D206 Performance Assessment

Part I: Research Question

Data File being used:

Medical_raw_data.csv

Describe **one** question or decision that you will address using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation.

Which key variables predict which patients are at high risk of readmission?

B. Describe the variables in the data set and indicate the specific type of data being described. Use examples from the data set that support your claims.

Variable types:

```
library(readr)

df_raw <- read.csv('C:/Users/Hydraconix/Desktop/DATA/medical_raw_data.csv')

str(df_raw)
```

Output:

```
'data.frame':	10000 obs. of  53 variables:
 $ X             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ CaseOrder     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Customer_id   : chr  "C412403" "Z919181" "F995323" "A879973" ...
 $ Interaction   : chr  "8cd49b13-f45a-4b47-a2bd-173ffa932c2f" "d2450b70-0337-4406-bdbb-bc1037f1734c" "a2057123-abf5-4a2c-abad-8ffe33512562" "1dec528d-eb34-4079-adce-0d7a40e82205" ...
 $ UID           : chr  "3a83ddb66e2ae73798bdf1d705dc0932" "176354c5eef714957d486009feabf195" "e19a0fa00aeda885b8a436757e889bc9" "cd17d7b6d152cb6f23957346d11c3f07" ...
 $ City          : chr  "Eva" "Marianna" "Sioux Falls" "New Richland" ...
 $ State         : chr  "AL" "FL" "SD" "MN" ...
 $ County        : chr  "Morgan" "Jackson" "Minnehaha" "Waseca" ...
 $ Zip           : int  35621 32446 57110 56072 23181 74423 44086 22641 32404 56362 ...
 $ Lat           : num  34.3 30.8 43.5 43.9 37.6 ...
 $ Lng           : num  -86.7 -85.2 -96.6 -93.5 -76.9 ...
 $ Population    : int  2951 11303 17125 2162 5287 981 2558 479 40029 5840 ...
 $ Area          : chr  "Suburban" "Urban" "Suburban" "Suburban" ...
 $ Timezone      : chr  "America/Chicago" "America/Chicago" "America/Chicago" "America/Chicago" ...
 $ Job           : chr  "Psychologist, sport and exercise" "Community development worker" "Chief Executive Officer" "Early years teacher" ...
 $ Children      : int  1 3 3 0 NA NA 0 7 NA 2 ...
 $ Age           : int  53 51 53 78 22 76 50 40 48 78 ...
 $ Education     : chr  "Some College, Less than 1 Year" "Some College, 1 or More Years, No Degree" "Some College, 1 or More Years, No Degree" "GED or Alternative Credential" ...
 $ Employment    : chr  "Full Time" "Full Time" "Retired" "Retired" ...
 $ Income        : num  86576 46806 14370 39741 1210 ...
 $ Marital       : chr  "Divorced" "Married" "Widowed" "Married" ...
 $ Gender        : chr  "Male" "Female" "Female" "Male" ...
 $ ReAdmis       : chr  "No" "No" "No" "No" ...
```

```
$ VitD_levels      : num  17.8 19 17.4 17.4 16.9 ...
$ Doc_visits       : int  6 4 4 4 5 6 6 7 6 7 ...
$ Full_meals_eaten  : int  0 2 1 1 0 0 0 2 3 1 ...
$ VitD_supp        : int  0 1 0 0 2 0 0 0 0 2 ...
$ Soft_drink       : chr  NA "No" "No" "No" ...
$ Initial_admin     : chr  "Emergency Admission" "Emergency Admission" "Elective Admission" "Elective Admission" ...
$ HighBlood        : chr  "Yes" "Yes" "Yes" "No" ...
$ Stroke           : chr  "No" "No" "No" "Yes" ...
$ Complication_risk : chr  "Medium" "High" "Medium" "Medium" ...
$ Overweight       : int  0 1 1 0 0 1 1 1 1 1 ...
$ Arthritis        : chr  "Yes" "No" "No" "Yes" ...
$ Diabetes         : chr  "Yes" "No" "Yes" "No" ...
$ Hyperlipidemia    : chr  "No" "No" "No" "No" ...
$ BackPain         : chr  "Yes" "No" "No" "No" ...
$ Anxiety          : int  1 NA NA NA 0 0 1 0 NA 0 ...
$ Allergic_rhinitis : chr  "Yes" "No" "No" "No" ...
$ Reflux_esophagitis: chr  "No" "Yes" "No" "Yes" ...
$ Asthma           : chr  "Yes" "No" "No" "Yes" ...
$ Services         : chr  "Blood Work" "Intravenous" "Blood Work" "Blood Work" ...
$ Initial_days      : num  10.59 15.13 4.77 1.71 1.25 ...
$ TotalCharge       : num  3191 4215 2178 2465 1886 ...
$ Additional_charges: num  17939 17613 17505 12993 3717 ...
$ Item1            : int  3 3 2 3 2 4 4 1 3 5 ...
$ Item2            : int  3 4 4 5 1 5 3 2 3 5 ...
$ Item3            : int  2 3 4 5 3 4 3 2 2 5 ...
$ Item4            : int  2 4 4 3 3 4 2 5 3 3 ...
$ Item5            : int  4 4 3 4 5 3 3 4 3 4 ...
$ Item6            : int  3 4 4 5 3 5 4 2 3 2 ...
```

```
$ Item7            : int  3 3 3 5 4 4 5 4 4 3 ...
$ Item8            : int  4 3 3 5 3 6 5 2 2 2 ...
```

Variable Name:	**Data Type:**	**Description:**
X	Integer	Index
CaseOrder:	Integer	Index ,a placeholder variable to preserve the original order of the raw data file
Customer_id	Character string	Unique patient ID
Interaction	Character string	Related to patient transactions, procedures and admissions
UID	Character string	Unique IDs related to patient transactions, procedures, and admissions
City	Character string	Patient’s city of residence as listed on the billing statement
State	Character string	Patient’s state of residence as listed on the billing statement
County	Character string	Patient’s county of residence as listed on the billing statement
Zip	Integer	Patient’s zip code of residence as listed on the billing statement
Lat	Continuous numeric	GPS coordinates indicating latitude of patient’s residence as listed on the billing statement
Lng	Continuous numeric	GPS coordinates indicating longitude of patient’s residence as listed on the billing statement
Population	Integer	Population within a mile radius of patient- based on census data
Area	Nominal categorical character string	Area type- based on census data
Timezone	Nominal categorical character string	Time zone of patient’s residence as provided by patient
Job	Nominal categorical character string	Patient’s (or primary insurance holder’s) job as provided by patient
Children	Integer	Number of children in patient’s household as provided by patient
Age	Integer	Patient’s age as provided by patient

Education	Nominal categorical character string	Patient's highest earned degree as provided by patient
Employment	Ordinal categorical character string	Indicating patient's employment status as provided by patient
Income	Numeric	Annual income of patient (or primary insurance holder) as provided by patient
Marital	Nominal categorical character string	Patient's (or primary insurance holder's) marital status as provided by patient
Gender	Binary categorical character string	Whether or not patient was readmitted within a month of release [Yes, No] *target variable
ReAdmis	Binary categorical character string	Whether or not patient was readmitted within a month of release [Yes, No] *target variable
VitD_levels	Continuous numeric	Indicating patient's vitamin D levels as measured in ng/mL
Doc_visits	Integer	Number of times the primary physician visited the patient during the initial hospitalization
Full_meals_eaten	Integer	Number of full meals eaten (partial meals count as 0) VitD_supp: integer indicating number of times that vitamin D supplements were administered to patient
Soft_drink	Binary categorical character string	Whether or not patient regularly drinks three or more sodas in a day [Yes, No]
Initial_admin	Nominal categorical character string	The means by which the patient was initially admitted into the hospital
HighBlood	Character string	Whether or not the patient has high blood pressure [Yes, No]
Stroke	Binary categorical character string	Whether or not the patient has had a stroke [Yes, No]
Complication_risk	Ordinal categorical character string	Level of complication risk [High, Medium, Low]
Overweight	Binary categorical character string	Whether (1) or not (0) the patient is overweight, as determined by age, gender, and height

Arthritis	Binary categorical character string	Whether or not the patient has arthritis [Yes, No]
Diabetes	Binary categorical character string	Whether or not the patient has diabetes [Yes, No]
Hyperlipidemia	Binary categorical character string	Whether or not the patient has hyperlipidemia [Yes, No]
BackPain	Binary categorical character string	Whether or not the patient has chronic backpain [Yes, No]
Anxiety	Binary categorical character string	Whether (1) or not (0) the patient has an anxiety disorder
Allergic_rhinitis	Binary categorical character string	Whether or not the patient has allergic rhinitis [Yes, No]
Reflux_esophagitis	Binary categorical character string	Whether or not the patient has reflux esophagitis [Yes, No]
Asthma	Binary categorical character string	Whether or not the patient has asthma [Yes, No]
Services	Nominal categorical character string	The primary service the patient received while hospitalized
Total # of data types identified: 8		

Part II: Data - Cleaning Plan

The approach for assessing the quality of the data will focus on the following data preparation tasks:

- Changing misleading values
- Adding an index field
- Reexpressing categorical data as numeric data
- Standardizing the numeric fields
- Identifying outliers

The raw data provided does not always come in the proper format. There are also variables within the raw data set that do not provide accurate representation of the data model. It is necessary to change the misleading values before I start the exploratory data analysis. Next would be to add my own index field. “Adding an index field serves two purposes: (i) it acts as an ID field for data sets without such a field and (ii) it tracks the sort order of the records in the database. In data science, we often repartition and re-sort the data; it is therefore helpful to have an index field, in order to recover the original sort order when desired.”(Chantal D. Larose, Daniel T. Larose, 2019) The current indexes in the raw data set are in currently in the data; I want the index to act as an ID field instead.

The next data preparation task would be to change the expression of categorical data as numeric. "To provide this information to our algorithms, we transform the data values into numeric values, where it is clear that one value is larger than another. (Chantal D. Larose, Daniel T. Larose, 2019) "Certain algorithms perform better when the numeric fields are standardized so that the field mean equals 0 and the field standard deviation equals 1. Positive z-values may be interpreted as representing the number of standard deviations above the mean the data value lies, while negative z-value represent the number of standard deviations below the mean. Some analysts standardize all their numeric fields as a matter of course." (Chantal D. Larose, Daniel T. Larose, 2019) According to Chantal D. Larose and Daniel T. Larose in Data Science Using Python and R(2019), "Once the numeric fields are standardized, one may use the z-values to identify outliers, which are record with extreme value along a particular dimension or dimensions. The data scientist should consult with the client regarding what he or she would like to do with the outliers. Outliers should not be automatically removed! Nor should they be automatically changed."

For the data cleaning process, I will be using R, a software for statistical computing, to implement my coding solutions, manipulating the data, and creating visual representations for the performance assessment. “In recent years, progress in statistical learning has been marked by the increasing availability of powerful and relatively user-friendly software, such as the popular and freely available R system.” (James, Witten, Hastie, Tibshirani, p. 6)

R lets me use packages that are built in the software to clean data and identify outliers. This is what makes this software a great tool for statistical analysis. The built-in packages that I will be using in the data cleaning process will be readr, caret, dplyr, ggplot2, mice, and FactoMineR.

“Readr prints out the column specification that gives the name and type of each column.”(Wickam, Grolemund, 2017) “Caret is a set of functions that attempt to streamline the process of creating predictive models."(Cran.project.org, paras. 1) I will be using Caret for encoding dummy variables. The dyplr functions will allow me to solve most of my data manipulation encounters. The mice package in R will help me input missing values. Lastly, I will be using the FactomineR for my exploratory data analysis.

Data cleaning Outline:

- The removal of irrelevant, and(or) misleading variables from the analysis
  - Variables:'X', 'Customer_id', 'Interaction_id', 'Job', 'Income', 'Marital'
- I am removing the job and the marital variables in the dataset due to possible inconsistencies of this data that can lead to inaccurate conclusions.
- Next will be renaming misleading variables. There was data that was collected on the income that was taken upon registration, this data will be renamed as total income. The data type will be changed from categoric data to numeric data as well.
- Reset index
- Changing character values to numeric values or separating values into separate variables using dummy variables.
  https://stackoverflow.com/questions/54602192/make-only-some-features-dummyvars
- Changing NULL values from the raw data to be reflected as '0' observation.
- Using the MICE imputation for all of the other NULL Values
  https://www.rdocumentation.org/packages/mice/versions/2.25/topics/mice
- Identifying Outliers
  - Summaries of univariate stats, searching for any flags
  - Visualization of potential outliers using graphs
  - Running hypothesis test on potential outliers (Using the Grubbs tests)
    https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h1.htm
  - Standardizing variables if necessary
- PCA
  - Identifying which variables I will be using in my analysis
    http://factominer.free.fr/factomethods/principal-components-analysis.html
  - Explaining how the organization can benefit from the results of the PCA

Note: The data cleaning process will not be deleting or altering any data. Only if the data has been verified to be a discrepancy.

