Part I. Research Question and Variables

A. Research Question

How do readmission rates vary across different levels of complication risk for specific medical conditions?

B. Variables in the data set

Variable Name	Data Type	Description	Example
CaseOrder	Quantitative	Placeholder used to preserve the initial sequence of the raw data file	25
Customer_id	Qualitative	An exclusive identifier assigned to each patient	Y563432
Interaction	Qualitative	Distinct IDs associated with patient transactions, procedures, and admissions	ca1e9204-516e-4145-869c-fdc66a5e3063
UID	Qualitative	Distinct IDs associated with patient transactions, procedures, and admissions	bf6eb8bbc86ade8a382121b6e07944ed
City	Qualitative	City where patient resides based on their billing statement	Quapaw
State	Qualitative	State where patient resides based on their billing statement	OK
County	Qualitative	County where patient resides based on their billing statement	Ottawa
Zip	Quantitative	Zip code where patient resides based on their billing statement	74363
Lat	Quantitative	GPS coordinates where patient resides based on their billing statement	36.95315
Lng	Quantitative	GPS coordinates where patient resides based on their billing statement	-94.7144
Population	Quantitative	Population residing within one-mile of the patient (from census data)	2879
Area	Qualitative	Type of area determined from unofficial census data	Rural
TimeZone	Qualitative	Time zone where patient resides based from their patient's sign-up information	America/Chicago
Job	Qualitative	Occupation of the patient or primary insurance holder as disclosed in the admissions form	Surveyor, minerals
Children	Quantitative	Indicates how many children in the patient's household	3
Age	Quantitative	Patient's age as disclosed in the admissions form	75

Education	Qualitative	The highest educational degree obtained by the patient as disclosed in the admissions form	Regular High School Diploma
Employment	Qualitative	Employment status of the patient as disclosed in the admissions form	Retired
Income	Quantitative	Yearly income of the patient or primary insurance holder as disclosed in the admissions form	32677.97
Marital	Qualitative	Marital status of the patient or primary insurance holder as disclosed in the admissions form	Separated
Gender	Qualitative	Self-identified gender of the patient as male, female or nonbinary	Male
ReAdmis	Qualitative	Indicates if the patient was readmitted within one month of release	No
VitD_levels	Quantitative	Vitamin D levels measured in ng/mL of the patient	16.49455
Doc_visits	Quantitative	Count of visits made by the primary physician during the initial hospitalization	6
Full_meals_eaten	Quantitative	Count of full meals consumed during hospitalization	1
VitD_supp	Quantitative	Frequency of administrating vitamin D supplements	0
Soft_drink	Qualitative	Indicates if the patient habitually consumes three or more sodas in a day	No
Initial_admin	Qualitative	How the patient's initial admission to the hospital occurred	Emergency Admission
HighBlood	Qualitative	Indicates if the patient has high blood pressure	Yes
Stroke	Qualitative	Indicates if the patient has experienced a stroke	No
Complication_risk	Qualitative	Degree of risk for complications evaluated through primary patient assessment	Medium
Overweight	Qualitative	Indicates if the patient is overweight based on age, gender, and height	1
Arthritis	Qualitative	Indicates if the patient has been diagnosed with arthritis	No
Diabetes	Qualitative	Indicates if the patient has been diagnosed with diabetes	Yes
Hyperlipidemia	Qualitative	Indicates if the patient has been diagnosed with hyperlipidemia	No
BackPain	Qualitative	Indicates if the patient experiences chronic back pain	Yes

