Milestone 1

Angel Waters

2022-06-05

# Data Loading and Cleanup

The necessary packages were loaded for the analysis.

library(ggplot2)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggQC)

Load the pre-cleaned up data and subset further.

##Loading the DK raw data subset for further cleaning and visualizing the data.  
rawData <-   
 read\_csv("SWANBaselineData\_ProfessorKSubset (1).csv")

## New names:  
## Rows: 3302 Columns: 33  
## -- Column specification  
## -------------------------------------------------------- Delimiter: "," chr  
## (18): HBCHOLE0, MIGRAIN0, ANEMIA0, LISTEN0, TAKETOM0, CONFIDE0, HELPSIC0... dbl  
## (15): ...1, SWANID, AGE0, HSWRKHR0, HOSPSTA0, PULSE0, SYSBP10, DIABP10, ...  
## i Use `spec()` to retrieve the full column specification for this data. i  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## \* `` -> `...1`

str(rawData)

## spec\_tbl\_df [3,302 x 33] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ...1 : num [1:3302] 1 2 3 4 5 6 7 8 9 10 ...  
## $ SWANID : num [1:3302] 10005 10046 10056 10092 10126 ...  
## $ AGE0 : num [1:3302] 48 52 51 45 48 51 46 47 46 47 ...  
## $ HBCHOLE0: chr [1:3302] "(1) No" "(1) No" "(1) No" "(1) No" ...  
## $ MIGRAIN0: chr [1:3302] "(1) No" "(1) No" "(1) No" "(1) No" ...  
## $ ANEMIA0 : chr [1:3302] "(1) No" "(1) No" "(2) Yes" "(2) Yes" ...  
## $ LISTEN0 : chr [1:3302] "(5) All Of The Time" "(5) All Of The Time" "(4) Most Of The Time" "(5) All Of The Time" ...  
## $ TAKETOM0: chr [1:3302] "(5) All Of The Time" "(5) All Of The Time" "(4) Most Of The Time" "(5) All Of The Time" ...  
## $ CONFIDE0: chr [1:3302] "(5) All Of The Time" "(5) All Of The Time" "(4) Most Of The Time" "(5) All Of The Time" ...  
## $ HELPSIC0: chr [1:3302] "(1) None Of The Time" "(5) All Of The Time" "(5) All Of The Time" "(5) All Of The Time" ...  
## $ COMMITE0: chr [1:3302] "(1) No" "(2) Yes" "(2) Yes" "(2) Yes" ...  
## $ BOTHER0 : chr [1:3302] "(3) Occasionally/Mod Amt Of The Time (3-4 Days)" "(2) Some/A Little Of The Time (1-2 Days)" "(2) Some/A Little Of The Time (1-2 Days)" "(2) Some/A Little Of The Time (1-2 Days)" ...  
## $ APPETIT0: chr [1:3302] "(4) Most/All Of The Time (5-7 Days)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" ...  
## $ BLUES0 : chr [1:3302] "(4) Most/All Of The Time (5-7 Days)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" ...  
## $ KEEPMIN0: chr [1:3302] "(3) Occasionally/Mod Amt Of The Time (3-4 Days)" "(1) Rarely/None Of The Time (< 1 Day)" "(3) Occasionally/Mod Amt Of The Time (3-4 Days)" "(3) Occasionally/Mod Amt Of The Time (3-4 Days)" ...  
## $ DEPRESS0: chr [1:3302] "(4) Most/All Of The Time (5-7 Days)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" ...  
## $ FAILURE0: chr [1:3302] "(3) Occasionally/Mod Amt Of The Time (3-4 Days)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" ...  
## $ HAPPY0 : chr [1:3302] "(1) Rarely/None Of The Time (< 1 Day)" "(3) Occasionally/Mod Amt Of The Time (3-4 Days)" "(4) Most/All Of The Time (5-7 Days)" "(3) Occasionally/Mod Amt Of The Time (3-4 Days)" ...  
## $ HSWRKHR0: num [1:3302] 18 30 60 2 16 15 15 40 14 49 ...  
## $ HOSPSTA0: num [1:3302] 0 0 0 0 0 0 0 0 0 0 ...  
## $ SMOKERE0: chr [1:3302] "(1) No" "(2) Yes" "(1) No" "(2) Yes" ...  
## $ INCOME0 : chr [1:3302] "(2) $20,000 to $49,999" "(3) $50,000 to $99,999" "(3) $50,000 to $99,999" "(3) $50,000 to $99,999" ...  
## $ PULSE0 : num [1:3302] 36 38 36 32 40 41 33 30 35 31 ...  
## $ SYSBP10 : num [1:3302] 114 120 92 108 98 120 82 88 118 120 ...  
## $ DIABP10 : num [1:3302] 80 58 60 70 72 80 64 62 80 72 ...  
## $ HEIGHT0 : num [1:3302] 151 156 162 167 164 ...  
## $ WEIGHT0 : num [1:3302] 49.5 67.7 54.4 88.9 77.2 ...  
## $ HDLRESU0: num [1:3302] 40 57 76 44 45 51 76 65 41 87 ...  
## $ GLUCRES0: num [1:3302] 102 100 88 114 93 88 81 92 90 86 ...  
## $ INSURES0: num [1:3302] 7.2 13.7 4.3 26.8 11.3 14.7 5.5 9.6 46.1 8.8 ...  
## $ LDLRESU0: num [1:3302] 73 136 85 136 151 142 109 148 149 137 ...  
## $ TRIGRES0: num [1:3302] 122 138 75 85 57 233 75 96 157 71 ...  
## $ RACE : chr [1:3302] "(5) Hispanic" "(2) Chinese/Chinese American" "(4) Caucasian/White Non-Hispanic" "(4) Caucasian/White Non-Hispanic" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ...1 = col\_double(),  
## .. SWANID = col\_double(),  
## .. AGE0 = col\_double(),  
## .. HBCHOLE0 = col\_character(),  
## .. MIGRAIN0 = col\_character(),  
## .. ANEMIA0 = col\_character(),  
## .. LISTEN0 = col\_character(),  
## .. TAKETOM0 = col\_character(),  
## .. CONFIDE0 = col\_character(),  
## .. HELPSIC0 = col\_character(),  
## .. COMMITE0 = col\_character(),  
## .. BOTHER0 = col\_character(),  
## .. APPETIT0 = col\_character(),  
## .. BLUES0 = col\_character(),  
## .. KEEPMIN0 = col\_character(),  
## .. DEPRESS0 = col\_character(),  
## .. FAILURE0 = col\_character(),  
## .. HAPPY0 = col\_character(),  
## .. HSWRKHR0 = col\_double(),  
## .. HOSPSTA0 = col\_double(),  
## .. SMOKERE0 = col\_character(),  
## .. INCOME0 = col\_character(),  
## .. PULSE0 = col\_double(),  
## .. SYSBP10 = col\_double(),  
## .. DIABP10 = col\_double(),  
## .. HEIGHT0 = col\_double(),  
## .. WEIGHT0 = col\_double(),  
## .. HDLRESU0 = col\_double(),  
## .. GLUCRES0 = col\_double(),  
## .. INSURES0 = col\_double(),  
## .. LDLRESU0 = col\_double(),  
## .. TRIGRES0 = col\_double(),  
## .. RACE = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

