D206: Data Cleaning

Lucas Benish

Western Governors University

Part I: Research Question

# 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.

What are the top 3 to 5 metrics that affect whether a patient may have a hospital readmission in the future?

# 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.

To determine and examine the variables within my dataset, summary(medical\_raw\_data) was called but didn’t appear to give as valuable of information as I was looking for at this time, and will be revisited in the future. Variable name was determined from calling colnames(medical\_raw\_data) performed in R. Variable description has been modified to a shorthand version based on the combination of observing the dataframe and information given in the data dictionary that was provided. Variable type was determined by calling str(medical\_raw\_data) performed in R.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Type of variable** | **Example** |
| X | Maintains order of original data | Integer | 1 |
| CaseOrder | Maintains order of original data | Integer | 1 |
| Customer\_id | Unique ID for each individual patient, combination of numbers/letters | Character | c412403 |
| Interaction | Unique ID for patient interactions, combination of numbers/letters | Character | 8cd49b13-f45a-4b47-a2bd-173ffa932c2f |
| UID | Unique ID for patient interactions/procedures | Character | 3a83ddb66e2ae73798bdf1d705dc0932 |
| City | City in which the patient lives, per billing | Character | Eva |
| State | State in which the patient lives, per billing | Character | AL |
| County | County in which the patient lives, per billing | Character | Morgan |
| ZIP | Zip code in which patient lives, per billing | Integer | 35621 |
| Lat | Latitude of patient’s address, per billing | Numeric | 34.3496 |
| Lng | Longitude of patient’s address, per billing | Numeric | -86.72508 |
| Population | Population within 1 mile radius of patient’s address | Integer | 2951 |
| Area | Type of home area – urban/rural/suburban | Character | Suburban |
| Timezone | Patient’s timezone | Character | America/Chicago |
| Job | Patient’s job | Character | Psychologist, sport and exercise |
| Children | Number of children patient has at home | Integer | 1 |
| Age | Patient’s age | Integer | 53 |
| Education | Highest level of patient’s education completed | Character | Some College, Less than 1 Year |
| Employment | Patient’s current employment status | Character | Full Time |
| Income | Patient’s/primary insurance holder’s current reported income | Number | 86575.93 |
| Marital | Patient’s marital status at time of admission | Character | Divorced |
| Gender | Gender that patient identifies as | Character | Male |
| ReAdmis | Has patient been readmitted within month of release? | Character | No |
| VitD\_levels | Patient’s vitamin D levels | Numeric | 17.80233 |
| Doc\_visits | Number of physician visits during initial stay | Integer | 6 |
| Full\_meals\_eaten | How many full meals were eaten by patient? | Integer | 0 |
| VitD\_supp | Number of vitamin D supplements were given to patient | Integer | 0 |
| Soft\_drink | Does patient drink 3+ sodas within single day? (yes/no) | Character | NA |
| Initial\_admin | What time of admission was initial admission? (emergent, elective, observe) | Character | Emergency Admission |
| HighBlood | Does the patient have high blood pressure? (yes/no) | Character | Yes |
| Stroke | Has the patient suffered a stroke? (yes/no) | Character | No |
| Complication\_risk | Patient’s risk stratification of complications (low, medium, high) | Character | Medium |
| Overweight | Is the patient considered overweight? (yes/no – answers are 0/1 though) | Integer | 0 |
| Arthritis | Does the patient have arthritis? (yes/no) | Character | Yes |
| Diabetes | Does the patient have diabetes? (yes/no) | Character | Yes |
| Hyperlipidemia | Does the patient have hyperlipidemia? (yes/no) | Character | No |
| BackPain | Does the patient have chronic back pain? (yes/no) | Character | Yes |
| Anxiety | Does the patient have anxiety? (yes/no, but integer answers) | Integer | 1 |
| Allergic\_rhinitis | Does patient have allergic rhinitis? (yes/no) | Character | Yes |
| Reflux\_esophagitis | Does patient have reflux? (yes/no) | Character | No |
| Asthma | Does patient have asthma? (yes/no) | Character | Yes |
| Services | What primary service has patient had during stay? (blood, IV, CT, MRI) | Character | Blood Work |
| Initial\_days | How many days did patient stay for initial stay? | Numeric | 10.58577 |
| TotalCharge | Average daily charge to patient based on total cost and number of days | Numeric | 3191.049 |
| Additional\_charges | Average charge for misc procedures/medicine/etc. | Numeric | 17939.4 |
| Item1 | Importance of timely admissions (Ranked, 1-8) | Integer | 3 |
| Item2 | Importance of timely treatments (Ranked, 1-8) | Integer | 3 |
| Item3 | Importance of timely visits (Ranked, 1-8) | Integer | 2 |
| Item4 | Importance of reliability (Ranked, 1-8) | Integer | 2 |
| Item5 | Importance of options (Ranked, 1-8) | Integer | 4 |
| Item6 | Importance of treatment hours (Ranked, 1-8) | Integer | 3 |
| Item7 | Importance of courteous staff (Ranked, 1-8) | Integer | 3 |
| Item8 | Importance of physician active listening (Ranked, 1-8) | Integer | 4 |

Highlighted information is information that may need correction in future. For example, the name of the variable TotalCharge relating to average daily charge rather than actual total charge, or recategorization of variables.

Part II: Data-Cleaning Plan

# Explain the plan for cleaning the data by doing the following:

## Propose a plan that includes the relevant techniques and specific steps needed to identify anomalies in the data set.

My plan for identifying anomalies in the data is as follows:

Examine the data set (summary, column names, definitions, etc.)