Anxiety	Qualitative	Indicates if the patient has been diagnosed with an anxiety disorder	0
Allergic_rhinitis	Qualitative	Indicates if the patient has been diagnosed with allergic rhinitis	Yes
Reflux_esophagitis	Qualitative	Indicates if the patient has been diagnosed with reflux esophagitis	Yes
Asthma	Qualitative	Indicates if the patient has been diagnosed with asthma	No
Services	Qualitative	Primary services to the patient during hospitalization	Blood Work
Initial_days	Quantitative	Duration of the patient's stay in the hospital during the initial visit	11.64407
TotalCharge	Quantitative	Daily charge to the patient, representing an average per patient, calculated by dividing the total charge by the number of days hospitalized	3532.603
Additional_charges	Quantitative	Average charged billed to the patient for various miscellaneous items	12876.79
Item1	Quantitative	Customer's response when asked to rate the importance of timely admission	4
Item2	Quantitative	Customer's response when asked to rate the importance of timely treatment	4
Item3	Quantitative	Customer's response when asked to rate the importance of timely visits	3
Item4	Quantitative	Customer's response when asked to rate the importance of reliability	4
Item5	Quantitative	Customer's response when asked to rate the importance of options	2
Item6	Quantitative	Customer's response when asked to rate the importance of hours of treatment	3
Item7	Quantitative	Customer's response when asked to rate the importance of courteous staff	6
Item8	Quantitative	Customer's response when asked to rate the importance of evidence of active listening from doctor	4

Part II. Data Cleaning Plan

C1. What methods were used to detect the data quality issues?

To find duplicates, use **duplicated().** Missing values are discovered using **is.na(), miss_var_summary()** and **miss_case_summary()**. Outliers are visualized using **geom_boxplot()**,and **ggplot().** For reexpressing variables, **as.numeric()** and **as.factor()** was utilized for this analysis.

C2. Why use the methods were used to detect the data quality issues?

To detect duplicates, I primarily used **duplicated().** Additional capabilities using **count()** to count occurrences of a specific variable where n is greater than 1, which indicates duplicates, were also used for this analysis.

To find missing values, the **is.na()** was vital to indicate if a data frame has missing values. This method will display a 'True' for missing values and a 'False' for non-missing values. Furthermore, **sum()** and **colSums()** count the total missing values in the whole dataset and each column. **miss_var_summary()** and **miss_case_summary()** from naniar are used to summarize information about missing values at the column (variable) and row (case) levels, respectively.

To find outliers, visualization was used to highlight the extreme values and data points that fall outside of the typical range. The **geom_boxplot()** is a simple way to provide a visual representation as a boxplot of the distribution of a variable. The **ggplot()** can be used to make a box plot of multiple items to compare distributions.

Lastly, to reexpress variables in this analysis, **as.numeric()** and **as.factor()** were used in 17 variables to create uniformity. The **round()** was also used to round out the values for days to the nearest whole number.

C3. Why use R?

This data analysis and cleaning was performed in R. The language provides a more straightforward way to do statistical analysis and data visualization compared to Python. I found it quite effortless to perform data cleaning using R. It was a seamless transformation to different data types and formats. Data manipulation was also made simple due to the powerful libraries that facilitate efficient filtering, sorting, and transforming data. The libraries and packages were easy to install and use in Rstudio. The following packages were utilized for this analysis: tidyverse, naniar, and base R libraries.

Firstly, tidyverse provides several packages that are essential for analysts. This analysis used dplyr for data manipulation, ggplot2 for data visualization, tidyr for data tidying, reshape2 for data reshaping and plyr for more data manipulation. Naniar, on the other hand, is used to discover missing data. This package provides a summary and visualization, both vital to analyzing missing data patterns. Lastly, the base R libraries, such as readxl and stats, are used to read Excel files and perform statistics calculations, respectively.

Part III. Data Cleaning Treatment

D1. Discuss what you found

Using the previous code attached, here are my findings. The dataset has no duplicates. On the other hand, there are 12955 missing values or NAs. Children have the most missing values, with 2588 or 25.9% missing values. Soft drink takes the second with 2467 or 24.7% missing values. Income is next with 2464 or 24.6% missing values. Age has 2414 or 24.1% missing values. Initial days have 1056 or 10.6% missing values. Anxiety has 984 or 9.8% missing values. Lastly, overweight has 982 or 9.8% missing values.