Data Cleaning Process:

head(df_raw)

Output:

```
  X CaseOrder Customer_id                          Interaction                              UID
1 1         1     C412403 8cd49b13-f45a-4b47-a2bd-173ffa932c2f 3a83ddb66e2ae73798bdf1d705dc0932
2 2         2     Z919181 d2450b70-0337-4406-bdbb-bc1037f1734c 176354c5eef714957d486009feabf195
3 3         3     F995323 a2057123-abf5-4a2c-abad-8ffe33512562 e19a0fa00aeda885b8a436757e889bc9
4 4         4     A879973 1dec528d-eb34-4079-adce-0d7a40e82205 cd17d7b6d152cb6f23957346d11c3f07
5 5         5     C544523 5885f56b-d6da-43a3-8760-83583af94266 d2f0425877b10ed6bb381f3e2579424a
6 6         6     S543885 e3b0a319-9e2e-4a23-8752-2fdc736c30f4 03e447146d4a32e1aaf75727c3d1230c
         City State       County   Zip      Lat       Lng Population     Area         Timezone
1         Eva    AL       Morgan 35621 34.34960 -86.72508       2951 Suburban  America/Chicago
2    Marianna    FL      Jackson 32446 30.84513 -85.22907      11303    Urban  America/Chicago
3 Sioux Falls    SD    Minnehaha 57110 43.54321 -96.63772      17125 Suburban  America/Chicago
4 New Richland    MN      Waseca 56072 43.89744 -93.51479       2162 Suburban  America/Chicago
5  West Point    VA King William 23181 37.59894 -76.88958       5287    Rural America/New_York
6      Braggs    OK     Muskogee 74423 35.67302 -95.19180        981    Urban  America/Chicago
                              Job Children Age                                Education
1 Psychologist, sport and exercise        1  53           Some College, Less than 1 Year
2     Community development worker        3  51 Some College, 1 or More Years, No Degree
3          Chief Executive Officer        3  53 Some College, 1 or More Years, No Degree
4             Early years teacher        0  78           GED or Alternative Credential
5      Health promotion specialist       NA  22             Regular High School Diploma
6              Corporate treasurer       NA  76             Regular High School Diploma
  Employment   Income       Marital Gender ReAdmis VitD_levels Doc_visits Full_meals_eaten
1  Full Time 86575.93      Divorced   Male      No    17.80233          6                0
2  Full Time 46805.99       Married Female      No    18.99464          4                2
3    Retired 14370.14       Widowed Female      No    17.41589          4                1
4    Retired 39741.49       Married   Male      No    17.42008          4                1
```

```
5  Full Time  1209.56      Widowed Female      No    16.87052         5             0
6    Retired       NA Never Married   Male      No    19.95614         6             0
  VitD_supp Soft_drink        Initial_admin HighBlood Stroke Complication_risk Overweight
1         0      <NA>   Emergency Admission       Yes     No            Medium          0
2         1        No   Emergency Admission       Yes     No              High          1
3         0        No    Elective Admission       Yes     No            Medium          1
4         0        No    Elective Admission        No    Yes            Medium          0
5         2       Yes    Elective Admission        No     No               Low          0
6         0        No Observation Admission        No     No            Medium          1
  Arthritis Diabetes Hyperlipidemia BackPain Anxiety Allergic_rhinitis Reflux_esophagitis
1       Yes      Yes             No      Yes       1               Yes                 No
2        No       No             No       No      NA                No                Yes
3        No      Yes             No       No      NA                No                 No
4       Yes       No             No       No      NA                No                Yes
5        No       No            Yes       No       0               Yes                 No
6       Yes      Yes             No      Yes       0               Yes                 No
  Asthma    Services Initial_days TotalCharge Additional_charges Item1 Item2 Item3 Item4 Item5
1    Yes  Blood Work    10.585770    3191.049          17939.403     3     3     2     2     4
2     No Intravenous    15.129562    4214.905          17612.998     3     4     3     4     4
3     No  Blood Work     4.772177    2177.587          17505.192     2     4     4     4     3
4    Yes  Blood Work     1.714879    2465.119          12993.437     3     5     5     3     4
5     No     CT Scan     1.254807    1885.655           3716.526     2     1     3     3     5
6     No  Blood Work     5.957250    2774.090          12742.590     4     5     4     4     3
  Item6 Item7 Item8
1     3     3     4
2     4     3     3
3     4     3     3
4     5     5     5
5     3     4     3
```

6 5 4 6

The summary of this step is to view the raw data. I will then start my analysis once going over all of the variables and referring back to the research question.

Removing irrelevant columns from data_raw

```
library(dplyr)
```

Output:

```
Attaching package: 'dplyr'
The following objects are masked from 'package:stats':
  filter, lag
The following objects are masked from 'package:base':
  intersect, setdiff, setequal, union
```

```
df <- df_raw[,c(6:53)]
```

(Removed X, Caseorder, Interaction, UID)

The summary of outcome for this step is that I have removed the indexes to create a new index that acts as an ID field without the field in the data set. The rest of the variables were not relevant to my exploratory analysis. I do not see any connection with variables job and marital to the research question. Therefore, they will be removed from the analysis.

```
df <- select(df, c(-Job, -Marital))
```

(Removed Job and Marital)

The outcome of this step is to remove the variables that will reduce the negative impacts to the model.

Renaming misleading variable names in df

```
names(df)[names(df) == 'Income'] <- 'Total_income'
names(df)[names(df) == 'Item1'] <- 'Survey_TimelyAdmin'
names(df)[names(df) == 'Item2'] <- 'Survey_TimelyTreatment'
names(df)[names(df) == 'Item3'] <- 'Survey_TimelyVisits'
names(df)[names(df) == 'Item4'] <- 'Survey_Reliability'
names(df)[names(df) == 'Item5'] <- 'Survey_Options'
names(df)[names(df) == 'Item6'] <- 'Survey_HoursTreatment'
names(df)[names(df) == 'Item7'] <- 'Survey_CourteousStaff'
names(df)[names(df) == 'Item8'] <- 'Survey_ActiveListening'
head(df)
```

```
         City State      County   Zip      Lat       Lng Population     Area        Timezone
1         Eva    AL      Morgan 35621 34.34960 -86.72508       2951 Suburban  America/Chicago
2    Marianna    FL     Jackson 32446 30.84513 -85.22907      11303    Urban  America/Chicago
3  Sioux Falls    SD   Minnehaha 57110 43.54321 -96.63772      17125 Suburban  America/Chicago
4 New Richland    MN      Waseca 56072 43.89744 -93.51479       2162 Suburban  America/Chicago
5   West Point    VA King William 23181 37.59894 -76.88958       5287    Rural America/New_York
6      Braggs    OK    Muskogee 74423 35.67302 -95.19180        981    Urban  America/Chicago
  Children Age                            Education Employment Total_Income Gender
1        1  53         Some College, Less than 1 Year  Full Time     86575.93   Male
2        3  51 Some College, 1 or More Years, No Degree  Full Time     46805.99 Female
3        3  53 Some College, 1 or More Years, No Degree    Retired     14370.14 Female
4        0  78          GED or Alternative Credential    Retired     39741.49   Male
5       NA  22            Regular High School Diploma  Full Time      1209.56 Female
6       NA  76            Regular High School Diploma    Retired           NA   Male
  ReAdmis VitD_levels Doc_visits Full_meals_eaten VitD_supp Soft_drink        Initial_admin
1      No    17.80233          6                0         0       <NA>  Emergency Admission
```

2	No	18.99464	4	2	1	No	Emergency Admission
3	No	17.41589	4	1	0	No	Elective Admission
4	No	17.42008	4	1	0	No	Elective Admission
5	No	16.87052	5	0	2	Yes	Elective Admission
6	No	19.95614	6	0	0	No	Observation Admission

	HighBlood	Stroke	Complication_risk	Overweight	Arthritis	Diabetes	Hyperlipidemia	BackPain
1	Yes	No	Medium	0	Yes	Yes	No	Yes
2	Yes	No	High	1	No	No	No	No
3	Yes	No	Medium	1	No	Yes	No	No
4	No	Yes	Medium	0	Yes	No	No	No
5	No	No	Low	0	No	No	Yes	No
6	No	No	Medium	1	Yes	Yes	No	Yes

	Anxiety	Allergic_rhinitis	Reflux_esophagitis	Asthma	Services	Initial_days	TotalCharge
1	1	Yes	No	Yes	Blood Work	10.585770	3191.049
2	NA	No	Yes	No	Intravenous	15.129562	4214.905
3	NA	No	No	No	Blood Work	4.772177	2177.587
4	NA	No	Yes	Yes	Blood Work	1.714879	2465.119
5	0	Yes	No	No	CT Scan	1.254807	1885.655
6	0	Yes	No	No	Blood Work	5.957250	2774.090

	Additional_charges	Survey_TimelyAdmin	Survey_TimelyTreatment	Survey_TimelyVisits
1	17939.403	3	3	2
2	17612.998	3	4	3
3	17505.192	2	4	4
4	12993.437	3	5	5
5	3716.526	2	1	3
6	12742.590	4	5	4

	Survey_Reliability	Survey_Options	Survey_HoursTreatment	Survey_CourteousStaff
1	2	4	3	3
2	4	4	4	3

```
3         4       3          4          3
4         3       4          5          5
5         3       5          3          4
6         4       3          5          4
 Survey_ActiveListening
1           4
2           3
3           3
4           5
5           3
6           6
```

The outcome of this step is to give the variables the correct names to distinguish them during the PCA. This will help improve the accuracy of the model.

Set index

```
Number_of_rows <- dim(df)[1]
row.names(df) <- c(1:num_rows)
head(df)
```

```
       City State      County   Zip      Lat       Lng Population     Area        Timezone
1       Eva    AL      Morgan 35621 34.34960 -86.72508       2951 Suburban  America/Chicago
2    Marianna    FL     Jackson 32446 30.84513 -85.22907      11303    Urban  America/Chicago
3  Sioux Falls    SD   Minnehaha 57110 43.54321 -96.63772      17125 Suburban  America/Chicago
4 New Richland    MN      Waseca 56072 43.89744 -93.51479       2162 Suburban  America/Chicago
5   West Point    VA King William 23181 37.59894 -76.88958       5287    Rural America/New_York
6      Braggs    OK    Muskogee 74423 35.67302 -95.19180        981    Urban  America/Chicago
 Children Age                         Education Employment total_Income Gender
1        1  53        Some College, Less than 1 Year  Full Time        86575.93   Male
```

```
2       3  51 Some College, 1 or More Years, No Degree  Full Time       46805.99 Female
3       3  53 Some College, 1 or More Years, No Degree    Retired       14370.14 Female
4       0  78           GED or Alternative Credential    Retired        39741.49   Male
5      NA  22             Regular High School Diploma  Full Time         1209.56 Female
6      NA  76             Regular High School Diploma    Retired              NA   Male
 ReAdmis VitD_levels Doc_visits Full_meals_eaten VitD_supp Soft_drink        Initial_admin
1      No   17.80233          6                0         0       <NA>  Emergency Admission
2      No   18.99464          4                2         1         No  Emergency Admission
3      No   17.41589          4                1         0         No    Elective Admission
4      No   17.42008          4                1         0         No    Elective Admission
5      No   16.87052          5                0         2        Yes    Elective Admission
6      No   19.95614          6                0         0         No Observation Admission
 HighBlood Stroke Complication_risk Overweight Arthritis Diabetes Hyperlipidemia BackPain
1       Yes     No            Medium          0       Yes      Yes             No      Yes
2       Yes     No              High          1        No       No             No       No
3       Yes     No            Medium          1        No      Yes             No       No
4        No    Yes            Medium          0       Yes       No             No       No
5        No     No               Low          0        No       No            Yes       No
6        No     No            Medium          1       Yes      Yes             No      Yes
 Anxiety Allergic_rhinitis Reflux_esophagitis Asthma    Services Initial_days TotalCharge
1       1               Yes                 No    Yes  Blood Work    10.585770    3191.049
2      NA                No                Yes     No Intravenous    15.129562    4214.905
3      NA                No                 No     No  Blood Work     4.772177    2177.587
4      NA                No                Yes    Yes  Blood Work     1.714879    2465.119
5       0               Yes                 No     No     CT Scan     1.254807    1885.655
6       0               Yes                 No     No  Blood Work     5.957250    2774.090
 Additional_charges Survey_TimelyAdmin Survey_TimelyTreatment Survey_TimelyVisits
1          17939.403                  3                      3                   2
2          17612.998                  3                      4                   3
```

```
3    17505.192       2       4       4
4    12993.437       3       5       5
5     3716.526       2       1       3
6    12742.590       4       5       4
 Survey_Reliability Survey_Options Survey_HoursTreatment Survey_CourteousStaff
1         2        4         3         3
2         4        4         4         3
3         4        3         4         3
4         3        4         5         5
5         3        5         3         4
6         4        3         5         4
 Survey_ActiveListening
1         4
2         3
3         3
4         5
5         3
6         6
```

This outcome of this step is to make an index that acts as an ID field that is not in the dataset. This will help repartitioning as well as re-sorting the data when I need to.

Changing expressions of categorical data as numeric data

State

```
x <- df[order(df$State),"State"]
unique(x)
```

```
[1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "IA" "ID" "IL" "IN" "KS" "KY"
```

```
[19] "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT" "NC" "ND" "NE" "NH" "NJ" "NM" "NV" "NY"
"OH"

[37] "OK" "OR" "PA" "PR" "RI" "SC" "SD" "TN" "TX" "UT" "VA" "VT" "WA" "WI" "WV" "WY"
```

```
library(plyr)
```

Output:

```
You have loaded plyr after dplyr - this is likely to cause problems.

If you need functions from both plyr and dplyr, please load plyr first, then dplyr:

library(plyr); library(dplyr)

------------------------------------------------------------------------------------------

Attaching package: 'plyr'

The following objects are masked from 'package:dplyr':

  arrange, count, desc, failwith, id, mutate, rename, summarise, summarize
```

```
new_data <- df$State

df_state_dict <- c(

  "AL" = 1, "AK" = 2, "AZ" = 3, "AR" = 4, "CA" = 5, "CO" = 6, "CT" = 7, "DE" = 8, "DC" = 9, "FL" = 10,

  "GA" = 11, "HI" = 12, "ID" = 13, "IL" = 14, "IN" = 15, "IA" = 16, "KS" = 17, "KY" = 18, "LA" = 19, "ME" = 20,

  "MD" = 21, "MA" = 22, "MI" = 23, "MN" = 24, "MS" = 25, "MO" = 26, "MT" = 27, "NE" = 28, "NV" = 29,
"NH" = 30,

  "NJ" = 31, "NM" = 32, "NY" = 33, "NC" = 34, "ND" = 35, "OH" = 36, "OK" = 37, "OR" = 38, "PA" = 39, "PR"
= 40,

  "RI" = 41, "SC" = 42, "SD" = 43, "TN" = 44, "TX" = 45, "UT" = 46, "VT" = 47, "VA" = 48, "WA" = 49, "WV" =
50,

  "WI" = 51, "WY" = 52)

Df_state_val <- revalue(x= new_data, replace = df_state_dict)

df$State <- as.numeric(df_state_val)
```

Area

```
unique(df$Area)
```

```
[1] "Suburban" "Urban"    "Rural"
```

```
new_data <- df$Area
df_area_dict <- c(
  "Rural" = 1,
  "Suburban" = 2,
  "Urban" = 3)
Df_area_val <- revalue(x= new_data, replace = df_area_dict)
df$Area <- as.numeric(df_area_val)
```

Timezone

```
unique(df$Timezone)
```

```
new_data <- df$Timezone
df_timezone_dict <- c(
  "America/Puerto_Rico" = -2,
  "America/Detroit" = -3,
  "America/Indiana/Indianapolis" = -3,
  "America/Indiana/Marengo" = -3,
  "America/Indiana/Vincennes" = -3,
  "America/Indiana/Vevay" = -3,
  "America/Indiana/Winamac" = -3,
  "America/Kentucky/Louisville" = -3,
  "America/New_York" = -3,
  "America/Toronto" = -3,
  "America/Chicago" = -4,
```

```
  "America/Indiana/Knox" = -4,
  "America/Indiana/Tell_City" = -4,
  "America/Menominee" = -4,
  "America/North_Dakota/Beulah" = -4,
  "America/North_Dakota/New_Salem" = -4,
  "America/Boise" = -5,
  "America/Denver" = -5,
  "America/Phoenix" = -5,
  "America/Los_Angeles" = -6,
  "America/Anchorage" = -7,
  "America/Nome" = -7,
  "America/Sitka" = -7,
  "America/Yakutat" = -7,
  "America/Adak" = -8,
  "Pacific/Honolulu" = -8)
Df_timezone_val <- revalue(x= new_data, replace = df_timezone_dict)
df$Timezone <- as.numeric(df_timezone_val)
```

