Final Project

Angel Waters

2022-06-30

Introduction

A subset of the SWAN data set was analyzed over the past six weeks to learn of any statistically significant relationships between the variables within the data set. This data represents women in their middle years and will be utilized in an attempt to better address the health needs of women during those important years (43-52 years is being represented in this data).

Research Questions

The research questions below were formulated from the initial **Exploratory Data Analysis**. These questions were assessed using hypothesis testing to understand if similarities or difference are statistically relevant. If they are, conclusions were made on both a practical and statistical level. The SWAN data set was treated as a population for the majority of the analyses unless otherwise stated. 1. Do women with anemia have a different pulse than those without an anemia diagnosis? 2. Do the proportion of women who smoke at age 45 the same as the proportion of all women who smoke in the SWAN data set? 3. Is there a positive linear relationship between Height and Weight? 4. Can Diastolic Blood Pressure be used to predict Systolic Blood Pressure? 5. Is there a significant linear relationship between weight and blood pressure?

Exploratory Data Analysis

Data Loading and Clean Up

A subset had been created by Professor Dan Koloski to have some selected variables of importance. This was further subsetted into a smaller data set that was cleaned up and analyzed to answer the specific questions outlined above (see **Research Questions**).

The cleaned up data was loaded into R and subsetted for relevent variables. These variables were updated to represent logical or adjust the case for character and categorical variables.

Additional Variables

Racial subdivisions were determined by understanding the frequency of the races identified. If a group was identified to have a frequency **>** 0.200 then it was considered a majority group, the rest were considered minorities (actual frequencies outlined in **Table 1** and a Pareto plot in **Fig. 1** in **Appendix**). This was determined because there are 5 races identified, so even distribution would be 20% of the data. Support score was calculated by combining the scores of the following questions, each on a scale of 1 to 5: * Have someone who listens? * Have someone to take to the doctor? * Have someone to confide in? * Have someone to do chores when sick?

Summary Statistics

Numerical data summary statistics were calculated and combined into a data table (see **Table 2** in **Appendix**).

Data Visualizations

Continuous variables were plotted on histograms (see **Fig. 2-6** in **Appendix**). This is to show how the data is distributed for this analysis and aided in how the **Research Questions** would be asked. Pulse and Height were normally distributed; Age and Weight were positively skewed where the skew is more defined for Weight and closer to normal for Age. Support score was also presented as histogram, where it was negatively skewed.

For the categorical variables, frequencies were plotted on bar charts to show if there was even sampling for those data sets (see **Fig. 7** and **Fig. 8** in **Appendix**). Patients who smoked had more patients identified as non-smokers than those identified as smokers; patients without an anemia diagnosis had a greater frequency than those who did not.

Numerical relationships were briefly examined to be used to potentially formulate questions about the significance of their relationships. There looks to be a positive linear relationship for height vs weight, which should be explored further to make any conclusions about that relationship (see **Fig. 9** in **Appendix**).

For the two categorical variables (anemia and smoking status), their relationships were explored to understand if there was any relationship between them and other categorical variables. Anemia status was compared to race to show if there were any susceptibility trends. It appears each race has the same density shown in the density plot in **Fig. 10**. Pulse in beats/30 seconds was also reviewed against smoking status (see **Fig. 11** in **Appendix**). The density was displayed and shows there is a similar peak for those who smoke and those who do not.

Hypothesis Testing

Hypothesis testing was used to compare and contrast relationships between the data's population parameters and sample statistics. The questions assessed through this analysis are stated in each section and the statistical conclusions can be viewed at the end of each section. All data sets assume the samples are randomly selected and the data is normally distributed.

Do women with an anemia diagnosis have a different average pulse than those who do not have an anemia diagnosis?

The claim for this question is that there is a statistical difference of pulse between those who were diagnosed with anemia and those who were not. Because of the wording of this claim the null hypothesis would be that the mean pulse of those with anemia is equal to the mean pulse of those without anemia; the alternative is that they are not equal to each other. The claim was tested at an alpha = 0.05 using a two-tailed hypothesis method. A sample of 100 for each anemia status from the SWAN data set was analyzed.

```
##
##  Welch Two Sample t-test
##
## data:  with_samp$PULSE0 and without_samp$PULSE0
## t = -0.25717, df = 197.13, p-value = 0.7973
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.473604  1.133604
## sample estimates:
## mean of x mean of y
##     34.36     34.53

##          t
## -0.2571732

## [1] "Do not reject Null Hypothesis"

## There is not enough evidence to support the claim: Women with anemia have
a different average pulse than women without it
```

The calculated t statistic was within the non-critical region of the t curve (see **Fig. 12** in **Appendix**). This resulted in a fail to reject the null hypothesis. *There is not enough evidence to support the claim that the pulse of women with an anemia diagnosis is different from the pulse women without an anemia diagnosis.*

Is the proportion of women who smoke at age 45 years the same as the proportion of all women who smoke?

Women at age 45 were believed to have the same proportion of women who smoke as the entire SWAN data set. To test this claim, all patients aged 45 were subsetted and calculated the proportion of women who smoked. That was compared to all women who smoked in the SWAN data set comparing their proportions. This was tested at an alpha of 0.05. The null hypothesis that aligns with the claim is the two proportions are equal, and the alternative is they are not equal.

```
## Null: p = 43 %

## Alternative: p neq 43 %

## z =  1.111843

## [1] "Do not reject Null Hypothesis"

## There is enough evidence to support the claim: The proportion of smokers a
t age 45 is equal to the proportion of smokers in the SWAN dataset
```

The z statistic falls in the non-critical regions of the z plot (see **Fig. 13** in **Appendix**).Based on the z score at an alpha of 0.05, the null hypothesis was not rejected. Because the claim aligns with the null hypothesis, *there was enough evidence to support the claim that the two proportions are statistically equivalent.*

Regression Analysis

For the regression analysis, numerical data was plotted on a Scatterplot Matrix to understand if there were any linear relationships that could be visualized in the data set (see **Fig. 14** in **Appendix**). Questions were formulated in an attempt to understand the relationships in continuous numerical data.

Is there a significant relationship between height and weight?

The relationship between height and weight was explored to see if the hypothesis that there is a significant positive linear relationship in the SWAN data set is an accurate claim. Data was visualized through a density scatterplot because the data was highly dense throughout the graph making it visually difficult to see any relationship (see **Fig. 15** in **Appendix**). The correlation coefficient was calculated and tested for significance to ensure the relationship is significant. This was tested at 0.95 confidence level.

From the analysis, the relationship was a weak positive, however based on the calculated p value, was significant (p<0.05). The null hypothesis was rejected and the claim was supported.

```
## 
##  Pearson's product-moment correlation
## 
## data:  milestone2_subset$HEIGHT0 and milestone2_subset$WEIGHT0
## t = 20.551, df = 3258, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3080083 0.3687967
## sample estimates:
##       cor 
## 0.3387559
```

Can Diastolic Blood Pressure be used to predict Systolic Blood Pressure?

