Milestone 1

Angel Waters

06/05/2022

Introduction

A study was conducted to understand various aspects of women's health during their middle age for health care professionals grasp female quality of life during this life stages (Sutton-Tyrell *et al.* 1997). The raw data was taken, and a subset was created for the analysis of this report. The specific subset pertains to the height, weight, and blood pressure of women of ages ranging from 42 to 53. Additionally, their status of anemia diagnosis and if they had ever smoked.

The questions trying to be answered are the following: How does smoking effect the blood pressure of women at this life stage? What is the relationship between anemia and race, are there races more susceptible to the disease? The purpose of this analysis is to familiarize with the data.

Key Findings

Data Cleaning

To make the data more manageable, a subset was taken of the full data set to capture the 7 fields necessary for this analysis (with additional patient specific identifier fields). The columns were then mutated to format the attributes for data presentation. See *Appendix* for details.

Descriptive Statistics

Descriptive statistics were calculated and tabulated in **Table 1**. The medians and means were very similar to each other, except weight, and all were within 1 standard deviation from each other.

Table 1. Descriptive table of the numeric values in the data set.

Continuous Data	Standard Deviation	Mean	Medians
Age (years)	2.689278	45.84956	46
Pulse (beats/30 sec)	4.811298	35.18725	35
Height (cm)	6.741504	162.3564	162.4
Weight (kg)	20.48696	74.8819	70.6

Data Visualization

The spread of the numerical data was plotted by histograms (see **Fig 1** through **Fig 4**). The spread for pulse and height were normally distributed; weight is positively skewed. Age distribution appears to be uniform with a slight positive skew. Categorical spread show there was an uneven collection for patients who had anemia or patients who smoked compared to those who didn't (see **Fig 5** and **Fig 6**). Because race had more than two results, a pareto was used to plot the frequency (see **Fig 7**). The largest groups sampled were Caucasian/White Non-Hispanic and Black/African American.

A jitter plot for the patient's anemia status by race was plotted. Despite having different sample sizes, each group had more patients with no history of anemia (see **Fig 8**). Density plot of pulse by smoking status shows similar curve shapes comparing to each other (see **Fig 9**). The height of the peaks is as expected because the sample size is smaller for the confirmed smoking population.

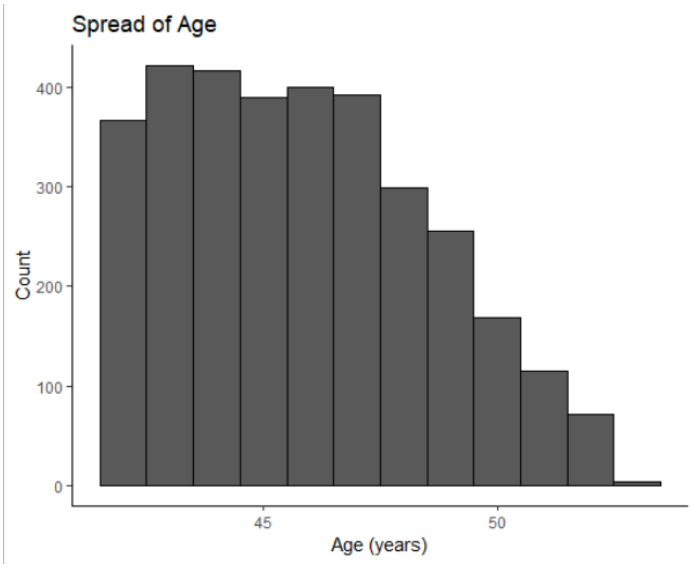


Figure 1. Histogram of patient age.

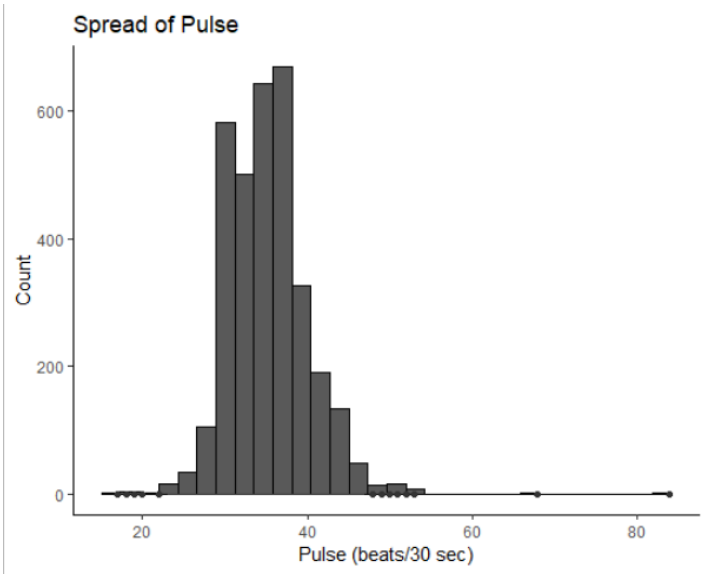


Figure 2. Histogram of patient pulse with an overlay of the outliers using box plot.

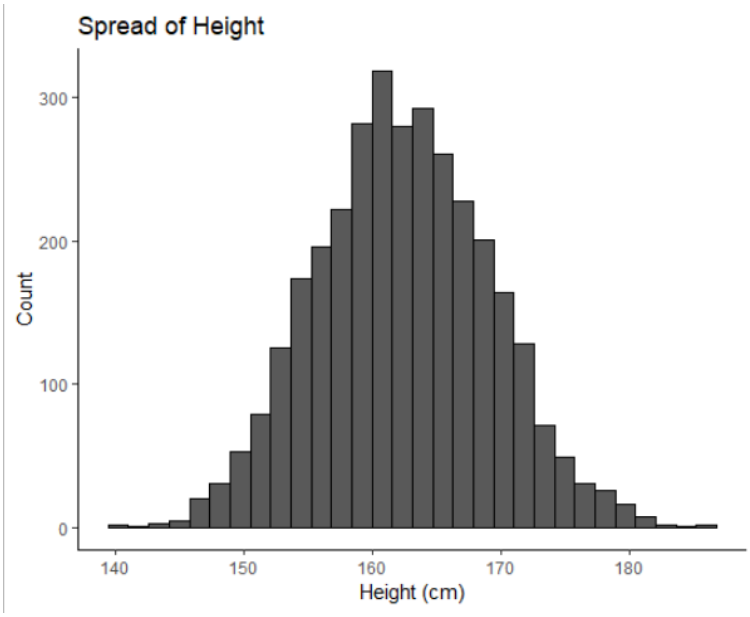


Figure 3. Histogram of patient height.

