D206 Code Utilized

library(visdat)

library(ggplot2)

library(tidyverse)

library(dplyr)

library(naniar)

library(plyr)

library(modeest)

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

**General visualization/background**:

summary(medical\_raw\_data)

colnames(medical\_raw\_data)

str(medical\_raw\_data)

**Presences of duplicates**:

sum(duplicated(medical\_raw\_data)) [produced 0 exact duplicates]

medical\_raw\_data\_copy <- distinct(medical\_raw\_data) [copy created for performing imputations, to maintain original data frame as well as remove any duplicates, of which there are none]

medical\_raw\_data\_copy <- medical\_raw\_data\_copy[, -1] [dropped initial column that was a duplicate]

**Presence of Missing Values**:

vis\_miss(medical\_raw\_data\_copy) [produced 2.4% missingness overall]

miss\_case\_table(medical\_raw\_data\_copy) [produced 7 x 3 tibble showing missingness by number of variables missing per case]

miss\_prop\_summary(medical\_raw\_data\_copy) [produced 1 x 3 tibble showing missingness of the dataframe as a whole (2.4%), of variables/columns (13.2%, or 7 columns), and of cases/rows (76.9%, or approximately 7690 rows missing at least 1 variable]

sum(is.na(medical\_raw\_data\_copy)) [produced 12,955 NA values]

summary(is.na(medical\_raw\_data\_copy)) [produced summary showing 7 columns that contain NAs (TRUE) with number of NAs, with remaining columns showing 10,000 values (FALSE)]

colSums(is.na(medical\_raw\_data\_copy)) [also produces total NA for each column, slightly cleaner than summary()]

**Missingness/subsetting**:

na\_columns <- c(“Children”, “Age”, “Income”, “Soft\_drink”, “Overweight”, “Anxiety”, “Initial\_days”)

vis\_miss(medical\_raw\_data\_copy[, na\_columns], cluster = TRUE) [visualization of only columns with missingness, looking for pattern to the missingness; missingness appears to be Missing Completely At Random (MCAR)]

vis\_miss(arrange(medical\_raw\_data\_copy[,na\_columns], “na\_column”))

Performed for all columns to visualize pattern of missingness/relation of missingness

Confirmed that appears to be missing at random

Performed with various combinations of 2-3 na\_columns as well, continued to find no pattern to missingness

**Histograms of all the columns with missing data**:

Prior to assessing Soft\_drinks:

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

soft\_drink\_numeric <- soft\_drink\_numeric[]-1 (made 1s and 2s into 0s/1s)

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

hist(medical\_raw\_data\_copy$Initial\_days) [bimodal distribution near 1st/3rd quartiles, 1st>3rd]

hist(medical\_raw\_data\_copy$Income) [positive skew]

hist(medical\_raw\_data\_copy$Anxiety) [Only two options, 0 is mode]

hist(medical\_raw\_data\_copy$Overweight) [Only two options, 1 is mode]

hist(medical\_raw\_data\_copy$Age) [relatively even, normal distribution]

hist(medical\_raw\_data\_copy$Children) [positive skew]

hist(soft\_drink\_numeric) [only two options, 0 is mode]

summary(medical\_raw\_data\_copy[na\_columns]) [summary of missing columns, providing quartiles, means, medians of numerical columns; reviewing specifically these columns vs all columns]

summary(soft\_drink\_numeric)

mlv(medical\_raw\_data\_copy$Initial\_days, na.rm = TRUE) [produced most likely value of 9.224275]

na\_rows <- which(unique(is.na(medical\_raw\_data[, na\_columns]))) [creates vector with row values that contain 1+ missing piece of information]

na\_col\_id <- c(na\_columns, “Customer\_id”, “CaseOrder”) [additional na\_column vector to ensure customer\_id and caseorder are maintained during any transformations/visualizations]

**Z-scores and outliers**

column\_name\_z <- scale(x = medical\_raw\_data\_copy$column\_name)

column\_name\_outliers <- which(medical\_raw\_data\_copy$column\_name\_z >3 | medical\_raw\_data\_copy$column\_name < -3) [produces list of rows with outliers, for Na\_columns, only Children and Income have outliers]

outlier\_rows <- unique(c(all column\_name\_outliers that **are not empty**))

Comparisons:

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”])