There were 21 quantitative variables to detect for outliers. Namely, the case order, zip, latitude, longitude, population, children, age, income, vitamin D supplement, vitamin D levels, doctor visits, full meals eaten, initial days, total charge, additional charges, item1, item2, item3, item4, item5, item6, item7 and item8. Out of the 19 variables, 16 had outliers. Population had the most outliers, with 855

outliers. Vitamin D levels is next with 534 outliers. Total charges has 466 outliers. The survey items 4, 1, 3, 5, 6, 8, 7, and 2 have 450, 449, 443, 443, 443, 442, 438, and 429 outliers, respectively. Additional charges have 424 outliers. Children have 303 outliers. Income has 252 outliers. Longitude has 237 outliers. Latitude has 150 outliers. Vitamin D supplement has 70 outliers. Lastly, full meals eaten has 8 outliers. Case order, zip, age, doctor visits and initial days has no outliers.

D2. Discuss what you did to treat and why you used the treatment method

Since there are no duplicates, no treatment methods for duplicates were used.

To treat missing variables, the **mice** (Multivariate Imputation by Chained Equations) package is used to perform the imputation of missing data. Predictive Mean Matching (PMM) was employed as the imputation method. Each imputation needs predictor values to perform the process. The goal is to use information from other observed variables to impute the missing values. There were 7 variables: children, soft drink, income, age, initial days, anxiety, and overweight that had missing values. The mice package was used for all these variables. Marital, age, and employment were used as predictor values for children. Age, diabetes, and overweight were used as predictor values for soft drink. Education, employment, age, marital and children were used to impute income. Education, employment, income, marital and children were used to impute age. Initial admin, doc visits, services, high blood, and diabetes were used to impute initial days. Age, gender, income, education, and services were used to impute anxiety. Age, gender, high blood, diabetes, hyperlipidemia, arthritis, back pain, asthma, and soft drink were used to impute overweight. The imputation model was used for all these variables since it preserves the sample size. Retaining data is crucial when 4 out of the 9 variables have approximately 25% missing data. Dropping these cases might result in a smaller sample size, leading to less precision.

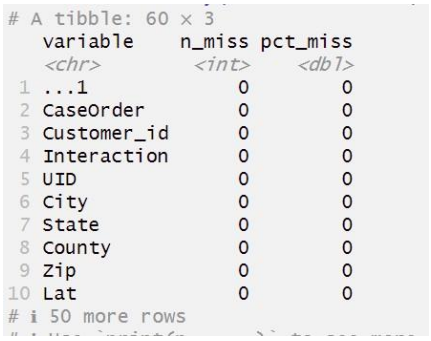
Winsorization, and retention methods were used to treat outliers. Using the dplyr package, winsorization was used in the variables population, age, income, vitamin D supplements, vitamin D levels, total charge, and additional charges. Winsorizing is replacing extreme values with less extreme ones using a specified data percentile. For population, the 91st percentile was used for the upper limit and the 9th percentile for the lower limit. For children, the 94th percentile was used for the upper limit, and the 6th percentile was used for the lower limit. For income, the 96th percentile was used for the upper limit, and the 4th percentile for the lower limit. For vitamin D supplements, the 99th percentile was used for the upper limit, and the 1st percentile was used for the lower limit. For vitamin D levels, the 94th percentile was used for the upper limit, and the 6th percentile was used for the lower limit. For the total charge, the 95th percentile was used for the upper limit, and the 5th percentile was used for the lower limit. For the additional charges, the 95th percentile was used for the upper limit, and the 5th percentile was used for the lower limit. Winsorizing helps address the outliers while preserving the sample size. Moreover, winsorizing was simple and flexible. Choosing the percentile to customize to the needs of the variable was a matter of trial and error. This process was made easier since there were visuals. For this analysis, the winsorized variables are in a new column to provide flexibility and maintain the integrity of the original data set. There were variables that I chose to retain. Variables like latitude and longitude were retained because these are geographical data. Item 1 to Item 8 was also retained since the outliers are expected for the data. The outlier values were within the range of values in the survey questions. Lastly, full meals eaten outliers were retained since there were only 8 cases. This is 0.08% of the cases, which I consider to be insignificant.

Multiple variables underwent reexpression. Children, age, income, and initial days were originally classified as char (character) data type. These variables were converted into numeric, with initial days rounded to the nearest whole number. Since these variables are numeric, they should be converted to numeric. Soft drink and medical conditions: high blood, stroke, overweight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, reflux esophagitis, and asthma were converted to num (numeric) data type. “Yes ” was revalued to 1, “No” was revalued to 0, and NA stayed as NA. Since these variables are Boolean variables, converting them to numeric values shows uniformity throughout the data set. Consistency and standardization make it easier for different analysts to consistently work with the same data. Complication risk was revalued as a factor with levels “Low,” “Medium,” and “High.” This reexpresion to a factor enables analysis to reflect the distinct categories. The levels also help convey the meaning of each category and improve the understanding of the data.