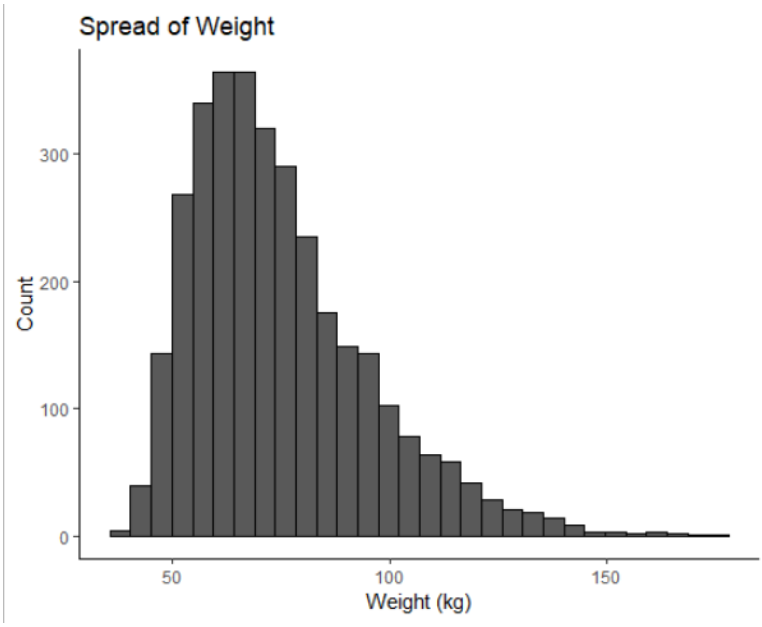


Figure 4. Histogram of patient weight.

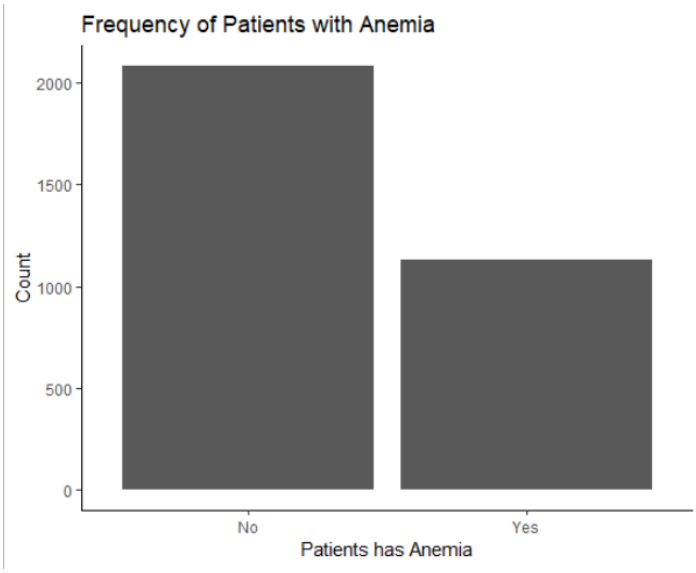


Figure 5. Frequency of patient anemia results.

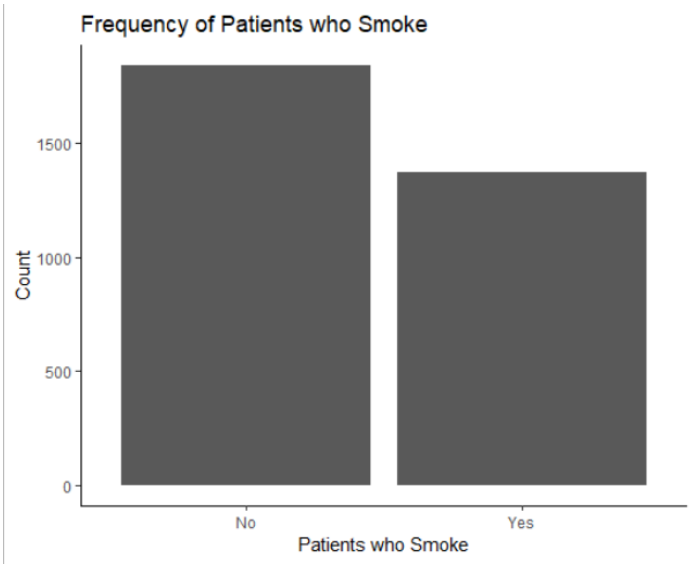


Figure 6. Frequency of patient smoke history results.

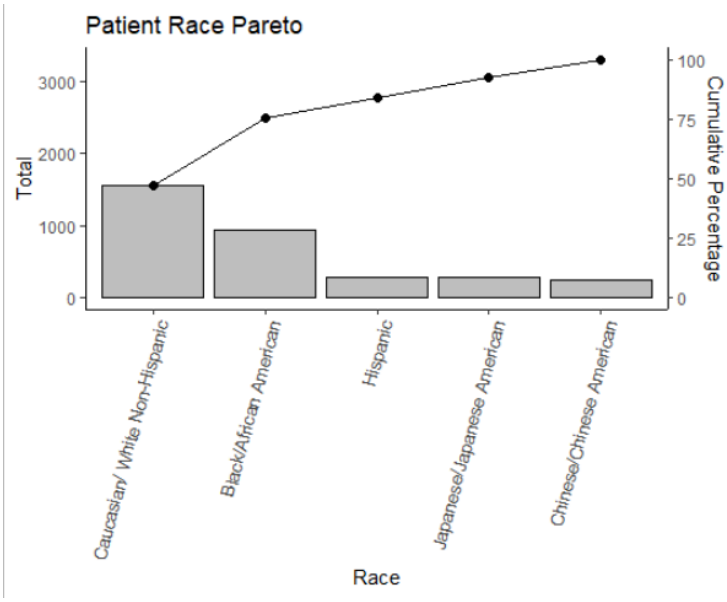


Figure 7. Pareto of race.