Education

```
unique(df$Education)
[1] "Some College, Less than 1 Year"        "Some College, 1 or More Years, No Degree"
 [3] "GED or Alternative Credential"        "Regular High School Diploma"
 [5] "Bachelor's Degree"                    "Master's Degree"
 [7] "Nursery School to 8th Grade"          "9th Grade to 12th Grade, No Diploma"
 [9] "Doctorate Degree"                     "Associate's Degree"
[11] "Professional School Degree"           "No Schooling Completed"
```

```
New_data <- df$Education
Df_education_dict <- c(
  "No Schooling Completed" = 0,
  "Nursery School to 8th Grade" = 8,
  "9th Grade to 12th Grade, No Diploma" = 12,
  "GED or Alternative Credential" = 12,
  "Regular High School Diploma" = 12,
  "Some College, Less than 1 Year" = 13,
  "Some College, 1 or More Years, No Degree" = 14,
  "Associate's Degree" = 15,
  "Bachelor's Degree" = 16,
  "Master's Degree" = 18,
  "Professional School Degree" = 20,
  "Doctorate Degree" = 24
)
Df_education_val <- revalue(x= new_data, replace = df_education_dict)
df$Education <- as.numeric(df_education_val)
```

Readmission

```
unique(df$ReAdmis)
```

```
[1] "No"  "Yes"
```

```
New_data <- df$ReAdmis
bi_dict <- c(
  "No" = 0,
  "Yes" = 1)
bi_val <- revalue(x= new_data, replace = bi_dict)
df$ReAdmis <- as.numeric(bi_val)
```

Soft Drink

```
unique(df$Soft_drink)
```

```
[1] NA    "No"  "Yes"
```

```
New_data <- df$Soft_drink
bi_val <- revalue(x= new_data, replace = bi_dict)
df$Soft_drink <- as.numeric(bi_val)
```

High blood pressure

```
 unique(df$HighBlood)
```

```
[1] "Yes" "No"
```

```
New_data <- df$HighBlood
bi_val <- revalue(x= new_data, replace = bi_dict)
df$HighBlood <- as.numeric(bi_val)
```

Stroke

```
New_data <- df$Stroke
bi_val <- revalue(x= newdata, replace = bi_dict)
df$Stroke <- as.numeric(bi_val)
```

Complication Risk

```
unique(df$Complication_risk)
```

```
[1] "Medium" "High"   "Low"
```

```
New_data <- df$Complication_risk
Df_comprisk_dict <- c(
  "Low" = 1,
  "Medium" = 2,
  "High" = 3)
Df_risk_val <- revalue(x= new_data, replace = df_risk_dict)
df$Complication_risk <- as.numeric(df_risk_val)
```

Arthritis

```
New_data <- df$Arthritis
bi_val <- revalue(x= new_data, replace = bi_dict)
df$Arthritis <- as.numeric(bi_val)
```

Diabetes

```
New_data <- df$Diabetes
bi_val <- revalue(x= new_data, replace = bi_dict)
df$Diabetes <- as.numeric(bi_val)
```

Hyperlipidemia

```
New_data <- df$Hyperlipidemia
bi_val <- revalue(x= new_data, replace = bi_dict)
df$Hyperlipidemia <- as.numeric(bi_val)
```

Back Pain

```
New_data <- df$BackPain
binary_val <- revalue(x= new_data, replace = bi_dict)
df$BackPain <- as.numeric(bi_val)
```

Allergic rhinitis

```
New_data <- df$Allergic_rhinitis
bi_val <- revalue(x= new_data, replace = bi_dict)
df$Allergic_rhinitis <- as.numeric(bi_val)
```

Reflux esophagitis

```
New_data <- df$Reflux_esophagitis
bi_val <- revalue(x= new_data, replace = bi_dict)
df$Reflux_esophagitis <- as.numeric(bi_val)
```

Asthma

```
New_data <- df$Asthma
bi_val <- revalue(x= new_data, replace = bi_dict)
df$Asthma <- as.numeric(bi_val)
```

Services

```
unique(df$Services)
```

```
[1] "Blood Work"  "Intravenous" "CT Scan"     "MRI"
```

```
New_data <- df$Services
Df_services_dict <- c(
  "Blood Work" = 1,
  "Intravenous" = 2,
```

```
  "CT Scan" = 3,
  "MRI" = 4)
Df_services_val <- revalue(x= new_data, replace = df_services_dict)
df$Services <- as.numeric(df_services_val)

library(caret)
```

```
Loading required package: lattice
Loading required package: ggplot2
```

Employment

```
unique(df$Employment)
```

```
[1] "Full Time"  "Retired"    "Unemployed" "Student"    "Part Time"
```

```
dmy <- dummyVars(" ~ Employment", data = df)
my_dummy <- data.frame(predict(dmy, newdata = df))
df$Employment_FullTime <- my_dummy$EmploymentFull.Time
df$Employment_PartTime <- my_dummy$EmploymentPart.Time
df$Employment_Retired <- my_dummy$EmploymentRetired
df$Student <- my_dummy$EmploymentStudent
df$Unemployed <- my_dummy$EmploymentUnemployed
df <- select(df, -Employment)
```

Gender

```
unique(df$Gender)
```

```
[1] "Male"              "Female"            "Prefer not to answer"
```

```
dmy <- dummyVars(" ~ Gender", data = df)

my_dummy <- data.frame(predict(dmy, newdata = df))

df$Female <- my_dummy$GenderFemale

df$Male <- my_dummy$GenderMale

df <- select(df, -Gender)
```

Initial Admission

```
unique(df$Initial_admin)
```

```
[1] "Emergency Admission"   "Elective Admission"    "Observation Admission"
```

```
dmy <- dummyVars(" ~ Initial_admin", data = df)

my_dummy <- data.frame(predict(dmy, newdata = df))

df$Admin_elective <- my_dummy$Initial_adminElective.Admission

df$Admin_observation <- my_dummy$Initial_adminObservation.Admission

df$Admin_emergency <- my_dummy$Initial_adminEmergency.Admission

df <- select(df, -Initial_admin)
```

The steps taken here was to change the categorical data to numeric data. This step will let the algorithms analyze the data that are being used in the data wrangling process. The caret package was used to create separate columns for each input. This step allows me to analyze the variable more effectively.

Imputation of NULL values

```
summary(df)
```

```
    City               State           County               Zip             Lat
 Length:10000       Min.   : 1.00   Length:10000       Min.   :  610   Min.   :17.97
 Class :character   1st Qu.:14.00   Class :character   1st Qu.:27592   1st Qu.:35.26
 Mode  :character   Median :26.00   Mode  :character   Median :50207   Median :39.42
                    Mean   :26.84                      Mean   :50159   Mean   :38.75
```

```
               3rd Qu.:39.00                  3rd Qu.:72412   3rd Qu.:42.04
               Max.   :52.00                  Max.   :99929   Max.   :70.56

     Lng           Population           Area          Timezone          Children
 Min.   :-174.21   Min.   :     0.0   Min.   :1.000   Min.   :-10.000   Min.   : 0.000
 1st Qu.: -97.35   1st Qu.:   694.8   1st Qu.:1.000   1st Qu.: -6.000   1st Qu.: 0.000
 Median : -88.40   Median :  2769.0   Median :2.000   Median : -6.000   Median : 1.000
 Mean   : -91.24   Mean   :  9965.2   Mean   :1.993   Mean   : -5.861   Mean   : 2.098
 3rd Qu.: -80.44   3rd Qu.: 13945.0   3rd Qu.:3.000   3rd Qu.: -5.000   3rd Qu.: 3.000
 Max.   : -65.29   Max.   :122814.0   Max.   :3.000   Max.   : -4.000   Max.   :10.000
                                                                        NA's   :2588
      Age          Education      Total_Income        ReAdmis         VitD_levels
 Min.   :18.0   Min.   : 0.00   Min.   :   154.1   Min.   :0.0000   Min.   : 9.519
 1st Qu.:35.0   1st Qu.:12.00   1st Qu.: 19450.8   1st Qu.:0.0000   1st Qu.:16.513
 Median :53.0   Median :14.00   Median : 33942.3   Median :0.0000   Median :18.081
 Mean   :53.3   Mean   :13.61   Mean   : 40484.4   Mean   :0.3669   Mean   :19.413
 3rd Qu.:71.0   3rd Qu.:16.00   3rd Qu.: 54075.2   3rd Qu.:1.0000   3rd Qu.:19.790
 Max.   :89.0   Max.   :24.00   Max.   :207249.1   Max.   :1.0000   Max.   :53.019
 NA's   :2414                   NA's   :2464
   Doc_visits    Full_meals_eaten   VitD_supp        Soft_drink        HighBlood
 Min.   :1.000   Min.   :0.000    Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 1st Qu.:4.000   1st Qu.:0.000    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
 Median :5.000   Median :1.000    Median :0.0000   Median :0.0000   Median :0.000
 Mean   :5.012   Mean   :1.001    Mean   :0.3989   Mean   :0.2581   Mean   :0.409
 3rd Qu.:6.000   3rd Qu.:2.000    3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000
 Max.   :9.000   Max.   :7.000    Max.   :5.0000   Max.   :1.0000   Max.   :1.000
                                                   NA's   :2467
     Stroke       Complication_risk   Overweight       Arthritis         Diabetes
 Min.   :0.0000   Min.   :1.000    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
```

```
1st Qu.:0.0000   1st Qu.:2.000    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :2.000    Median :1.0000   Median :0.0000   Median :0.0000
Mean   :0.1993   Mean   :2.123    Mean   :0.7091   Mean   :0.3574   Mean   :0.2738
3rd Qu.:0.0000   3rd Qu.:3.000    3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :1.0000   Max.   :3.000    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
                 NA's   :982
Hyperlipidemia     BackPain        Anxiety      Allergic_rhinitis Reflux_esophagitis
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000    Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000    1st Qu.:0.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000    Median :0.0000
Mean   :0.3372   Mean   :0.4114   Mean   :0.3223   Mean   :0.3941    Mean   :0.4135
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000    3rd Qu.:1.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000    Max.   :1.0000
                 NA's   :984
    Asthma          Services      Initial_days     TotalCharge    Additional_charges
Min.   :0.0000   Min.   :1.000   Min.   : 1.002   Min.   : 1257   Min.   : 3126
1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 7.912   1st Qu.: 3253   1st Qu.: 7986
Median :0.0000   Median :1.000   Median :34.447   Median : 5852   Median :11574
Mean   :0.2893   Mean   :1.672   Mean   :34.432   Mean   : 5892   Mean   :12935
3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:61.125   3rd Qu.: 7615   3rd Qu.:15626
Max.   :1.0000   Max.   :4.000   Max.   :71.981   Max.   :21524   Max.   :30566
                                 NA's   :1056
Survey_TimelyAdmin Survey_TimelyTreatment Survey_TimelyVisits Survey_Reliability
Min.   :1.000      Min.   :1.000          Min.   :1.000       Min.   :1.000
1st Qu.:3.000      1st Qu.:3.000          1st Qu.:3.000       1st Qu.:3.000
Median :4.000      Median :3.000          Median :4.000       Median :4.000
Mean   :3.519      Mean   :3.507          Mean   :3.511       Mean   :3.515
3rd Qu.:4.000      3rd Qu.:4.000          3rd Qu.:4.000       3rd Qu.:4.000
Max.   :8.000      Max.   :7.000          Max.   :8.000       Max.   :7.000
```

```
Survey_Options  Survey_HoursTreatment Survey_CourteousStaff Survey_ActiveListening
Min.   :1.000   Min.   :1.000         Min.   :1.000         Min.   :1.00
1st Qu.:3.000   1st Qu.:3.000         1st Qu.:3.000         1st Qu.:3.00
Median :3.000   Median :4.000         Median :3.000         Median :3.00
Mean   :3.497   Mean   :3.522         Mean   :3.494         Mean   :3.51
3rd Qu.:4.000   3rd Qu.:4.000         3rd Qu.:4.000         3rd Qu.:4.00
Max.   :7.000   Max.   :7.000         Max.   :7.000         Max.   :7.00

Employment_FullTime Employment_PartTime Employment_Retired    Student        Unemployed
Min.   :0.0000      Min.   :0.0000      Min.   :0.000      Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000      1st Qu.:0.0000      1st Qu.:0.000      1st Qu.:0.0000   1st Qu.:0.0000
Median :1.0000      Median :0.0000      Median :0.000      Median :0.0000   Median :0.0000
Mean   :0.6029      Mean   :0.0991      Mean   :0.098      Mean   :0.1017   Mean   :0.0983
3rd Qu.:1.0000      3rd Qu.:0.0000      3rd Qu.:0.000      3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000      Max.   :1.0000      Max.   :1.000      Max.   :1.0000   Max.   :1.0000

    Female           Male        Admin_elective   Admin_observation Admin_emergency
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000    Min.   :0.000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000    1st Qu.:0.000
Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000    Median :1.000
Mean   :0.5018   Mean   :0.4768   Mean   :0.2504   Mean   :0.2436    Mean   :0.506
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000    3rd Qu.:1.000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000    Max.   :1.000
```

There are Null values found in 7 columns:

Age – Anxiety – Children – Total_Income – Initial_days – Overweight – Soft_drink

* Children, Soft_drink, and Anxiety will be converted to 0, the variables consist of yes/no data. It is assumed that these were left blank and did not apply.

```
var <- df$Children
df$Children <- replace(var, is.na(var), 0)
var <- df$Soft_drink
df$Soft_drink <- replace(var, is.na(var), 0)
var <- df$Anxiety
df$Anxiety <- replace(var, is.na(var), 0)
```

The outcome of this step was to replace the values that were determined by analyzing the variables that they were "0" value observations.