There are two types of blood pressure measurements in the SWAN data set. Deciding which measure to use for the final comparison was determined by this analysis. If they were found to not be equivalent, both would be analyzed; however, if they are found to be equivalent either would be chosen. The significance of the correlation was tested at a confidence of 0.95 that there is a significant positive relationship between diastolic and systolic blood pressure measurements.

Based on the analysis outputs from the correlation significance test, the p value was less than the alpha 0.05 therefore the null hypothesis can be rejected. There is enough evidence to support the claim that there is a significant relationship between the two blood pressure measurements.

```
## 
##  Pearson's product-moment correlation
## 
```

```
## data:  milestone2_subset$DIABP10 and milestone2_subset$SYSBP10
## t = 54.833, df = 3286, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6729525 0.7086728
## sample estimates:
##       cor
## 0.6912347
```

A linear model was calculated and used to predict values for the systolic blood pressure. The model was evaluated using tools in R (see **Fig. 16** in **Appendix**). The diagnostics show the regression model passes for the majority with slight skewness at the higher blood pressure measurements (see **Normal Q-Q** in **Fig. 16**). The residuals plots do fall in the normal locations (around the zero line), however some fall close to and outside the Cook's Distance (see **Residuals vs Fitted** and **Residuals vs Leverage** in **Fig. 16**) These calculated predictions were plotted on the scatter plot to show the relative locations of the predicted measurements in red (see **Fig. 17** in **Appendix**).

```
##
## Call:
## lm(formula = SYSBP10 ~ DIABP10, data = milestone2_subset)
##
## Coefficients:
## (Intercept)      DIABP10
##      34.870        1.103

##   DIABP10
## 1      60
## 2      70
## 3      80
## 4      90
## 5     100
```

Is there a significant linear relationship between Weight and Blood Pressure

The relationship between weight and blood pressure was explored. The hypothesis is there was a significant positive relationship between weight and blood pressure (see **Fig. 18** in **Appendix**). This correlation was tested at a confidence of 0.95. The calculated correlation coefficient was a very weak positive. Despite the weak positive, the relationship was calculated to be a significant relationship.

```
## [1] 0.1672174

##
##  Pearson's product-moment correlation
##
## data:  milestone2_subset$WEIGHT0 and milestone2_subset$DIABP10
## t = 9.7032, df = 3273, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.1337325 0.2003210
## sample estimates:
##       cor
## 0.1672174
```

A sub-question emerged from this analysis. **How does racial subdivision effect the weight vs blood pressure relationship?** Because the relationship between the weight and blood pressure of patience was trying to be explored to determine if subdivision effected the correlation as a whole, the subdivisions were subsetted and not set as dummy variables, and each correlation was tested at a confidence of 0.95. Both subdivisions were significant and positive, however, minorities had a higher correlation and therefore a stronger relationship between weight and Diastolic Blood Pressure (see **Fig. 19** in **Appendix**).

```
## 
##  Pearson's product-moment correlation
## 
## data:  minority$WEIGHT0 and minority$DIABP10
## t = 10.838, df = 811, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2941105 0.4143010
## sample estimates:
##       cor
## 0.3556754

## 
##  Pearson's product-moment correlation
## 
## data:  majority$WEIGHT0 and majority$DIABP10
## t = 9.3132, df = 2460, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1461089 0.2224304
## sample estimates:
##       cor
## 0.1845478
```

Conclusion

The data subset from the SWAN data was analyzed to test relationships between specific variables. The data distributions were normal for continuous variables and even for the categorical variables as explored in the **Exploratory Data Analysis**.

Hypothesis testing showed that there was no significant difference between the pulses of women who have been diagnosed with anemia and those who have not in the SWAN data set. Additionally, the proportion of women who smoked at age 45 were not statistically significantly different than the proportion of women who smoke in the SWAN data set.

There was not enough evidence to support either of the claims that there were differences in the samples obtained from the data set.

The linear relationships explored in this report were found to be statistically significant, however the relationship between the height and weight and weight and blood pressure were very weak linear relationships. Weight vs Diastolic Blood pressure is slightly effected by racial subdivisions, where minorities have almost twice the correlation coefficient value than majorities.

This analysis can be used to look at diagnostics and health products targeting women in the years ranging from 43 to 52 years old. This small subset can be a guide for analysis in the larger data set that has been compiled over the years.

Appendix

References

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

Data Visualizations

```
## Table 1. Race Frequency Data
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

## Table 2a. Summary Statistics
sumStats

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

## Table 2b. Summary Statistics
summary(milestone2_subset)

##      SWANID          AGE0        ANEMIA0             LISTEN0
##  Min.   :10005   Min.   :42.00   Length:3302        Min.   :1.000
##  1st Qu.:31808   1st Qu.:44.00   Class :character   1st Qu.:4.000
##  Median :54230   Median :46.00   Mode  :character   Median :4.000
##  Mean   :54362   Mean   :45.85                      Mean   :4.206
##  3rd Qu.:76745   3rd Qu.:48.00                      3rd Qu.:5.000
##  Max.   :99992   Max.   :53.00                      Max.   :5.000
##                  NA's   :5                          NA's   :5
##     TAKETOM0        CONFIDE0       SYSBP10         DIABP10
##  Min.   :1.000   Min.   :1.00   Min.   : 70.0   Min.   : 40.00
##  1st Qu.:4.000   1st Qu.:4.00   1st Qu.:108.0   1st Qu.: 68.00
##  Median :5.000   Median :4.00   Median :116.0   Median : 76.00
##  Mean   :4.174   Mean   :4.19   Mean   :118.2   Mean   : 75.57
##  3rd Qu.:5.000   3rd Qu.:5.00   3rd Qu.:128.0   3rd Qu.: 80.00
##  Max.   :5.000   Max.   :5.00   Max.   :230.0   Max.   :140.00
##  NA's   :6       NA's   :5      NA's   :9       NA's   :13
##     HELPSIC0       SMOKERE0             PULSE0         HEIGHT0
##  Min.   :1.000   Length:3302        Min.   :17.00   Min.   :140.5
##  1st Qu.:3.000   Class :character   1st Qu.:32.00   1st Qu.:157.8
##  Median :4.000   Mode  :character   Median :35.00   Median :162.4
##  Mean   :3.746                      Mean   :35.19   Mean   :162.4
##  3rd Qu.:5.000                      3rd Qu.:38.00   3rd Qu.:167.0
##  Max.   :5.000                      Max.   :84.00   Max.   :186.2
##  NA's   :5                          NA's   :7       NA's   :32
##     WEIGHT0           RACE           Subdivision         SupportScore
##  Min.   : 37.60   Length:3302        Length:3302        Min.   : 4.00
##  1st Qu.: 59.60   Class :character   Class :character   1st Qu.:15.00
##  Median : 70.60   Mode  :character   Mode  :character   Median :17.00
##  Mean   : 74.88                                         Mean   :16.32
##  3rd Qu.: 85.50                                         3rd Qu.:19.00
##  Max.   :175.40                                         Max.   :20.00
##  NA's   :14                                             NA's   :6
##    SupportAvg
##  Min.   :1.000
##  1st Qu.:3.750
##  Median :4.250
##  Mean   :4.079
##  3rd Qu.:4.750
##  Max.   :5.000
##  NA's   :6
```