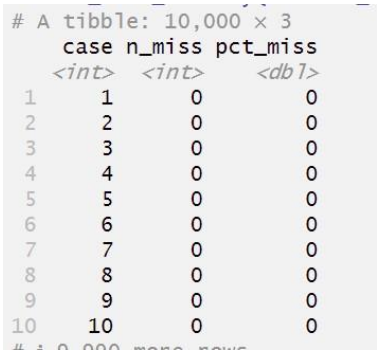
D3. Summarize all the work performed

No duplicates were found.

Missing values were imputed using predictor values. Using the mice package, imputation was performed for several variables: children, soft drink, income, age, initial days, anxiety, and overweight. This iterative process is repeated for each variable until all missing data is handled. Using the naniar package to show missing data, the result is now 0 missing and 0% missing.



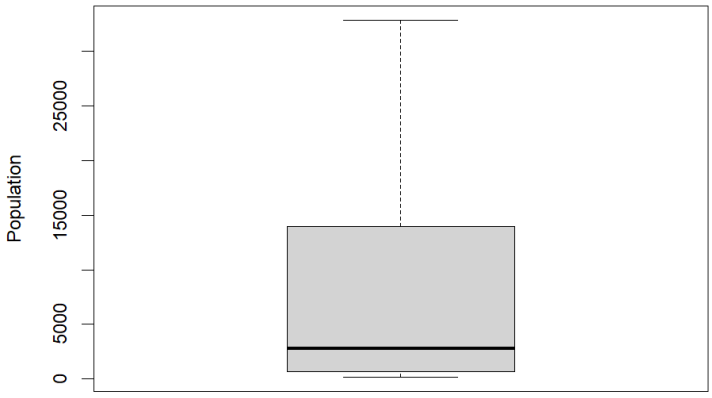
```
# A tibble: 60 × 3
   variable    n_miss pct_miss
   <chr>        <int>    <dbl>
 1 ...1             0        0
 2 CaseOrder        0        0
 3 Customer_id      0        0
 4 Interaction      0        0
 5 UID              0        0
 6 City             0        0
 7 State            0        0
 8 County           0        0
 9 Zip              0        0
10 Lat              0        0
# i 50 more rows
```



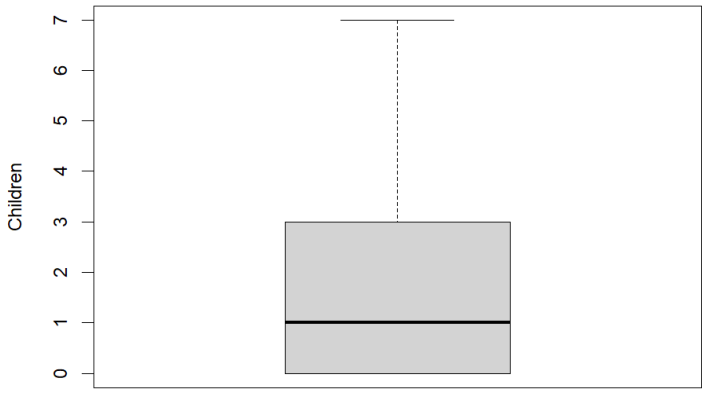
```
# A tibble: 10,000 × 3
    case n_miss pct_miss
   <int>  <int>    <dbl>
 1     1      0        0
 2     2      0        0
 3     3      0        0
 4     4      0        0
 5     5      0        0
 6     6      0        0
 7     7      0        0
 8     8      0        0
 9     9      0        0
10    10      0        0
```

Outliers were winsorized or retained. Using the dplyr package, winsorizing was performed for several variables, such as population, children, income, vitamin D supplement, vitamin D levels, total charge, and additional charges. For the winsorized variables, boxplots show that there are no outliers. Winsorizing involves setting lower and upper limits for each variable, which are then used to replace the extreme values. The winsorized variables are in new columns to maintain data integrity. Boxplots for the new winsorized column show no outliers (see below). Variables that were retained remained the same.