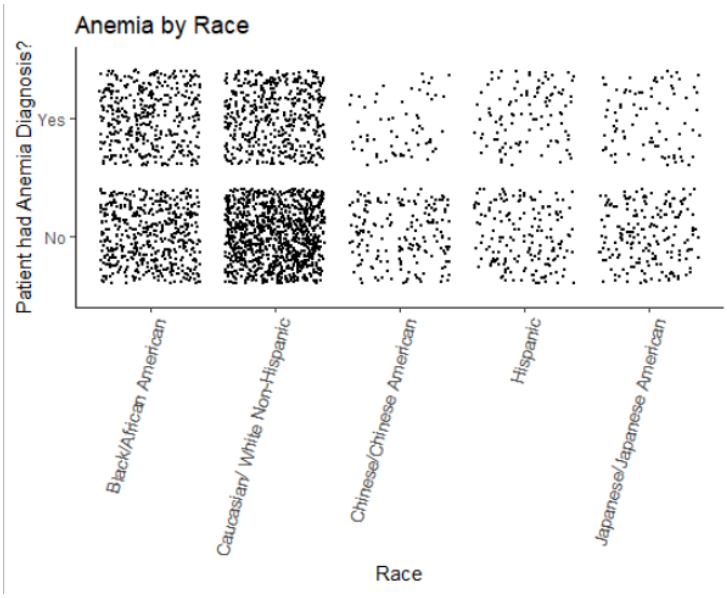


Figure 8. Density of a patient's anemia status by Race.

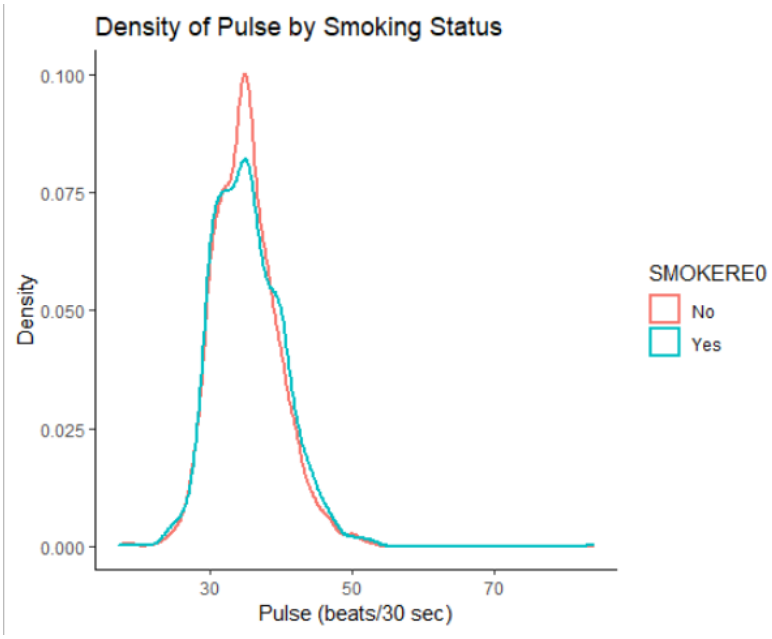


Figure 9. Density plot of pulse by smoking status.

Conclusion

The study surveyed a total of 3302 women for this subset. The top two race identified were Caucasian/White Non-Hispanic and Black/African American (respectively). Smokers and women with anemia in their medical history were fewer than those without. Height and Pulse were evenly distributed, the other two numerical fields were positively skewed, with Weight being more skewed than Age. Even though each race has a different sample size, they all did not have any trend to having anemia, further analysis may be necessary to compare proportions. Pulse showed there was no difference between those who smoked and those who did not.

Bibliography

Sutton-Tyrrell, Kim, Selzer, Faith, Sowers, MaryFran, R. (Mary Frances Roy), Neer, Robert, Powell, Lynda, Gold, Ellen B., … McKinlay, Sonja. Study of Women's Health Across the Nation (SWAN): Baseline Dataset, [United States]. (1997). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-05-15. https://doi.org/10.3886/ICPSR28762.v5

Appendix

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_conflict
s() --
## 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 dat
a.
rawData <-
  read_csv("SWANBaselineData_ProfessorKSubset (1).csv")

