Part I: Research Question

1. Are patients with specific characteristics, whether personal or medical, more likely to be readmitted than others?
2. Below, I have provided a table detailing each variable included in the dataset "medical\_raw\_data.csv" along with the description as provided in the data dictionary. In order to discover the information present in the "DATA TYPE" and "COMMENTS" columns, the csv was loaded into R Studio as a data frame and the function summary called upon that data frame to gather summary statistics and data-type information. Next, the function unique was called on the character type variables in order to determine how many categories existed within the variable and if any of them were masks for null, NA, values. Finally, the examples found in the "EXAMPLES" column were taken from the first, non-null entry in the data frame for each column, primarily the first row of data when sorted by CaseOrder.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| VARIABLE NAME | DESCRIPTION | DATA TYPE | COMMENTS | EXAMPLE |
| CaseOrder | A column provided in order to preserve the initial ordering of the dataset. | Numerical, Discrete | This variable serves no purpose other than for sorting the data into its original order. | 1 |
| Customer\_id | Unique patient ID. | Character, Indexing | This is a suitable candidate for an index column. | "C412403" |
| Interaction | Unique IDs related to patient transactions, procedures, and admissions. | Character, Indexing | This column is a unique identifier, but the format is much more complex to work with than the customer ID; recommend dropping. | "8cd49b13-f45a-4b47-a2bd-173ffa932c2f" |
| UID | Unique IDs related to patient transactions, procedures, and admissions. | Character, Indexing | Same as Interaction. | "3a83ddb66e2ae  73798bdf1d705dc0932" |
| City | Patient city of residence as listed on the billing statement | Character, Categorical | There are 6072 unique categories in this variable, making it very poor as a categorical data point. | "Eva" |
| State | Patient state of residence as listed on the billing statement | Character, Categorical | There are 52 states, indicating some issue, or out-of-country states, in the data. Requires further analysis. | "AL" |
| County | Patient county of residence as listed on the billing statement | Character, Categorical | With 1607 unique entries, this is a much better categorical variable when requiring location. | "Morgan" |
| Zip | Patient zip code of residence as listed on the billing statement | Numerical, Categorical | With 8612 zip-codes included, this is a poor categorical variable. Though it is easier to use than the city variable, as the values are numerical, it would still require encoding. | 35621 |
| Lat | GPS coordinates of patient residence as listed on the billing statement | Numerical, Continuous | This variable is useful for visualization; specifically, we can pair this variable with Lng to create a heat map of patient locations. | 34.34960 |
| Lng | GPS coordinates of patient residence as listed on the billing statement | Numerical, Continuous | This variable is useful for visualization; specifically, we can pair this variable with Lat to create a heat map of patient locations. | - 86.72508 |
| Population | Population within a mile radius of patient, based on census data | Numerical, Discrete | There are no null values in this variable, but the minimum is zero, which is impossible. Requires further investigation. | 2951 |
| Area | Area type (rural, urban, suburban), based on unofficial census data | Character, Categorical | This variable contains no null values and only 3 categories. | "Suburban" |
| Timezone | Time zone of patient residence based on patient's sign-up information | Character, Categorical | This variable contains no null values and 26 categories. Recommend checking that included time zones exist. | "America/Chicago" |
| Job | Job of the patient (or primary insurance holder) as reported in the admissions information | Character, Categorical | With 639 categories, and little relevance to the medical field, this variable is unlikely to be useful in analysis unless we specifically designed the question to include it. | "Psychologist, sport and exercise" |
| Children | Number of children in the patient's household as reported in the admissions information | Numerical, Discrete | There are 2588 null values in the column and the max, 10, is significantly larger than the 3rd quartile, 3, implying outliers. | 1 |
| Age | Age of the patient as reported in admissions information | Numerical, Discrete | This variable includes 2414 null values and no apparent outliers or anomalies upon first look. The minimum age is 18, meaning they included no minors in the report. | 53 |
| Education | Highest earned degree of patient as reported in admissions information | Character, Ordinal | This variable contains no nulls and 12 unique categories. | "Some College, Less than 1 Year" |
| Employment | Employment status of patient as reported in admissions information | Character, Categorical | This variable contains no nulls and five categories. | "Full Time" |
| Income | Annual income of the patient (or primary insurance holder) as reported at time of admission | Numerical, Continuous | This variable includes 2464 nulls. The minimum and maximum are both clear outliers. | 86575.93 |
| Marital | Marital status of the patient (or primary insurance holder) as reported on admission information | Character, Categorical | This variable contains no null values and five categories. | "Divorced" |
| Gender | Customer self-identification as male, female, or nonbinary | Character, Categorical | This variable contains no nulls and three categories. One option is "Prefer not to answer," which could be used as a placeholder for nulls. | "Male" |
| ReAdmis | Whether the patient was readmitted within a month of release or not (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "No" |
| VitD\_levels | The patient's vitamin D levels as measured in ng/mL | Numerical, Continuous | This variable appears to have outliers on the high-end as the maximum is well above the 3rd quartile. | 17.80233 |
| Doc\_visits | Number of times the primary physician visited the patient during the initial hospitalization | Numerical, Discrete | This variable has no null values and appears to be without outliers upon initial inspection. | 6 |
| Full\_meals\_eaten | Number of full meals the patient ate while hospitalized (partial meals count as 0, and some patients had more than three meals in a day if requested) | Numerical, Discrete | Counting partial meals as zero immediately presents a problem and potentially skews the data. This variable also contains potential outliers at the high-end. | 0 |
| VitD\_supp | The number of times that vitamin D supplements were administered to the patient | Numerical, Discrete | This variable's median value is zero, with a maximum of five. This variable will need considerable attention to be made fruitful. | 0 |
| Soft\_drink | Whether the patient habitually drinks three or more sodas in a day (yes, no) | Character, Categorical | This variable contains null values. | "No" |
| Initial\_admin | The means by which the patient was admitted into the hospital initially (emergency admission, elective admission, observation) | Character, Categorical | This variable contains no null values and three categories. | "Emergency Admission" |
| HighBlood | Whether the patient has high blood pressure (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "Yes" |
| Stroke | Whether the patient has had a stroke (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "No" |
| Complication\_risk | Level of complication risk for the patient as assessed by a primary patient assessment (high, medium, low) | Character, Ordinal | This variable contains three categories, "Low," "Medium," and "High", and no nulls. | "Medium" |
| Overweight | Whether the patient is considered overweight based on age, gender, and height (yes, no) | Numerical, Discrete, Categorical | This variable contains 982 nulls and the categories are already numerically encoded. | 0 |
| Arthritis | Whether the patient has arthritis (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "Yes" |
| Diabetes | Whether the patient has diabetes (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "Yes" |
| Hyperlipidemia | Whether the patient has hyperlipidemia (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "No" |
| BackPain | Whether the patient has chronic back pain (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "Yes" |
| Anxiety | Whether the patient has an anxiety disorder (yes, no) | Numerical, Discrete, Categorical | 984 nulls, categories already numerically encoded. | 1 |
| Allergic\_rhinitis | Whether the patient has allergic rhinitis (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "Yes" |
| Reflux\_esophagitis | Whether the patient has reflux esophagitis (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "No" |
| Asthma | Whether the patient has asthma (yes, no) | Character, Categorical | This variable is a binary categorical variable with no null values. | "Yes" |
| Services | Primary service the patient received while hospitalized (blood work, intravenous, CT scan, MRI) | Character, Categorical | This variable contains no nulls and four categories. | "Blood Work" |
| Initial\_days | The number of days the patient stayed in the hospital during the initial visit | Numerical, Continuous | This variable contains 1056 null values. | 10.585770 |
| TotalCharge | The amount charged to the patient daily. This value reflects an average per patient based on the total charge divided by the number of days hospitalized. This amount reflects the typical charges billed to patients, not including specialized treatments | Numerical, Continuous | There are apparent outliers on the high- and low-ends of this variable. It contains no null values. | 3191.049 |
| Additional\_charges | The average amount charged to the patient for miscellaneous procedures, treatments, medicines, anesthesiology, etc. | Numerical, Continuous | This variable displays a significant range of values with potential outliers on the high-end as the max is nearly double the 3rd quartile value. | 17939.403 |
| Item1 | On a scale of 1 to 8 (1 = most important, 8 = least important): Timely Admission | Numerical, Discrete, Ordinal | The scale is set up so that smaller numbers represent greater importance. Reversing this scale will be important for calculation and visualization. | 3 |
| Item2 | On a scale of 1 to 8 (1 = most important, 8 = least important): Timely Treatment | Numerical, Discrete, Ordinal | The scale is set up so that smaller numbers represent greater importance. Reversing this scale will be important for calculation and visualization. | 3 |
| Item3 | On a scale of 1 to 8 (1 = most important, 8 = least important): Timely Visits | Numerical, Discrete, Ordinal | The scale is set up so that smaller numbers represent greater importance. Reversing this scale will be important for calculation and visualization. | 2 |
| Item4 | On a scale of 1 to 8 (1 = most important, 8 = least important): Reliability | Numerical, Discrete, Ordinal | The scale is set up so that smaller numbers represent greater importance. Reversing this scale will be important for calculation and visualization. | 2 |
| Item5 | On a scale of 1 to 8 (1 = most important, 8 = least important): Options | Numerical, Discrete, Ordinal | The scale is set up so that smaller numbers represent greater importance. Reversing this scale will be important for calculation and visualization. | 4 |
| Item6 | On a scale of 1 to 8 (1 = most important, 8 = least important): Hours of Treatment | Numerical, Discrete, Ordinal | The scale is set up so that smaller numbers represent greater importance. Reversing this scale will be important for calculation and visualization. | 3 |
| Item7 | On a scale of 1 to 8 (1 = most important, 8 = least important): Courteous Staff | Numerical, Discrete, Ordinal | The scale is set up so that smaller numbers represent greater importance. Reversing this scale will be important for calculation and visualization. | 3 |
| Item8 | On a scale of 1 to 8 (1 = most important, 8 = least important): Evidence of active listening from doctor | Numerical, Discrete, Ordinal | The scale is set up so that smaller numbers represent greater importance. Reversing this scale will be important for calculation and visualization. | 4 |