Examine the presence or absence of duplicate values (or columns) within the dataset

Visualize the data for missingness

Hypothesize appropriate treatment technique of missing data based upon % missing, skewness of variable, and after noting presence of outliers

Skewness could be determined from looking at a histogram

Determine what categories may need to be recategorized as numerical

Determine presence of outliers in numeric variables

Hypothesize appropriate treatment of outliers

This could be visualized by a histogram/boxplot, or utilizing z-scores

My plan for treatment later will be as follows:

Treat outliers based upon findings (% outliers, feasibility of actual values)

Treat null values using appropriate measure of central tendency, again based on skew

Recategorize variables to numeric as possible to perform principal component analysis (PCA)

Determine number of principal components to keep via PCA

## Justify your approach for assessing the quality of the data, including:

### Characteristics of the data being assessed

### The approach used to assess the quality

I am utilizing the above steps, as at a cursory glance I can see the following in my dataset: it is a large dataset with some missing variables, it is likely to have at least a few outliers with a sample that large, and many variables that may play a role in patient readmission rates are categorical variables, and PCA needs to be performed on numeric variables.

On top of the characteristics of my dataset, per Dr. Middleton’s 2nd webinar, it is important to follow a specific “order of operations” when dealing with dirty data (Getting Started with Missing Data and Outliers, 2021). This includes examining the dataset, detecting duplicate/missing values/outliers, re-expressing categorical variables, and performing the PCA.

## Justify your selected programming language and any libraries and packages that will support the data-cleaning process.

My choice of programming language for this performance assessment is to utilize R over python. My rationale for this is that I had an easier time grasping the content of the R DataCamp lectures and practice sessions for cleaning data and spent more time with R as there were 6 courses vs. 5 courses for Python. DataCamp has also provided an infographic to discuss the pros and cons of each language, and I agree with their sentiment that R is slightly easier to pick up for somebody with minimal to no coding experience (2020). I fall into this category of a programming novice. It is also possible to write a functional code in various ways with the assistance of the packages that are provided in R, thus demonstrating the flexibility and adaptability of the language (DataCamp Team, 2020).

I believe that either language would be appropriate for performing this task, but I also intend to visualize and quantify the missingness of my data using visdat and naniar, and R handles visualizations slightly better than Python does (DataCamp Team, 2020). I will also likely be using ggplot2 for boxplots and/or histograms to assist with treatment of outliers and missing values. I will be using the built-in read\_csv() function to upload my data into R Studio. I also may possibly use dplyr to assist with piping throughout my work, via the %>% pipe within the package. Modeest will be used to assist with determining the mode of any appropriate outliers/missing values. The library plyr will be utilized for revaluing and replacing values when transforming categorical variables. And finally, factoextra will be utilized for performing PCA on the cleaned dataset. Naniar, visdat, ggplot2, and dplyr were known packages from completing DataCamp activities throughout the last month. Modeest, plyr, and factoextra were packages that were brought to my attention by Dr. Middleton’s Webinar 2, “Getting Started with Missing Data and Outliers,” Webinar 3, “Getting Started with Re-expression of Categorical Variables” and Webinar 4, “Understanding PCA,” respectively (2021).

## Provide the code you will use to identify the anomalies in the data.

**Upload data/setup environment**

medical\_raw\_data <- read.csv("C:/Users/lgben/OneDrive/Desktop/MSDA/D206 - Data Cleaning/Medical\_Raw\_Data.csv")

library(visdat)

library(ggplot2)

library(tidyverse)

library(dplyr)

library(naniar)

library(modeest)

library(plyr)

library(factoextra)

**General visualization/background**:

summary(medical\_raw\_data)

colnames(medical\_raw\_data)

str(medical\_raw\_data)

head(medical\_raw\_data)

**Presence of duplicates**:

sum(duplicated(medical\_raw\_data))

medical\_raw\_data\_copy <- distinct(medical\_raw\_data)

medical\_raw\_data\_copy <- medical\_raw\_data\_copy[, -1]

**Presence of Missing Values**:

vis\_miss(medical\_raw\_data\_copy)

miss\_case\_table(medical\_raw\_data\_copy)

miss\_prop\_summary(medical\_raw\_data\_copy)

sum(is.na(medical\_raw\_data\_copy))

summary(is.na(medical\_raw\_data\_copy))

colSums(is.na(medical\_raw\_data\_copy))

*Histograms of variables with missingness*

hist(medical\_raw\_data\_copy$Age)

hist(medical\_raw\_data\_copy$Children)

hist(medical\_raw\_data\_copy$Income)

hist(medical\_raw\_data\_copy$Soft\_drink)

hist(medical\_raw\_data\_copy$Overweight)

hist(medical\_raw\_data\_copy$Anxiety)

hist(medical\_raw\_data\_copy$Initial\_days)

*Creating numeric soft\_drink\_category*

soft\_drink\_numeric <- as.numeric(as.factor(medical\_raw\_data\_copy$Soft\_drink))

soft\_drink\_numeric <- soft\_drink\_numeric[]-1

medical\_raw\_data\_copy <- cbind(medical\_raw\_data\_copy, soft\_drink\_numeric)

hist(medical\_raw\_data\_copy$soft\_drink\_numeric)