## New names:
## Rows: 3302 Columns: 33
## -- Column specification
## ------------------------------------------------------- Delimiter: "," ch
r
## (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 mes
sage.
## * `` -> `...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) Rarel
y/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Ra
rely/None Of The Time (< 1 Day)" ...
##  $ BLUES0  : chr [1:3302] "(4) Most/All Of The Time (5-7 Days)" "(1) Rarel
y/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Ra
rely/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 Ti
me (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) Rarel
y/None Of The Time (< 1 Day)" "(1) Rarely/None Of The Time (< 1 Day)" "(1) Ra
rely/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 D
ay)" "(1) Rarely/None Of The Time (< 1 Day)" ...
##  $ HAPPY0  : chr [1:3302] "(1) Rarely/None Of The Time (< 1 Day)" "(3) Occ
asionally/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) N
o",
                                                              "No", ifelse(AN
EMIA0=="(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 America
n",
                   ifelse(RACE=="(2) Chinese/Chinese American",
                          "Chinese/Chinese American",
                          ifelse(RACE=="(3) Japanese/Japanese American",
                                 "Japanese/Japanese American",
                                 ifelse(RACE=="(4) Caucasian/White Non-Hispan
ic",
                                        "Caucasian/ White Non-Hispanic",
                                        ifelse(RACE=="(5) Hispanic", "Hispani
c", 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"=s
ds, "Mean"=means, "Medians"=medians)
desc_table
```

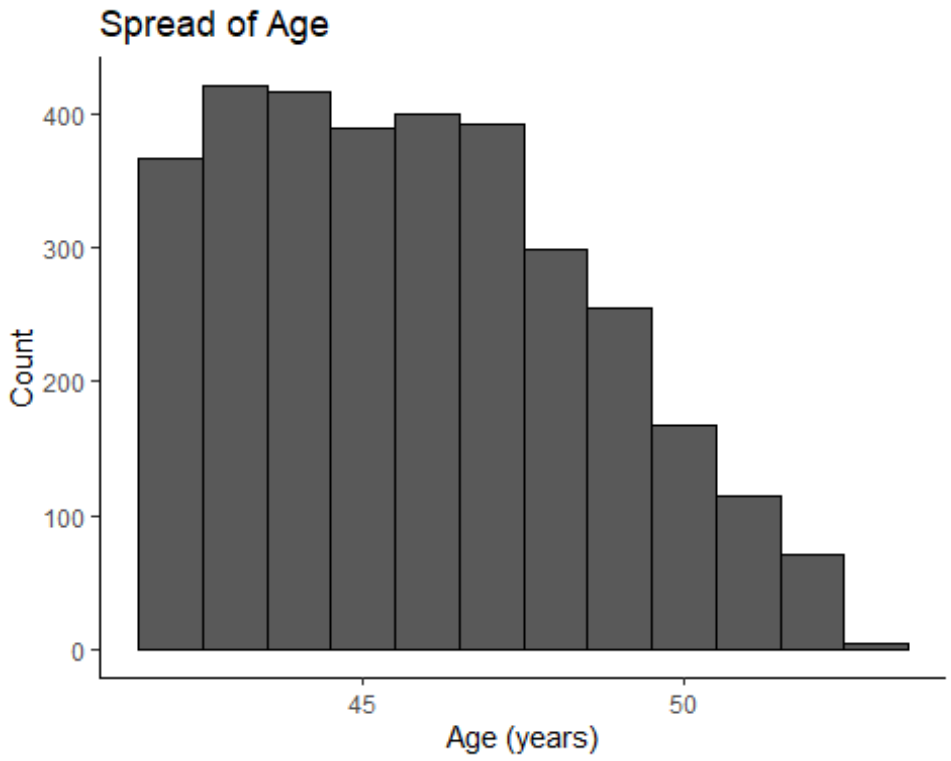
```
##   Continuous.Data Standard.Deviation      Mean Medians
## 1             Age           2.689278  45.84956    46.0
## 2           Pulse           4.811298  35.18725    35.0
## 3          Height           6.741504 162.35645   162.4
## 4          Weight          20.486960  74.88190    70.6
```

```
#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")
```

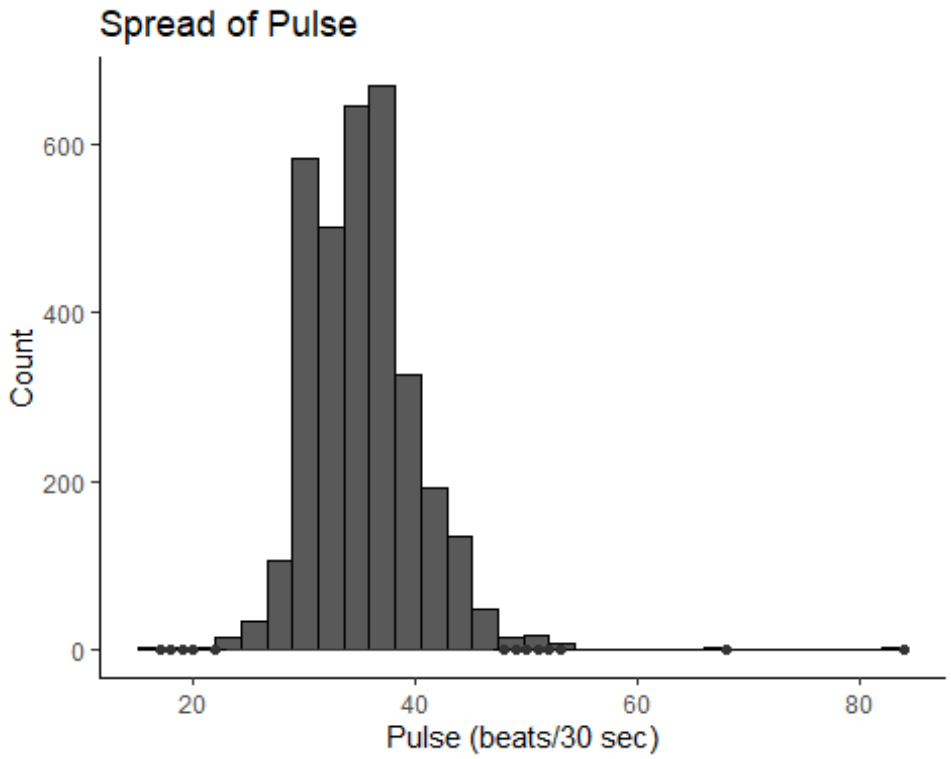
Data Visualization

Understand the spread of the numerical data.

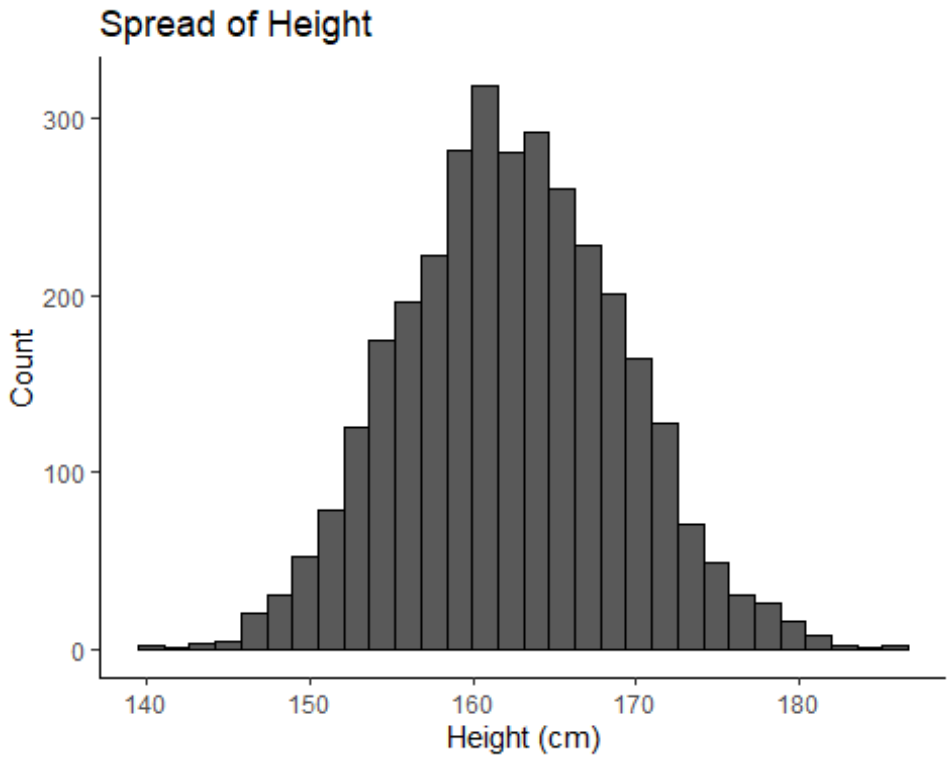
```
age_hist <- ggplot(milestone1_subset)+
  geom_histogram(mapping=aes(AGE0), na.rm = TRUE, binwidth = 1, color="black"
)+
  theme_classic()+
  labs(title="Spread of Age", x="Age (years)", y="Count")
age_hist
```



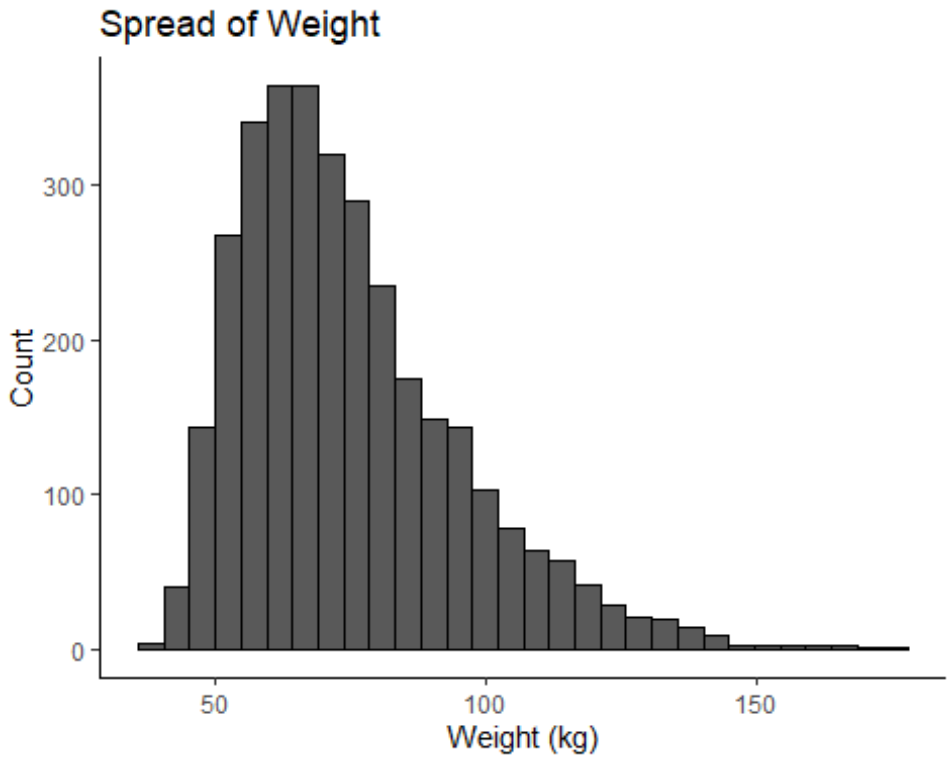