##Further subsets the data to meet the requirements for the final  
milestone1\_subset <- subset(rawData, select = c(  
 SWANID,  
 AGE0,  
 ANEMIA0,  
 SMOKERE0,  
 PULSE0,  
 HEIGHT0,  
 WEIGHT0,  
 RACE)  
 )  
#view the different columns and their structures  
view(milestone1\_subset)  
str(milestone1\_subset)

## tibble [3,302 x 8] (S3: tbl\_df/tbl/data.frame)  
## $ SWANID : num [1:3302] 10005 10046 10056 10092 10126 ...  
## $ AGE0 : num [1:3302] 48 52 51 45 48 51 46 47 46 47 ...  
## $ ANEMIA0 : chr [1:3302] "(1) No" "(1) No" "(2) Yes" "(2) Yes" ...  
## $ SMOKERE0: chr [1:3302] "(1) No" "(2) Yes" "(1) No" "(2) Yes" ...  
## $ PULSE0 : num [1:3302] 36 38 36 32 40 41 33 30 35 31 ...  
## $ HEIGHT0 : num [1:3302] 151 156 162 167 164 ...  
## $ WEIGHT0 : num [1:3302] 49.5 67.7 54.4 88.9 77.2 ...  
## $ RACE : chr [1:3302] "(5) Hispanic" "(2) Chinese/Chinese American" "(4) Caucasian/White Non-Hispanic" "(4) Caucasian/White Non-Hispanic" ...