Figure 1: Patient Race Pareto

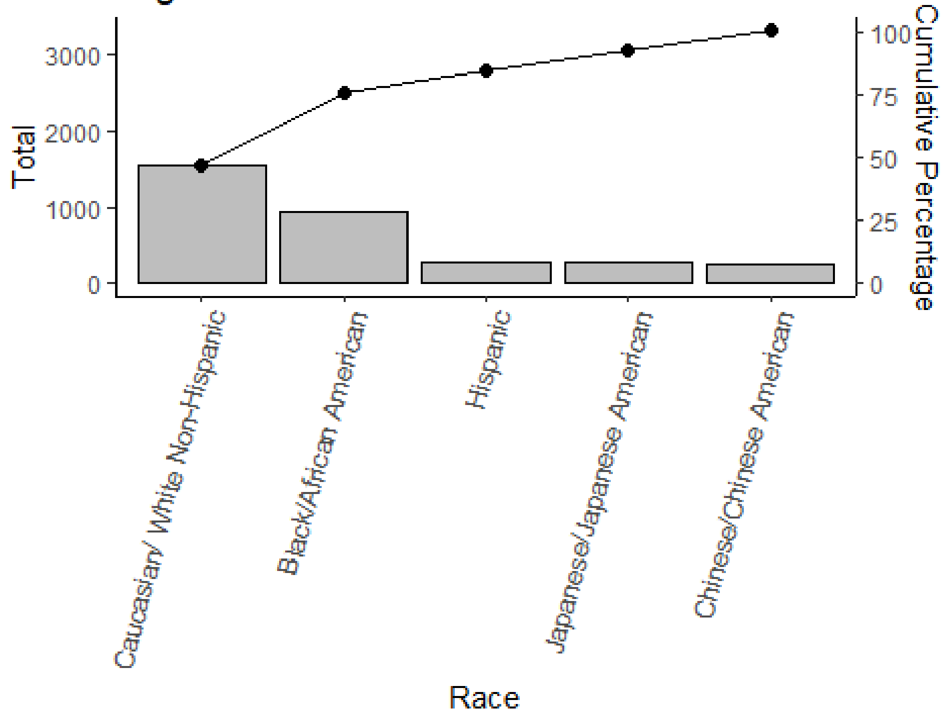


Figure 2: Spread of Age

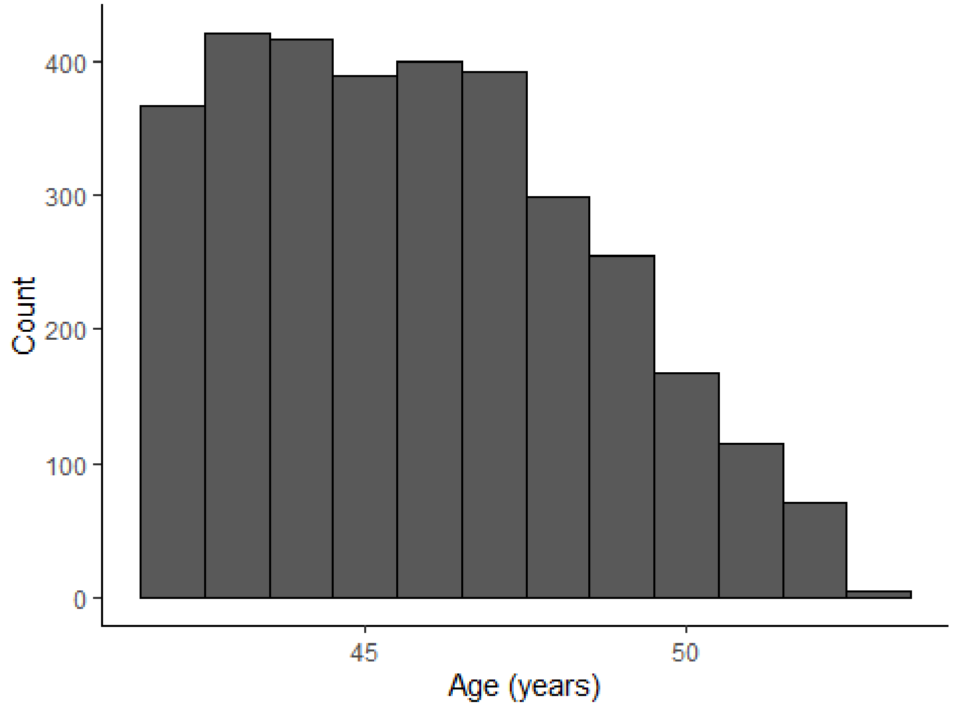


Figure 3: Spread of Pulse

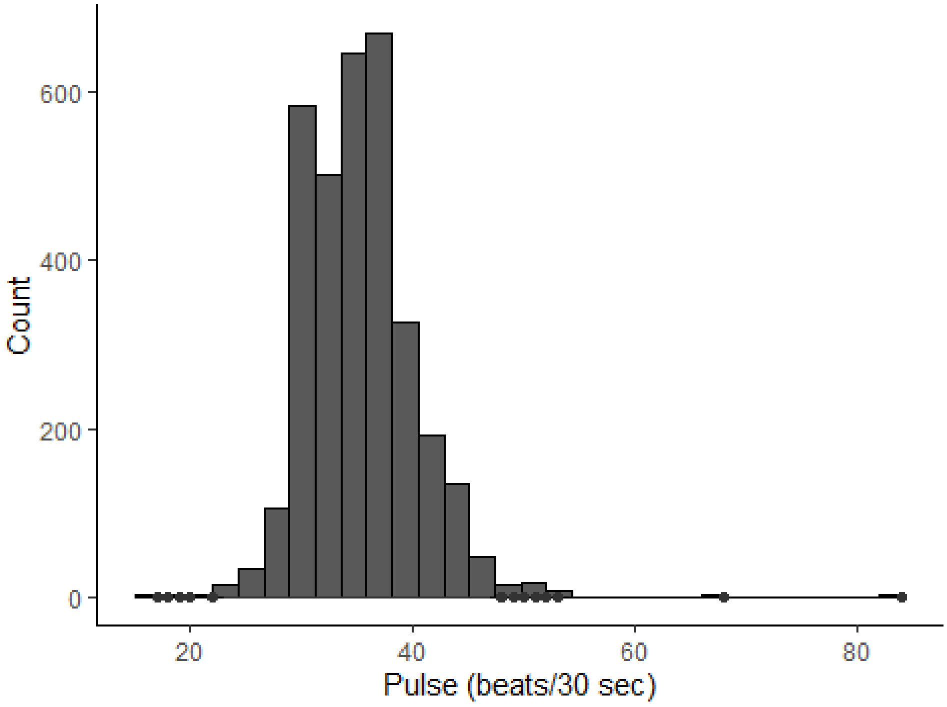


Figure 4: Spread of Height