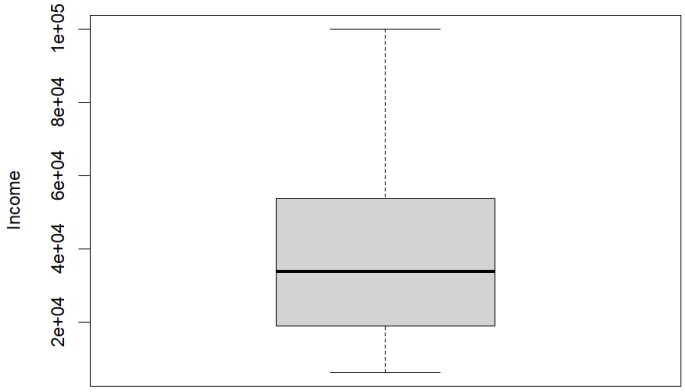
Box Plot of Population (winsorized)



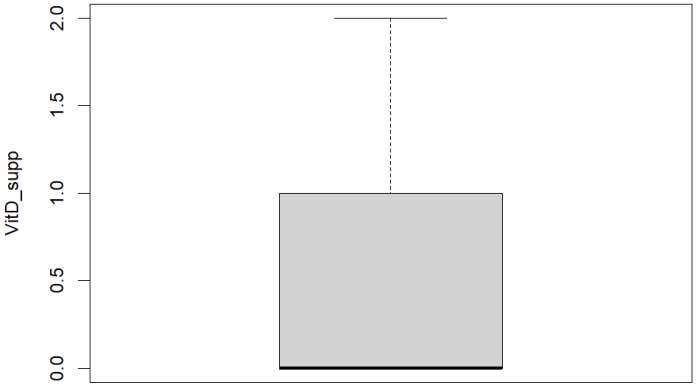
Box Plot of Children (winsorized)



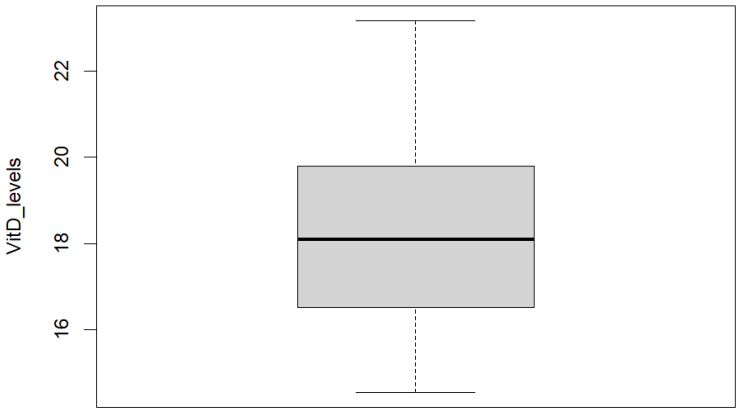
Box Plot of Income (winsorized)



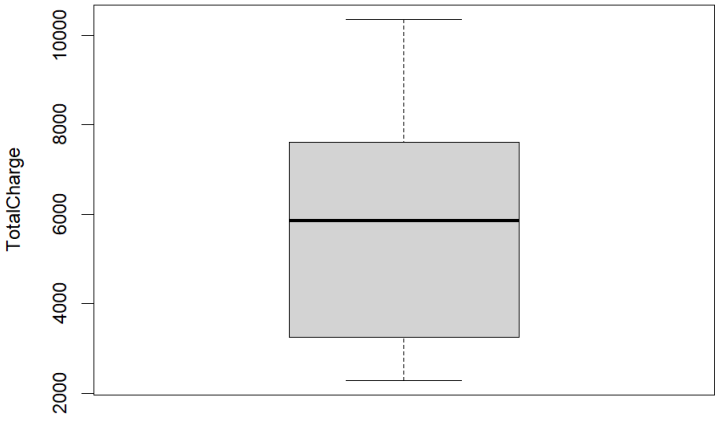
Box Plot of Vitamin D supplement (winsorized)

