications.
  2. Missing Values  
     Following the extensive output from info(), I employed isnull().sum() to provide a clear overview of missing values per column, enhancing their visibility within the dataset. To address these missing values, I initially utilized the Missingno package to visually assess their distribution across the entire dataset, subsequently complementing this analysis with histogram visualizations, particularly on quantitative columns to discern any patterns in their distributions. Before imputing the missing values, I used the describe() and median() function to check the value to fill the null values on each column to avoid the confusion if the graph trend differs from original. I utilized the median() function for Children and Income columns with skewed graphs, mean() function for Age displaying a relatively uniform distribution, another median() function for Initial\_days column for non-symmetric shaped histogram, and mode() function from statistics package for Soft\_drink Overweight, and Anxiety to fill the null values because they were categorical values.
  3. Outliers  
     A total of 14 quantitative columns were identified within the dataset. Among those, CaseOrder was excluded due to its sequential nature, serving solely as an identifier for row count purposes, while the Lat and Lng columns were omitted as they present geographical coordinates in GPS. Notably, Age, Docs\_visits, and Initial\_days were retained for analysis since they had no outliers.   
     Population column, despite containing approximately 8% ouliers, was deemed significant to the dataset’s integrity and thus retained.   
     Children exhibited around 5% outliers, their impact on the dataset was deemed negligible, warranting retention and treatment of outliers using the median.   
     Income, despite approximately 9% outliers, retained its validity and relevance within the dataset.  
     Outliers within the VitD\_levels column, appearing both below the minimum and above the maximum thresholds, were imputed with the median to ensure data consistency.  
     The presence of only 8 outliers for Full\_meals\_eaten and 70 for VitD\_supp, their minimal impact on the dataset led to their outlier imputation using the median.   
     Both TotalCharge and Additional\_charges were retained despite outliers accounting for approximately 5% of each column, as their significance to the dataset outweighed the presence of outliers, given their substantial numerical values.
  4. Re-expression and others  
     A comprehensive validation process was conducted utilizing the unique() method across the columns earmarked for categorical values. This step aimed to verify against potential issues such as misspelling, leading spaces, zero values, and capitalization errors.   
     Following are the 22 columns that were re-expressed using dictionary: Area, Timezone, Education, Employment, Marital, Gender, ReAdmis, Soft\_drink, Initial\_admin, HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, and Service.

D3. Summary

In summarizing the undertaken tasks, post CSV file ingestion, through checks for duplicates, missing values ,and outliers were conducted. Notably, no duplicates were detected, e ven upon examination of CaseOrder, Customer\_id, Interactions, and UID columns individually. Addressing missing values, a tailored approach was employed, utilizing median, mode, and mean imputation techniques, guided by graph type and the categorical nature of respective columns. The decision to retain outliers were based on their continued validity and significant impact on the dataset, either due to their prevalence or substantial numerical presentation. Outliers that did not meet these criteria were imputed with the median to maintain dataset integrity. Furthermore, 22 columns were re-categorized numerically for enhanced analytical clarity. The subsequent result from info() function and Missingno graph visually attest to the completion of data cleansing, with all null values effectively addressed.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A bar code with text

Description automatically generated

D4. See code attached (Saemi Ramirez D206 Data Cleaning Treatment.ipynb)

D5. See csv attached (Saemi Ramirez D206 Result.csv)

D6. Limitations

By imputing the null values and outliers, the limitations are loss of variability, impact on relationships between correlated variables, and information loss. Bias in summary statistics, misleading visualizations, reduced model performance, and misleading correlations (Pant, 2023) can also be limitations for retaining the outliers in the dataset.

D7. How the Limitations Affected the Analysis

The missing values in columns such as Children, Age, Income, and Initial\_days were addressed by imputing either the mean or median. Given the considerable range of missing values observed, from as low as 1056 to as high as 2588, the presence of a singular imputed value in these columns is notable and could significantly influence visualizations and subsequent analyses. Such imputation practices may obscure potential relationships between variables, potentially leading to erroneous conclusions or interpretations. Additionally, there is a risk of information loss, where important insights or trends inherent in the dataset might be masked as a result of this imputation strategy.  
Columns containing outliers, notably Population, Income, TotalCharge, and Additional\_charges, were retained despite their potential to skew statistical analysis and modeling outcomes, thereby introducing bias and inaccuracies. Furthermore, the presence of outliers can distort visualizations and summary statistics, leading to potentially misleading insights and interpretations.  
Regarding columns where outliers were imputed, such as Children, VitD\_levels, Full\_meals\_eaten, and vitD\_supp, the imputation process may have resulted in the loss of meaningful data points. Imputing outliers can obscure important patterns or trends within the dataset, potentially diminishing the overall quality and reliability of analyses and interpretations.

E1. PCA variables and loading matrix

I utilized the variables Lat, Lng, Income, VitD\_levels, Initial\_days, TotalCharge, and Additional\_charges for PCA. The subsequent screenshot depicts the PCA loading matrix.

A table with numbers and symbols

Description automatically generated

E2. Which PCs should be retained and why?

After generating a scree plot chart, it shows that PC1, PC2, PC3, and PC4 are important because their eigenvalues are greater or equal to 1.

**A graph with lines and numbers

Description automatically generated**

E3. **Benefit from PCA**

Business organizations can derive substantial benefits from integrating PCA models into their operations. In the realm of data collection and analysis, businesses often contend with a multitude of variables, yet not all are equally vital to their growth trajectory. Dimensionality reduction, a core function of PCA, categorically addresses this challenge by streaming the feature space. This reduction can be achieved through two primary methodologies: feature elimination and feature extraction (Brems, 2017). Feature elimination involves the removal of variables deemed nonessential to business growth. For instance, in the context of a clothing store, variables like open hours, locations, and neighborhood demographics may be pruned. Conversely, feature extraction entails the creation of new features derived from existing variables, such as delineating specific materials with particular design attributes.

F. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7d531d0d-7fac-4330-b64b-b163000b4691>

G. No third-party code references were used.

H. References

Brems, M. (2017, April 17). *A One-Stop Shop for Principal Component Analysis*. Medium. Retrieved April 30, 2024, from <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>

Pant, D. (2023, September 23). *Use and Downsides of Outliers.* Medium. Retrieved April 30, 2024, from <https://medium.com/@dhirajpant.ad/use-and-downsides-of-outliers-b0e5c074e3b5#:~:text=Outliers%20can%20significantly%20skew%20summary,of%20central%20tendency%20and%20spread> s