**Z-scores and outliers**

*Creating z-score columns*

medical\_raw\_data\_copy$age\_z <- scale(x = medical\_raw\_data\_copy$Age)

medical\_raw\_data\_copy$children\_z <- scale(x = medical\_raw\_data\_copy$Children)

medical\_raw\_data\_copy$income\_z <- scale(x = medical\_raw\_data\_copy$Income)

medical\_raw\_data\_copy$vit\_d\_levels\_z <- scale(x = medical\_raw\_data\_copy$VitD\_levels)

medical\_raw\_data\_copy$doc\_visits\_z <- scale(x = medical\_raw\_data\_copy$Doc\_visits)

medical\_raw\_data\_copy$full\_meals\_z <- scale(x = medical\_raw\_data\_copy$Full\_meals\_eaten)

medical\_raw\_data\_copy$vit\_d\_supp\_z <- scale(x = medical\_raw\_data\_copy$VitD\_supp)

medical\_raw\_data\_copy$soft\_drink\_numeric\_z <- scale(x = medical\_raw\_data\_copy$soft\_drink\_numeric)

medical\_raw\_data\_copy$overweight\_z <- scale(x = medical\_raw\_data\_copy$Overweight)

medical\_raw\_data\_copy$anxiety\_z <- scale(x = medical\_raw\_data\_copy$Anxiety)

medical\_raw\_data\_copy$initial\_days\_z <- scale(x = medical\_raw\_data\_copy$Initial\_days)

medical\_raw\_data\_copy$total\_charge\_z <- scale(x = medical\_raw\_data\_copy$TotalCharge)

medical\_raw\_data\_copy$additional\_charges\_z <- scale(x = medical\_raw\_data\_copy$Additional\_charges)

medical\_raw\_data\_copy$item\_1\_z <- scale(x = medical\_raw\_data\_copy$Item1)

medical\_raw\_data\_copy$item\_2\_z <- scale(x = medical\_raw\_data\_copy$Item2)

medical\_raw\_data\_copy$item\_3\_z <- scale(x = medical\_raw\_data\_copy$Item3)

medical\_raw\_data\_copy$item\_4\_z <- scale(x = medical\_raw\_data\_copy$Item4)

medical\_raw\_data\_copy$item\_5\_z <- scale(x = medical\_raw\_data\_copy$Item5)

medical\_raw\_data\_copy$item\_6\_z <- scale(x = medical\_raw\_data\_copy$Item6)

medical\_raw\_data\_copy$item\_7\_z <- scale(x = medical\_raw\_data\_copy$Item7)

medical\_raw\_data\_copy$item\_8\_z <- scale(x = medical\_raw\_data\_copy$Item8)

*Creating outlier vectors containing cases with outliers*

age\_outliers <- which(medical\_raw\_data\_copy$age\_z >3 | medical\_raw\_data\_copy$age\_z < -3)

children\_outliers <- which(medical\_raw\_data\_copy$children\_z > 3 | medical\_raw\_data\_copy$children\_z < -3)

income\_outliers <- which(medical\_raw\_data\_copy$income\_z >3 | medical\_raw\_data\_copy$income\_z < -3)

vit\_d\_levels\_outliers <- which(medical\_raw\_data\_copy$vit\_d\_levels\_z > 3 | medical\_raw\_data\_copy$vit\_d\_levels\_z < -3)

doc\_visits\_outliers <- which(medical\_raw\_data\_copy$doc\_visits\_z > 3 | medical\_raw\_data\_copy$doc\_visits\_z < -3)

full\_meals\_outliers <- which(medical\_raw\_data\_copy$full\_meals\_z > 3 | medical\_raw\_data\_copy$full\_meals\_z < -3)

vit\_d\_supp\_outliers <- which(medical\_raw\_data\_copy$vit\_d\_supp\_z > 3 | medical\_raw\_data\_copy$vit\_d\_supp\_z < -3)

soft\_drink\_outliers <- which(medical\_raw\_data\_copy$soft\_drink\_numeric\_z > 3 | medical\_raw\_data\_copy$soft\_drink\_numeric\_z < -3)

overweight\_outliers <- which(medical\_raw\_data\_copy$overweight\_z > 3 | medical\_raw\_data\_copy$overweight\_z < -3)

anxiety\_outliers <- which(medical\_raw\_data\_copy$anxiety\_z > 3 | medical\_raw\_data\_copy$anxiety\_z < -3)

initial\_days\_outliers <- which(medical\_raw\_data\_copy$initial\_days\_z > 3 | medical\_raw\_data\_copy$initial\_days\_z < -3)

total\_charge\_outliers <- which(medical\_raw\_data\_copy$total\_charge\_z > 3 | medical\_raw\_data\_copy$total\_charge\_z < -3)

additional\_charges\_outliers <- which(medical\_raw\_data\_copy$additional\_charges\_z >3 | medical\_raw\_data\_copy$additional\_charges\_z < -3)

item\_1\_outliers <- which(medical\_raw\_data\_copy$item\_1\_z > 3 | medical\_raw\_data\_copy$item\_1\_z < -3)

item\_2\_outliers <- which(medical\_raw\_data\_copy$item\_2\_z > 3 | medical\_raw\_data\_copy$item\_2\_z < -3)

item\_3\_outliers <- which(medical\_raw\_data\_copy$item\_3\_z > 3 | medical\_raw\_data\_copy$item\_3\_z < -3)

item\_4\_outliers <- which(medical\_raw\_data\_copy$item\_4\_z > 3 | medical\_raw\_data\_copy$item\_4\_z < -3)

item\_5\_outliers <- which(medical\_raw\_data\_copy$item\_5\_z > 3 | medical\_raw\_data\_copy$item\_5\_z < -3)

item\_6\_outliers <- which(medical\_raw\_data\_copy$item\_6\_z > 3 | medical\_raw\_data\_copy$item\_6\_z < -3)

item\_7\_outliers <- which(medical\_raw\_data\_copy$item\_7\_z > 3 | medical\_raw\_data\_copy$item\_7\_z < -3)

item\_8\_outliers <- which(medical\_raw\_data\_copy$item\_8\_z > 3 | medical\_raw\_data\_copy$item\_8\_z < -3)

*Viewing actual values of outliers*

medical\_raw\_data\_copy[children\_outliers, “Children”]

medical\_raw\_data\_copy[doc\_visits\_outliers, “Doc\_visits”]

medical\_raw\_data\_copy[full\_meals\_outliers, “Full\_meals\_eaten”]

medical\_raw\_data\_copy[income\_outliers, “Income”]

medical\_raw\_data\_copy[total\_charge\_outliers, “TotalCharge”]

medical\_raw\_data\_copy[vit\_d\_levels\_outliers, “VitD\_levels”]

medical\_raw\_data\_copy[vit\_d\_supp\_outliers, “VitD\_supp”]