The rest of the NULL values will be replaced using MICE

```
library(mice)
```

```
Attaching package: 'mice'
The following object is masked from 'package:stats':
    filter
The following objects are masked from 'package:base':
    cbind, rbind
```

```
micedata <- df

micedata$Overweight=as.factor(micedata$Overweight)


mymice =
mice(micedata,m=5,method=c("","","","","","","","","","","pmm","","pmm","","","","","","","","","","logr
eg","","","","","","","","","","pmm","","","","","","","","","","","","","","","","","","","","","","","","",'') ,maxit=20)

summary(micedata$Age)

mymice$imp$Age

mymicecomplete <- complete(mymice, 2)

df<- mymicecomplete


summary(df)
```

*Change back Overweight to numeric

```
df$Overweight=as.numeric(df$Overweight)


md.pattern(df)
```

	City	State	County	Zip	Lat	Lng	Population	Area	Timezone	Children	Education	ReAdmis
4618	1	1	1	1	1	1	1	1	1	1	1	1
1496	1	1	1	1	1	1	1	1	1	1	1	1
1481	1	1	1	1	1	1	1	1	1	1	1	1
467	1	1	1	1	1	1	1	1	1	1	1	1
535	1	1	1	1	1	1	1	1	1	1	1	1
187	1	1	1	1	1	1	1	1	1	1	1	1
165	1	1	1	1	1	1	1	1	1	1	1	1
69	1	1	1	1	1	1	1	1	1	1	1	1
509	1	1	1	1	1	1	1	1	1	1	1	1
166	1	1	1	1	1	1	1	1	1	1	1	1

153 1 1 1 1 1 1 1 1 1 1 1 1

54 1 1 1 1 1 1 1 1 1 1 1 1

53 1 1 1 1 1 1 1 1 1 1 1 1

22 1 1 1 1 1 1 1 1 1 1 1 1

22 1 1 1 1 1 1 1 1 1 1 1 1

3 1 1 1 1 1 1 1 1 1 1 1 1

0 0 0 0 0 0 0 0 0 0 0 0

VitD_levels Doc_visits Full_meals_eaten VitD_supp Soft_drink HighBlood Stroke

4618 1 1 1 1 1 1 1

1496 1 1 1 1 1 1 1

1481 1 1 1 1 1 1 1

467 1 1 1 1 1 1 1

535 1 1 1 1 1 1 1

187 1 1 1 1 1 1 1

165 1 1 1 1 1 1 1

69 1 1 1 1 1 1 1

509 1 1 1 1 1 1 1

166 1 1 1 1 1 1 1

153 1 1 1 1 1 1 1

54 1 1 1 1 1 1 1

53 1 1 1 1 1 1 1

22 1 1 1 1 1 1 1

22 1 1 1 1 1 1 1

3 1 1 1 1 1 1 1

0 0 0 0 0 0 0

Complication_risk Arthritis Diabetes Hyperlipidemia BackPain Anxiety Allergic_rhinitis

4618 1 1 1 1 1 1 1

1496 1 1 1 1 1 1 1

1481 1 1 1 1 1 1 1

467	1	1	1	1	1	1	1
535	1	1	1	1	1	1	1
187	1	1	1	1	1	1	1
165	1	1	1	1	1	1	1
69	1	1	1	1	1	1	1
509	1	1	1	1	1	1	1
166	1	1	1	1	1	1	1
153	1	1	1	1	1	1	1
54	1	1	1	1	1	1	1
53	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1
22	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1
	0	0	0	0	0	0	0

	Reflux_esophagitis	Asthma	Services	TotalCharge	Additional_charges	Survey_TimelyAdmin
4618	1	1	1	1	1	1
1496	1	1	1	1	1	1
1481	1	1	1	1	1	1
467	1	1	1	1	1	1
535	1	1	1	1	1	1
187	1	1	1	1	1	1
165	1	1	1	1	1	1
69	1	1	1	1	1	1
509	1	1	1	1	1	1
166	1	1	1	1	1	1
153	1	1	1	1	1	1
54	1	1	1	1	1	1
53	1	1	1	1	1	1
22	1	1	1	1	1	1

22	1	1	1	1	1	1
3	1	1	1	1	1	1
	0	0	0	0	0	0

	Survey_TimelyTreatment	Survey_TimelyVisits	Survey_Reliability	Survey_Options
4618	1	1	1	1
1496	1	1	1	1
1481	1	1	1	1
467	1	1	1	1
535	1	1	1	1
187	1	1	1	1
165	1	1	1	1
69	1	1	1	1
509	1	1	1	1
166	1	1	1	1
153	1	1	1	1
54	1	1	1	1
53	1	1	1	1
22	1	1	1	1
22	1	1	1	1
3	1	1	1	1
	0	0	0	0

	Survey_HoursTreatment	Survey_CourteousStaff	Survey_ActiveListening	Employment_FullTime
4618	1	1	1	1
1496	1	1	1	1
1481	1	1	1	1
467	1	1	1	1
535	1	1	1	1
187	1	1	1	1
165	1	1	1	1

69 1 1 1 1

509 1 1 1 1

166 1 1 1 1

153 1 1 1 1

54 1 1 1 1

53 1 1 1 1

22 1 1 1 1

22 1 1 1 1

3 1 1 1 1

0 0 0 0

Employment_PartTime Employment_Retired Student Unemployed Female Male Admin_elective

4618 1 1 1 1 1 1 1

1496 1 1 1 1 1 1 1

1481 1 1 1 1 1 1 1

467 1 1 1 1 1 1 1

535 1 1 1 1 1 1 1

187 1 1 1 1 1 1 1

165 1 1 1 1 1 1 1

69 1 1 1 1 1 1 1

509 1 1 1 1 1 1 1

166 1 1 1 1 1 1 1

153 1 1 1 1 1 1 1

54 1 1 1 1 1 1 1

53 1 1 1 1 1 1 1

22 1 1 1 1 1 1 1

22 1 1 1 1 1 1 1

3 1 1 1 1 1 1 1

0 0 0 0 0 0 0

Admin_observation Admin_emergency Overweight Initial_days Age total_income

4618	1	1	1	1	1	1	0
1496	1	1	1	1	1	0	1
1481	1	1	1	1	0	1	1
467	1	1	1	1	0	0	2
535	1	1	1	0	1	1	1
187	1	1	1	0	1	0	2
165	1	1	1	0	0	1	2
69	1	1	1	0	0	0	3
509	1	1	0	1	1	1	1
166	1	1	0	1	1	0	2
153	1	1	0	1	0	1	2
54	1	1	0	1	0	0	3
53	1	1	0	0	1	1	2
22	1	1	0	0	1	0	3
22	1	1	0	0	0	1	3
3	1	1	0	0	0	0	4
	0	0	982	1056	2414	2464	6916

I chose to use the MICE(Multivariate Imputation by Chained Equations) because i felt like it was the most simplified approach to impute missing data. "The mice package in R, helps you imputing missing values with plausible data values. These plausible values are drawn from a distribution specifically designed for each missing datapoint."(Imputing Missing Data with R; MICE package, 2014)

I felt that the best approach to handle the missing data was to use the built-in models that are provided for the continuous data(predictive mean matching) and binary data(logistic regression.

The outcome of this step was that I used plausible data values that was generated using complete datasets. I used the mean-substitution and logistic regression within the MICE package. I have now ensured that all missing values been identified and have been mitigated using the approach best seemed fit.

Identifying Outliers

```
df <- df[,c(14, 1, 3, 2, 4:12, 44, 45, 46, 48, 47, 49, 50, 13, 15, 18, 16, 17, 19:32, 51:53, 33:43)]
```

(Setting dataset to identify outliers)

```
Head(df)
```

```
  ReAdmis        City      County State   Zip      Lat       Lng Population Area Timezone
1       0         Eva      Morgan     1 35621 34.34960 -86.72508       2951    2       -6
2       0    Marianna     Jackson    10 32446 30.84513 -85.22907      11303    3       -6
3       0 Sioux Falls   Minnehaha    43 57110 43.54321 -96.63772      17125    2       -6
4       0 New Richland     Waseca    24 56072 43.89744 -93.51479       2162    2       -6
5       0  West Point King William   48 23181 37.59894 -76.88958       5287    1       -5
6       0      Braggs    Muskogee    37 74423 35.67302 -95.19180        981    3       -6
  Children Age Education Employment_FullTime Employment_PartTime Employment_Retired
Unemployed
1        1  53        13                   1                   0                  0        0
2        3  51        14                   1                   0                  0        0
3        3  53        14                   0                   0                  1        0
4        0  78        12                   0                   0                  1        0
5        0  22        12                   1                   0                  0        0
6        0  76        12                   0                   0                  1        0
```

	Student	Female	Male	total_income	VitD_levels	VitD_supp	Doc_visits	Full_meals_eaten
1	0	0	1	86575.93	17.80233	0	6	0
2	0	1	0	46805.99	18.99464	1	4	2
3	0	1	0	14370.14	17.41589	0	4	1
4	0	0	1	39741.49	17.42008	0	4	1
5	0	1	0	1209.56	16.87052	2	5	0
6	0	0	1	NA	19.95614	0	6	0

	Soft_drink	HighBlood	Stroke	Complication_risk	Overweight	Arthritis	Diabetes	Hyperlipidemia
1	0	1	0	2	0	1	1	0
2	0	1	0	3	1	0	0	0
3	0	1	0	2	1	0	1	0
4	0	0	1	2	0	1	0	0
5	1	0	0	1	0	0	0	1
6	0	0	0	2	1	1	1	0

	BackPain	Anxiety	Allergic_rhinitis	Reflux_esophagitis	Asthma	Services	Admin_elective
1	1	1	1	0	1	1	0
2	0	0	0	1	0	2	0
3	0	0	0	0	0	1	1
4	0	0	0	1	1	1	1
5	0	0	1	0	0	3	1
6	1	0	1	0	0	1	0

	Admin_observation	Admin_emergency	Initial_days	TotalCharge	Additional_charges
1	0	1	10.585770	3191.049	17939.403
2	0	1	15.129562	4214.905	17612.998
3	0	0	4.772177	2177.587	17505.192
4	0	0	1.714879	2465.119	12993.437
5	0	0	1.254807	1885.655	3716.526
6	1	0	5.957250	2774.090	12742.590

Survey_TimelyAdmin Survey_TimelyTreatment Survey_TimelyVisits Survey_Reliability

1	3	3	2	2
2	3	4	3	4
3	2	4	4	4
4	3	5	5	3
5	2	1	3	3
6	4	5	4	4

	Survey_Options	Survey_HoursTreatment	Survey_CourteousStaff	Survey_ActiveListening
1	4	3	3	4
2	4	4	3	3
3	3	4	3	3
4	4	5	5	5
5	5	3	4	3
6	3	5	4	6

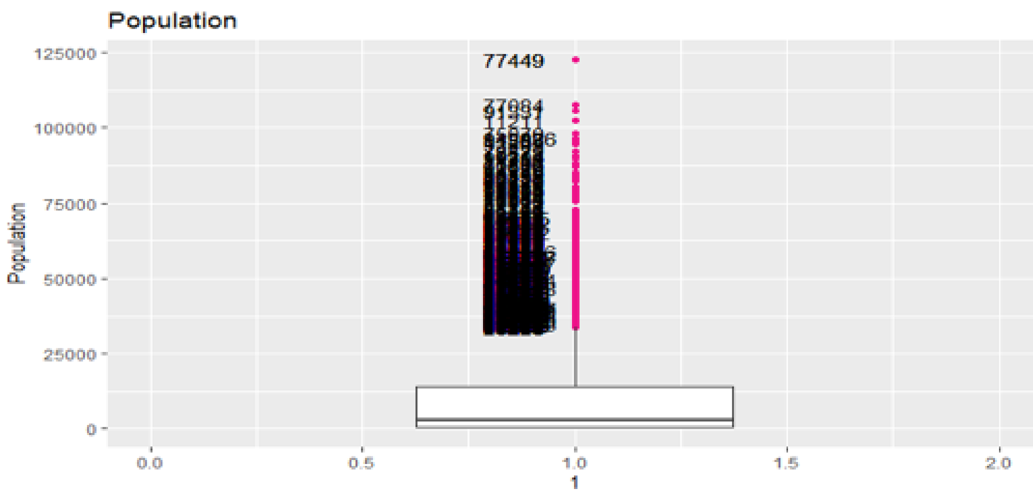
The summary for this step was to re-organize the data so that it is easier to view when analyzing each variable.

Checking for outliers using boxplots

Population

```
library(ggplot2)

graph1 <- qplot(data = df, y= Population, x=1,
        geom='boxplot',
        outlier.color='deeppink2',
        xlim=c(0,2),
        main='Population') +
  geom_text(aes(label=ifelse(Population %in% boxplot.stats(Population)$out,
```

```
                    as.character(Zip), "")), hjust = 1.5)

graph1
```



When reviewing the top zip codes, it looks like the populations are accurate; however, there are still outliers that could create a discrepancy in the effectiveness of the model. So this variable will be standardized.

The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

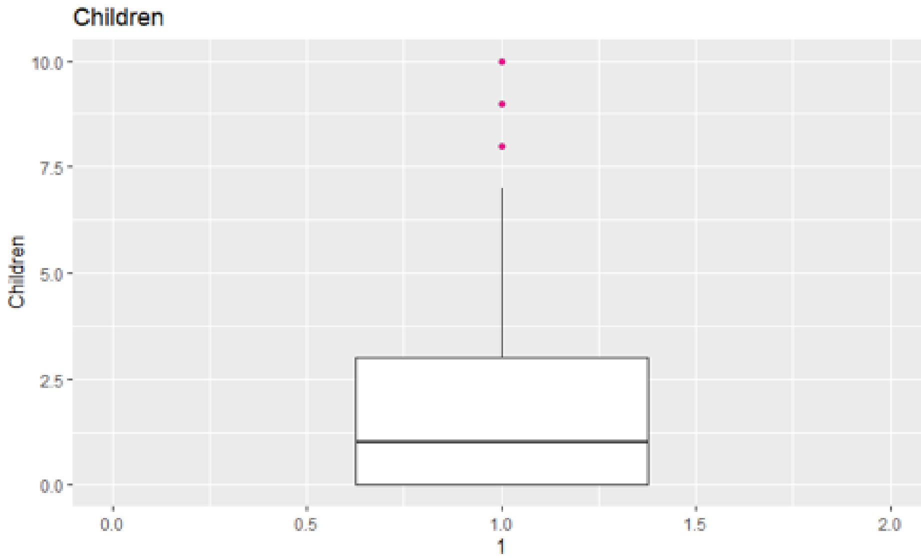
```
df$Population <- scale(x = df$Population)
```

This outcome of this step is to standardize the variable so that the variable is able to be analyzed more efficiently in the model.

Children

```
Graph2 <- qplot(data = df, y= Children, x=1,
            geom='boxplot',
            outlier.color='deeppink2',
            xlim=c(0,2),
```

```
        main='Children')
graph2
```



The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

Performing a Grubbs test on these outliers

```
library(outliers)

x <- df$Children
grubbs.test(x)
```

Grubbs test for one outlier

data: x

G = 4.07815, U = 0.99834, p-value = 0.2254

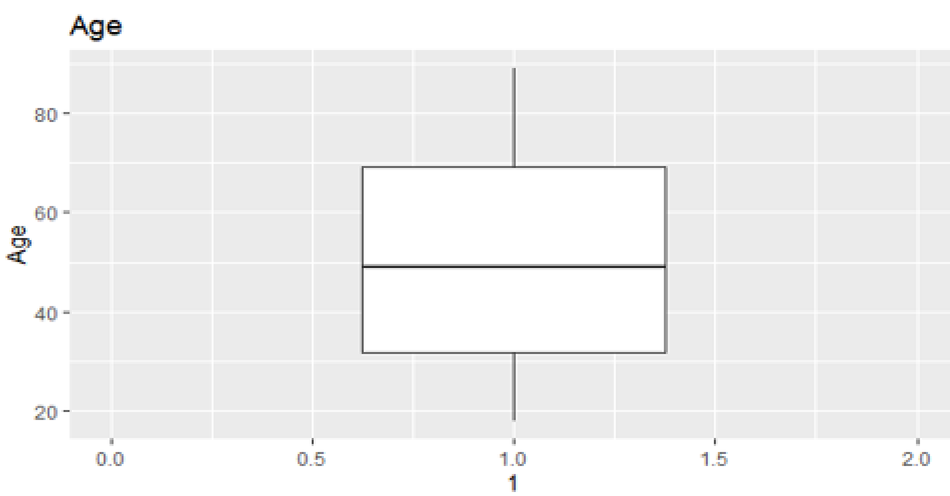
alternative hypothesis: highest value 10 is an outlier

The P-value is greater than 0.05, these values will remain the same.