The above rows highlighted in green are the variables which I believe best apply to the question presented in Part A.

**Part II: Data-Cleaning Plan**

1. Explain the plan for cleaning the data by doing the following:
   1. The primary aim is to identify data points that have missing or null values. We will identify null values by using the summary function from base R. The output of this function will help us identify null values in the data by counting the NA values in each numeric variable and displaying the categorical values within each categorical variable. However, the summary cannot show us everything, specifically within categorical variables, and more investigation will be required.

There are three particular cases in which a variable will require additional analysis: first, all numerical variables, with or without NA values, should be examined for outliers through the use of box plots; second, any categorical variables that display values that are questionable, or have too many categories to visualize with the summary command, will need to be explored further through use of the levels function; finally, some variables appear to have issues in their definition, determinable through reading the data dictionary, such as the Item variables which require having their scale inverted to be properly interpretable, or could otherwise benefit from data-wrangling to better fit the data to the specific question being asked. The latter of this final category will be identified and noted in Part D but will not be wrangled as this process is not designed to ensure the data is model ready for any specific question, but ready to be explored and wrangled once a question has been decided upon or assigned.

In order to identify the NA or missing values, the summary function will be used on the whole dataset. The output of this function will allow us to see what variables are numeric, and if these numeric variables contain NA values, or if they are categorical variables, and what the categorical values within are. This summary will also give us an idea of what variables need further analysis. For example, the Population variable containing values less than one is illogical; additionally, the County variable has too many factors to visualize in the summary and must be investigated with the levels function.

* 1. By setting the parameter stringsAsFactors equal to TRUE, we can load the data in such a way that any character type variables will be a factor, recording each unique occurrence into a level of the variable it occurs in. This will make detecting anomalous values easier as we will only need to sort through each unique occurrence for categorical variables rather than the whole dataset. Another advantage of loading character variables as factors is that the summary function will show us the levels of these variables in its output. For those variables with a considerable number of levels, the summary will not list all the levels but identify the variables meeting this criterion and for these we will have to use the levels function to identify each unique option, again only possible when the character data is loaded as a factor.

The summary function is especially useful when evaluating numerical variables, as the summary output includes the minimum and maximum values, the median value, the mean value, and the first and third quartile values. With this information, we can quickly identify potential errors in the data, as well as identify variables with potential outliers using the inter-quartile range and the maximum and minimum values.

These steps allow us to gather a strong baseline of what variables require inspection, have NA values, or are otherwise noteworthy. It provides a cursory overview of the distribution of both numeric and categorical variables, which allows us some insight into the data and what relationships may be worth investigating in later phases of the analysis.