*Creating vectors with* ***outlier cases***

outlier\_rows <- unique(c(children\_outliers, doc\_visits\_outliers, full\_meals\_outliers, income\_outliers, total\_charge\_outliers, vit\_d\_levels\_outliers, vit\_d\_supp\_outliers))

item\_outliers <- unique(c(item\_1\_outliers, item\_2\_outliers, item\_3\_outliers, item\_4\_outliers, item\_5\_outliers, item\_6\_outliers, item\_7\_outliers, item\_8\_outliers))

multiple\_item\_outliers <- c(item\_1\_outliers, item\_2\_outliers, item\_3\_outliers, item\_4\_outliers, item\_5\_outliers, item\_6\_outliers, item\_7\_outliers, item\_8\_outliers)

multiple\_item\_outliers <- subset(multiple\_item\_outliers, duplicated(multiple\_item\_outliers))

*Viewing actual values for outliers for “Item outliers”*

medical\_raw\_data\_copy[item\_outliers, c(“Item1”, “Item2”, “Item3”, “Item4”, “Item5”, “Item6”, “Item7”, “Item8”)]

medical\_raw\_data\_copy[multiple\_item\_outliers, c(“Item1”, “Item2”, “Item3”, “Item4”, “Item5”, “Item6”, “Item7”, “Item8”)]

*Creating histograms of outliers to visualize the data, and comparing with histograms of full dataset*

hist(medical\_raw\_data\_copy$Children)

hist(medical\_raw\_data\_copy[children\_outliers, “Children”])

hist(medical\_raw\_data\_copy$Doc\_visits)

hist(medical\_raw\_data\_copy[doc\_visits\_outliers, “Doc\_visits”])

hist(medical\_raw\_data\_copy$Full\_meals\_eaten)

hist(medical\_raw\_data\_copy[full\_meals\_outliers, “Full\_meals\_eaten”])

hist(medical\_raw\_data\_copy$Income)

hist(medical\_raw\_data\_copy[income\_outliers, “Income”])

hist(medical\_raw\_data\_copy$TotalCharge)

hist(medical\_raw\_data\_copy[total\_charge\_outliers, “TotalCharge”])

hist(medical\_raw\_data\_copy$VitD\_levels)

hist(medical\_raw\_data\_copy[vit\_d\_levels\_outliers, “VitD\_levels”])

hist(medical\_raw\_data\_copy$VitD\_supp)

hist(medical\_raw\_data\_copy[vit\_d\_supp\_outliers, “VitD\_supp”])

hist(medical\_raw\_data\_copy$Item1)

hist(medical\_raw\_data\_copy$Item2)

hist(medical\_raw\_data\_copy$Item3)

hist(medical\_raw\_data\_copy$Item4)

hist(medical\_raw\_data\_copy$Item5)

hist(medical\_raw\_data\_copy$Item6)

hist(medical\_raw\_data\_copy$Item7)

hist(medical\_raw\_data\_copy$Item8)

*Comparing outliers to other relevant medical data*

medical\_raw\_data\_copy[children\_outliers, c(“Age”, “Income”, “Children”)]

medical\_raw\_data\_copy[income\_outliers, c(“Age”, “Income”, “Children”)]

medical\_raw\_data\_copy[doc\_visits\_outliers, c(“Doc\_visits”, “Initial\_days”)]

medical\_raw\_data\_copy[full\_meals\_outliers, c(“Full\_meals\_eaten”, “Initial\_days”)]

medical\_raw\_data\_copy[total\_charge\_outliers, c(“TotalCharge”, “Initial\_days”)]

Part III: Data Cleaning

# Summarize the data-cleaning process by doing the following:

## Describe the findings, including all anomalies, from the implementation of the data-cleaning plan from part C.

**Duplicates**:

It was noted during the initial observation of the dataset, during the str() call, that the two first columns had the exact same information in them. There were no exact duplicate rows, but the duplicate columns were noted. The first column, “X”, contained the same information as the second column, “CaseOrder.”

**Missingness**:

The dataset is missing approximately 2.4% of its data, or 12,955 missing values. 7690 rows, or approximately 77% of cases, are missing at least 1 variable. There are 7 categories (13%) that are missing information, including: Children, Age, Income, Soft\_drinks, Overweight, Anxiety, and Initial\_days. Children is missing 2588 values. Age is missing 2414 values. Income is missing 2464 values. Soft\_drink is missing 2467 values. Overweight is missing 982 values. Anxiety is missing 984 values. Initial days is missing 1056 values. The missingness is noted to be MCAR, or missing completely at random.

**Outliers**:

Outliers were found by utilizing z-scores for each numerical category. The code for this was adapted from Dr. Middleton’s Webinar 2, “Getting Started with Missing Data and Outliers” (2021). Outliers were not found for all numeric columns, and the columns without outliers were as follows: Initial\_days, Overweight, Anxiety, and soft\_drink\_numeric.

Outliers were found for the following columns, totaling 915 cases:

Children contained 148 outliers. Doc\_visits contained 8 outliers. Full\_meals\_eaten contained 33 outliers. Income contained 113 outliers. TotalCharges contained 276 outliers. VitD\_levels contained 500 outliers. VitD\_supp contained 70 outliers. Each survey item (Item1 through Item 8) contained between 10-13 outliers. This total number of outliers is equivalent to slightly greater than 9% of the entire sample.