The outcome of this step is to determine whether to outliers are creating a negative impact on the model. The approach in this step is to analyze the p-values as well as the mean to determine if the variable will be standardized.

Age

```
Graph3 <- qplot(data = df, y= Age, x=1,
        geom='boxplot',
        outlier.color='deeppink2',
        xlim=c(0,2),
        main='Age')
Graph3
```

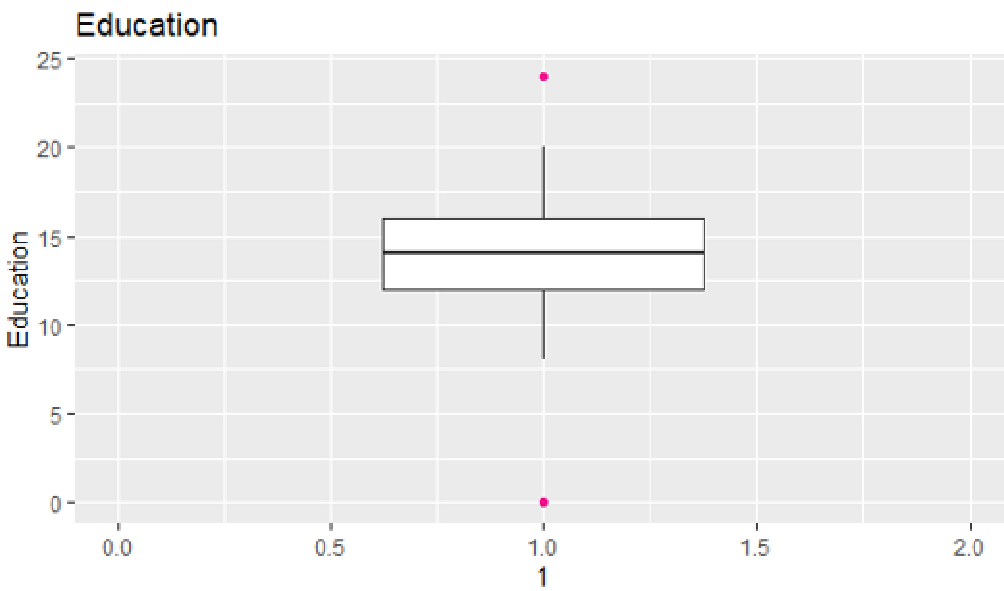


There are no outliers within this variable

The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

Education

```
Graph4 <- qplot(data = df, y= Education, x=1,
        geom='boxplot',
        outlier.color='deeppink2',
        xlim=c(0,2),
        main='Education')
graph4
```



There are very low and very high levels of education values that are outliers, so this variable will be standardized, to reduce inaccuracies in the model.

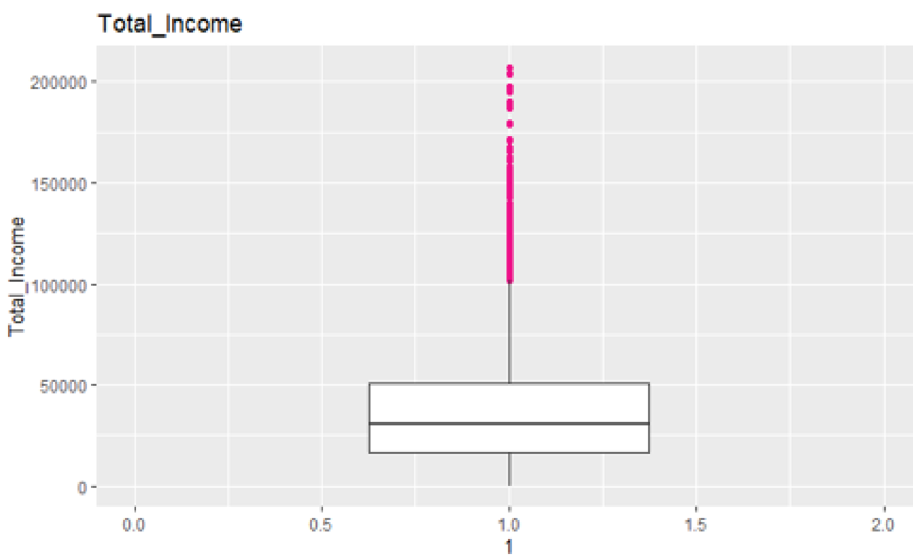
The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

```
df$Education <- scale(x = df$Education)
```

The outcome of this step is to determine whether to outliers are creating a negative impact on the model. The approach in this step is to analyze the p-values as well as the mean to determine if the variable will be standardized.

Total_Income

```
Graph5 <- qplot(data = df, y= Total_Income, x=1,
            geom='boxplot',
            outlier.color='deeppink2',
            xlim=c(0,2),
            main='Total_Income')
graph5
```



The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

```
x <- df$Total_Income

grubbs.test(x)
```

```
        Grubbs test for one outlier

data:  x
G = 5.81774, U = 0.99551, p-value = 2.164e-05
alternative hypothesis: highest value 207249.13 is an outlier
```

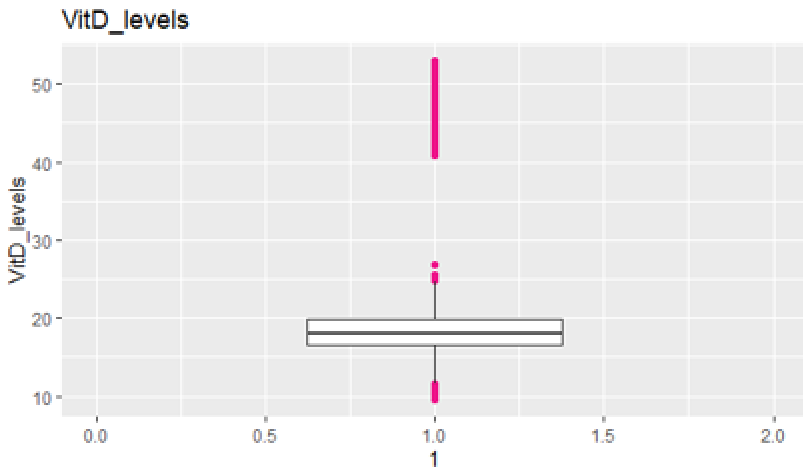
These values are very off from the mean. So they will be standardized.

```
df$Total_Income <- scale(x = df$Total_Income)
```

The outcome of this step is to determine whether to outliers are creating a negative impact on the model. The approach in this step is to analyze the p-values as well as the mean to determine if the variable will be standardized.

VitD_levels

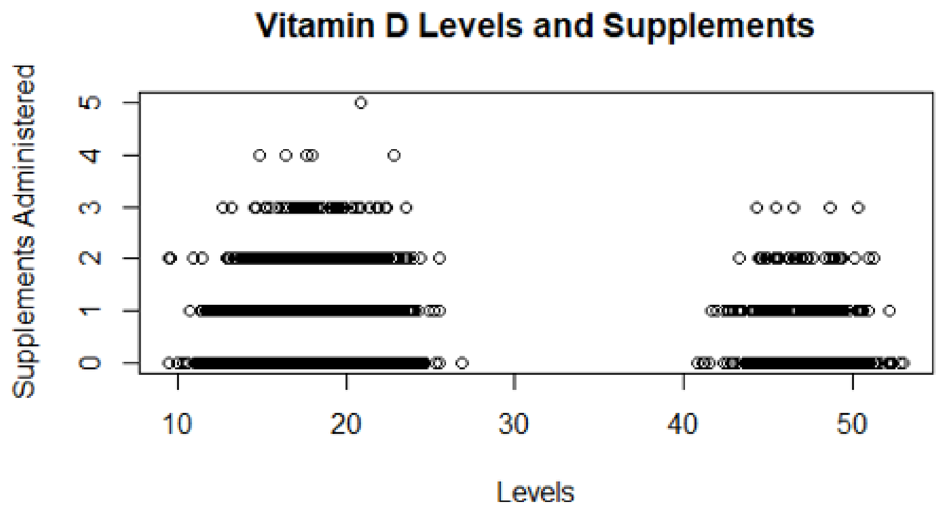
```
Graph6 <- qplot(data = df, y= VitD_levels, x=1,
        geom='boxplot',
        outlier.color='deeppink2',
        xlim=c(0,2),
        main='VitD_levels')
graph6
```



The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

Check for correlation between VitD_levels and VitD_supp

```
plot(df$VitD_levels, df$VitD_supp,
    main ='Vitamin D Levels and Supplements',
    xlab ='Levels',
    ylab = 'Supplements Administered')
```



When analyzing the graph, there looks to be two separate groups of data. The group consisting of patients with high Vitamin D levels will now be checked against potential reasons for the supplements administered of Vitamin D for patients with normal levels of Vitamin D.

```
high_VitD <- which(df$VitD_levels > 30 & df$VitD_supp>1)

houtput <- df[high_VitD,] ; houtput
```

```
ReAdmis            City      County State  Zip     Lat       Lng  Population Area
95        0         Lincoln      Benton   26 65338 38.36077  -93.28146 -0.48710768    3
838       0        Lawrence     Douglas   17 66049 38.98240  -95.34463  1.40600915    2
1070      0         Whitney Westmoreland   39 15693 40.25315  -79.40764 -0.65776813    2
1083      0        Powderly       Lamar   45 75473 33.81433  -95.48986 -0.42450970    3
1380      0 Elko New Market       Scott   24 55054 44.57008  -93.35030 -0.49547207    2
1446      0      Smithville        Clay   26 64089 39.39226  -94.56261  0.19863704    2
1813      0         Withams    Accomack   48 23488 37.95241  -75.60823 -0.65055048    2
1973      0    Philadelphia Philadelphia   39 19139 39.96144  -75.22981  2.35449002    2
2090      0           Elgin       Union   38 97827 45.58792 -117.84525 -0.51159375    3
2345      0           Grady       Curry   32 88120 34.87865 -103.45422 -0.65729595    3
```

```
2373     0        Spreckels     Monterey     5 93962 36.62471 -121.64649 -0.64623338    2
2496     0             Mina      Mineral    29 89422 38.17367 -118.41485 -0.66471597    2
2670     0            Grant        Allen    19 70644 30.79160  -92.94369 -0.65702613    1
3043     0            Boise          Ada    13 83709 43.54978 -116.28929  3.27666355    2
3714     0           Spring   Montgomery    45 77382 30.19805  -95.54607  2.01660931    3
4003     0            Waldo    Sheboygan    51 53093 43.65997  -87.94222 -0.53533781    1
4029     0      Mount Olive     Macoupin    14 62069 39.08833  -89.73938 -0.46390326    3
4200     0 Fort George G Meade Anne Arundel    21 20755 39.10578  -76.74679  0.02237785    2
     Timezone Children Age  Education Employment_FullTime Employment_PartTime
95         -6        0  59 -4.4033282                   1                   0
838        -6        0  34 -0.5206366                   0                   1
1070       -5        1  21  3.3620550                   1                   0
1083       -6        0  82  0.4500363                   1                   0
1380       -6        4  50 -0.5206366                   1                   0
1446       -6        2  60 -0.5206366                   1                   0
1813       -5        1  83 -1.8148671                   1                   0
1973       -5        1  69  0.1264787                   1                   0
2090       -8        1  70 -0.5206366                   1                   0
2345       -7        2  NA -0.5206366                   1                   0
2373       -8        0  71  1.4207092                   0                   0
2496       -8        0  84  0.1264787                   0                   0
2670       -6        0  50  0.7735939                   1                   0
3043       -7        0  84 -0.1970790                   0                   0
3714       -6        0  NA -0.5206366                   1                   0
4003       -6        0  31  0.1264787                   0                   0
4029       -6        0  47 -0.1970790                   1                   0
4200       -5        1  41 -0.5206366                   1                   0
     Employment_Retired Unemployed Student Female Male total_income VitD_levels VitD_supp
95                    0          0       0      1    0   -0.2586776    49.25631         2
```

838	0	0	0	0	1	-0.7929269	44.32063	3
1070	0	0	0	1	0	NA	45.51117	2
1083	0	0	0	0	1	-0.6585808	46.27789	2
1380	0	0	0	0	1	0.9195290	48.32009	2
1446	0	0	0	0	1	NA	46.22136	2
1813	0	0	0	1	0	NA	48.96519	2
1973	0	0	0	1	0	0.5548327	44.53013	2
2090	0	0	0	0	1	-0.2317255	47.20976	2
2345	0	0	0	1	0	-0.6514634	50.88236	2
2373	0	1	0	1	0	-1.0437999	49.24153	2
2496	0	0	1	1	0	2.3218030	46.54305	2
2670	0	0	0	1	0	NA	46.67834	2
3043	0	0	1	1	0	-0.5201280	45.66015	2
3714	0	0	0	1	0	-0.7591039	43.25143	2
4003	0	0	1	0	1	-0.2294844	50.25739	3
4029	0	0	0	1	0	-0.4539052	44.92039	2
4200	0	0	0	0	1	-0.2194153	45.67685	2

	Doc_visits	Full_meals_eaten	Soft_drink	HighBlood	Stroke	Complication_risk	Overweight
95	5	1	0	1	0	2	NA
838	5	1	0	0	0	2	0
1070	6	1	0	0	0	2	1
1083	5	1	0	0	0	2	1
1380	6	3	1	0	0	1	1
1446	6	2	0	0	0	3	1
1813	6	1	0	1	0	2	1
1973	5	2	1	1	0	3	1
2090	4	1	1	0	0	2	0
2345	4	0	0	1	0	1	1
2373	6	1	0	0	0	1	0

2496	7	2	0	0	0	2	1
2670	5	0	0	1	0	2	1
3043	4	1	1	0	0	3	1
3714	2	0	0	0	0	3	1
4003	5	1	0	0	0	3	0
4029	5	0	0	0	0	2	1
4200	6	1	0	0	0	2	NA

	Arthritis	Diabetes	Hyperlipidemia	BackPain	Anxiety	Allergic_rhinitis	Reflux_esophagitis
95	1	0	0	0	0	0	1
838	0	0	0	0	1	0	1
1070	1	0	0	1	1	0	1
1083	0	0	1	0	1	1	0
1380	0	1	0	0	0	0	0
1446	0	0	0	1	1	0	0
1813	0	1	0	0	0	1	1
1973	0	0	0	0	1	1	0
2090	0	0	0	1	0	0	0
2345	0	0	0	1	1	1	1
2373	0	0	1	0	0	0	0
2496	0	0	0	0	0	0	0
2670	0	0	1	0	0	0	1
3043	0	1	0	0	0	0	1
3714	0	0	1	0	0	1	0
4003	0	0	0	0	1	0	0
4029	0	0	0	0	0	0	0
4200	0	1	0	0	1	0	0