```
##put a box plot in to highlight the outliers in this chart
pulse_hist <- ggplot(milestone1_subset,mapping=aes(PULSE0), na.rm = TRUE)+
  geom_histogram(color="black")+
  geom_boxplot()+
  theme_classic()+
  labs(title="Spread of Pulse", x="Pulse (beats/30 sec)", y="Count")
pulse_hist

## `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`.
```

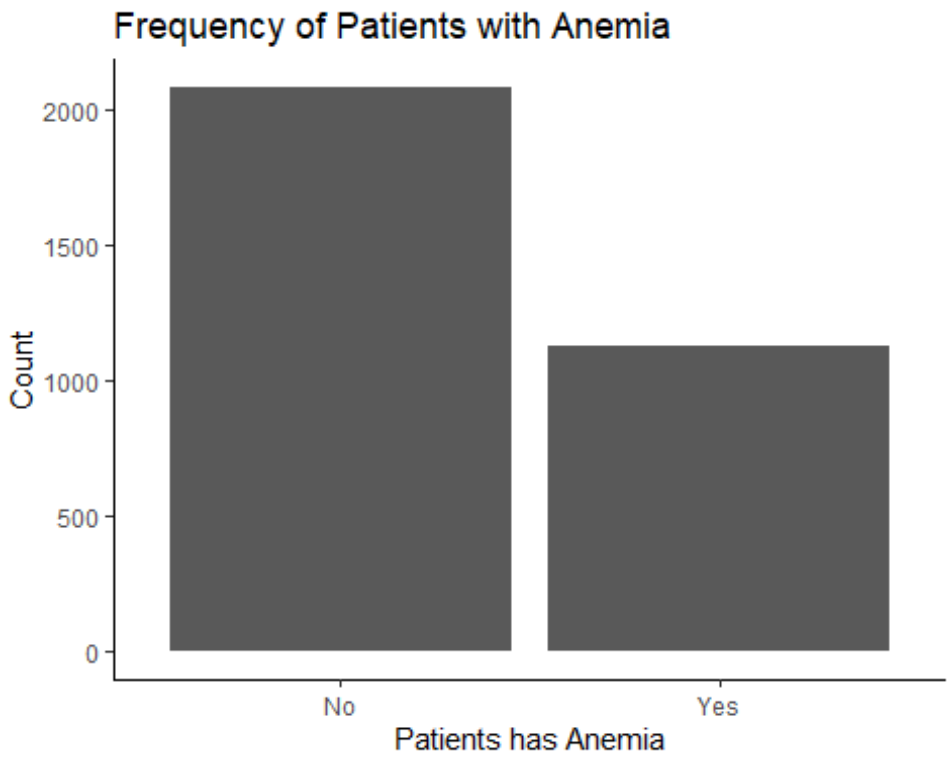



```
weight_hist <- ggplot(milestone1_subset)+
  geom_histogram(mapping=aes(WEIGHT0), na.rm = TRUE, color="black")+
  theme_classic()+
  labs(title="Spread of Weight", x="Weight (kg)", y="Count")
weight_hist