## Justify your methods for mitigating each type of discovered anomaly in the data set.

**Dealing with duplicates**:

The first column, “X” was dropped from the dataset as it did not provide any relevant/new information over the CaseOrder column. These columns contained identical information and would therefore qualify as a duplicate column. This could be seen with the snippet of each column provided via str().

**Dealing with outliers**:

I felt that with my dataset, there was not a one size fits all method for treating the outliers. Some of the outlier values are not so unreasonable that I believe of them to be inaccurate or entry errors. Based upon the total number of cases of outliers, dropping them all is not a valid solution as nearly 10% of the cases would be lost. For this reason, I decided that some may be dropped, while others are imputed, while others are left as is. I believe that viewing the data in specific contexts helped to shed light on what appropriate treatment may be for each category (Birkett, 2019).

Upon looking at the outliers for Children, it is seen that some of these individuals fall in the 18-25 age range and it is more unlikely than not that these entries are errors, as 9-10 children is slightly unrealistic for that age group. Rather than dropping the Children outliers, these will be imputed with the median. Most of the ages noted for the individuals within the Income outliers were 35+ years old which makes it believable that their incomes may be in the 6 figure range. Therefore Income will remain as it is for treatment.

The outliers for Doc\_visits were those individuals that had either 1 or 9 visits, however those with 9 doctor visits had Initial\_days of under 5 days, while those with 1 visit all had 35+ initial day stay. This seems odd from a healthcare standpoint thus these cases will be dropped from the dataset.

Most of the outliers for full meals eaten, which fell between 5 and 7 full meals, had double digit stays for their “Initial\_days” category and it would again not be unreasonable to have eaten multiple full meals when staying for 10+ days. There was a total of 2 patients within the full\_meals\_eaten\_outliers (cases 959, 2920) who had 5+ meals however for a single day initial stay. These two cases will be dropped, while the remainder will stay as they are for treatment.

TotalCharge had 276 outliers. It has already been noted from the data dictionary that this category represents the average daily charge, rather than the total, for each case. Outliers for this category were all values above $16,053.46. Some of these individuals had stays in the range of 20-40 days, but many of these individuals again had stays in the 50–70-day range. Given this information it is not unreasonable to see how these patients could accumulate large total charges, and therefore large average daily charges, and they will be left as they are for treatment.

Outliers for the VitD\_levels category were well above the rest of the distribution and were in a normal distribution of their own. There was a total of 500 outliers for this category, accounting for 5% of the overall sample. Because of this, removal would likely not be a good option for these outliers and these values will be imputed as the median value.

Outliers for the VitD\_supp category did not account for much of the dataset as there were only 70 of these, accounting for 0.7% of the sample. Because the total number of cases where outliers are present is low, I feel comfortable with dropping these cases.

Outliers for all the survey Items contained between 10-13 outliers. All the outliers were survey responses that answered the question with a 7 or 8. A minority of these cases responded to multiple survey items unfavorably. Because there are only 86 unique cases within “Item\_outliers”, a vector that was created to group all the Item1 through Item8 outliers together, which accounts for 0.86% of the overall sample, I also feel comfortable dropping these cases.

As stated above I did not feel comfortable completely removing all the outliers due to accounting for >9% of the overall data set. I also felt that the context of some of the outliers made sense, as such I felt it an appropriate treatment to keep the outliers of the following categories: Income, most of Full\_meals\_eaten, and TotalCharge. VitD\_levels and Children were the only categories in which the outliers will have the median imputed, but VitD\_levels accounted for most of the outliers. Finally, the outlier cases of Doc\_visits, VitD\_supp, the 2 patients from Full\_meals\_eaten (959, 2920) that had 5+ full meals but a single day stay, and all survey “Item” outliers will be removed. This accounts for 166 total cases being removed, or 1.66% of the overall dataset.

**Dealing with missingness**:

Based upon the summary() of the missing columns and the histograms that were visualized, the following options appear to be the best options for imputed values.

*Medians*

Children will be imputed with the median of 1, as the histogram demonstrated a positive skew meaning that is a better measure of central tendency. Age will be imputed with the median of 53, but the mean could also be acceptable secondary to the normal distribution and mean being less than a year different (53.27). Income will be imputed with the median value of $33,945.90 as the histogram demonstrates a positive skew, once again demonstrating that the mean is not the most appropriate measure of central tendency.

*Modes*

Soft drink will be imputed with the mode of 0, which is also the median (only two options of 1 or 0). Overweight will be imputed with the mode of 1, which is again also the median of the two given options. Anxiety will be imputed with the mode of 0, which is again also the median of the two given options. Initial\_days will be imputed with the “most likely value” (MLV) of 9.224404. The most likely value is an “estimate of the mode” provided through the modeest package (Poncet, 2019). The histogram for Initial\_days demonstrates a slightly bimodal distribution, with the most values being noted in the 5-10 bucket. The MLV that was found falls within this bucket.

**Recategorization**:

In order to perform a principal component analysis, categorical variables need to be recategorized as numeric. First, the column name TotalCharge was renamed AverageCharge, as per the data dictionary this column actually gives the value of each average daily charge for the patient. This column title is more accurate. Next, the categorical columns with pertinent patient information, that are not geographical in nature, include: Education, Employment, Marital, Gender, ReAdmis, Initial\_admin, HighBlood, Stroke, Complication\_risk, Arthritis, Diabetes, Hyperlipidemia, BackPain, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, and Services. These will all be recategorized as numeric for the purposes of PCA performance.

The way that each category was visualized was using the table() function to demonstrate all of the categorical options for each column (StackOverflow, 2018). These categories were then given a numeric label and placed in vectors that would later be bound to the dataframe (Middleton, 2022). Following the recategorization of these columns and binding the new columns to the dataframe, the original columns will be removed from the dataset to reduce the dimensionality of the data (Bhalla, n.d.).