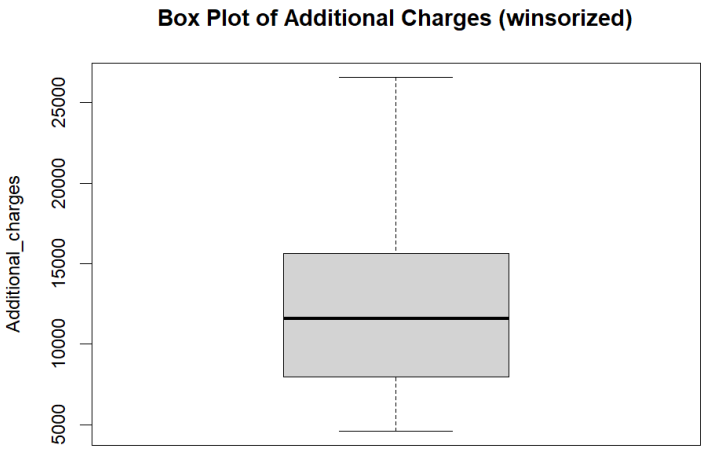


Box Plot of Vitamin D levels (winsorized)



Box Plot of Total Charge (winsorized)






```
> # Print or inspect the outlier counts
> print(outlier_counts)
                      Variable Count
1         Population_winsorized     0
2           Children_winsorized     0
3             income_winsorized     0
4          VitD_supp_winsorized     0
5        VitD_levels_winsorized     0
6        TotalCharge_winsorized     0
7 Additional_charges_winsorized     0
>
```

Before treatment of the missing data and outliers, variables need to be reexpressed into their correct data type. Variables such as children, age, income, and initial days were first to be reexpressed. These variables were converted to numeric format, ensuring null values stay null. The variables soft drink and medical conditions such as high blood, stroke, arthritis, diabetes, hyperlipidemia, back pain, allergic rhinitis, reflux esophagitis, and asthma were converted to numeric format with Yes as 1, No as 0, and NA as NA. This reexpression enhances the dataset's consistency by ensuring numeric formats for numeric variables. The final dataset is now more standardized and uniform. Below is the str() function that shows the reexpressed variables' final data type, length, components, and attributes.

```
> str(subset_df)
tibble [10,000 × 4] (S3: tbl_df/tbl/data.frame)
 $ Children    : num [1:10000] 1 3 3 0 0 1 0 7 0 2 ...
 $ Age         : num [1:10000] 53 51 53 78 22 76 50 40 48 78 ...
 $ Income      : num [1:10000] 86576 46806 14370 39741 1210 ...
 $ Initial_days: num [1:10000] 11 15 5 2 1 6 9 63 6 2 ...
>
```

```
> str(subset_df)
tibble [10,000 × 13] (S3: tbl_df/tbl/data.frame)
 $ Soft_drink       : num [1:10000] 0 0 0 0 1 0 0 0 0 0 ...
 $ HighBlood        : num [1:10000] 1 1 1 0 0 0 1 0 0 1 ...
 $ Stroke           : num [1:10000] 0 0 0 1 0 0 0 0 0 0 ...
 $ Complication_risk : Factor w/ 3 levels "Low","Medium",..: 2 3 2 2 1 2 1 2 1 3 ...
 $ Overweight       : num [1:10000] 0 1 1 0 0 1 1 1 1 1 ...
 $ Arthritis        : num [1:10000] 1 0 0 1 0 1 1 0 0 0 ...
 $ Diabetes         : num [1:10000] 1 0 1 0 0 1 1 0 0 0 ...
 $ Hyperlipidemia   : num [1:10000] 0 0 0 0 1 0 1 0 1 0 ...
 $ BackPain         : num [1:10000] 1 0 0 0 0 1 1 0 0 0 ...
 $ Anxiety          : num [1:10000] 1 1 0 0 0 0 1 0 0 0 ...
 $ Allergic_rhinitis : num [1:10000] 1 0 0 0 1 1 0 0 0 1 ...
 $ Reflux_esophagitis: num [1:10000] 0 1 0 1 0 0 1 0 0 1 ...
 $ Asthma           : num [1:10000] 1 0 0 1 0 0 0 0 0 1 ...
>
```

D6. Disadvantages of the methods used

There are drawbacks to the imputation method in handling missing data. Firstly, the method assumes that there is a linear relationship between variables. If it is nonlinear, imputed values may be inaccurate to the true pattern in the data. It is also important to note that the imputed values are estimates. It is important to note that the process introduces uncertainty, which does not fully capture the complexity of the missing data.