* 1. We will complete this task using R in R Studio with the additional packages dplyr for quick, easy management of data.frame objects (the select and inner\_join functions specifically), as well as factoextra for the Principal Component Analysis. R is a robust language for statistical analysis and the inclusion of the dplyr package allows for vectors and matrices to be easily moved between data.frame objects and computed on.
  2. The code used to implement this plan is provided below and we provide an interpretation of the resulting output at each step in Part D.

med\_raw\_data <- read.csv(file="medical\_raw\_data.csv", stringsAsFactors = TRUE)

summary(med\_raw\_data)

options(max.print=5000)

length(levels(med\_raw\_data$City))

levels(med\_raw\_data$City)[1:5000]

levels(med\_raw\_data$City)[5001:6072]

levels(med\_raw\_data$State)

levels(med\_raw\_data$County)

levels(med\_raw\_data$Timezone)

levels(med\_raw\_data$Job)

levels(med\_raw\_data$Education)

**Part III: Data Cleaning**

1. Summarize the data-cleaning process by doing the following:
   1. To begin, we load the required packages for examining and cleaning the data, dplyr, and performing the Principal Component Analysis, factoextra. Next, we direct the environment to set the working directory to the local folder containing our data, script, and eventually the output file for the cleaned dataset. Now, we load the raw csv data into a data.frame named "med\_raw\_data" and display a summary for each variable, which details summary statistics of numerical variables and the categorical values available to character variables. We explicitly loaded these character variables as the factor data type with the stringsAsFactors=TRUE parameter for easier visualization in the summary function and easier numerical coding in later steps.

The output of the summary output shows us what variables are included and a general overview of their characteristics. Starting from the first variable and moving through each, we find a variable X that exactly replicates the CaseOrder variable which we know from the data dictionary provided to be a placeholder variable used to preserve the initial ordering of the raw data and appears to have been generated from the read.csv function as an index. As the variable X is not present in the original data but is, rather, generated by the read.csv function, it can be safely discarded. Moving further along, we find that the Customer\_id, Interaction, and UID variables each contain 10,000 unique factors. These variables are unique and more meaningful to identifying the case and we should choose one as an index as opposed to the X or CaseOrder variables, which contain unique values for each observation but contain no information useful to identifying the case itself. Looking at the values for these variables, we see that the Customer\_id variable is a six-digit code beginning with a letter and followed by five numbers, while the Interaction and UID variables are composed of complex, alphanumeric character values; it is easy to see that the Customer\_id variable is easier to read and write and is our ideal choice for an index variable.

The City, State, and County variables come next in the summary, and we see that each has many categories stored within them that will need to be examined with the use of the levels function, performed later in the code.

Next up in our summary is the Zip variable, which stores zip codes numerically. This is not the best way to store these values, as the zip code '00610' is both a real zip code and a value in our dataset, though it is currently represented numerically as 610, which is not a valid Zip. If the research question chosen depends on information within the Zip variable, it may be necessary to make this change to avoid misinterpretation. Despite this, the variable exhibits no NA values and requires no intervention.

The Lat and Lng variables represent the patient's home location in terms of latitude and longitude, respectively. In order to examine these values, we will later plot these variables against one another to see if there are any values physically out of place, though the variable contains no NA values and appears properly distributed at first glance.

In the Population variable, we run into our first issue of missing or incorrectly recorded data. We see in the summary that the minimum value is zero, an impossible value as even someone living alone in a house with no neighbors within a mile radius should have a population of one, as the customer represented must be counted as well. These zero values must be dropped or replaced.

The summary shows us that the Area variable only has three factors and no NA values, telling us that the variable does not require any intervention. The following variables, Timezone and Job, however, are not as easily cleared as there appear to be many factors that will need to be examined more closely through application of the level function later in the code.

Looking now at the Children and Age variables, we see that both are numeric variables with many NA values. These NA values can be replaced with zero for the Children variable as a non-response to the question of having children can be interpreted as not having any children, hence why the number of children is uncountable. We should replace the NA values for Age with the median age of the dataset. We can identify potential outliers in the Children variable as the maximum value of ten is significantly greater than the third-quartile value of three. We also notice that the Age variable does not contain any values less than 18, meaning our research question from Part A must be edited to account for the lack of data on adolescent patients, becoming: "Are *adult* patients with specific characteristics, whether personal or medical, more likely to be readmitted than others?"

The Education variable is shown to have many factors which will require investigation through the levels function and the Employment variable shows itself to have a clear set of factors which can be arranged hierarchically, full-time employment being considered better than part-time which is considered better than being unemployed. This variable contains no null values.

Income is shown to contain NA values and outliers, both at the bottom and top of the distribution. The minimum, $154.10, is considerably less than the first-quartile value, $19450.80, and the maximum, $207249.10, is well above the third-quartile value, $54075.20.

Next, we encounter three categorical variables with no NA values and factors that are sensible and few. These variables are Marital, Gender, and ReAdmis. Regarding the factor of "Prefer not to answer" in the Gender variable, this response will not be considered as an NA value as it represents the group of persons who do not identify within the binary system of gender and is, therefore, a valid response to the question.

VitD\_levels show no NA values. There are potential outliers on the top end because the maximum value, 53.019, is significantly higher than the third-quartile value, 19.790. The next variable in the summary, Doc\_visits, similarly shows no NA values or significant outliers.

Next, we encounter another problematic variable: Full\_meals\_eaten. According to the data dictionary, this variable only records *complete* meals eaten. This means it will record a patient who consistently eats half of each meal they receive as having eaten zero meals, as will a patient who refuses to eat at all. The data has many entries of zero as well as clearly identifiable outliers in the top end, as the maximum of seven dramatically out scales the third-quartile value of two. It is difficult to contend with the problems presented by a variable like this, where information is recorded accurately but calculated misleadingly; the best recommendation is to not use such a variable in later analysis based on this data or to use the data to create a more useful variable such as meals eaten per day of stay.

The VitD\_supp variable records how many supplements of vitamin D were administered, and the summary shows that the median is zero, meaning at least half of all patients received none. While not technically incorrect, this variable is a poor indicator variable, as it is so poorly distributed.