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

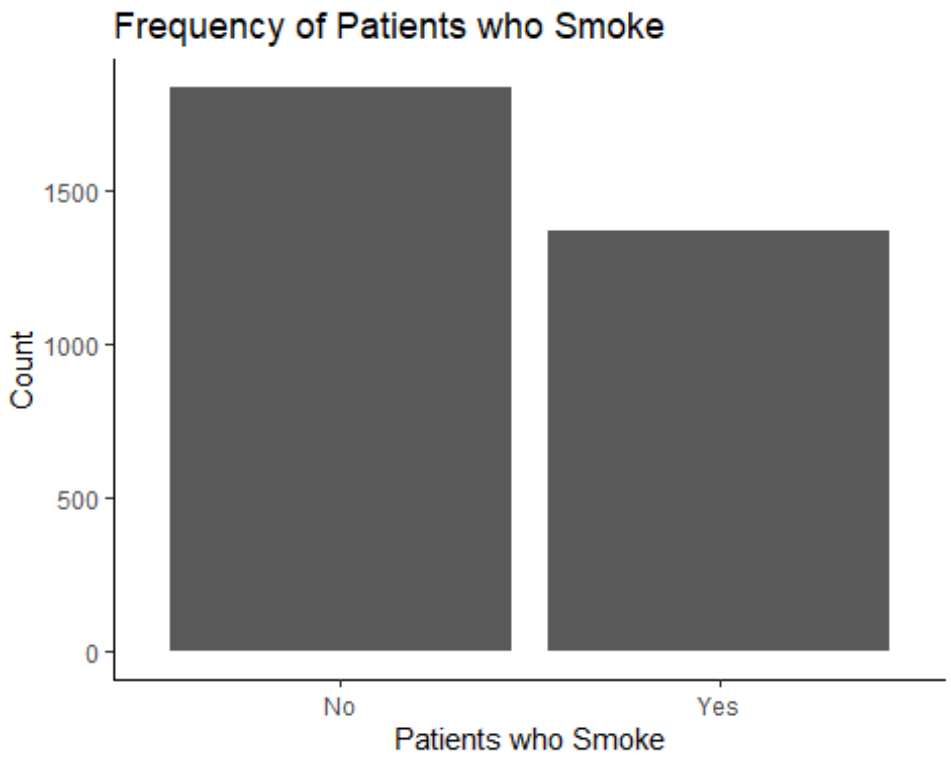


To understand the spread of the categorical data.

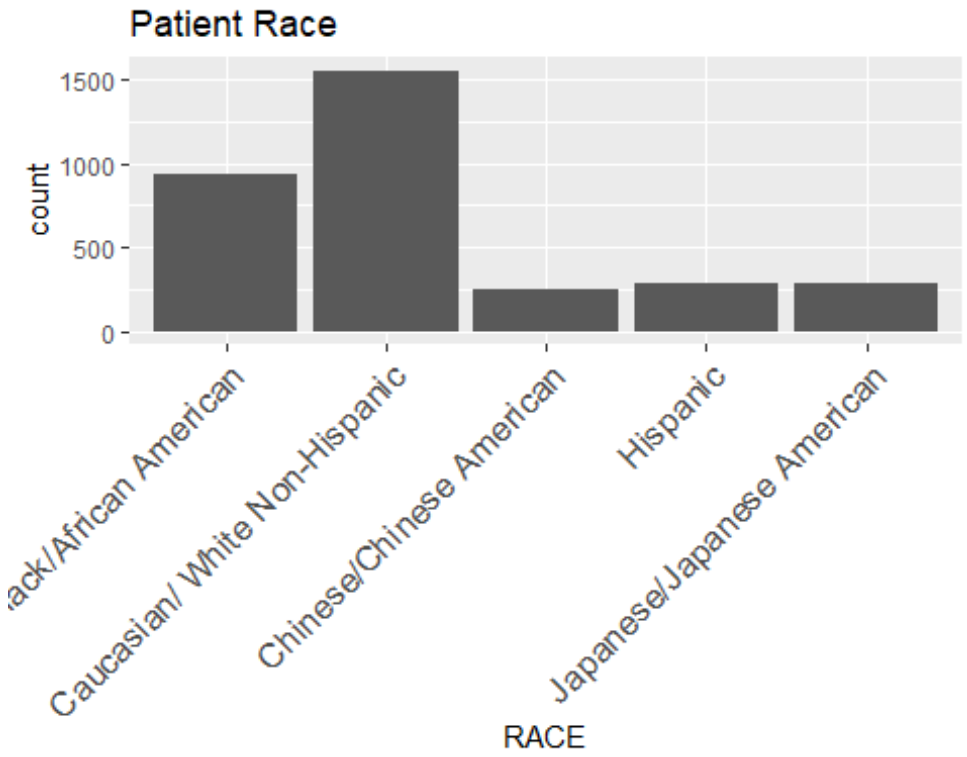
```
##NA's were not being removed with na.rm=TRUE, therefore they were removed be
fore calling the ggplot functions
anem_freq <- milestone1_subset %>% drop_na() %>%
  ggplot()+
  geom_bar(mapping=aes(ANEMIA0))+
  theme_classic()+
  labs(title= "Frequency of Patients with Anemia", y= "Count",
       x="Patients 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
```