Winsorizing can be beneficial, but it has its disadvantages. Firstly, the process assumes a symmetric distribution of data. It may introduce a distortion in the distribution and affect statistical interpretations. There is also a risk that valid extreme values will be replaced. If genuine outliers are present, this process could result in the loss of important cases.

Furthermore, reexpression is subjective and dependent on me. Different choices may lead to different results and interpretations, especially for converting the medical conditions into numeric. This may also introduce a loss of information. Extreme transformation can obscure patterns that might be important for the dataset.

D7. Challenges from the now-cleaned data

There are potential challenges and limitations to using my cleaned data.

Firstly, imputing missing values involves making assumptions about the relationship between variables. if the assumptions are inaccurate, imputed values could introduce bias into the analysis. Moreover, winsorizing may have altered the distribution of variables. The variability might cause distortion and affect the statistical properties of the variables when analyzing variables. In addition, variable conversion loss and data loss through NAs handling may lead to oversimplification of the information they originally carried. This could impact the identification of nuanced relationships between variables.

Overall, it is crucial for the data analyst to assess the data themselves. They should read the report on data cleaning procedures and how it will impact the research question.

Part IV. PCA

E1. Variables used for PCA and a screenshot of the PCA loadings matrix

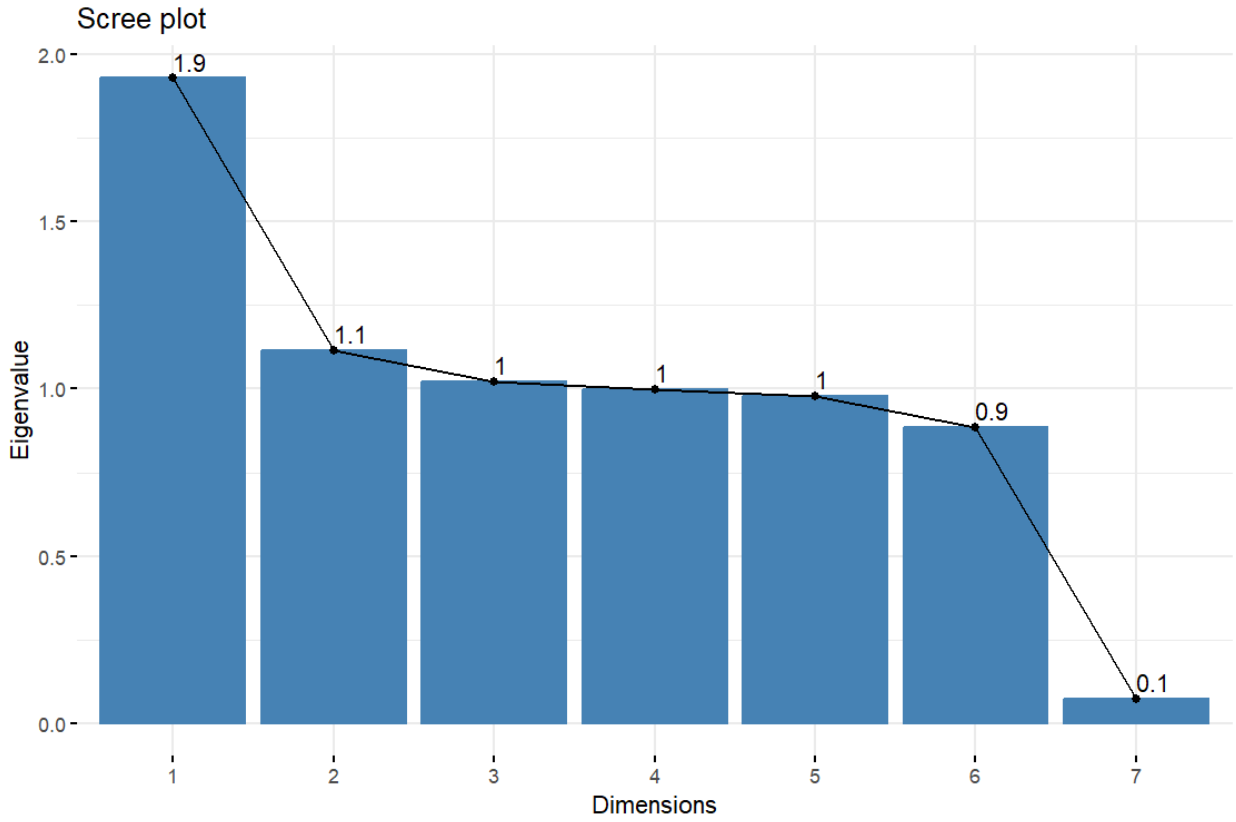
Variables used for Principal Component Analysis should be continuous quantitative variables. The following are the variables used for PCA: lat, lng, income, vitamin D levels, initial days, total charge and additional charges.

```
> print(loadings_df)
                        PC1           PC2           PC3           PC4
Lat                0.014546951 -0.7019559400  0.0841109701 -0.035234710
Lng                0.008196176  0.6982531946 -0.0295750878  0.123204420
Income             0.014950139  0.1314372560  0.2632209918 -0.895366758
VitD_levels       -0.555991423  0.0243477430  0.4257961198 -0.046353917
Initial_days      -0.436278615 -0.0374290050 -0.5699978301  0.019730326
TotalCharge       -0.706625609 -0.0004031842 -0.0005267956 -0.009942033
Additional_charges -0.026691926  0.0207527491  0.6454186462  0.423391946
                        PC5          PC6          PC7
Lat                0.072164237 -0.70251149 -0.000180710
Lng                0.108144496 -0.69614352  0.003201432
Income            -0.322240183 -0.08770220 -0.001846086
VitD_levels        0.440530174  0.06257330  0.555740453
Initial_days      -0.532392381 -0.09567077  0.436324060
TotalCharge        0.001949093 -0.01341152 -0.707387656
Additional_charges -0.633834422 -0.03036353  0.019049342
>
```