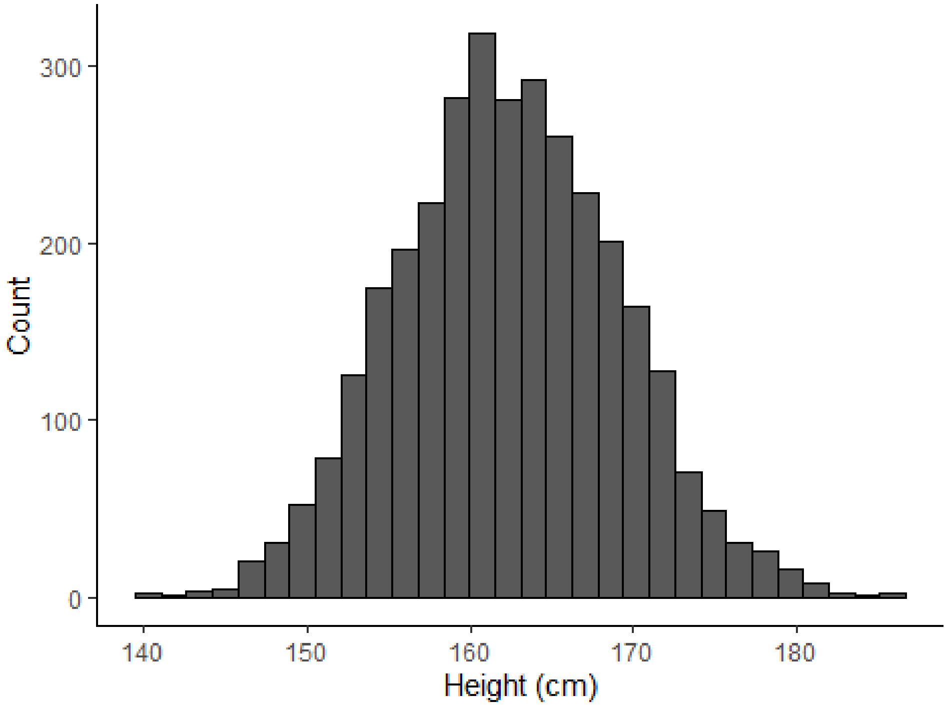


Figure 5: Spread of Weight

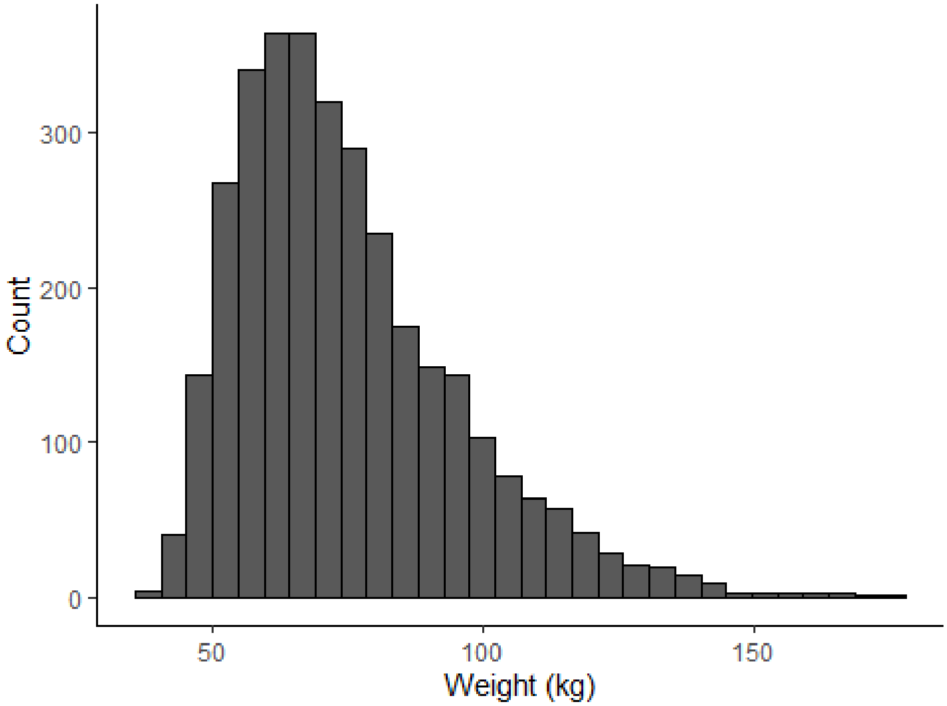


Figure 6: Spread of Support Score

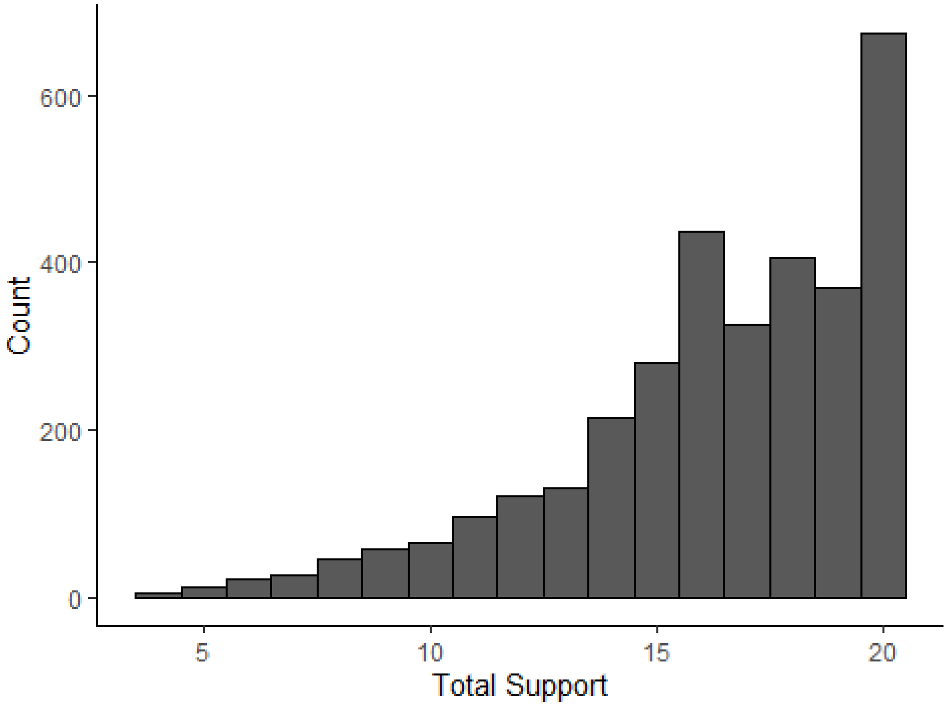


Figure 7: Frequency of Patients with Anemia