The variables from Soft\_drink to Services are all categorical variables with a low number of potential values. Variables in this section that have NA values include Soft\_drink, Overweight, and Anxiety. Responses of "No" can replace these NA values, represented by 0 in Overweight and Anxiety which are numerically encoded yes or no responses. Each of the other variables in this section exhibit no signs of anomalous values.

Initial\_days is a numerical variable shown to have NA values by the summary output and with a wide range of values. TotalCharge is shown to contain no NA values but displays potential outliers on the top end as the maximum, 21524, greatly exceeds the value of the third-quartile, 7615. Additional\_charges display a lack of NA values as well.

Finally, we come to our final eight variables: the customer feedback responses. In the data dictionary, they clarify that they score the survey like golf, meaning a lower value represents a more important response. Unfortunately, this will be difficult for our computer to understand and will need to be adjusted so that the range is preserved but inverted.

Moving away from our summary, we return to our code to examine variables identified in the summary as requiring further investigation. First, we will expand the maximum print output to 5000. The first variable to be examined is the City variable. We examine the length of the levels of the variable to discover that there are 6072 unique cities represented here. As this value is above 5000, we will have to separate the output of the levels into two commands. Examining the 6072 city names, no anomalies present themselves and we can conclude that the values are ready for encoding at a later step. We will perform this level analysis for the variables State, County, Timezone, Job, and Education as we identified them above as having many levels of Factors.

The State variable shows no discrepancies but, interestingly, includes Puerto Rico and the District of Columbia as states, leading to 52 values in the State variable. The County variable has some easily identifiable values that have some discrepancies. The following county names are not valid and will need to be scrubbed from the dataset later: 'AÃ±asco', 'BayamÃ³n', 'DoÃ±a Ana', 'MayagÃ¼ez', 'PeÃ±uelas', 'RincÃ³n'. We could clean the Timezone variable to not include so much extra information as all but "Pacific/Honolulu" contain "America/" at the beginning. The variable may be more useful for computation if it were encoded as the offset from +00:00 Greenwich time or encoded numerically, depending on the needs of the research question. The Job variable has a similar occurrence to that observed in Timezone where there appear to be categories and subcategories, for instance: "Therapist, drama" and "Therapist, music." These values could be divided and trimmed depending on the level of specificity desired by the research question. Finally, the Education variable exhibits no anomalous values.

* + 1. Once our numerical data has been cleansed of NA values and standardized, we can examine the numeric variables to identify outliers through the creation of box plots. The box plots created are included below.

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

In the first plot, we can see the latitude and longitude plotted against one another, providing us with a visualization of the physical distribution of our dataset. The plot shows us that all our patients live in the United States, including Hawaii and Alaska, or the U.S. territory of Puerto Rico. From this, we can assert that there are no values in the Lat or Lng variables that are incorrectly entered or impossible.

The remaining box plots show us that there are a significant number of outliers in some of our numeric variables. Most notably, the Population and Income variables appear to have a significant number and distribution of outliers. Interestingly, within the VitD\_levels variable, there is a separation between two groups, patients with low values and patients with high values.

Outliers are a source of much discussion in statistical analysis, recent studies showing that their removal can be detrimental to the final analysis by creating an increase in Type I errors, commonly referred to as false positives (Gress, Denvir, &Shapiro, 2). For this reason, we will not be removing outliers at this stage of the cleaning. It may be pertinent to address these outliers in later stages of the analysis if the research question would benefit from such wrangling, so it is still important to identify and acknowledge these outliers early in the data cleaning process.

* 1. In order to clean the data more efficiently, we will divide the data into categories depending on the data type of the variables within and the required cleaning performed. In line 20 we create our first division, assigning the variables Zip, Lat, Lng, Population, Children, Age, Income, VitD\_levels, Doc\_visits, Full\_meals\_eaten, VitD\_supp, Initial\_days, TotalCharge, and Additional\_charges to the data.frame med\_data\_num with the Customer\_id variable included as an index.

In our above analysis of the data, we saw that the Population variable included several incorrect values as values of 0. For R to understand that these values are inaccurate, we must label them as NA values using the na\_if function from the dplyr package.

We are now ready to impute values for our NA values in the categorical variables. As many of these variables have a high number of NA values, being defined here as a variable with over 10% of its values as NA, it would be detrimental to drop every patient with an NA value in one of these variables, as we could quickly lose a tremendous fraction of our observations. Instead of this approach, we will impute the median value of the variable in which the NA value exists and replace that value with the imputed value. We have selected the median here as opposed to the mean, as the median is less affected by outliers (Gelman & Hill, 532) within the variable and several numerical variables were detected as having potential outliers in the previous analysis. We see the left skew of these data points in the box plots presented in Part D.1.a.

Now that the NA values have been addressed we can standardize the numerical values, ensuring that the minimum and maximum values are zero and one, respectively. Standardization prevents the PCA from placing extra weight on variables with higher values by ensuring each numerical value has the same range, minimum, and maximum (Portland State University). The "standardize" function computes the standardization for each variable. We apply this function to each variable in the table of numerical variables through a for loop.

The next section of the raw data to be cleaned is the variables Item1 through Item8, which will be stored in the data.frame med\_data\_items with the Customer\_id variable as an index.

As discussed in Part C, the Item variables suffer from an inverted scale in which low values have a higher significance. Because a computer cannot understand this, we must invert the values within these eight variables by subtracting the current value from nine, meaning an initial value of eight becomes one and an initial value of one becomes eight. We will also standardize these values using a slight variation of our standardization function from the numerical variables.

Finally, we move on to our last subsection of the variables, the categorical variables. The data.frame med\_data\_cat includes the categorical variables Overweight, Anxiety, Complication\_risk, Education, Employment, City, State, County, Area, Timezone, Job, Marital, Gender, Initial\_admin, Services, ReAdmis, Soft\_drink, HighBlood, Stroke, Arthritis, Diabetes, Hyperlipidemia, BackPain, Allergic\_rhinitis, Reflux\_esophagitis, and Asthma, accompanied by the Customer\_id variable as an index.

While we could clean the other sections of the data as a batch, many of these variables will need individual consideration and cleaning. The first of these variables to be cleaned is the County variable, which has values that are nonsensical, as identified in Part C, that must be removed. We remove the offending county names with a for loop. This loop identifies and replaces the offending values with NA values which are then identified by the is.na function. Any rows identified as containing an NA in this column are dropped from the data.frame. We can drop the NA values from this column because of the relatively low number of values that are involved.