## Summarize the outcome from the implementation of *each* data-cleaning step.

From the steps taken above, I was able to first remove a duplicate column. This column did not provide any new, pertinent information to the dataset. Next, some specific outliers were removed, reducing the overall size of the dataframe slightly while others were imputed or left alone. Null values were then imputed utilizing the medians, modes, or most likely value, depending on the column that was being treated. Throughout these steps, histograms were visualized to ensure that the variables maintained their original distribution. Finally, categorical columns were made into numeric columns and their original columns were removed.

The outcome of this was a dataframe with a majority numeric columns/variables outside of geographic and personal information, no null values, and only a few outliers remaining, allowing for future performance of PCA with little concern over the results being biased.

## Provide the code used to mitigate anomalies.

**Treating outliers**

outliers\_rows\_drop <- unique(c(doc\_visits\_outliers, full\_meals\_outliers[c(3, 12)], vit\_d\_supp\_outliers, item\_outliers))

summary(medical\_raw\_data\_copy[-vit\_d\_levels\_outliers, “VitD\_levels”])

min(medical\_raw\_data\_copy[vit\_d\_levels\_outliers, “VitD\_levels”])

medical\_raw\_data\_copy$vitD\_levels <- replace(medical\_raw\_data\_copy$VitD\_levels, medical\_raw\_data\_copy$VitD\_levels > 40, 17.936)

hist(medical\_raw\_data\_copy$VitD\_levels)

medical\_raw\_data\_copy$Children <- replace(medical\_raw\_data\_copy$Children, medical\_raw\_data\_copy$Children > 8, 1)

hist(medical\_raw\_data\_copy$Children)

summary(medical\_raw\_data\_copy[, c(“Item1”, “Item2”, “Item3”, “Item4”, “Item5”, “Item6”, “Item7”, “Item8”)])

*New DF without outliers*

mrdc\_no\_outliers <- medical\_raw\_data\_copy[-outliers\_rows\_drop, ]

mrdc\_no\_outliers = subset(mrdc\_no\_outliers, select = -c(children\_z, anxiety\_z, overweight\_z, age\_z, initial\_days\_z, soft\_drink\_numeric\_z, vit\_d\_levels\_z, doc\_visits\_z, full\_meals\_z, vit\_d\_supp\_z, totalcharge\_z, additional\_charges\_z, income\_z, item\_1\_z, item\_2\_z, item\_3\_z, item\_4\_z, item\_5\_z, item\_6\_z, item\_7\_z, item\_8\_z))

summary(mrdc\_no\_outliers[, c("Item1", "Item2", "Item3", "Item4", "Item5", "Item6", "Item7", "Item8")])

**Treating Missingness**

*Removing previously added “Soft\_drink\_numeric column” and correcting true Soft\_drink column*

mrdc\_no\_outliers$Soft\_drink <- as.numeric(as.factor(mrdc\_no\_outliers$Soft\_drink))-1

mrdc\_no\_outliers = subset(mrdc\_no\_outliers, select = -c(soft\_drink\_numeric))

summary(mrdc\_no\_outliers[na\_columns])

*Histograms, again*

hist(mrdc\_no\_outliers$Initial\_days)

hist(mrdc\_no\_outliers$Income)

hist(mrdc\_no\_outliers$Anxiety)

hist(mrdc\_no\_outliers$Overweight)

hist(mrdc\_no\_outliers$Age)

hist(mrdc\_no\_outliers$Children)

hist(mrdc\_no\_outliers$Soft\_drink)

*Creating another extra DF in case imputing goes wrong*

mrdc\_copy <- mrdc\_no\_outliers

*Replacing nulls*

mrdc\_no\_outliers$Children <- replace\_na(mrdc\_no\_outliers$Children, median(mrdc\_no\_outliers$Children, na.rm = TRUE))

mrdc\_no\_outliers$Age <- replace\_na(mrdc\_no\_outliers$Age, median(mrdc\_no\_outliers$Age, na.rm = TRUE))

mrdc\_no\_outliers$Income <- replace\_na(mrdc\_no\_outliers$Income, median(mrdc\_no\_outliers$Income, na.rm = TRUE))

mrdc\_no\_outliers$Soft\_drink <- replace\_na(mrdc\_no\_outliers$Soft\_drink, median(mrdc\_no\_outliers$Soft\_drink, na.rm = TRUE))

mrdc\_no\_outliers$Overweight <- replace\_na(mrdc\_no\_outliers$Overweight, median(mrdc\_no\_outliers$Overweight, na.rm = TRUE))

mrdc\_no\_outliers$Anxiety <- replace\_na(mrdc\_no\_outliers$Anxiety, median(mrdc\_no\_outliers$Anxiety, na.rm = TRUE))

mlv(mrdc\_no\_outliers$Initial\_days, na.rm = TRUE) (Poncet, 2018)

mrdc\_no\_outliers$Initial\_days <- replace\_na(mrdc\_no\_outliers$Initial\_days, 9.224275)

*Histograms, again*

hist(mrdc\_no\_outliers$Initial\_days)

hist(mrdc\_no\_outliers$Income)

hist(mrdc\_no\_outliers$Anxiety)

hist(mrdc\_no\_outliers$Overweight)

hist(mrdc\_no\_outliers$Age)

hist(mrdc\_no\_outliers$Children)

hist(mrdc\_no\_outliers$Soft\_drink)

vis\_miss(mrdc\_no\_outliers)