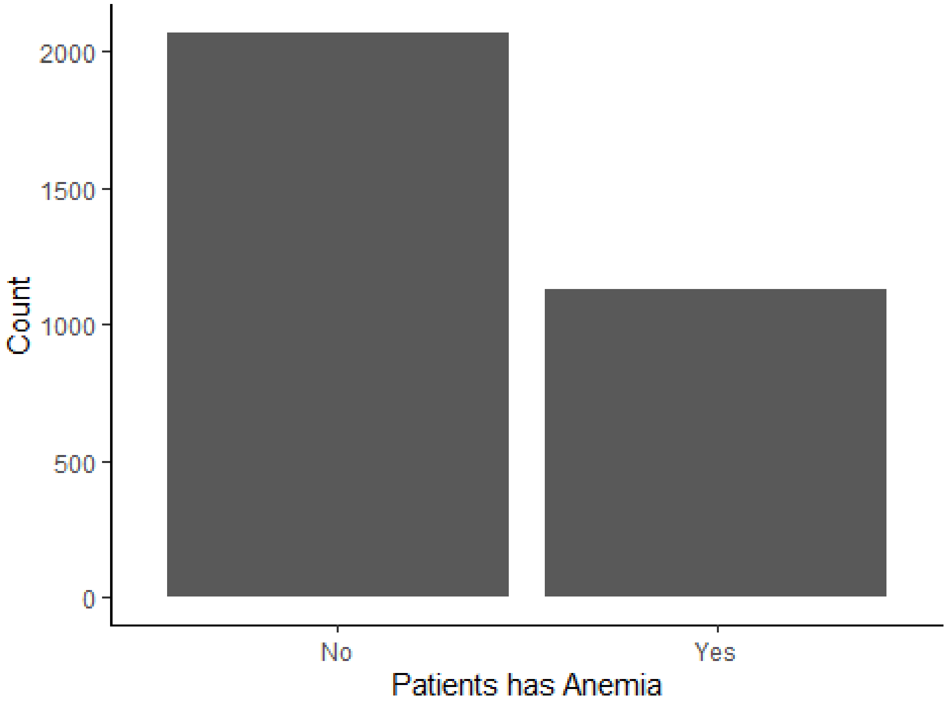


Figure 8: Frequency of Patients who Smoke

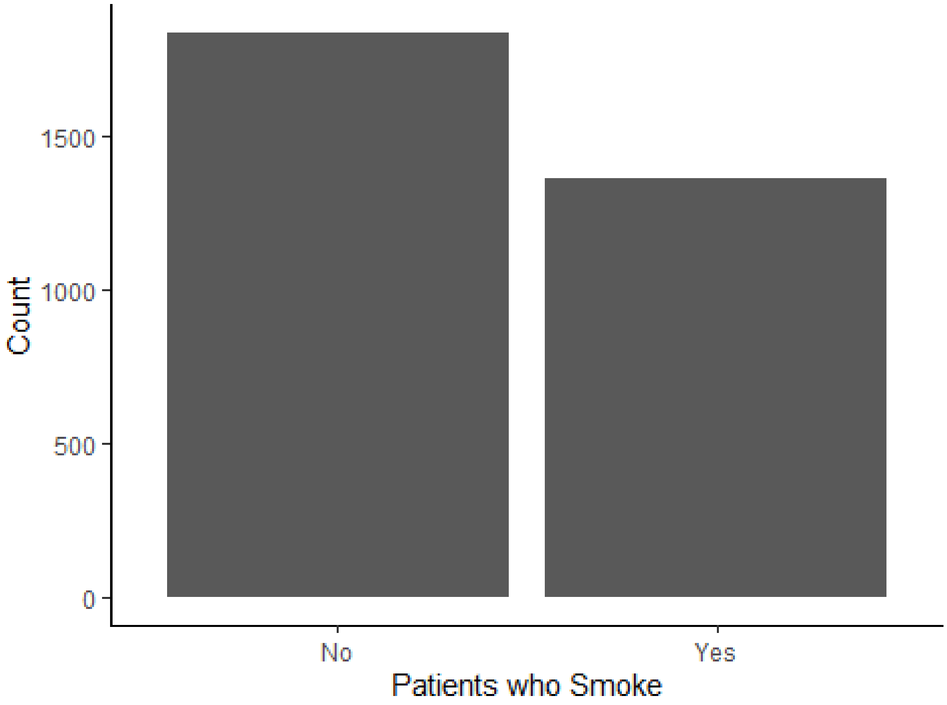


Figure 9: Height vs Weight Scatterplot

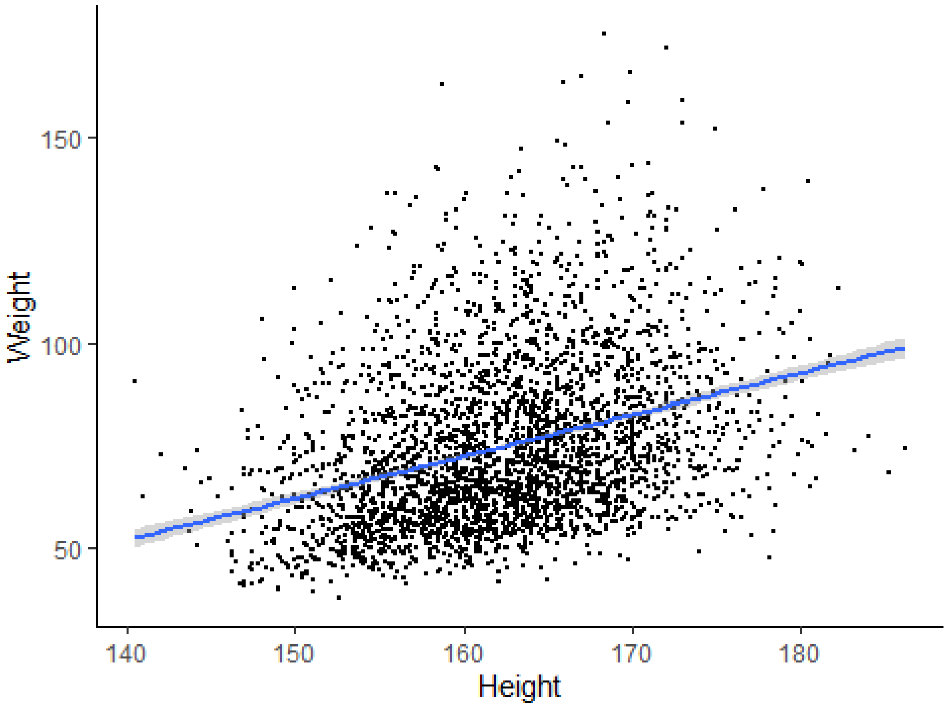