The other variables in this set that contain NA values are Anxiety, Overweight, and Soft\_drink. The Overweight and Anxiety variables contain less than 1000 NA values, as shown by a summary of the data in its current state, but the Soft\_drink variable contains 2467. In order to avoid trimming our dataset too much, we will replace NA values similarly to how we did with the numerical values discussed previously; however, we will not be using the median value but a value equivalent to a response of "No." We do this by interpreting a lack of a response as a response of "No" because of the yes or no nature of the values. With the diagnostic variables, Overweight and Anxiety, having not received a diagnosis, whether as an explicit response of "No" to the diagnosis or no response, indicates that there is no diagnosis. Likewise, if it did not explicitly state a patient as having been habitually drinking soft drinks, it is safest to assume that they do not drink them habitually. We will be numerically encoding our categorical variables for easier interpretation. The Overweight and Anxiety variables are already numerically encoded, but the Soft\_drink variable requires encoding. Because of this, the Overweight and Anxiety variables can simply have any NA values replaced with 0; Soft\_drinks, however, will need to have values of "Yes" encoded as 1 and all other values, "No" or NA, encoded as 0. For now, we will replace NA values in Soft\_drink with "No" and encode these values in a batch with the other yes or no variables requiring encoding.

Before moving on to the plethora of yes or no variables present in the set, we must first contend with the three ordinal variables in our set: Complication\_risk, Education, and Employment. For each of these variables, we will use the factor function from the dplyr package to create a numerical value for each level within the variable and ensure it is returned as a number by using the as.numeric function from base R. We will arrange the levels such that a higher value represents a higher amount of the categorical variable; for example, a complication risk of "High" indicates more risk and is therefore given a higher value than a patient with a "Low" risk. We explored the levels for these variables in the initial summary and following levels functions in the initial analysis conducted in Part C.

For the remaining categorical variables with responses other than yes or no (City, State, County, Area, Timezone, Job, Marital, Gender, Initial\_admin, and Services) for which the values do not belong in any specific order, we perform a similar operation as we did for the ordinal variables but allowing the levels to be organized in the order that they first appear in the data.

We also encoded the yes or no variables with "No" values being assigned a value of 1 and "Yes" values being assigned a value of 2 which will be converted to 0 and 1 respectively when the values are standardized in the next step.

Now that all our variables in each of the three data.frames have been cleaned and encoded, it is time to reunite our data. First, we must not forget the variables we left out for the cleaning process: CaseOrder, Interaction, and UID. We create a data.frame containing these two variables and the Customer\_id and use the dplyr package's inner\_join function to join this data.frame with the med\_data\_num data.frame using Customer\_id as the join index. Next, we join the med\_data\_items and med\_data\_cat data.frames using Customer\_id as the join index. Finally, we inner\_join the data.frames created by these joins with one another to finish uniting all our variables into one cleaned dataset.

* 1. In the summary output from Line 35, we see that the Population variable contains values between 1 and 122,814 with 109 NA values which were imputed in place of the anomalous values of zero. The output summary from Line 46 shows us that all the NA values in the med\_data\_num data.frame have been replaced successfully. In line 55, the output shows us that the numerical variables are all successfully standardized.

Line 78 provides a summary showing that the Item variables all contain values between one and eight, standardized to a range of zero to one after being inverted. The results are better summarized in the following box plot of the Item variables:

Chart, box and whisker chart

Description automatically generated

The box plot above, created in Line 80, shows us the distribution of the Item variables, identifying the mean of each as a score of either five or six, with most responses being between four and seven.

Line 110 summarizes the remaining data after dropping the rows containing anomalous values in the County variable. This summary shows 9986 remaining values, meaning we lost only 14 observations through this process.

A summary of the data.frame med\_data\_cat performed on Line 135 reveals that none of the variables within contain any NA values and are ready to be encoded.

The summaries on Lines 139, 154, 162, 167, and 172 confirm that we have correctly encoded the categorical variables on a scale relevant to the number of levels in the variable. Line 177 summarizes evidence that each of the encoded categorical variables are now standardized.

The final summary from Line 190 shows that all variables have been successfully reunited and that all variables are numerical and standardized, completing the cleaning process for the data, which is now ready for export. We export the data to a csv file by using the function write.csv on our cleaned dataset.

* 1. The R code used to complete the cleaning process is included here as well as an attachment to the submission.

#Packages

library(dplyr)

library(factoextra)

#Setting the working Directory

#Replace with your directory where the medical\_raw\_data.csv file is located

setwd("D:/Documents/WGU\_MSDA/NUM2/Task\_1\_Data")

#Loading the data

med\_raw\_data <- read.csv(file="medical\_raw\_data.csv", stringsAsFactors = TRUE)

#Part D.1

summary(med\_raw\_data)

options(max.print=5000)

length(levels(med\_raw\_data$City))

levels(med\_raw\_data$City)[1:5000]

levels(med\_raw\_data$City)[5001:6072]

levels(med\_raw\_data$State)

levels(med\_raw\_data$County)

levels(med\_raw\_data$Timezone)

levels(med\_raw\_data$Job)

levels(med\_raw\_data$Education)

#Part D.2

med\_data\_num <- select(med\_raw\_data, Customer\_id, Zip, Lat,

Lng, Population, Children, Age, Income, VitD\_levels, Doc\_visits, Full\_meals\_eaten, VitD\_supp, Initial\_days, TotalCharge, Additional\_charges)

summary(med\_data\_num)

for (x in 1:9999) {

med\_data\_num[x,5] <- na\_if(med\_data\_num[x,5], 0)

}

summary(med\_data\_num['Population'])

for (j in 2:13) {

i <- 0

for (x in is.na(med\_data\_num[,j])) {

i <- i + 1

if(x == TRUE) {

med\_data\_num[i,j] <- median(med\_data\_num[!is.na(med\_data\_num[,j]),j])

}

}

}

summary(med\_data\_num)

standardize <- function(x){

(x-min(x))/(max(x)-min(x))

}

for (x in 2:15) {

med\_data\_num[,x] <- standardize(med\_data\_num[,x])

}

summary(med\_data\_num)

#Part D.1.a

plot(med\_data\_num$Lng, med\_data\_num$Lat)

boxplot(med\_data\_num[,c(2,5:7)])

boxplot(med\_data\_num[,8:11])

boxplot(med\_data\_num[,12:15])

#Part D.2 resumes

med\_data\_items <- select(med\_raw\_data, Customer\_id, Item1, Item2, Item3, Item4,