**Recategorization**

*Renaming variables/creating numeric vectors*

table(mrdc\_no\_outliers$Education)

education\_num <- revalue(x = mrdc\_no\_outliers$Education, replace = c(

"No Schooling Completed" = 0,

"Nursery School to 8th Grade" = 1,

"9th Grade to 12th Grade, No Diploma" = 2,

"Regular High School Diploma" = 3,

"GED or Alternative Credential" = 3,

"Professional School Degree" = 4,

"Some College, Less than 1 Year" = 5,

"Some College, 1 or More Years, No Degree" = 6,

"Associate's Degree" = 7,

"Bachelor's Degree" = 8,

"Master's Degree" = 9,

"Doctorate Degree" = 10))

table(mrdc\_no\_outliers$Marital)

marital\_num <- revalue(x = mrdc\_no\_outliers$Marital, replace = c("Divorced" = 1, "Married" = 2, "Never Married" = 3, "Separated" = 4, "Widowed" = 5))

table(mrdc\_no\_outliers$Employment)

employment\_num <- revalue(x = mrdc\_no\_outliers$Employment, replace = c(

"Unemployed" = 0,

"Student" = 1,

"Part Time" = 2,

"Full Time" = 3,

"Retired" = 4))

table(mrdc\_no\_outliers$Gender)

gender\_num <- revalue(x = mrdc\_no\_outliers$Gender, replace = c(

“Female” = 1,

“Male” = 2,

“Prefer not to answer” = 3))

table(mrdc\_no\_outliers$ReAadmis)

readmis\_num <- as.numeric(as.factor(mrdc\_no\_outliers$ReAdmis))-1

table(mrdc\_no\_outliers$Initial\_admin)

initial\_admin\_num <- revalue(x = mrdc\_no\_outliers$Initial\_admin, replace = c(

“Elective Admission” = 1,

“Emergency Admission” = 2,

“Observation Admission” = 3))

hbp\_num <- as.numeric(as.factor(mrdc\_no\_outliers$HighBlood))-1

stroke\_num <- as.numeric(as.factor(mrdc\_no\_outliers$Stroke))-1

table(mrdc\_no\_outliers$Complication\_risk)

complication\_risk\_num <- revalue(x = mrdc\_no\_outliers$Complication\_risk, replace = c(

“Low” = 1,

“Medium” = 2,

“High” = 3))

arthritis\_num <- as.numeric(as.factor(mrdc\_no\_outliers$Arthritis))-1

diabetes\_num <- as.numeric(as.factor(mrdc\_no\_outliers$Diabetes))-1

hyperlipidemia\_num <- as.numeric(as.factor(mrdc\_no\_outliers$Hyperlipidemia))-1

back\_pain\_num <- as.numeric(as.factor(mrdc\_no\_outliers$BackPain))-1

allergic\_rhin\_num <- as.numeric(as.factor(mrdc\_no\_outliers$Allergic\_rhinitis))-1

reflux\_num <- as.numeric(as.factor(mrdc\_no\_outliers$Reflux\_esophagitis))-1

asthma\_num <- as.numeric(as.factor(mrdc\_no\_outliers$Asthma))-1

table(mrdc\_no\_outliers$Services)

services\_num <- revalue(x = mrdc\_no\_outliers$Services, replace = c(

“Blood Work” = 1,

“CT Scan” = 2,

“Intravenous” = 3,

“MRI” = 4))

*Binding numeric vectors to table*

mrdc\_no\_outliers <- cbind(mrdc\_no\_outliers, allergic\_rhin\_num, arthritis\_num, asthma\_num, back\_pain\_num, diabetes\_num, hbp\_num, hyperlipidemia\_num, readmis\_num, reflux\_num, stroke\_num)

mrdc\_no\_outliers$complication\_risk\_num <- as.numeric(complication\_risk\_num)

mrdc\_no\_outliers$education\_num <- as.numeric(education\_num)

mrdc\_no\_outliers$employment\_num <- as.numeric(employment\_num)

mrdc\_no\_outliers$gender\_num <- as.numeric(gender\_num)

mrdc\_no\_outliers$initial\_admin\_num <- as.numeric(initial\_admin\_num)

mrdc\_no\_outliers$marital\_num <- as.numeric(marital\_num)

mrdc\_no\_outliers$services\_num <- as.numeric(services\_num)

*Another copy dataframe, prior to removal of categorical variables*

mrdc\_no\_outliers2 <- mrdc\_no\_outliers

*Dropping previously “character” columns*

mrdc\_no\_outliers <- subset(mrdc\_no\_outliers, select = -c(Education, Employment, Marital, Gender, ReAdmis, Initial\_admin, HighBlood, Stroke, Complication\_risk, Arthritis, Diabetes, Hyperlipidemia, BackPain, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, Services))

mrdc\_no\_outliers$AverageCharge <- mrdc\_no\_outliers$TotalCharge

mrdc\_no\_outliers <- subset(mrdc\_no\_outliers, select = -c(TotalCharge))

*Creating CSV file of cleaned dataset*

write.csv(mrdc\_no\_outliers, “C:\\Users\\lgben\\OneDrive\\Desktop\\MSDA\\D206 – Data Cleaning\\Cleaned\_Med\_Data.csv”)

## Provide a copy of the cleaned data set.

The cleaned data set will be attached as a .csv file.

## Summarize the limitations of the data-cleaning process.

One limitation of the data cleaning process performed on this dataset is that some outliers were maintained within the dataset and the same treatment was not utilized for all outliers. This allows for some subjective bias in the treatment.

I decided to keep many of these cases as they were because the outliers made sense in the context of other variables. I also know that removal of all outliers would have taken a large portion of the dataset. I also felt that imputing all the outliers would have vastly changed some of the values, for example incomes in the $150,000 range much larger than $30,000-40,000 and missing values were already going to be imputed at this same value. I didn’t feel that introducing this level of imputation for a single variable was a good choice.

Another limitation is that some of the variables that required imputation, whether due to missingness or outliers, were not in a normalized distribution. Income had a bimodal distribution with one mode larger than the other, while Children and VitD\_levels both had positively skewed distributions.