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

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

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

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

summary(medical\_raw\_data\_copy[-vit\_d\_levels\_outliers, “VitD\_levels”]) [median 17.926, mean 17.959]

min(medical\_raw\_data\_copy[vit\_d\_levels\_outliers, “VitD\_levels”]) [min of 40.84167]

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

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

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

medical\_raw\_no\_outliers <- medical\_raw\_data[-outlier\_rows\_drop, -1]

mrdc\_no\_outliers <- medical\_raw\_data\_copy[-outlier\_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))

**Filling 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))

mrdc\_no\_outliers$Soft\_drink <- mrdc\_no\_outliers$Soft\_drink[]-1

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

summary(mrdc\_no\_outliers[na\_columns]) [summary of missing columns, providing quartiles, means, medians of numerical columns; reviewing specifically these columns vs all columns]

*Histograms, again*

hist(mrdc\_no\_outliers$Initial\_days) [bimodal distribution near 1st/3rd quartiles, 1st>3rd]

hist(mrdc\_no\_outliers$Income) [positive skew]

hist(mrdc\_no\_outliers$Anxiety) [Only two options, 0 is mode]

hist(mrdc\_no\_outliers$Overweight) [Only two options, 1 is mode]

hist(mrdc\_no\_outliers$Age) [relatively even, normal distribution]

hist(mrdc\_no\_outliers$Children) [positive skew]

hist(mrdc\_no\_outliers$Soft\_drink) [only two options, 0 is mode]

mrdc\_copy <- mrdc\_no\_outliers [extra DF in case imputing goes wrong]

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) [provided answer of 9.224404]

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

create all previously seen histograms

vis\_miss(mrdc\_no\_outliers) [no missing data noted]

**Recategorization**

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, 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

strok\_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))

**Websites utilized**

<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%20or%20python&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=504158803099&utm_targetid=aud-299261629654:kwd-348649097661&utm_loc_interest_ms=&utm_loc_physical_ms=9018981&gclid=CjwKCAiA6Y2QBhAtEiwAGHybPYzYHwR5UaJ5egOZz5qWt32Q_xPSuphivhAgFtCCcbggdvttOZnIJRoCHFQQAvD_BwE>

-used to help with determining which programming language to utilize

<https://statisticsglobe.com/print-all-rows-of-dplyr-tibble-to-console-in-r>

-used to help print full tibble while looking for NA values

<https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html>

-used to subset for dropping rows that contained outliers and creating new DF

<https://cran.r-project.org/web/packages/modeest/modeest.pdf>

-used to learn more on mlv() [most likely value] mode estimation of initial\_days

<https://cxl.com/blog/outliers/>

-used with assisting with appropriate outlier treatment

<https://stackoverflow.com/questions/46017812/r-get-all-categories-in-column>

-used to get table() function to determine all categories of education, marriage, etc.

**Powerpoint:**

2 (Missing Data and Outliers)

“Focus on those attributes and variables that are specifically related to the patient.” (43:33)

“Don’t be too concerned with some of the geographical elements of the dataset.” (43:43)

Order of operations for cleaning data (56:21)

“These are the lines of code so you can successfully extract your clean data set. “ (58:10)

“Example really just means essentially randomly selecting the observation for that variable.” (1:01:23)

3 (Recategorization)

“For the purposes of D206, what I would think would be the most beneficial to you would be for you to find variables that can be encoded ordinally.” (31:38)

“Just try to stay away from anything that’s not ordinal or can’t be essentially assigned some ordinal encoding.” (32:10)

Code for recategorizing factors (43:52)

4 (Data considerations/PCA)

“You need to make sure you carry this with you, particularly when you get to D212” (19:50)

“I typically encourage students, for the purposes of PCA in D206, to only use native numeric variables” (21:06)

“And if you decide to use a variable that you re-expressed, like a categorical variable that you re-expressed numerically, I would try to stay away from variables that you re-expressed that are just zeroes and ones like those dichotomous variables.” (21:19)

“Based on Kaiser rule, only going to keep those PCs whose eigenvalues are greater than 1” (51:58)

Additional sources (53:21)

**Textbook**:

<https://eds.p.ebscohost.com/eds/ebookviewer/ebook/bmxlYmtfXzIwOTEzNzFfX0FO0?sid=4eab6ad3-75e6-492d-b113-e5eb6980db3c@redis&vid=0&format=EB&rid=1>

Page 40 – identifying outliers