Item5, Item6, Item7, Item8)

summary(med\_data\_items)

standardize\_items <- function(x){

(x)/(8)

}

for (x in 2:9) {

med\_data\_items[,x] <- 9 - med\_data\_items[,x]

med\_data\_items[,x] <- standardize\_items(med\_data\_items[,x])

}

summary(med\_data\_items)

boxplot(med\_data\_items[,2:9])

med\_data\_cat <- select(med\_raw\_data, Customer\_id, Overweight,

Anxiety, Complication\_risk, Education, Employment, City, State, County, Area, Timezone, Job, Marital, Gender, Initial\_admin, Services, ReAdmis, Soft\_drink, HighBlood, Stroke, Arthritis, Diabetes, Hyperlipidemia, BackPain, Allergic\_rhinitis, Reflux\_esophagitis, Asthma)

summary(med\_data\_cat)

i <- 0

for (x in med\_data\_cat$County) {

i <- i + 1

if (x == 'AÃ±asco') {

med\_data\_cat[i,9] <- NA

} else if (x == 'BayamÃ³n') {

med\_data\_cat[i,9] <- NA

} else if (x == 'DoÃ±a Ana') {

med\_data\_cat[i,9] <- NA

} else if (x == 'MayagÃ¼ez') {

med\_data\_cat[i,9] <- NA

} else if (x == 'PeÃ±uelas') {

med\_data\_cat[i,9] <- NA

} else if (x == 'RincÃ³n') {

med\_data\_cat[i,9] <- NA

}

}

med\_data\_cat <- med\_data\_cat[!is.na(med\_data\_cat[,9]),]

summary(med\_data\_cat)

i <- 0

for (x in is.na(med\_data\_cat$Anxiety)) {

i <- i + 1

if(x == TRUE) {

med\_data\_cat[i, "Anxiety"] <- 0

}

}

i <- 0

for (x in is.na(med\_data\_cat$Overweight)) {

i <- i + 1

if(x == TRUE) {

med\_data\_cat[i, "Overweight"] <- 0

}

}

i <- 0

for (x in is.na(med\_data\_cat$Soft\_drink)) {

i <- i + 1

if(x == TRUE) {

med\_data\_cat[i, "Soft\_drink"] <- "No"

}

}

summary(med\_data\_cat)

med\_data\_cat$Complication\_risk <- as.numeric(factor(med\_data\_cat$Complication\_risk,

levels=c("Low", "Medium", "High")))

summary(med\_data\_cat$Complication\_risk)

med\_data\_cat$Education <- as.numeric(factor(med\_data\_cat$Education,

levels=c('No Schooling Completed',

'Nursery School to 8th Grade',

'9th Grade to 12th Grade, No Diploma',

'GED or Alternative Credential',

'Regular High School Diploma',

'Some College, Less than 1 Year',

'Some College, 1 or More Years, No Degree',

'Associate\'s Degree',

'Professional School Degree',

'Bachelor\'s Degree',

'Master\'s Degree',

'Doctorate Degree')))

summary(med\_data\_cat$Education)

med\_data\_cat$Employment <- as.numeric(factor(med\_data\_cat$Employment,

levels=c('Unemployed',

'Student',

'Retired',

'Part Time',

'Full Time')))

summary(med\_data\_cat$Employment)

for (x in 7:16) {

med\_data\_cat[,x] <- as.numeric(factor(med\_data\_cat[,x], levels=unique(med\_data\_cat[,x]), exclude=NULL))

}

summary(med\_data\_cat)

for (x in 17:27) {

med\_data\_cat[,x] <- as.numeric(factor(med\_data\_cat[,x], levels=c("No", "Yes")))

}

summary(med\_data\_cat)

for (x in 2:27) {

med\_data\_cat[,x] <- standardize(med\_data\_cat[,x])

}

summary(med\_data\_cat)

med\_data\_unique <- select(med\_raw\_data, CaseOrder, Customer\_id, Interaction, UID)

med\_data\_cleaned1 <- inner\_join(med\_data\_unique, med\_data\_num, by="Customer\_id")

med\_data\_cleaned2 <- inner\_join(med\_data\_items, med\_data\_cat, by="Customer\_id")

med\_data\_cleaned <- inner\_join(med\_data\_cleaned1, med\_data\_cleaned2, by="Customer\_id")

summary(med\_data\_cleaned)

write.csv(med\_data\_cleaned, "medical\_data\_cleaned.csv")

#Part E

med\_data\_cleaned.pca <- prcomp(med\_data\_cleaned[5:26], center=TRUE, scale=TRUE)

summary(med\_data\_cleaned.pca)

fviz\_eig(med\_data\_cleaned.pca, choice="eigenvalue", addlabels=TRUE, ncp=35)

med\_data\_cleaned.pca

* 1. A copy of the cleaned dataset in .csv format is provided as an attachment and can be easily recreated by executing the code provided as an attachment in ".R" format.
  2. While thorough in its investigation and elimination of NA values, the data cleaning process does not make every change that may be necessary to construct an effective model from the cleaned data. Specifically, there are variables that could benefit from wrangling and manipulation. We could parse the Jobs and Timezone categories to provide more specific values that may be more useful depending on their importance to the research question. The Zip variable could prove more useful as a categorical variable than a numeric one if being used as a locator. The State variable includes values that are not states but territories. Finally, the individual diagnoses have a wide range of detriment to the patient diagnosed and could benefit from being compressed into one categorical variable with ordinal ranking; for instance, having been diagnosed with allergic rhinitis, an allergic response in the nose, is a diagnosis, but not at all comparable to diabetes.

Additionally, the data cleaning process is limited by the individuals, or systems, which record the information themselves. It is nearly impossible to determine whether a patient's information was recorded properly and some variables that are calculated based on information in other variables, Overweight and Population, are not accompanied by the data used to calculate them, such as the patient's height and weight, making validation difficult.

Finally, the imputation of many NA values with median values or assumptions of "No" responses could prove to skew the results of any analysis performed on the data. While imputing allows us to preserve the quantity of observations that make analysis fruitful, it can also have unforeseen circumstances if the NA values were recorded out of error rather than purposefully or if the true value of the data points were outliers.