## Discuss how the limitations in part D6 affect the analysis of the question or decision from part A.

One way that the limitations above may affect my research question is that outliers that remained in the dataset could potentially have a larger bearing on readmission rates than they would have if the outliers were imputed, as some outliers were left as they were for treatment. The inverse could also be said for VitD\_levels that were imputed for the median, that they may have previously affected more of the variance in the dataset prior to imputation and their values being changed. It is obvious that there is no perfect way to perform the cleaning and analysis processes, but multiple analyses could be performed if the time allowed it to paint a clearer picture of how the variables affected the likelihood of readmission, i.e., with and without outliers, with and without imputation.

# Apply principal component analysis (PCA) to identify the significant features of the data set by doing the following:

## List the principal components in the data set.

The following is the output of the mrdc\_pca$rotation code:

Text

Description automatically generated

Text

Description automatically generated

It can be seen there are 38 principal components, as I utilized all the re-expressed categorical variables. In the format above not all rows are visible, but when viewing only those PCs that are relevant (as will be noted below), all rows will be visible.

## Describe how you identified the principal components of the data set.

The following code was adapted from Dr. Middleton’s “Webinar 4, Understanding PCA” to create the principal components and skree plot for the present dataset (2021). Principal components were found by first performing the PCA on the dataframe, with all columns starting from Children, using the following code:

mrdc\_pca <- prcomp(mrdc\_no\_outliers[,15:52], center = TRUE, scale = TRUE)

The components themselves were then visualized using the following code:

mrdc\_pca$rotation

This provided 38 principal components, as each numerical variable was utilized, including those that had been recategorized. Following this, a skree plot was created to determine which principal components were the most useful for analysis. This was done with the following code:

fviz\_eig(mrdc\_pca, choice = “eigenvalue”, addlabels = TRUE)

Chart, bar chart, histogram

Description automatically generated

From this skree plot, it was seen that the first four PCs were likely the most valuable, as the remaining PCs had eigenvalues of 1.1 or lower. According to the Kaiser rule, it is beneficial to keep PCs whose eigenvalue is greater than 1 (Middleton, 2021). I believe for my data set that since there were multiple eigenvalues at 1.1, these likely would not provide as much insight since they were so close to 1 and there were multiple PCs sharing this value. It could also be seen that the “elbow” of the plot began at those values whose eigenvalue was 1.1. The main PC loadings were then viewed with the following lines of code. The first gave each factor along with the loadings, while the second was better for visualizing which factors were positive or negative.

mrdc\_pca$rotation[, 1:4]

print(as\_tibble(mrdc\_pca$rotation[,1:4]), n = 38)

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## Describe how the organization can benefit from the results of the PCA.

One benefit to the organization could be improved tracking of particular values for each patient and assessing if there is a correlation between certain factors and a patient’s potential risk of readmission. Those patients who are not currently admitted to the organization as a readmission could be given a stratification of “at-risk” and could potentially benefit from further follow-up to reduce the risk or readmission. It is also seen with the prominent principal components that the survey responses have some bearing on the variance of the data. Improving on specific questions/measures may benefit patient care and satisfaction while hospitalized and improve overall function of the facility.

Part IV: Supporting Documents

# Provide a Panopto recording that demonstrates the warning- and error-free functionality of the code used to support the discovery of anomalies and the data cleaning process and summarizes the programming environment.

Please see the attached Panopto video.

# Reference the web sources used to acquire segments of third-party code to support the application. Be sure the web sources are reliable.

1. Bhalla, D. (n.d.). *R : Keep / drop columns from data frame*. ListenData. Retrieved February 16, 2022, from <https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html>
2. Middleton, K. (2021, December). *Webinar 2: Getting Started with "Missing Data and Outliers"*. *D206 - Data Cleaning*.
3. Middleton, K. (2021, December). *Webinar 3: Getting Started with Re-expression of Categorical Variables. D206 - Data Cleaning.*
4. Middleton, K. (2021, December). *Webinar 4: Understanding PCA*. *D206 - Data Cleaning.*
5. Poncet, P. (2019, November 18). *Mode estimation [R package modeest version 2.4.0]*. The Comprehensive R Archive Network. Retrieved February 18, 2022, from <https://cran.r-project.org/web/packages/modeest/index.html>
6. Stack Overflow. (2018, September). *R get all categories in column*. Stack Overflow. Retrieved February 20, 2022, from <https://stackoverflow.com/questions/46017812/r-get-all-categories-in-column>

# Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

1. DataCamp Team. (2020, January 9). *Choosing python vs. R for Data Analysis? An Infographic*. DataCamp. Retrieved February 14, 2022, from <https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis?utm_source=adwords_ppc&utm_medium=cpc&utm_campaignid=12492439679&utm_adgroupid=122563407961&utm_device=c&utm_keyword=r+or+python&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=504158803099&utm_targetid=aud-299261629654%3Akwd-348649097661&utm_loc_interest_ms=&utm_loc_physical_ms=9018981&gclid=CjwKCAiA6Y2QBhAtEiwAGHybPYzYHwR5UaJ5egOZz5qWt32Q_xPSuphivhAgFtCCcbggdvttOZnIJRoCHFQQAvD_BwE>
2. Birkett, A. (2019, August 24). How to Deal with Outliers in Your Data [web log]. Retrieved February 18, 2022, from <https://cxl.com/blog/outliers/>.
3. Middleton, K. (2021, December). *Webinar 2: Getting Started with "Missing Data and Outliers"*. *D206 - Data Cleaning*.
4. Middleton, K. (2021, December). *Webinar 3: Getting Started with Re-expression of Categorical Variables. D206 - Data Cleaning.*
5. Middleton, K. (2021, December). *Webinar 4: Understanding PCA*. *D206 - Data Cleaning.*

# Demonstrate professional communication in the content and presentation of your submission.