Clean up the categorical data for data analysis and presentation.

milestone1\_subset <- mutate(milestone1\_subset, ANEMIA0=ifelse(ANEMIA0=="(1) No",  
 "No", ifelse(ANEMIA0=="(2) Yes", "Yes", NA)),  
 SMOKERE0=ifelse(SMOKERE0=="(1) No", "No", ifelse(SMOKERE0=="(2) Yes", "Yes", NA)),  
 RACE=ifelse(RACE=="(1) Black/African American", "Black/African American",  
 ifelse(RACE=="(2) Chinese/Chinese American",   
 "Chinese/Chinese American",  
 ifelse(RACE=="(3) Japanese/Japanese American",  
 "Japanese/Japanese American",  
 ifelse(RACE=="(4) Caucasian/White Non-Hispanic",   
 "Caucasian/ White Non-Hispanic",   
 ifelse(RACE=="(5) Hispanic", "Hispanic", NA)  
 )))))

# Descriptive Statistics

Calculating various descriptive statistics of each of the variables.

summary(milestone1\_subset)

## SWANID AGE0 ANEMIA0 SMOKERE0   
## Min. :10005 Min. :42.00 Length:3302 Length:3302   
## 1st Qu.:31808 1st Qu.:44.00 Class :character Class :character   
## Median :54230 Median :46.00 Mode :character Mode :character   
## Mean :54362 Mean :45.85   
## 3rd Qu.:76745 3rd Qu.:48.00   
## Max. :99992 Max. :53.00   
## NA's :5   
## PULSE0 HEIGHT0 WEIGHT0 RACE   
## Min. :17.00 Min. :140.5 Min. : 37.60 Length:3302   
## 1st Qu.:32.00 1st Qu.:157.8 1st Qu.: 59.60 Class :character   
## Median :35.00 Median :162.4 Median : 70.60 Mode :character   
## Mean :35.19 Mean :162.4 Mean : 74.88   
## 3rd Qu.:38.00 3rd Qu.:167.0 3rd Qu.: 85.50   
## Max. :84.00 Max. :186.2 Max. :175.40   
## NA's :7 NA's :32 NA's :14

age\_sd <- sd(milestone1\_subset$AGE0, na.rm = TRUE)  
age\_mean <- mean(milestone1\_subset$AGE0, na.rm = TRUE)  
pulse\_sd <- sd(milestone1\_subset$PULSE0, na.rm = TRUE)  
pulse\_mean <- mean(milestone1\_subset$PULSE0, na.rm = TRUE)  
height\_sd <- sd(milestone1\_subset$HEIGHT0, na.rm = TRUE)  
height\_mean <- mean(milestone1\_subset$HEIGHT0, na.rm=TRUE)  
weight\_sd <- sd(milestone1\_subset$WEIGHT0, na.rm = TRUE)  
weight\_mean <- mean(milestone1\_subset$WEIGHT0, na.rm=TRUE)  
age\_med <- median(milestone1\_subset$AGE0, na.rm=TRUE)  
pulse\_med <- median(milestone1\_subset$PULSE0, na.rm=TRUE)  
height\_med <- median(milestone1\_subset$HEIGHT0, na.rm=TRUE)  
weight\_med <- median(milestone1\_subset$WEIGHT0, na.rm=TRUE)

Create a csv file to use in the report for descriptive statistics.

clmn\_names <- c("Age", "Pulse", "Height", "Weight")  
sds <- c(age\_sd, pulse\_sd, height\_sd, weight\_sd)  
means <- c(age\_mean, pulse\_mean, height\_mean, weight\_mean)  
medians <- c(age\_med, pulse\_med, height\_med, weight\_med)  
desc\_table <- data.frame("Continuous Data"=clmn\_names, "Standard Deviation"=sds, "Mean"=means, "Medians"=medians)  
desc\_table