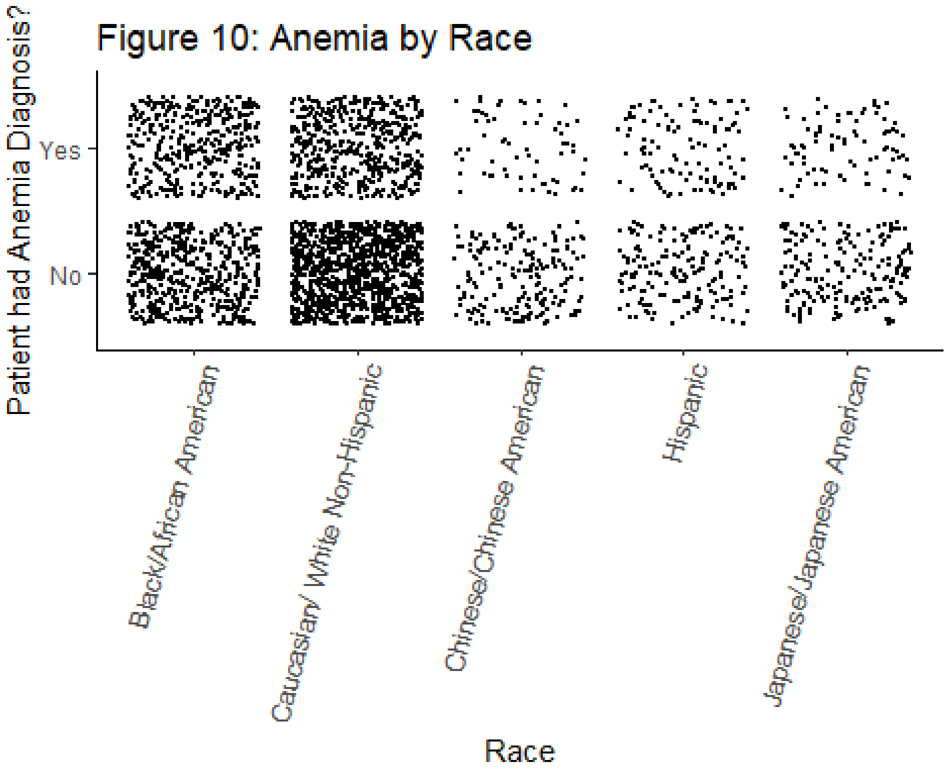
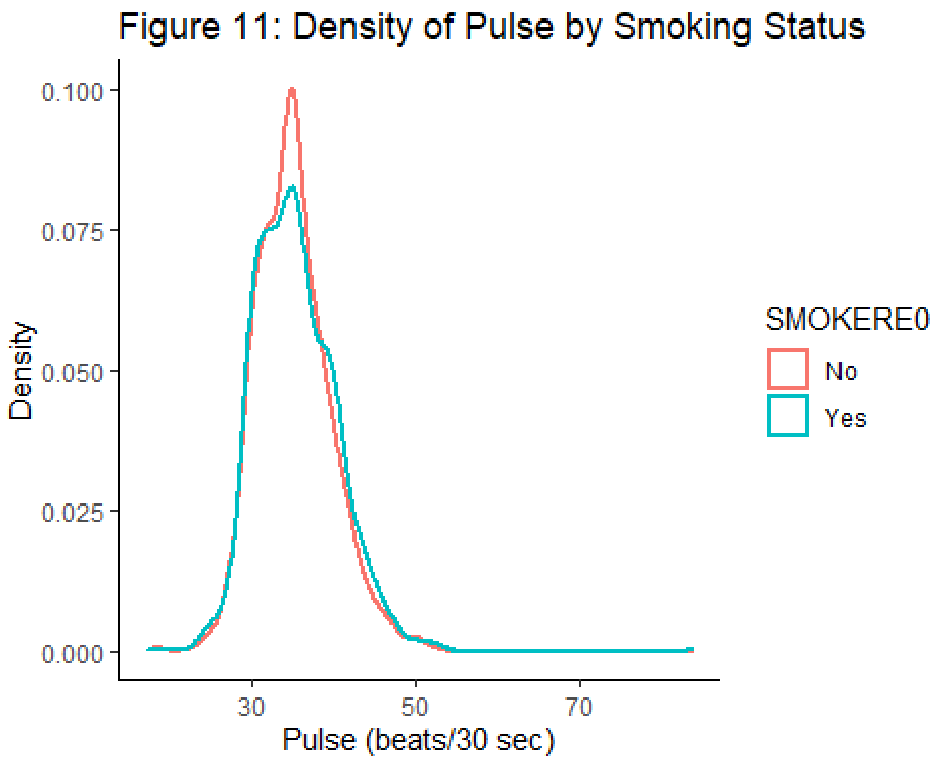
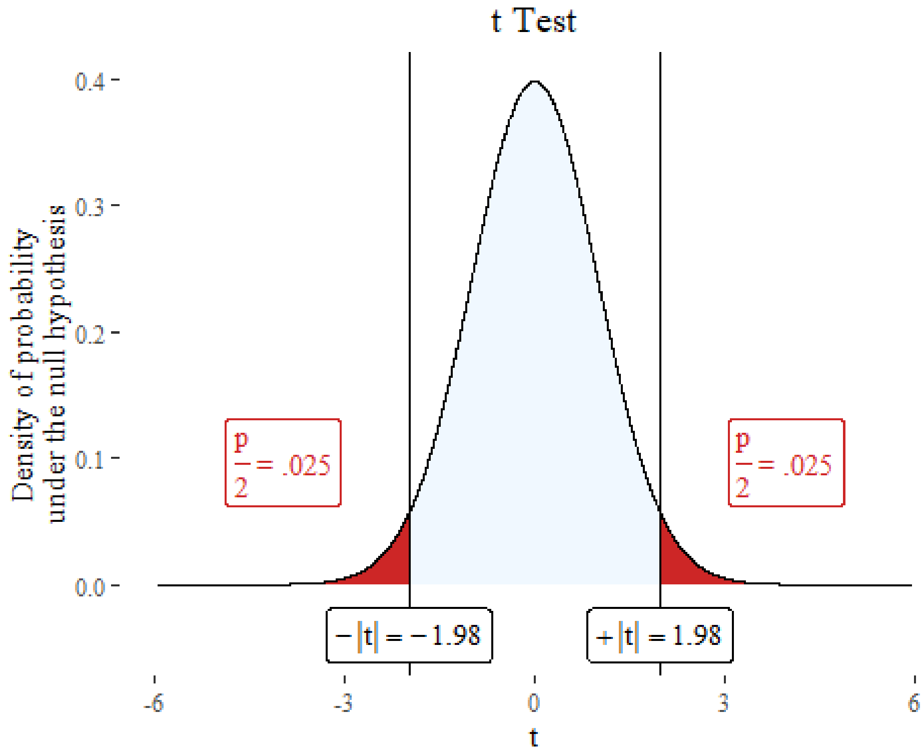


Figure 10: Anemia by Race





```
#Figure 12: t Plot
tplot
```



```
#Figure 13: z Plot
zplot
```

