Homework 3

Jacob Thielemier

7 April 2024

Question 1

Part (a and b)

- The names of the variables in the data set are:
  - male
  - age
  - education
  - currentSmoker
  - cigsPerDay
  - BPMeds
  - prevalentStroke
  - prevalentHyp
  - diabetes
  - totChol
  - sysBP
  - diaBP
  - BMI
  - heartRate
  - glucose
  - TenYearCHD

Part (c)

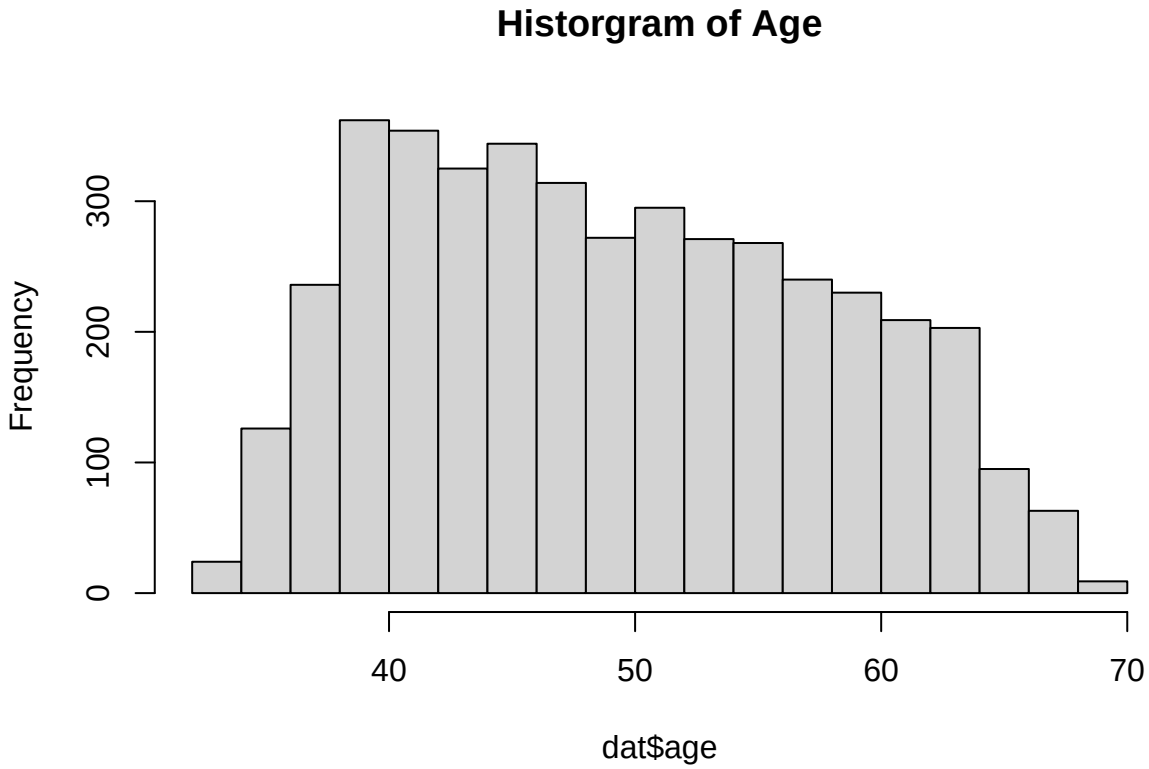
- There are 4240 observations.

Part (d)