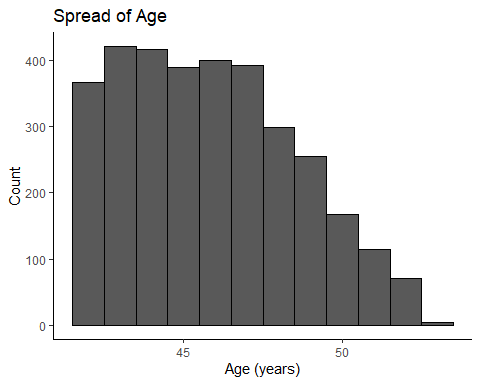
## Continuous.Data Standard.Deviation Mean Medians  
## 1 Age 2.689278 45.84956 46.0  
## 2 Pulse 4.811298 35.18725 35.0  
## 3 Height 6.741504 162.35645 162.4  
## 4 Weight 20.486960 74.88190 70.6

#write the table into a format that can be copied over into the report  
write.csv(desc\_table, "C:/Users/12072/OneDrive/Desktop/ALY 6010/Final Project\\Milestone1\_Descriptive.csv")

# Data Visualization

Understand the spread of the numerical data.

age\_hist <- ggplot(milestone1\_subset)+  
 geom\_histogram(mapping=aes(AGE0), na.rm = TRUE, binwidth = 1, color="black")+  
 theme\_classic()+  
 labs(title="Spread of Age", x="Age (years)", y="Count")  
age\_hist

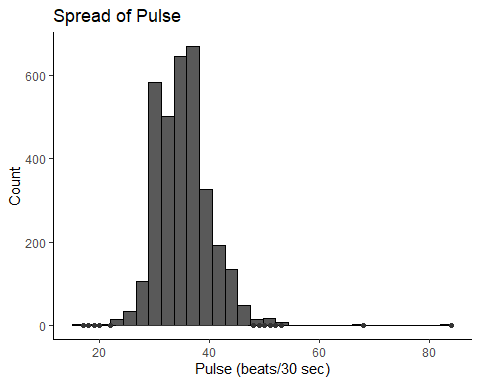


##put a box plot in to highlight the outliers in this chart  
pulse\_hist <- ggplot(milestone1\_subset,mapping=aes(PULSE0), na.rm = TRUE)+  
 geom\_histogram(color="black")+  
 geom\_boxplot()+  
 theme\_classic()+  
 labs(title="Spread of Pulse", x="Pulse (beats/30 sec)", y="Count")  
pulse\_hist

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

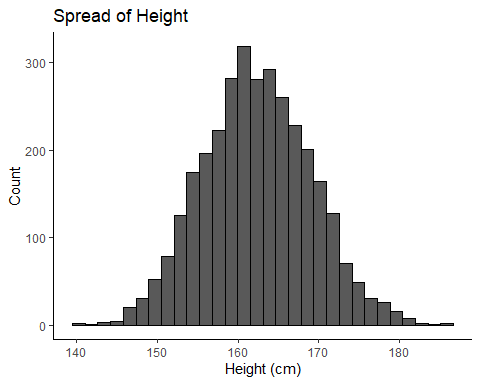
## Warning: Removed 7 rows containing non-finite values (stat\_bin).

## Warning: Removed 7 rows containing non-finite values (stat\_boxplot).



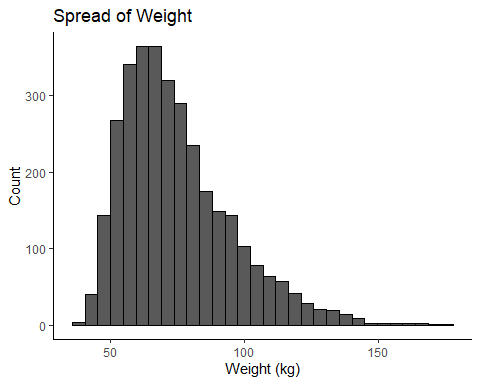
height\_hist <- ggplot(milestone1\_subset)+  
 geom\_histogram(mapping=aes(HEIGHT0), na.rm = TRUE, color="black")+  
 theme\_classic()+  
 labs(title="Spread of Height", x="Height (cm)", y="Count")  
height\_hist