* 1. The limitations discussed above limit the accuracy, otherwise known as the precision and recall, of any model built upon the cleaned data. This may lead to the model delivering incorrect assumptions more frequently or even being unable to address the research question appropriately if high accuracy is required. For example, using the research question from Part A, it could prove detrimental to the hospital if the model predicts that a certain group of patients are unlikely to be readmitted and recommends focusing extra time and resources to another group, leading to the group assumed to not readmit being readmitted, essentially making the initial problem worse.

1. Apply principal component analysis (PCA) to identify the significant features of the data set by doing the following:
   1. List the principal components in the data set.

PC1 PC2 PC3 PC4

Zip 0.0098161275 -0.083326510 0.6970981896 -0.023587219

Lat -0.0089864907 0.016246127 0.0549739160 0.010929879

Lng -0.0054519427 0.083961624 -0.7013723900 0.020447267

Population -0.0089884515 -0.028904092 0.0378259186 0.020715024

Children -0.0025227933 -0.003640862 0.0155329492 0.018432225

Age -0.0007639018 -0.081865905 -0.0406953375 -0.013507894

Income 0.0024852638 0.004009715 0.0120411567 -0.020583872

VitD\_levels 0.0095851498 -0.532247291 -0.0727462055 0.052229872

Doc\_visits -0.0072350095 0.003298889 0.0009912908 -0.008767683

Full\_meals\_eaten 0.0010018599 0.008906551 0.0184803641 0.017249430

VitD\_supp 0.0053740366 -0.032832656 0.0045560631 0.010960762

Initial\_days 0.0209486008 -0.440830179 -0.0323369458 0.067516805

TotalCharge 0.0188851005 -0.692246117 -0.0781693504 0.082844992

Additional\_charges -0.0045768062 -0.084575654 -0.0296752472 -0.023111075

Item1 0.4546547369 -0.022881126 -0.0134522684 -0.295084533

Item2 0.4283985108 -0.022203120 -0.0178482906 -0.291492549

Item3 0.3952108715 -0.022944993 -0.0123921101 -0.294955216

Item4 0.1521491789 0.060417564 0.0209913815 0.553473035

Item5 -0.1902218156 -0.068164286 -0.0232901933 -0.578660651

Item6 0.4102041344 0.031828836 -0.0058887288 0.161470749

Item7 0.3565423158 0.037387176 0.0034560096 0.169729007

Item8 0.3123416242 0.027955756 0.0161486390 0.165075651

PC5 PC6 PC7 PC8

Zip 0.027272770 -0.0510927915 1.599726e-02 -0.002652178

Lat -0.002284243 0.7005602674 -1.497775e-01 -0.002516208

Lng -0.023130353 -0.0393641630 -3.260409e-03 0.001264139

Population -0.023855575 -0.6938237082 4.364258e-02 0.004859832

Children 0.014097572 -0.0242560681 -1.222406e-01 -0.079536320

Age 0.699731351 -0.0118506717 -1.609784e-02 -0.022923262

Income -0.005404123 -0.0492218746 -1.778213e-01 0.616204672

VitD\_levels -0.054772122 0.0698970555 3.158852e-01 0.266694984

Doc\_visits 0.011022165 -0.0149659302 -1.614937e-01 0.624531713

Full\_meals\_eaten 0.037152679 0.1012639148 5.948355e-01 0.157846884

VitD\_supp 0.012040696 -0.0402997634 -5.643217e-01 0.161316248

Initial\_days -0.072162719 -0.0439604362 -3.586456e-01 -0.314565205

TotalCharge -0.078841089 0.0265609318 1.463391e-02 0.003413306

Additional\_charges 0.700215814 -0.0092624847 -1.280978e-02 -0.003389610

Item1 -0.009474394 0.0036462507 -7.410765e-03 -0.005187329

Item2 -0.011141057 0.0161312949 1.489040e-02 0.007708017

Item3 -0.009163389 -0.0154332893 4.368394e-03 0.007574360

Item4 0.024296836 0.0139266841 1.866347e-05 0.037372656

Item5 -0.022681712 0.0071973509 -5.919818e-03 -0.011568483

Item6 0.015217240 -0.0060732570 -1.539172e-02 -0.011362745

Item7 -0.000534246 -0.0023783828 -1.025491e-02 0.012647415

Item8 0.016974417 0.0007440152 2.726637e-02 -0.018956711

PC9 PC10 PC11 PC12

Zip -0.017511541 0.005548576 -0.0034650897 0.010430469

Lat 0.003100530 -0.063676829 -0.0182941717 0.022255394

Lng 0.014527102 0.005395757 0.0100304134 -0.019301104

Population -0.022547475 -0.049091670 -0.0581854284 -0.028471443

Children 0.897747593 -0.160504513 -0.3755401173 -0.010219719

Age -0.014816989 -0.007256110 0.0194395006 0.023376075

Income 0.308131539 0.371718420 0.4862033292 0.331989428

VitD\_levels -0.075950556 0.114235288 -0.4030522436 0.226965838

Doc\_visits -0.118611290 -0.696511899 -0.0531339723 -0.271093602

Full\_meals\_eaten 0.210325923 0.209877997 0.2033710587 -0.684406263

VitD\_supp -0.144841869 0.511243089 -0.3908057071 -0.462851091

Initial\_days 0.102451153 -0.152158753 0.5052983126 -0.266134143

TotalCharge 0.003503847 -0.010163444 0.0138110050 0.006968747

Additional\_charges -0.001875712 -0.004997925 0.0090779951 0.016596421

Item1 0.002973182 -0.009936376 0.0136646757 -0.008950397

Item2 -0.011202721 0.001739392 -0.0006742486 -0.003343721

Item3 0.006135725 -0.027370932 -0.0006074281 0.009866225

Item4 -0.029546560 -0.030640872 0.0394809934 0.032213261

Item5 0.008629136 -0.006923577 0.0073683349 -0.024424348

Item6 0.003057529 0.005301850 -0.0014531822 0.017194484

Item7 -0.011928793 0.026619036 -0.0195105328 0.032019957

Item8 0.031820804 -0.005888137 -0.0349155307 -0.073494182

PC13 PC14 PC15 PC16

Zip -0.115907136 -0.010003753 -0.0123209563 0.002313063

Lat 0.675051164 0.030578212 0.0844717816 0.025487567

Lng -0.023965792 0.005077969 -0.0090907199 -0.008075433

Population 0.704687114 0.018931780 0.0836598636 0.018355762

Children -0.030124972 -0.053183576 0.0004991621 -0.014233064

Age -0.005036566 0.003219973 0.0236192088 0.005259109

Income 0.047097844 0.057843424 -0.0120899620 0.007057713

VitD\_levels 0.009110764 0.009503435 -0.0078278971 0.005398575