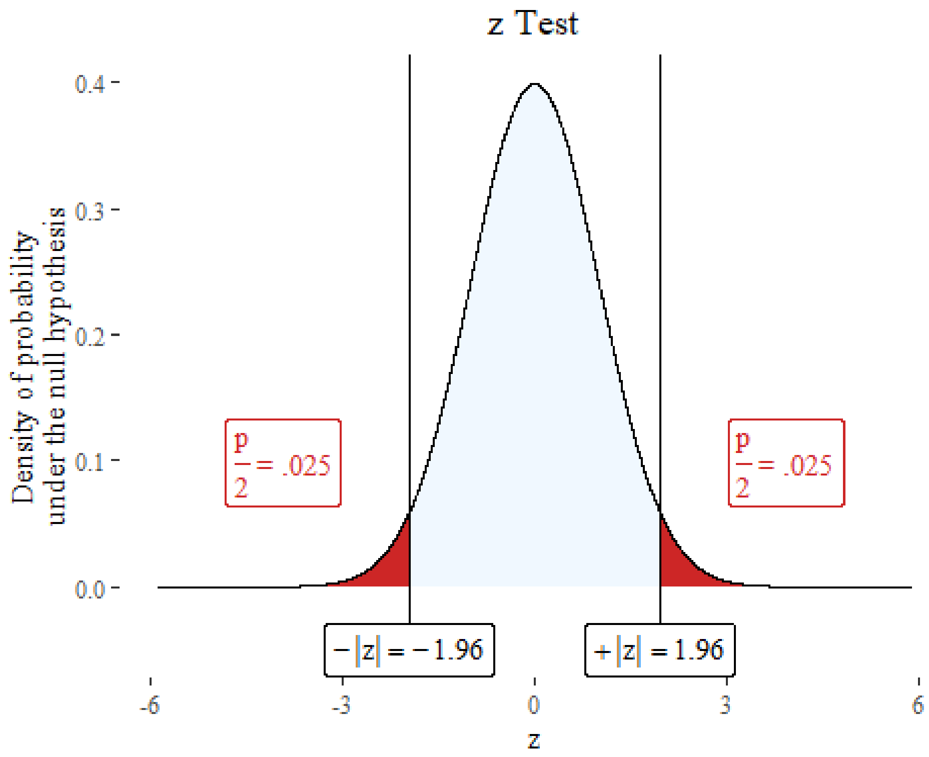
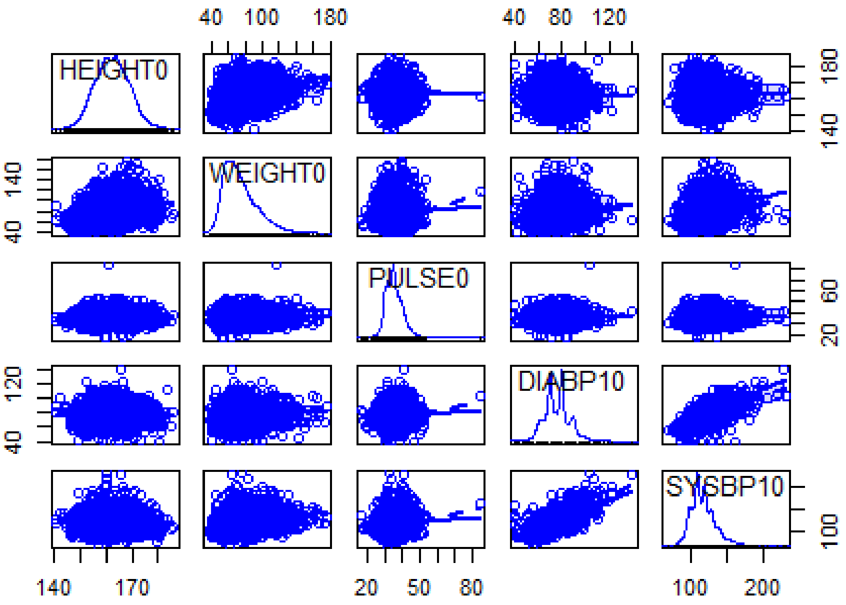
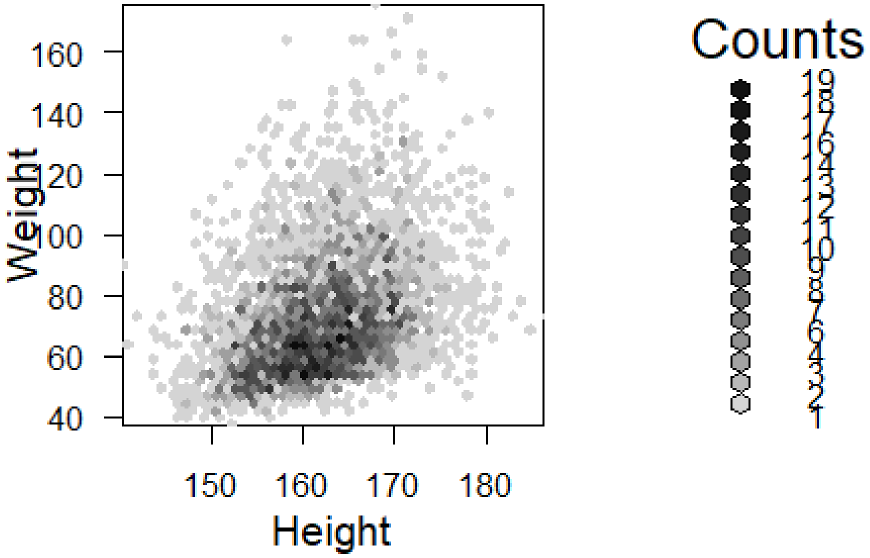
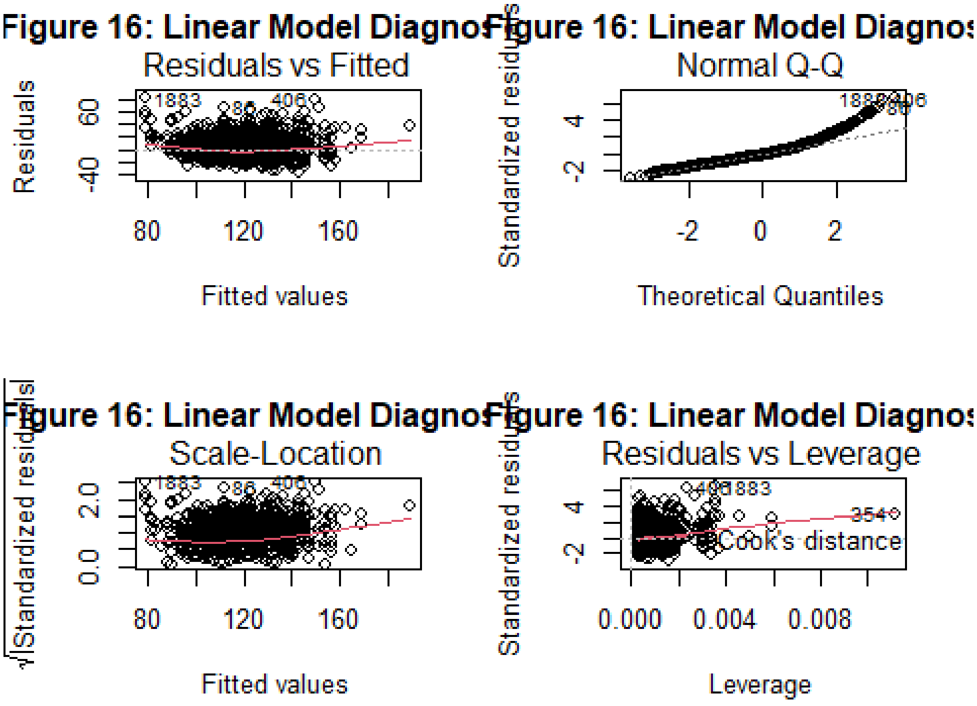