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



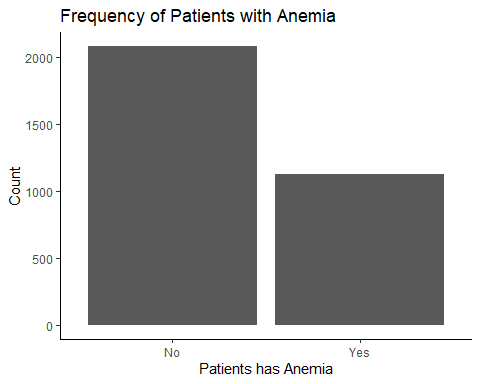
weight\_hist <- ggplot(milestone1\_subset)+  
 geom\_histogram(mapping=aes(WEIGHT0), na.rm = TRUE, color="black")+  
 theme\_classic()+  
 labs(title="Spread of Weight", x="Weight (kg)", y="Count")  
weight\_hist

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

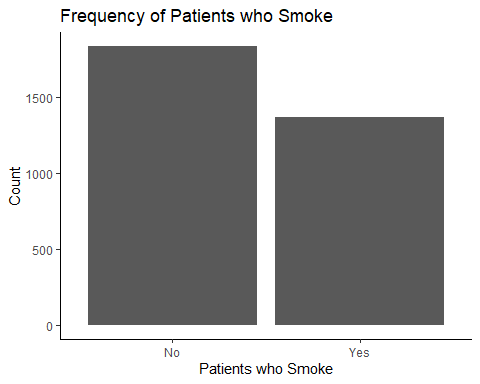


To understand the spread of the categorical data.

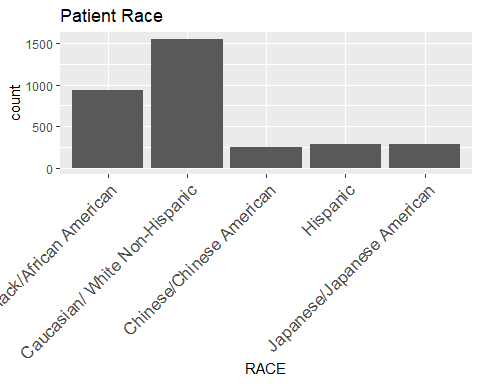
##NA's were not being removed with na.rm=TRUE, therefore they were removed before calling the ggplot functions  
anem\_freq <- milestone1\_subset %>% drop\_na() %>%   
 ggplot()+  
 geom\_bar(mapping=aes(ANEMIA0))+  
 theme\_classic()+  
 labs(title= "Frequency of Patients with Anemia", y= "Count",   
 x="Patients has Anemia")  
anem\_freq



smoke\_freq <- milestone1\_subset %>% drop\_na() %>%   
 ggplot()+  
 geom\_bar(mapping=aes(SMOKERE0))+  
 theme\_classic()+  
 labs(title= "Frequency of Patients who Smoke", y= "Count",   
 x="Patients who Smoke")  
smoke\_freq

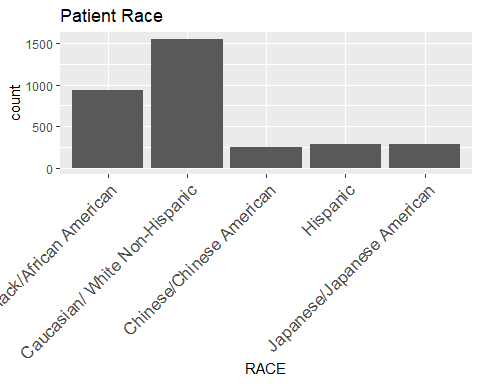


##Analysis of RACE  
ggplot(milestone1\_subset)+  
 geom\_bar(mapping=aes(RACE))+  
 labs(title="Patient Race")+  
 theme(axis.text.x=element\_text(size=13, angle=45, hjust=1, vjust=1))



A separate spread analysis for race was necessary because it has multiple possible results.