E2. PCs that should be retained and why

Use the **fviz_eig()** to display the eigenvalues associated with each principal component. Analysts can easily make informed decisions about how many principal components to include by visually inspecting the scree plot. The Kaiser rule suggests retaining values with eigenvalues greater or equal to 1. These components with higher values are considered to have more variance than an individual original variable, so make sure to retain them.

In the scree plot below, principal components 1, 2, 3, 4, and 5 should be kept. PC1 has the highest score out of all the 7 PCs and should be further analyzed.



E3. Organization benefits

Principal Component Analysis is a powerful tool in data analysis. It can simplify the intricacies of even the most complex data sets loaded with variables. Moreover, PCA is very versatile and practical. Reducing dimensionality helps improve the efficiency of machine learning models, combating the "curse of dimensionality" (Bigabid). It can also deal with high-dimensional data that might overwhelm models. Not only that, PCA aids in noise reduction and provides a means to visualize and understand complex data structures.

Overall, incorporating PCA into data analytics can improve workflow. Organizations can enhance algorithm performance and derive valuable insights from complex data sets.

References

Kabacoff, Robert I. Ph.D. (2017). Reshaping Data. *Quick-R by DataCamp.* https://www.statmethods.net/management/reshape.html

quantile: Sample Quantiles. (n.d.). *RDocumentation.* *https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile*

PCA: What, How, and Why. (n.d.). Bigabid. https://www.bigabid.com/what-is-pca-and-how-can-i-use-it/

Middleton, Keiona. (n.d.). Getting Started with D206 | Principal Component Analysis (PCA) [PowerPoint Slides]. D206 Data Cleaning, Western Governors University.

Radečić ,Dario. (2023, January 10). Imputation in R: Top 3 Ways for Imputing Missing Data. *Appsilon.* https://appsilon.com/imputation-in-r/