	Asthma	Services	Admin_elective	Admin_observation	Admin_emergency	Initial_days	TotalCharge
95	0	3	0	0	1	4.879928	14977.48
838	0	1	1	0	0	1.783260	13333.47

1070	0	2	0	0	1	9.961665	15137.32
1083	0	1	0	0	1	1.731035	14727.20
1380	0	2	1	0	0	8.273022	13861.09
1446	0	3	0	0	1	6.518810	15171.92
1813	0	1	0	1	0	3.950573	14729.64
1973	1	1	0	0	1	NA	15491.18
2090	0	1	0	1	0	NA	14394.84
2345	0	1	0	1	0	6.280303	14266.64
2373	0	1	0	0	1	18.724093	15262.74
2496	0	1	0	1	0	NA	13090.41
2670	0	1	0	0	1	13.580691	15522.65
3043	0	4	0	0	1	5.583115	14837.60
3714	0	1	0	1	0	NA	14717.59
4003	0	1	0	1	0	3.230261	14818.70
4029	1	1	0	1	0	8.209031	13372.92
4200	1	3	0	1	0	14.573414	14664.68

	Additional_charges	Survey_TimelyAdmin	Survey_TimelyTreatment	Survey_TimelyVisits
95	19669.392	3	4	3
838	5854.828	4	4	4
1070	4101.760	4	3	4
1083	13948.709	5	5	5
1380	8459.387	5	4	5
1446	10887.557	4	4	6
1813	27044.905	4	4	4
1973	23312.854	2	2	2
2090	11639.026	4	5	5
2345	7335.478	4	4	4
2373	12029.279	3	4	4
2496	13774.091	4	3	3

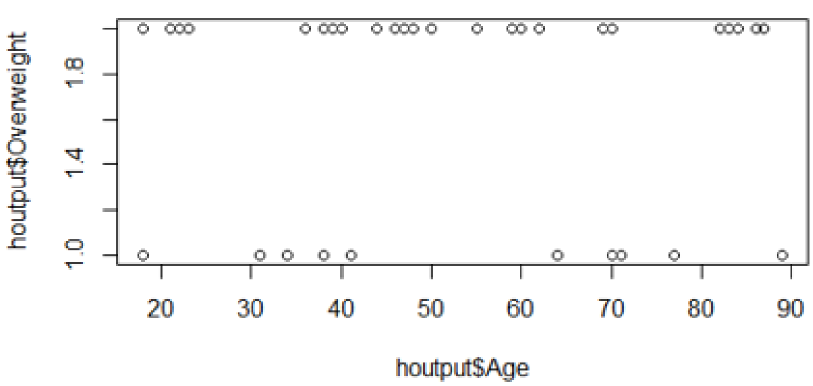
2670	16787.041	3	3	2
3043	14669.599	4	4	4
3714	9366.721	3	3	3
4003	5906.286	4	4	4
4029	7832.590	5	4	5
4200	7054.125	3	3	3

	Survey_Reliability	Survey_Options	Survey_HoursTreatment	Survey_CourteousStaff
95	4	5	3	4
838	4	4	3	4
1070	4	2	5	5
1083	4	2	6	4
1380	1	4	4	2
1446	4	4	4	3
1813	4	3	3	3
1973	3	3	3	5
2090	3	2	4	6
2345	4	3	3	4
2373	4	4	3	2
2496	4	2	3	4
2670	4	4	3	2
3043	4	4	3	3
3714	2	4	3	3
4003	3	3	4	4
4029	2	4	3	4
4200	4	3	3	3

	Survey_ActiveListening
95	3
838	4
1070	4

1083	5
1380	3
1446	5
1813	3
1973	4
2090	4
2345	4
2373	4
2496	5
2670	3
3043	5
3714	4
4003	2
4029	3
4200	4

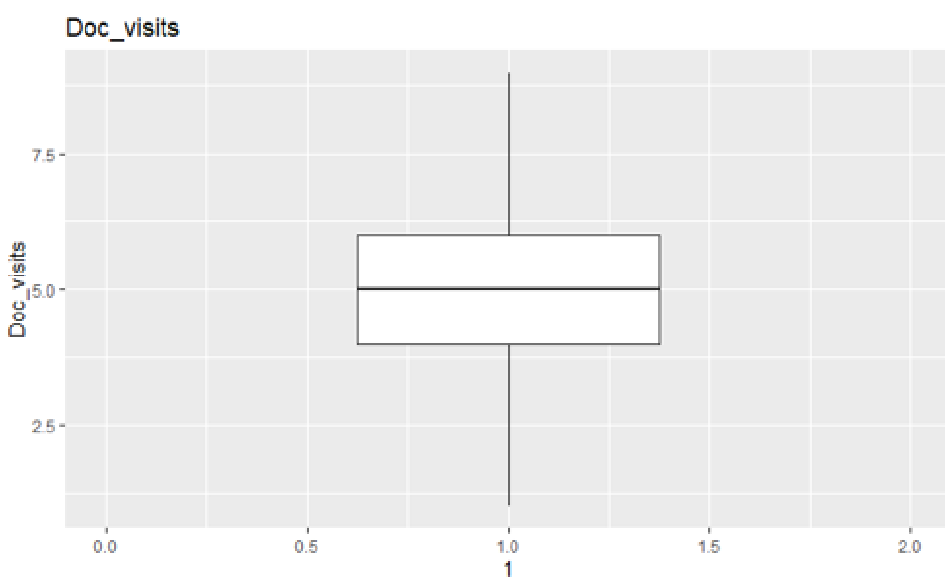
```
plot(houtput$Age, houtput$Overweight)
```



When analyzing the graph that shows the patient's age and weight. It is understandable that given the patients over 50 and/or overweight are at higher risk for health problems that would require vitamin supplements(Vitamin D). These would be considered normal outliers and are necessary for analysis, because it is likely that these patients are more likely to be readmitted if they are not given the vitamin supplements necessary. These values will remain.

Doc_visits

```
Graph7 <- qplot(data = df, y= Doc_visits, x=1,
          geom='boxplot',
          outlier.color='deeppink2',
          xlim=c(0,2),
          main='Doc_visits') +
 geom_text(aes(label=ifelse(Doc_visits %in% boxplot.stats(Doc_visits)$out,
               as.character(ReAdmis), "")), hjust = 1.5)
graph7
```

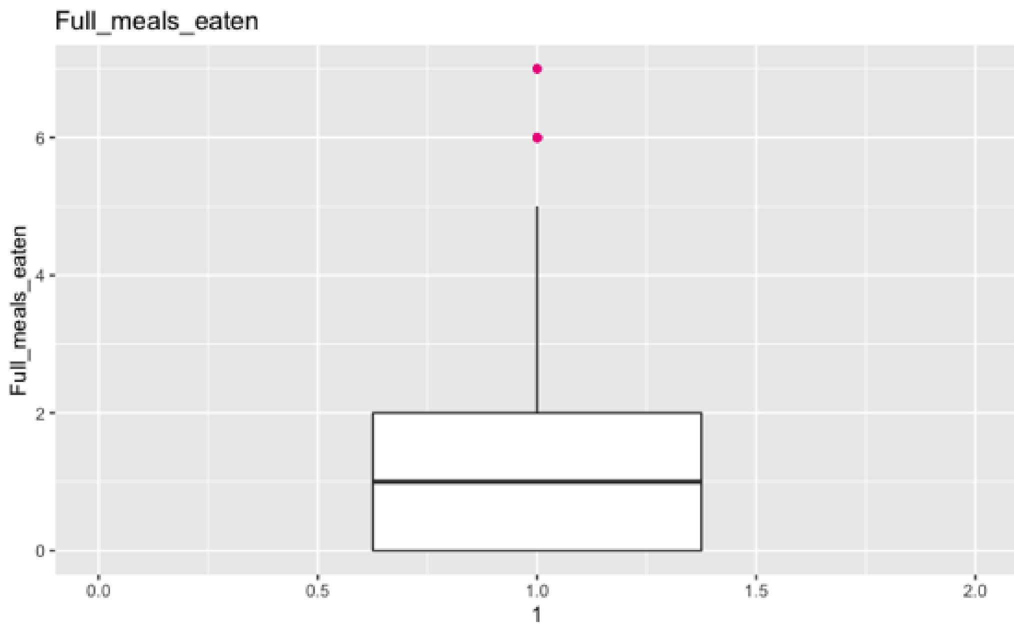


The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

There are no outliers within this variable

Full Meals Eaten

```
Graph8 <- qplot(data = df, y= Full_meals_eaten, x=1,
        geom='boxplot',
        outlier.color='deeppink2',
        xlim=c(0,2),
        main='Full_meals_eaten')
graph8
```



The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

Hypothesis test

```
x <- df$Full_meals_eaten

grubbs.test(x)
```

```
data:  x

G = 5.95030, U = 0.99646, p-value = 1.297e-05

alternative hypothesis: highest value 7 is an outlier
```

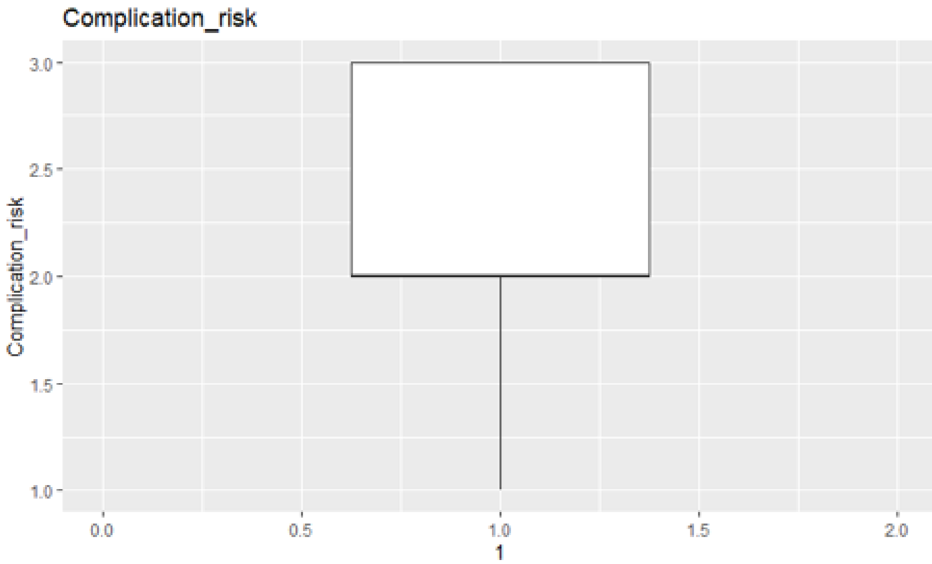
Need to standardize values

```
df$Full_meals_eaten <- scale(x = df$Full_meals_eaten)
```

The outcome of this step is to determine whether to outliers are creating a negative impact on the model. The approach in this step is to analyze the p-values as well as the mean to determine if the variable will be standardized.

Complication_risk

```
Graph9 <- qplot(data = df, y= Complication_risk, x=1,
        geom='boxplot',
        outlier.color='deeppink2',
        xlim=c(0,2),
        main='Complication_risk') +
  geom_text(aes(label=ifelse(Complication_risk %in% boxplot.stats(Complication_risk)$out,
            as.character(ReAdmis), "")), hjust = 1.5)

graph9
```

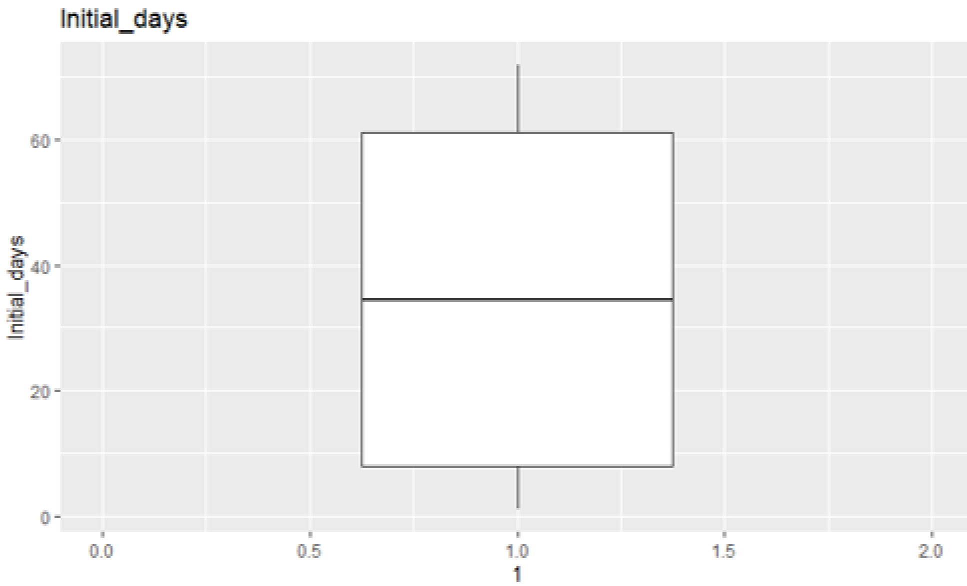


There are no outliers within this variable.