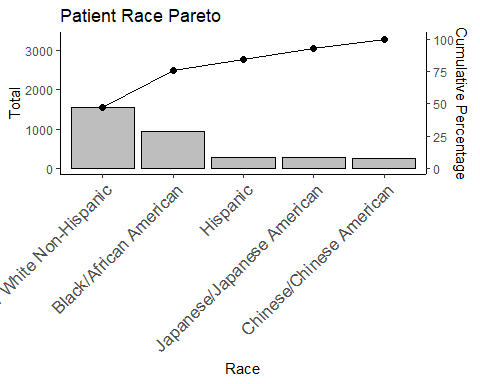
ggplot(milestone1\_subset)+  
 geom\_bar(mapping=aes(RACE))+  
 labs(title="Patient Race")+  
 theme(axis.text.x=element\_text(size=13, angle=45, hjust=1, vjust=1))



##creating frequency and pareto charts to plot the race data  
##Creates a tibble with the total counts for each race identified  
race <- milestone1\_subset %>% group\_by(RACE) %>% summarise(Total=n())  
##Calculates the frequency of each race  
race <- mutate(race, Frequency=Total/sum(Total))  
race

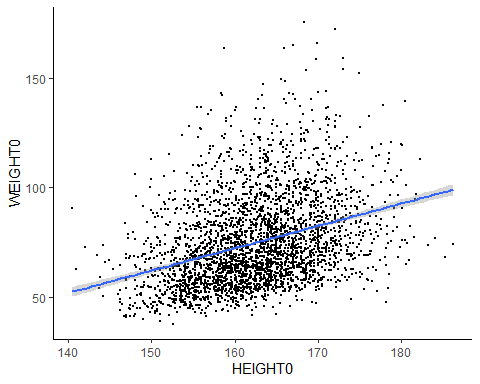
## # A tibble: 5 x 3  
## RACE Total Frequency  
## <chr> <int> <dbl>  
## 1 Black/African American 934 0.283   
## 2 Caucasian/ White Non-Hispanic 1552 0.470   
## 3 Chinese/Chinese American 250 0.0757  
## 4 Hispanic 285 0.0863  
## 5 Japanese/Japanese American 281 0.0851

race\_pareto <- ggplot(race, mapping=aes(x=RACE, y=Total))+  
 stat\_pareto(bars.fill="gray")+  
 labs(title="Patient Race Pareto", x="Race")+  
 theme\_classic()+  
 theme(axis.text.x=element\_text(size=13, angle=45, hjust=1, vjust=1))  
race\_pareto



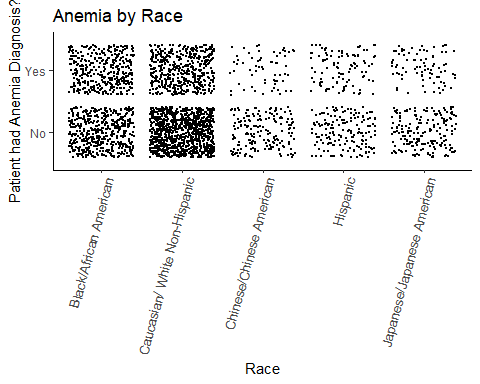
Height and weight relationship using a scatterplot.

hvw\_scatter <- ggplot(milestone1\_subset)+  
 geom\_point(mapping=aes(HEIGHT0, WEIGHT0), na.rm=TRUE, size=0.5)+  
 geom\_smooth(mapping=aes(HEIGHT0, WEIGHT0), method=lm, formula=y~x, na.rm=TRUE)+  
 theme\_classic()  
hvw\_scatter



Is there a relationship between race and anemia?

avr\_jitter <- milestone1\_subset %>% drop\_na() %>% ggplot()+  
 geom\_jitter(mapping=aes(RACE, ANEMIA0), na.rm=TRUE, size=0.5)+  
 labs(title="Anemia by Race", x="Race", y="Patient had Anemia Diagnosis?")+  
 theme\_classic()+  
 theme(axis.text.x=element\_text(size=10, angle=75, hjust=1, vjust=1))  
avr\_jitter



Is there a relationship between smoking and measured pulse?

pulse\_density <- milestone1\_subset %>% drop\_na() %>% ggplot()+  
 geom\_density(mapping=aes(PULSE0, color=SMOKERE0), size=1)+  
 theme\_classic()+  
 labs(title="Density of Pulse by Smoking Status", x= "Pulse (beats/30 sec)", y="Density")  
pulse\_density