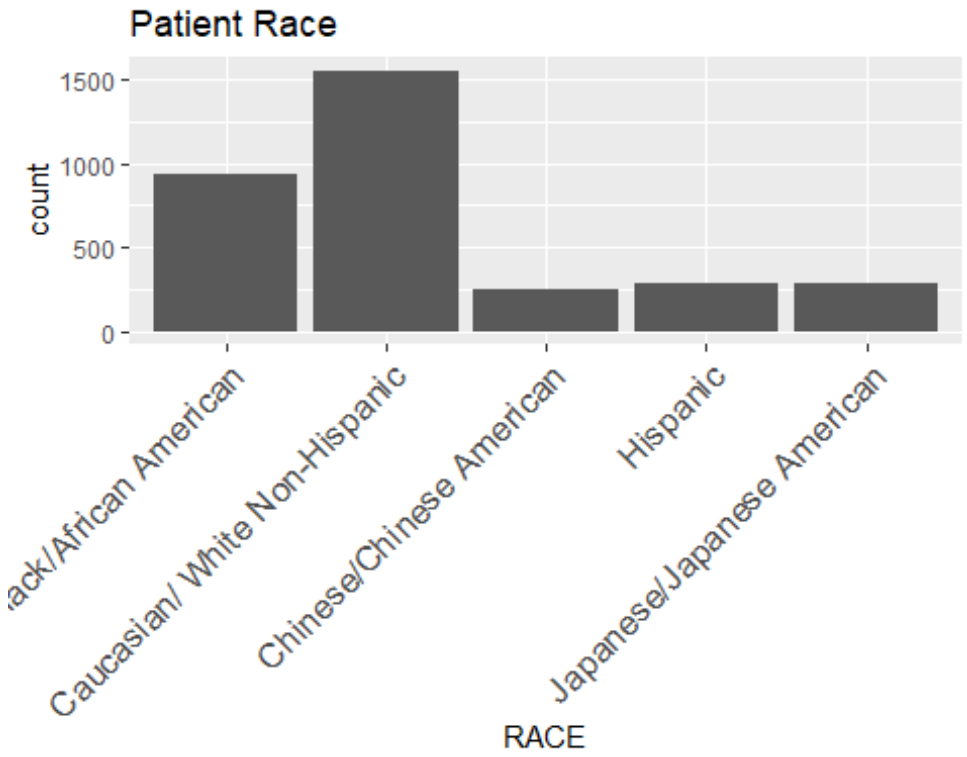



```
##Analysis of RACE
ggplot(milestone1_subset)+
  geom_bar(mapping=aes(RACE))+
  labs(title="Patient Race")+
  theme(axis.text.x=element_text(size=13, angle=45, hjust=1, vjust=1))
```



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

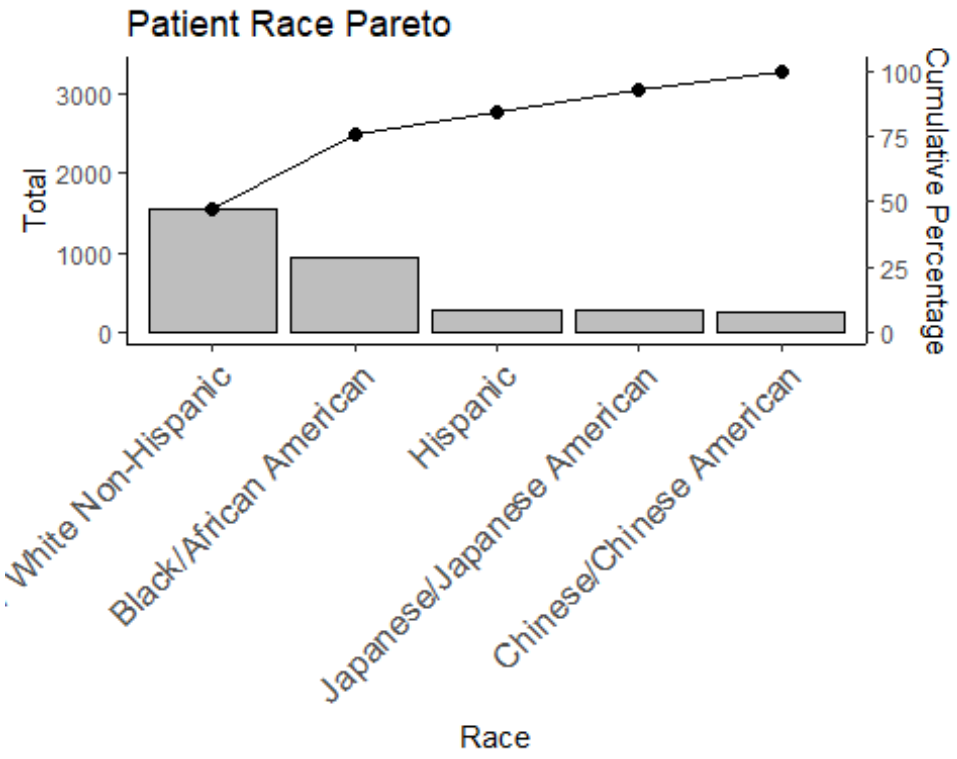
```
ggplot(milestone1_subset)+
  geom_bar(mapping=aes(RACE))+
  labs(title="Patient Race")+
  theme(axis.text.x=element_text(size=13, angle=45, hjust=1, vjust=1))
```




```
##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
```