Doc\_visits -0.080019150 0.010829201 0.0242533672 0.024257436

Full\_meals\_eaten 0.068860221 -0.061056977 0.0432649943 0.013107292

VitD\_supp -0.022188214 -0.039741121 -0.0279726891 -0.012301432

Initial\_days -0.016966013 0.001314611 0.0223636128 -0.010985919

TotalCharge -0.005094645 0.001734638 0.0038615939 0.005382348

Additional\_charges 0.032293508 -0.005200937 -0.0057464247 -0.014358114

Item1 0.026841264 -0.097021978 -0.0707186971 -0.009833445

Item2 0.019858540 -0.149580117 -0.1286341153 -0.060839364

Item3 0.063010209 -0.206730638 -0.2038555196 -0.241083046

Item4 0.057450183 -0.366659167 -0.3510775492 -0.390214418

Item5 0.006993528 0.125851741 0.0578527936 -0.133350457

Item6 -0.023609206 -0.047313280 0.0529648837 0.795372040

Item7 -0.091788228 0.049598519 0.8418868166 -0.329988838

Item8 0.008263740 0.870257488 -0.2824692775 -0.154558132

PC17 PC18 PC19 PC20

Zip -3.704157e-04 -0.006440538 -0.001163004 -0.0106363304

Lat 1.350884e-02 0.003048076 -0.015311392 -0.0073461379

Lng 5.444677e-03 -0.005792888 0.007573926 -0.0061690369

Population -1.485806e-02 0.026825002 -0.014060871 -0.0066919573

Children -1.617886e-02 0.014406889 -0.008426165 -0.0031044662

Age 3.140819e-02 0.055626204 -0.701667918 -0.0490185700

Income 8.156493e-03 0.003578380 -0.006728675 -0.0032584259

VitD\_levels -8.690199e-03 -0.009993756 -0.024477948 0.0094520569

Doc\_visits 8.961405e-03 0.008999731 -0.003669228 -0.0015247909

Full\_meals\_eaten 3.620417e-04 -0.014883550 -0.010087341 0.0007170550

VitD\_supp -3.703176e-03 -0.009751747 -0.006693483 -0.0004585361

Initial\_days 1.335207e-02 0.002397455 -0.003845488 -0.0191021791

TotalCharge -7.408355e-05 0.001937358 0.021550241 0.0027609206

Additional\_charges -2.669555e-02 -0.049271197 0.702154680 0.0480340457

Item1 -8.155724e-02 0.187145857 -0.048544889 0.8034584142

Item2 -9.304097e-02 0.624605600 0.073683058 -0.5316354388

Item3 4.229685e-01 -0.625379449 -0.027142313 -0.1938416656

Item4 -4.839824e-01 -0.102957384 -0.040525770 0.0096604775

Item5 -6.983779e-01 -0.293394792 -0.048598332 -0.0952052291

Item6 -2.727585e-01 -0.273694461 -0.007922265 -0.1266696635

Item7 -7.051359e-02 -0.062436841 0.013870683 -0.0498881268

Item8 -4.050980e-02 -0.034697066 -0.003689242 -0.0332741368

PC21 PC22

Zip -0.6985934922 0.0224045253

Lat -0.1161371964 0.0027592011

Lng -0.7044897033 0.0236001290

Population -0.0275460535 -0.0008464491

Children -0.0013991930 -0.0023381341

Age -0.0006811786 -0.0164349721

Income -0.0013953338 -0.0009723594

VitD\_levels 0.0123783263 0.5438217257

Doc\_visits 0.0026110314 -0.0002299013

Full\_meals\_eaten -0.0021255768 -0.0013106661

VitD\_supp 0.0055324111 -0.0015622550

Initial\_days 0.0185443275 0.4508741575

TotalCharge -0.0247717089 -0.7062581447

Additional\_charges 0.0033055779 0.0259339530

Item1 -0.0110236746 0.0044870685

Item2 0.0029238023 0.0031832167

Item3 0.0101305596 -0.0067053889

Item4 -0.0005193217 -0.0007946022

Item5 -0.0003653945 -0.0039198550

Item6 0.0025655494 0.0019991071

Item7 -0.0010186637 -0.0053013299

Item8 0.0024585937 -0.0050961988

* 1. In order to perform the PCA on the dataset, we performed the prcomp function from base R on variables 5-26. We have only selected these variables as variables one through four are unique identifiers, not predictor variables, and the remaining variables are categorical variables which are not well suited for a PCA but would be beneficial in an MCA. Once the PCA was performed, we use the factoextra package's fviz\_eig function with the choice parameter set to "eigenvalue" to visualize the eigenvalues for each principal component. The code to achieve this is included below.

med\_data\_cleaned.pca <- prcomp(med\_data\_cleaned[5:26], center=TRUE, scale=TRUE)

summary(med\_data\_cleaned.pca)

fviz\_eig(med\_data\_cleaned.pca, choice="eigenvalue", addlabels=TRUE, ncp=35)

The resulting scree plot from the fviz\_eig function is as follows:

Chart, histogram

Description automatically generated

The scree plot tells us that the first 12 principal components have an eigenvalue that is approximately one or higher, implying that these components are the most significant principal components.

* 1. The PCA can help the organization by providing a more workable dataset upon which effective models can be created. While difficult to interpret, the PCA allows the analyst to understand what variables provide the greatest amount of variation in the dataset and, more importantly, reduce the dimensionality of the dataset so that the modeling process and end deliverable are faster and more efficient.

Part IV. Supporting Documents

1. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e91fba21-e675-481e-961b-ad87015d1250>
2. No outside sources were used to generate code for this project and all code was written personally by the student.
3. Bibliography

Gelman, A., & Hill, J. (2006). Missing-data imputation. In Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research, pp. 529-544). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790942.031

Gress, T. W., Denvir, J., & Shapiro, J. I. (n.d.). Effect of removing outliers on statistical inference: implications to interpretation of experimental data in medical research. https://mds.marshall.edu/cgi/viewcontent.cgi?article=1170&amp;context=mjm.

Portland State University. (n.d.). 11.5 - Alternative: Standardize the variables. https://online.stat.psu.edu/stat505/book/export/html/675.