The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

Initial_days

```
graph10 <- qplot(data = df, y= Initial_days, x=1,
             geom='boxplot',
             outlier.color='deeppink2',
             xlim=c(0,2),
             main='Initial_days') +
  geom_text(aes(label=ifelse(Initial_days %in% boxplot.stats(Initial_days)$out,
                  as.character(Initial_days), "")), hjust = 1.5)
graph10
```

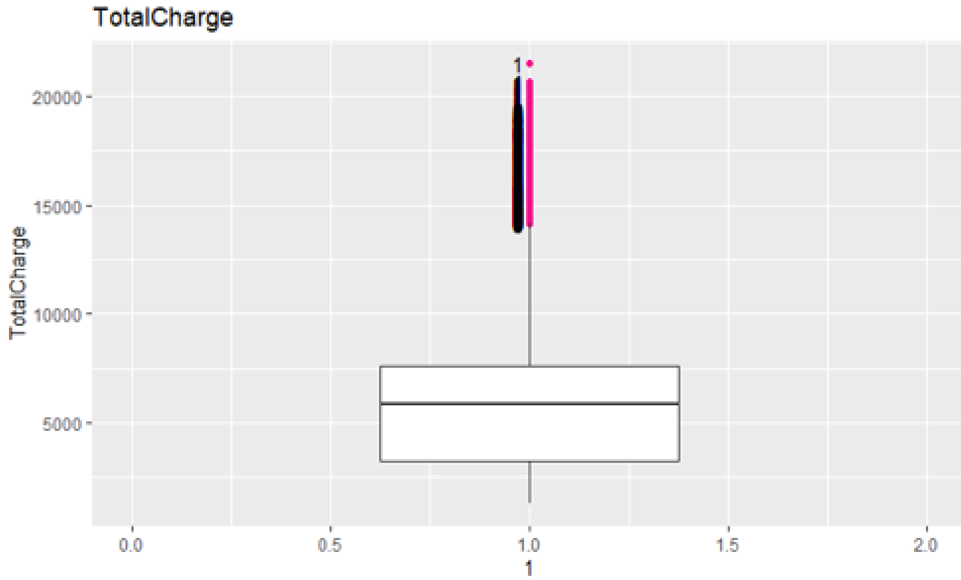


There are no outliers within this variable

The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

TotalCharge

```
graph11 <- qplot(data = df, y= TotalCharge, x=1,
             geom='boxplot',
             outlier.color='deeppink2',
             xlim=c(0,2),
             main='TotalCharge') +
  geom_text(aes(label=ifelse(TotalCharge %in% boxplot.stats(TotalCharge)$out,
                     as.character(ReAdmis), "")), hjust = 1.5)
graph11
```

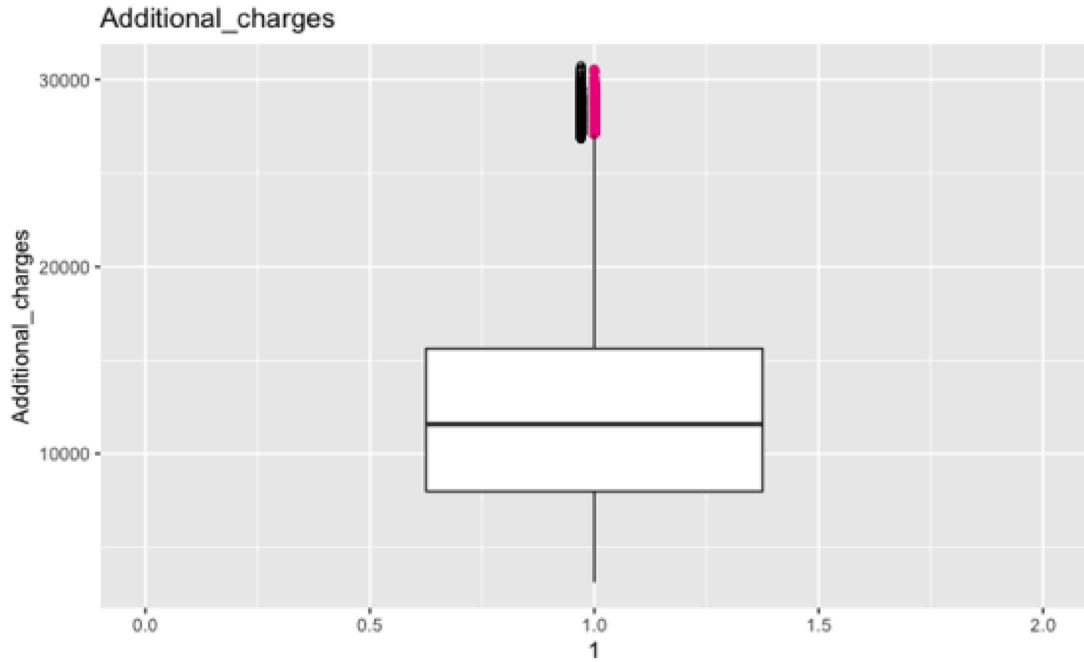


The visual is displaying that there are outliers within the TotalCharge variable. However, it looks like there is a correlation with both the TotalCharge and Readmission variables. So this data will not need changing.

The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

Additional_charges

```
Graph12 <- qplot(data = df, y= Additional_charges, x=1,

             geom='boxplot',

             outlier.color='deeppink2',

             xlim=c(0,2),

             main='Additional_charges') +

  geom_text(aes(label=ifelse(Additional_charges %in% boxplot.stats(Additional_charges)$out,

                      as.character(ReAdmis), "")), hjust = 1.5)

graph12
```



The visual is showing that the additional charges variable has outliers. I cannot determine if there is a correlation with the readmission variable so we will standardize this variable.

The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

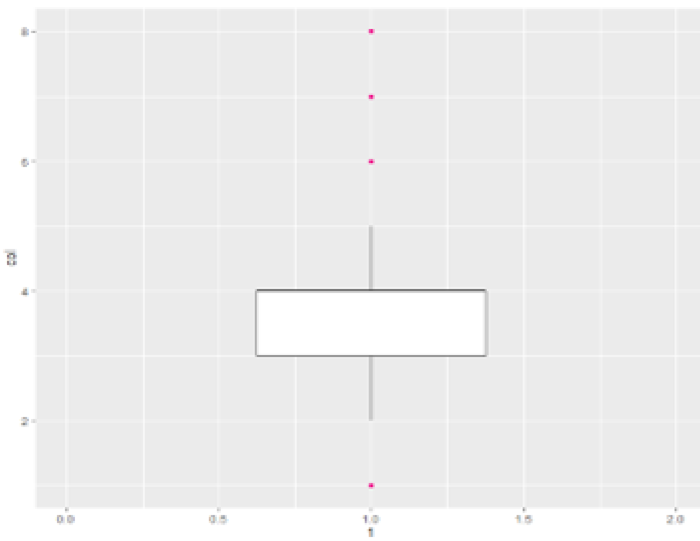
```
df$Additional_charges <- scale(x = df$Additional_charges)
```

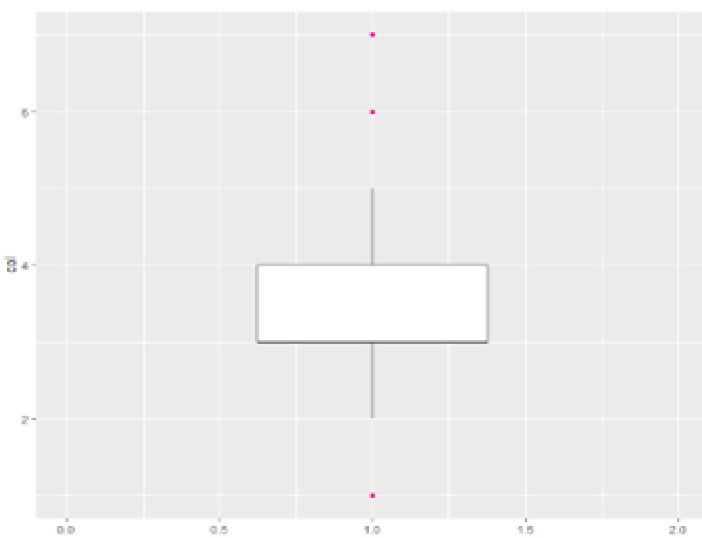
The outcome of this step is to determine whether to outliers are creating a negative impact on the model. The approach in this step is to analyze the p-values as well as the mean to determine if the variable will be standardized.

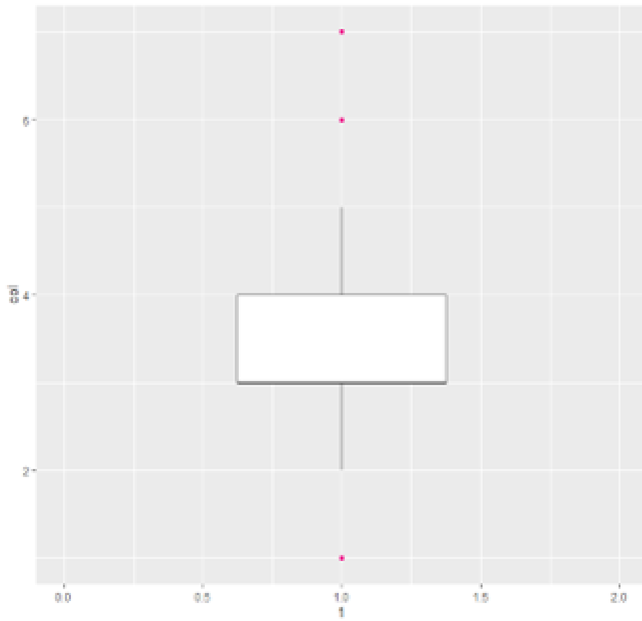
Survey Results

```
survey_results <- df[,46:53]

for (col in survey_results){

  graph13 <- qplot(data = survey_results, y= col, x=1,
```

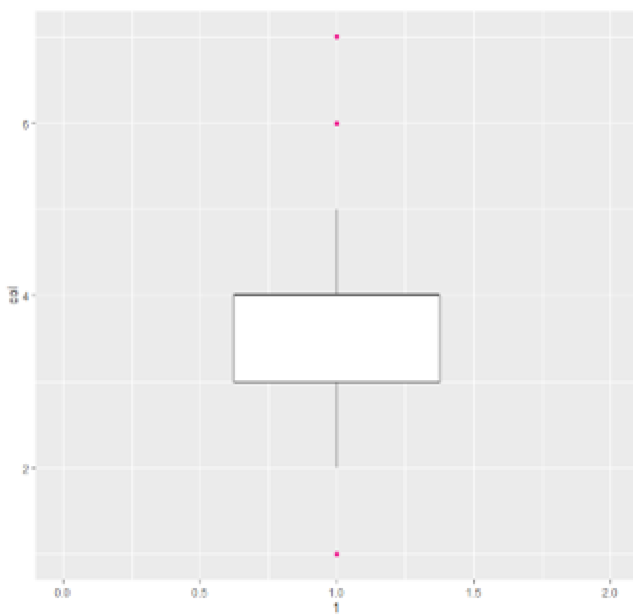
```
            geom='boxplot',
            outlier.color='deeppink2',
            xlim=c(0,2))
print(graph13)}
```



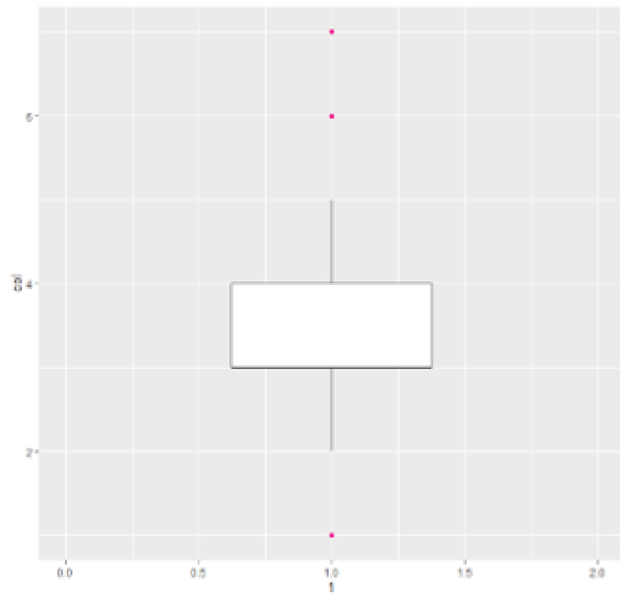


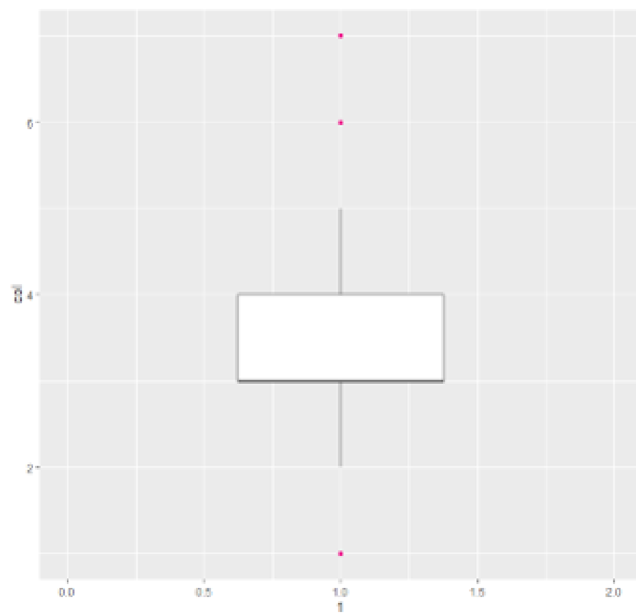


6
4
2
col
0.0
0.5
1.0
1.5
2.0
t



6
4
2
col
0.0
0.5
1.0
1.5
2.0
t





The outcome of this step is that I have created a visualization to analyze the outliers within the variables. This step helps me look closer on how the variables are measured and determine the accuracy of the model.

```
for (col in survey_results){
  x <- col
  print(grubbs.test(x))}
```

```
	Grubbs test for one outlier

data:  x
G = 4.34239, U = 0.99811, p-value = 0.06985
alternative hypothesis: highest value 8 is an outlier



	Grubbs test for one outlier

data:  x
G = 3.37574, U = 0.99886, p-value = 1
alternative hypothesis: highest value 7 is an outlier



	Grubbs test for one outlier

data:  x
G = 4.34653, U = 0.99811, p-value = 0.06854
alternative hypothesis: highest value 8 is an outlier



	Grubbs test for one outlier

data:  x
```

G = 3.36289, U = 0.99887, p-value = 1

alternative hypothesis: highest value 7 is an outlier

Grubbs test for one outlier

data: x

G = 3.40043, U = 0.99884, p-value = 1

alternative hypothesis: highest value 7 is an outlier

Grubbs test for one outlier

data: x

G = 3.36844, U = 0.99887, p-value = 1

alternative hypothesis: highest value 7 is an outlier

Grubbs test for one outlier

data: x

G = 3.43253, U = 0.99882, p-value = 1

alternative hypothesis: highest value 7 is an outlier

Grubbs test for one outlier

data: x

G = 3.34861, U = 0.99888, p-value = 1

alternative hypothesis: highest value 7 is an outlier

These values will remain.

The outcome of this step is to determine whether to outliers are creating a negative impact on the model. The approach in this step is to analyze the p-values as well as the mean to determine if the variable will be standardized.

The Limitations of the Analysis

Considering that this data was provided by a third party rather than the individual, this has been identified as a major limitation. Another limitation that was identified was how the variables are being measured. An example would be when analyzing the Vitamin D supplement intake for patients, it is not clear if the number of supplements were being taken daily, or throughout the stay of the patient's visit. There was also no indication of the dosage amount being taken. Discrepancies such as these examples could skew results of the analysis.

After the data cleaning process, the strengths that this clean data set has is that I have made almost all the values numeric. I have also cleaned the null values that appeared in the data set. This allows the data set to provide more accurate calculations and proper visualizations. One of the limitations recognized when cleaning the data set, is that it is difficult to interpret all the data types being used in the raw data.