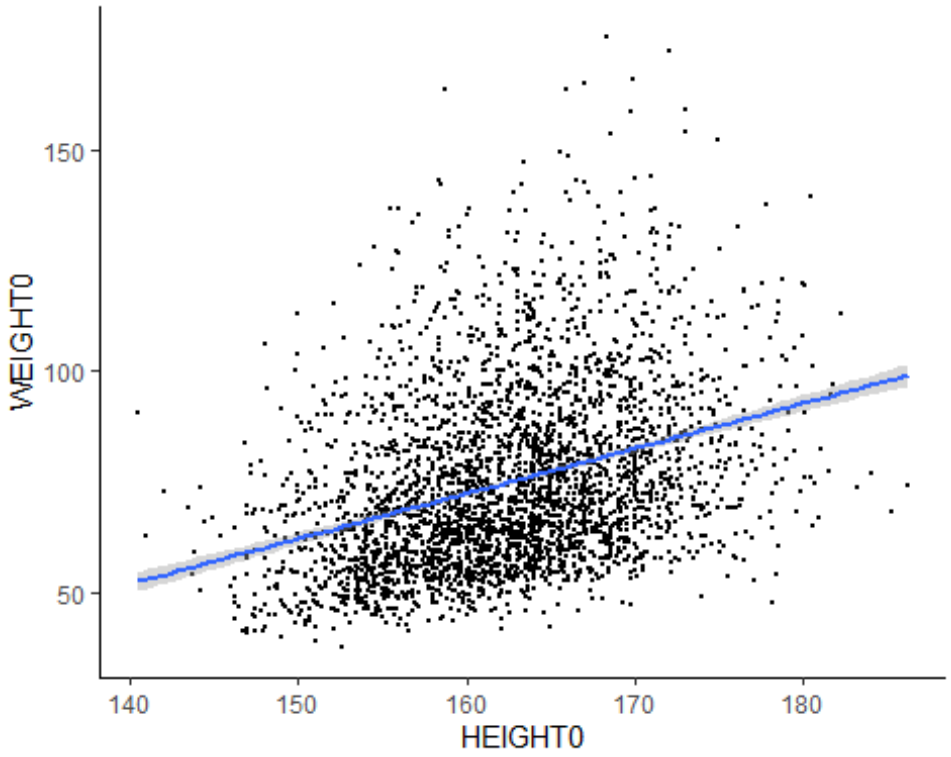
```
## # A tibble: 5 x 3
##   RACE                          Total Frequency
##   <chr>                         <int>     <dbl>
## 1 Black/African American          934    0.283
## 2 Caucasian/ White Non-Hispanic  1552    0.470
## 3 Chinese/Chinese American        250    0.0757
## 4 Hispanic         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



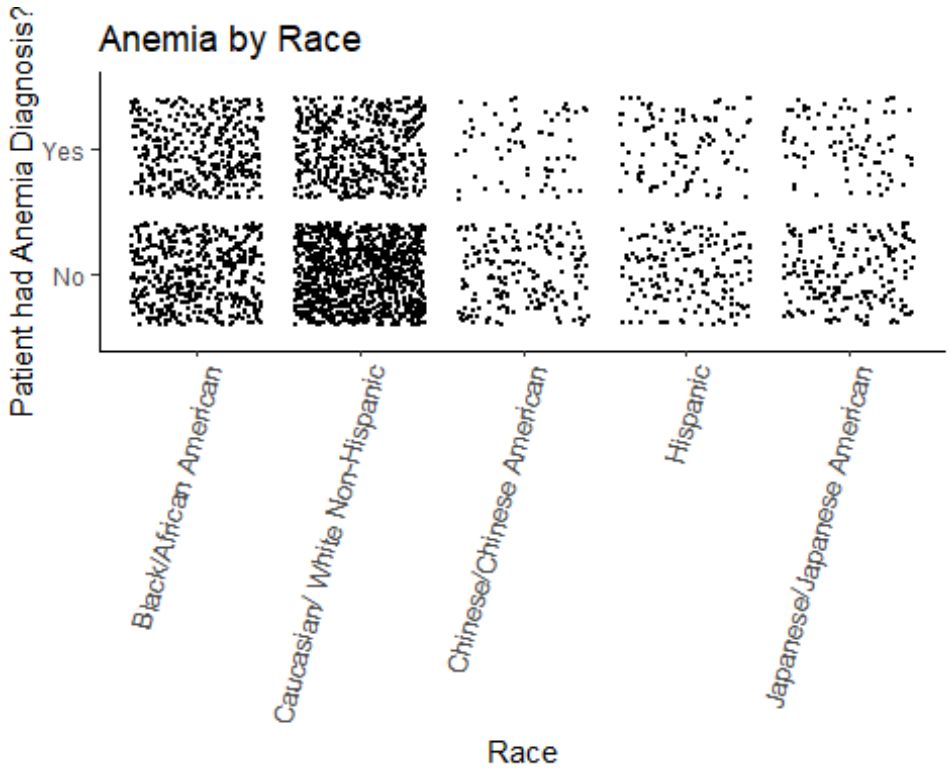
Height and weight relationship using a scatterplot.

```
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=TR
UE)+
  theme_classic()
hvw_scatter
```



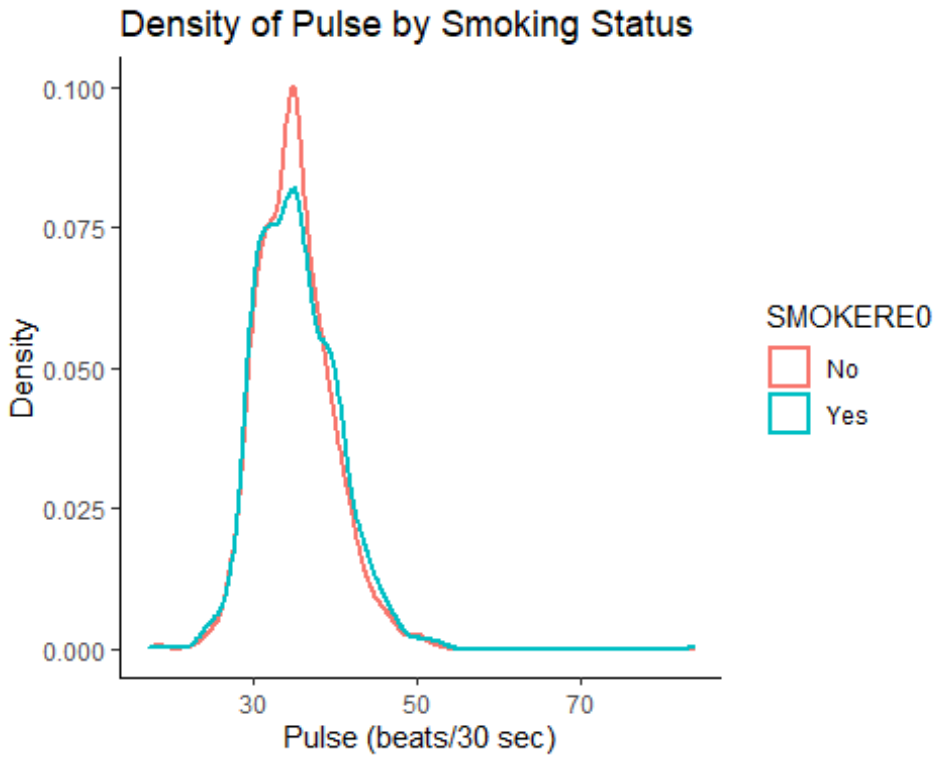
Is there a relationship between race and anemia?

```
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
```


Density of Pulse by Smoking Status
0.100
0.075
0.050
0.025
0.000
Density
30
50
70
Pulse (beats/30 sec)
SMOKERE0
No
Yes