Figure 14: Scatterplot Matrix



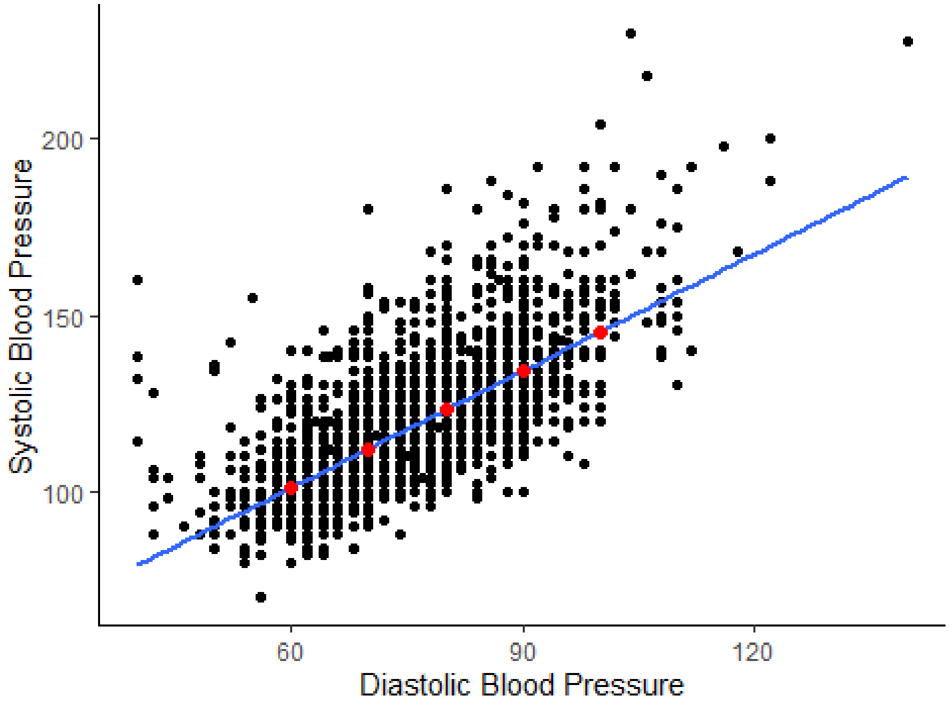
15: Hexagonal Binning of Height vs Weight



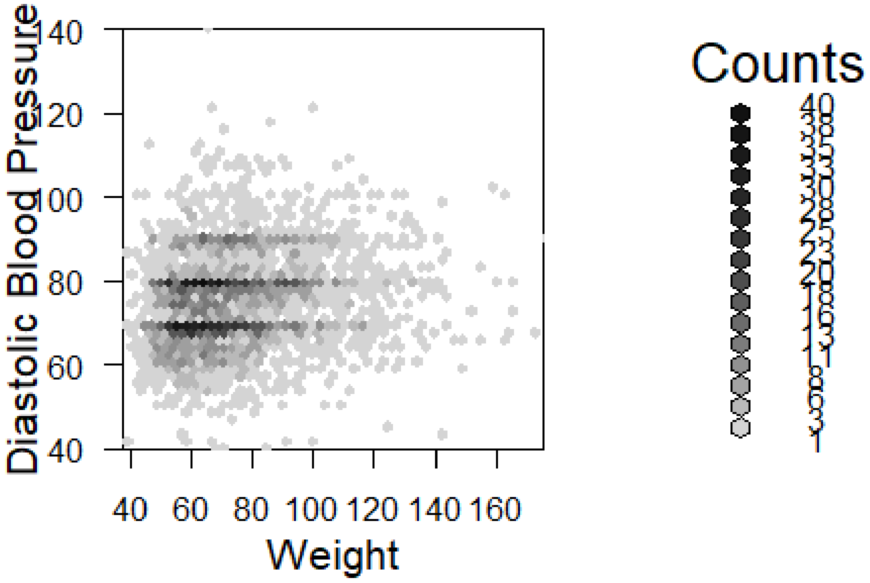


```
## `geom_smooth()` using formula 'y ~ x'
```

Figure 17: Blood Pressure Regression

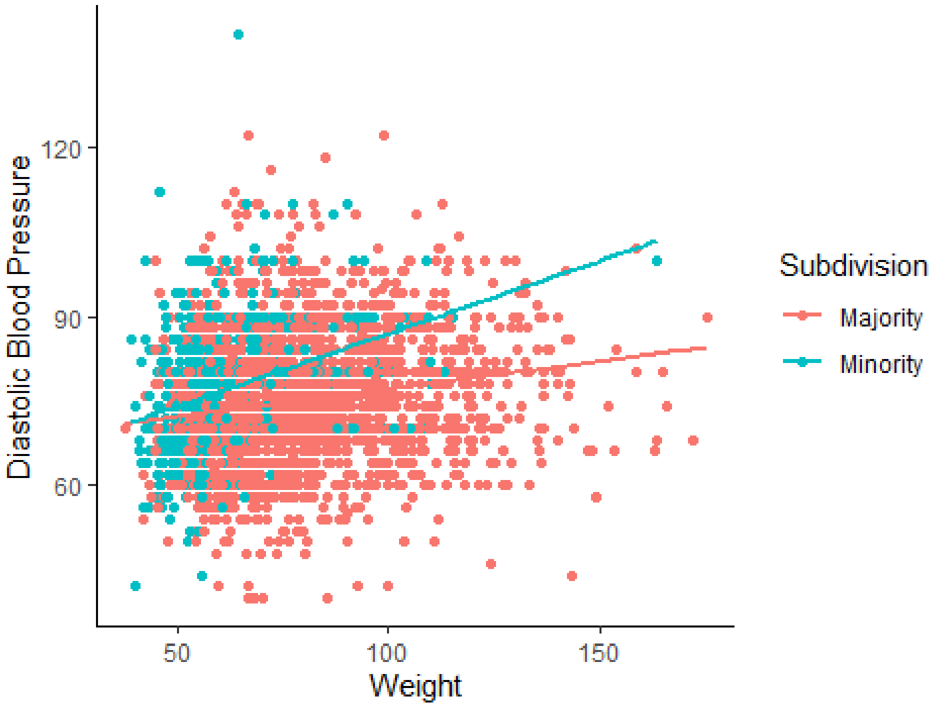


Hexagonal Binning of Weight vs Blood Press



```
## `geom_smooth()` using formula 'y ~ x'
```

Figure 19: Weight vs Blood Pressure


Diastolic Blood Pressure
Weight
Subdivision
Majority
Minority