Principle Component Analysis

```
head(df)
```

```
ReAdmis        City       County State   Zip      Lat       Lng Population Area Timezone
1       0         Eva       Morgan     1 35621 34.34960 -86.72508 -0.4731446    2       -6
2       0    Marianna      Jackson    10 32446 30.84513 -85.22907  0.0902373    3       -6
3       0 Sioux Falls    Minnehaha    43 57110 43.54321 -96.63772  0.4829587    2       -6
4       0 New Richland      Waseca    24 56072 43.89744 -93.51479 -0.5263663    2       -6
5       0  West Point King William    48 23181 37.59894 -76.88958 -0.3155703    1       -5
6       0      Braggs    Muskogee    37 74423 35.67302 -95.19180 -0.6060304    3       -6
  Children Age  Education Employment_FullTime Employment_PartTime Employment_Retired
Unemployed
1        1  53 -0.1970790                   1                   0                  0          0
2        3  51  0.1264787                   1                   0                  0          0
3        3  53  0.1264787                   0                   0                  1          0
4        0  78 -0.5206366                   0                   0                  1          0
```

```
5     0  22 -0.5206366          1           0           0       0
6     0  76 -0.5206366          0           0           1       0
 Student Female Male total_Income VitD_levels VitD_supp Doc_visits Full_meals_eaten
1    0    0   1    1.60794401   17.80233      0      6    -0.993337188
2    0    1   0    0.22053314   18.99464      1      4     0.990559733
3    0    1   0   -0.91102128   17.41589      0      4    -0.001388728
4    0    0   1   -0.02591843   17.42008      0      4    -0.001388728
5    0    1   0   -1.37014019   16.87052      2      5    -0.993337188
6    0    0   1           NA    19.95614      0      6    -0.993337188
 Soft_drink HighBlood Stroke Complication_risk Overweight Arthritis Diabetes Hyperlipidemia
1       0       1     0             2        0      1      1            0
2       0       1     0             3        1      0      0            0
3       0       1     0             2        1      0      1            0
4       0       0     1             2        0      1      0            0
5       1       0     0             1        0      0      0            1
6       0       0     0             2        1      1      1            0
 BackPain Anxiety Allergic_rhinitis Reflux_esophagitis Asthma Services Admin_elective
1     1     1             1             0     1      1          0
2     0     0             0             1     0      2          0
3     0     0             0             0     0      1          1
4     0     0             0             1     1      1          1
5     0     0             1             0     0      3          1
6     1     0             1             0     0      1          0
 Admin_observation Admin_emergency Initial_days TotalCharge Additional_charges
1              0              1   10.585770    3191.049      0.764967085
2              0              1   15.129562    4214.905      0.715077864
3              0              0    4.772177    2177.587      0.698600372
4              0              0    1.714879    2465.119      0.009003875
5              0              0    1.254807    1885.655     -1.408920097
```

```
6          1         0    5.957250    2774.090      -0.029336751
 Survey_TimelyAdmin Survey_TimelyTreatment Survey_TimelyVisits Survey_Reliability
1          3            3            2           2
2          3            4            3           4
3          2            4            4           4
4          3            5            5           3
5          2            1            3           3
6          4            5            4           4
 Survey_Options Survey_HoursTreatment Survey_CourteousStaff Survey_ActiveListening
1        4            3            3            4
2        4            4            3            3
3        3            4            3            3
4        4            5            5            5
5        5            3            4            3
6        3            5            4            6
>
```

I will not use the target variable (ReAdmis), the qualititative variables (City and County), or the redundant variables (State, Lat, Lng, Timezone)

The outcome of this step is to create a new dataset that I will use for the PCA.

Standardization

```
df_sub <- scale(x = df[,c(5, 8, 9, 11:53)])
head(df_sub)
```

```
Zip Population       Area   Children      Age  Education Employment_FullTime
1 -0.5292516 -0.4731446  0.008079945 -0.2681162 -0.0143121 -0.1970790          0.8115319
2 -0.6448340  0.0902373  1.232313962  0.6977202 -0.1111214  0.1264787          0.8115319
```

```
3  0.2530317  0.4829587  0.008079945  0.6977202 -0.0143121  0.1264787         -1.2321143
4  0.2152444 -0.5263663  0.008079945 -0.7510343  1.1958036 -0.5206366         -1.2321143
5 -0.9821161 -0.3155703 -1.216154073 -0.7510343 -1.5148555 -0.5206366          0.8115319
6  0.8832923 -0.6060304  1.232313962 -0.7510343  1.0989943 -0.5206366         -1.2321143
  Employment_PartTime Employment_Retired Unemployed    Student     Female       Male
1          -0.3316476         -0.3296006 -0.3301597 -0.3364558 -1.0035563  1.0474759
2          -0.3316476         -0.3296006 -0.3301597 -0.3364558  0.9963566 -0.9545805
3          -0.3316476          3.0336712 -0.3301597 -0.3364558  0.9963566 -0.9545805
4          -0.3316476          3.0336712 -0.3301597 -0.3364558 -1.0035563  1.0474759
5          -0.3316476         -0.3296006 -0.3301597 -0.3364558  0.9963566 -0.9545805
6          -0.3316476          3.0336712 -0.3301597 -0.3364558 -1.0035563  1.0474759
  total_Income VitD_levels  VitD_supp  Doc_visits Full_meals_eaten Soft_drink  HighBlood
1   1.60794401 -0.23951785 -0.6346809  0.94459928     -0.993337188 -0.4912094  1.2020163
2   0.22053314 -0.06217740  0.9563968 -0.96793217      0.990559733 -0.4912094  1.2020163
3  -0.91102128 -0.29699603 -0.6346809 -0.96793217     -0.001388728 -0.4912094  1.2020163
4  -0.02591843 -0.29637274 -0.6346809 -0.96793217     -0.001388728 -0.4912094 -0.8318523
5  -1.37014019 -0.37811197  2.5474746 -0.01166644     -0.993337188  2.0355880 -0.8318523
6           NA  0.08083369 -0.6346809  0.94459928     -0.993337188 -0.4912094 -0.8318523
      Stroke Complication_risk Overweight  Arthritis   Diabetes Hyperlipidemia   BackPain
1 -0.4988811        -0.1688644 -1.5613384  1.3408228  1.6285072      -0.713232  1.1960691
2 -0.4988811         1.2006768  0.6404051 -0.7457362 -0.6139979      -0.713232 -0.8359885
3 -0.4988811        -0.1688644  0.6404051 -0.7457362  1.6285072      -0.713232 -0.8359885
4  2.0042853        -0.1688644 -1.5613384  1.3408228 -0.6139979      -0.713232 -0.8359885
5 -0.4988811        -1.5384057 -1.5613384 -0.7457362 -0.6139979       1.401928 -0.8359885
6 -0.4988811        -0.1688644  0.6404051  1.3408228  1.6285072      -0.713232  1.1960691
     Anxiety Allergic_rhinitis Reflux_esophagitis     Asthma   Services Admin_elective
1  1.5623419         1.2398683         -0.8396186  1.5672823 -0.8069574     -0.5779372
2 -0.6400008        -0.8064566          1.1908979 -0.6379833  0.3938721     -0.5779372
3 -0.6400008        -0.8064566         -0.8396186 -0.6379833 -0.8069574      1.7301187
```

4	-0.6400008	-0.8064566	1.1908979	1.5672823	-0.8069574	1.7301187
5	-0.6400008	1.2398683	-0.8396186	-0.6379833	1.5947016	1.7301187
6	-0.6400008	1.2398683	-0.8396186	-0.6379833	-0.8069574	-0.5779372

	Admin_observation	Admin_emergency	Initial_days	TotalCharge	Additional_charges
1	-0.5674677	0.9880217	-0.9071506	-0.7995390	0.764967085
2	-0.5674677	0.9880217	-0.7342977	-0.4964039	0.715077864
3	-0.5674677	-1.0120223	-1.1283086	-1.0995966	0.698600372
4	-0.5674677	-1.0120223	-1.2446130	-1.0144664	0.009003875
5	-0.5674677	-1.0120223	-1.2621148	-1.1860294	-1.408920097
6	1.7620385	-1.0120223	-1.0832266	-0.9229888	-0.029336751

	Survey_TimelyAdmin	Survey_TimelyTreatment	Survey_TimelyVisits	Survey_Reliability
1	-0.5027299	-0.4896481	-1.4631734	-1.4620544
2	-0.5027299	0.4766991	-0.4948898	0.4679230
3	-1.4717544	0.4766991	0.4733939	0.4679230
4	-0.5027299	1.4430463	1.4416775	-0.4970657
5	-1.4717544	-2.4223426	-0.4948898	-0.4970657
6	0.4662946	1.4430463	0.4733939	0.4679230

	Survey_Options	Survey_HoursTreatment	Survey_CourteousStaff	Survey_ActiveListening
1	0.4883553	-0.5061140	-0.4836475	0.4703965
2	0.4883553	0.4625253	-0.4836475	-0.4890090
3	-0.4823371	0.4625253	-0.4836475	-0.4890090
4	0.4883553	1.4311645	1.4744395	1.4298020
5	1.4590477	-0.5061140	0.4953960	-0.4890090
6	-0.4823371	1.4311645	0.4953960	2.3892076

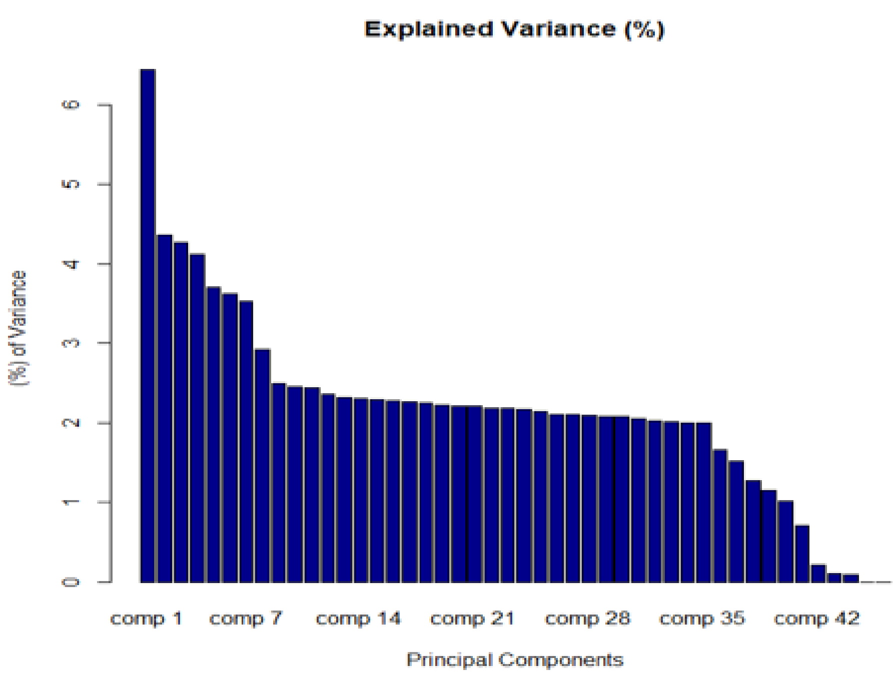
Principal Component Analysis

```
library(FactoMineR)

df_sub.pca <- PCA(df_sub, scale.unit=TRUE, graph=F)
```

The outcome of this step is that I applied the FactoMineR to the new dataset. This package allows me to reduce the dimensionality of the dataset so that it is summarized.

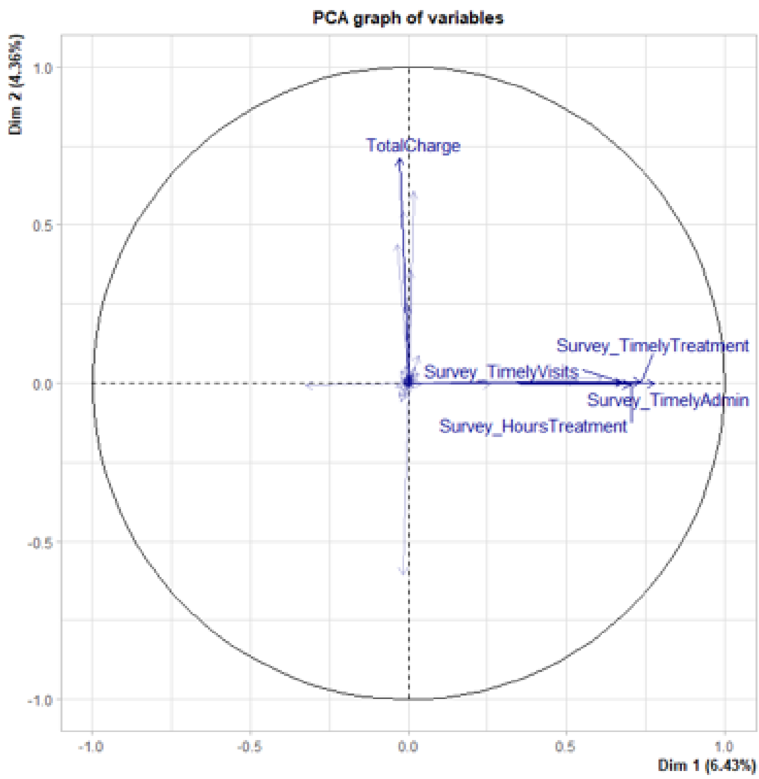
Scree Plot

```
eig.val <- df_sub.pca$eig

barplot(eig.val[, 2],

     main = "Explained Variance (%)",

     xlab = "Principal Components",

     ylab = "(%) of Variance",

     col = "darkblue")
```



Graph of Variables

```
plot(df_sub.pca, choix = "var", autoLab = "auto", col.var="darkblue", label="var", graph.type = "ggplot", select="cos2 0.40")
```



To determine the principal components, I used the package FactoMineR to run the PCA and used eigenvalues for my scree plot. The PCA graph of variables let me reduce the data to the important variables. The variables with cos2 over 0.4 were those that I identified as "important," and they were variables:

- TotalCharge - Additional_charges - Survey_HoursTreatment - Survey_TimelyVisits - Survey_TimelyAdmin - Survey_TimelyTreatment

Hospitals can benefit from researching these components and analyzing the correlation between these principal components and readmission rates. From the results of an in-depth analysis, the organization can extract insights and attempt to reduce readmission rates.

References:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). In An introduction to statistical learning: with applications in R (pp. 6–6). essay, Springer.

Wickham, H., & Grolemund, G. (2017). R for data science: import, tidy, transform, visualize and model data. O'Reilly.

1.3.5.17.1. Grubbs' Test for Outliers. (n.d.). https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h1.htm.

A Short Introduction to the caret Package. (n.d.). https://cran.r-project.org/web/packages/caret/vignettes/caret.html.

Doug FirDoug Fir 15k3939 gold badges122122 silver badges229229 bronze badges, & Julius VainoraJulius Vainora 44k99 gold badges7979 silver badges9696 bronze badges. (1967, October 1). *Make only some features dummyVars*. Stack Overflow. https://stackoverflow.com/questions/54602192/make-only-some-features-dummyvars.

mice: Multivariate Imputation by Chained Equations (MICE). RDocumentation. (n.d.). https://www.rdocumentation.org/packages/mice/versions/2.25/topics/mice.

Sébastien Lê, G. F. F. H. (n.d.). *Principal Components Analysis*. FactoMineR. http://factominer.free.fr/factomethods/principal-components-analysis.html.

Alice, M. (2018, May 14). Imputing Missing Data with R; MICE package. DataScience+. https://datascienceplus.com/imputing-missing-data-with-r-mice-package/.