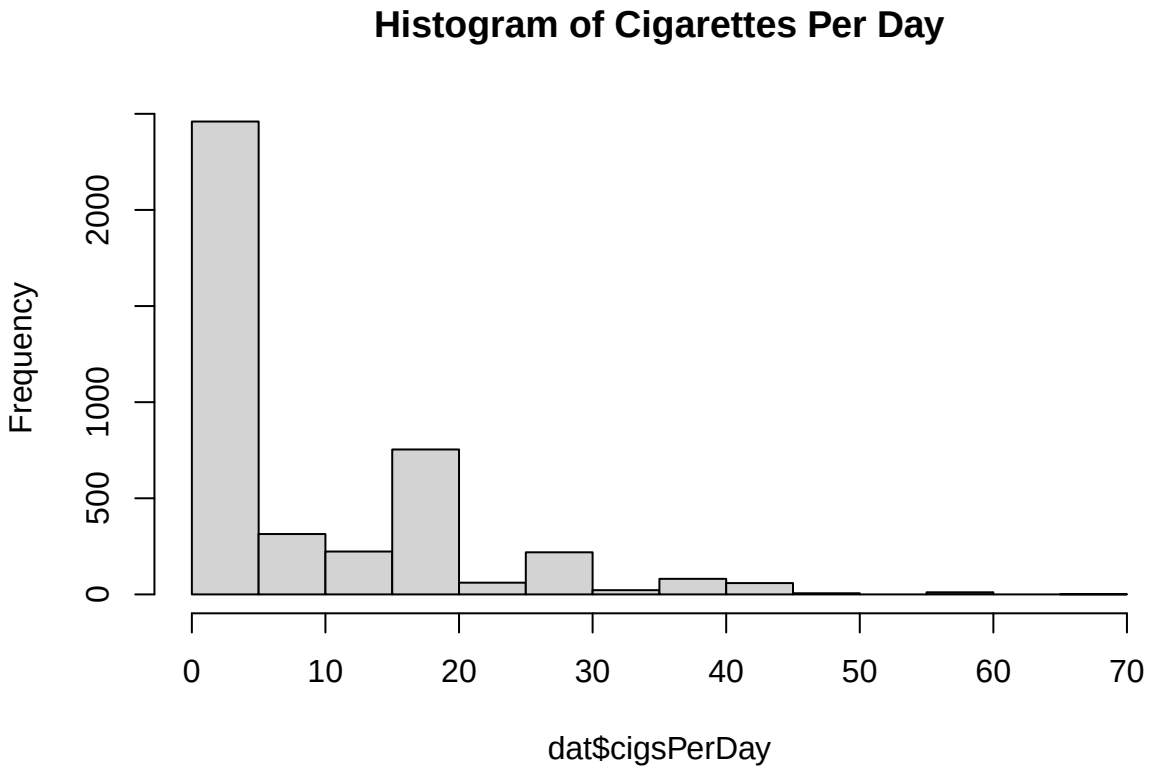
- Proposed model: sysBP ~ totChol + male + age + diabetes + diaBP + BMI + currentSmoker + cigsPerDay + glucose
- The following are possible confounders that are associated with sysBP and totChol:
  - male
  - age
  - diabetes
  - diaBP
  - BMI
- The following are precision variables that will help our estimates:

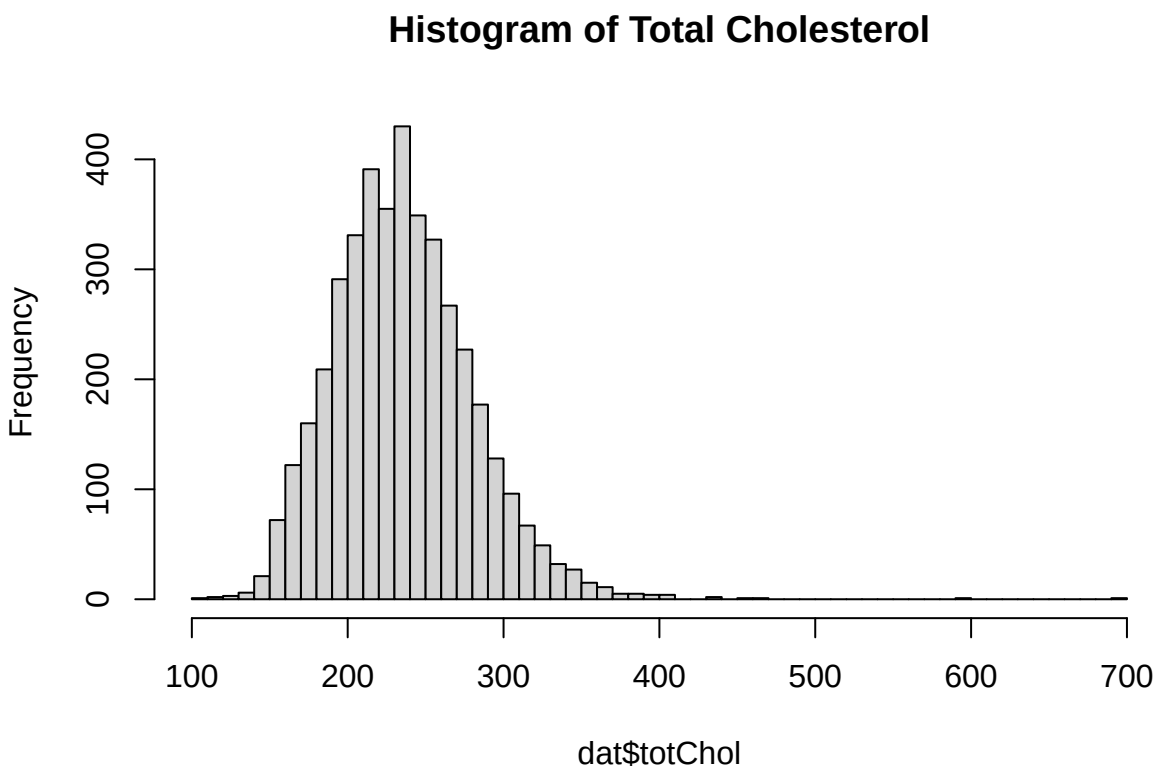
- currentSmoker
- cigsPerDay
- glucose

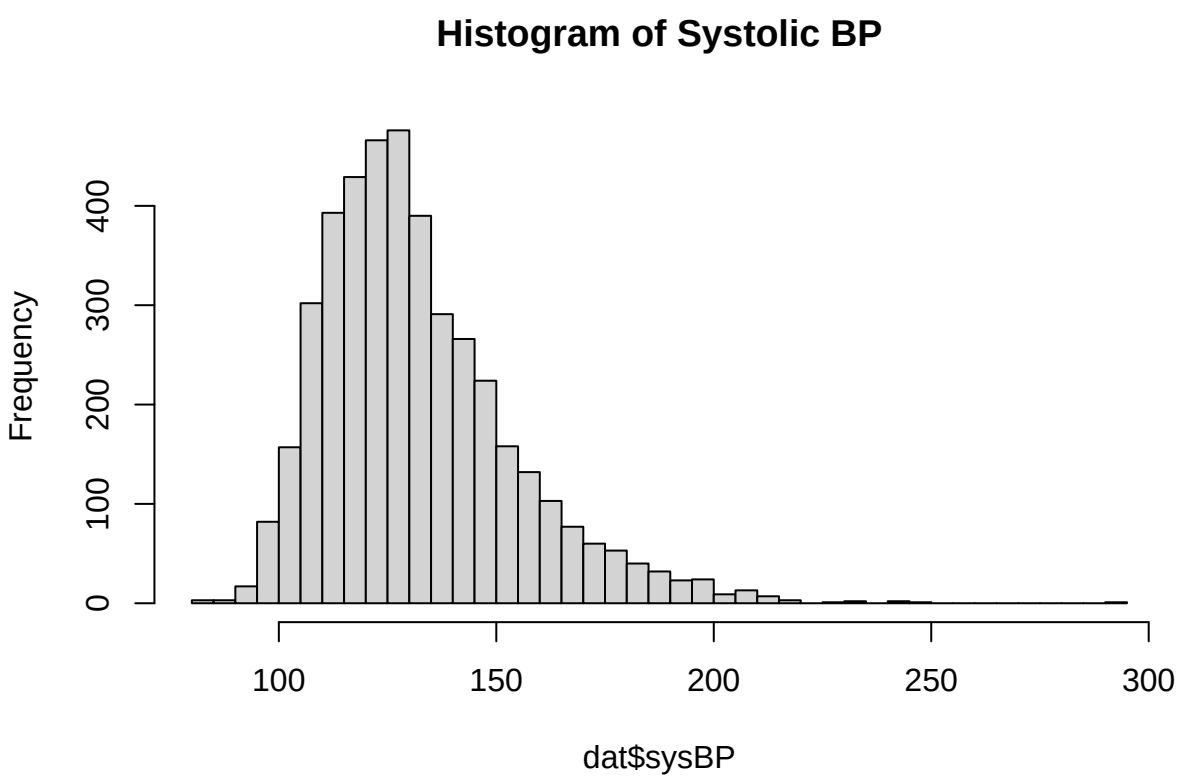
Question 2

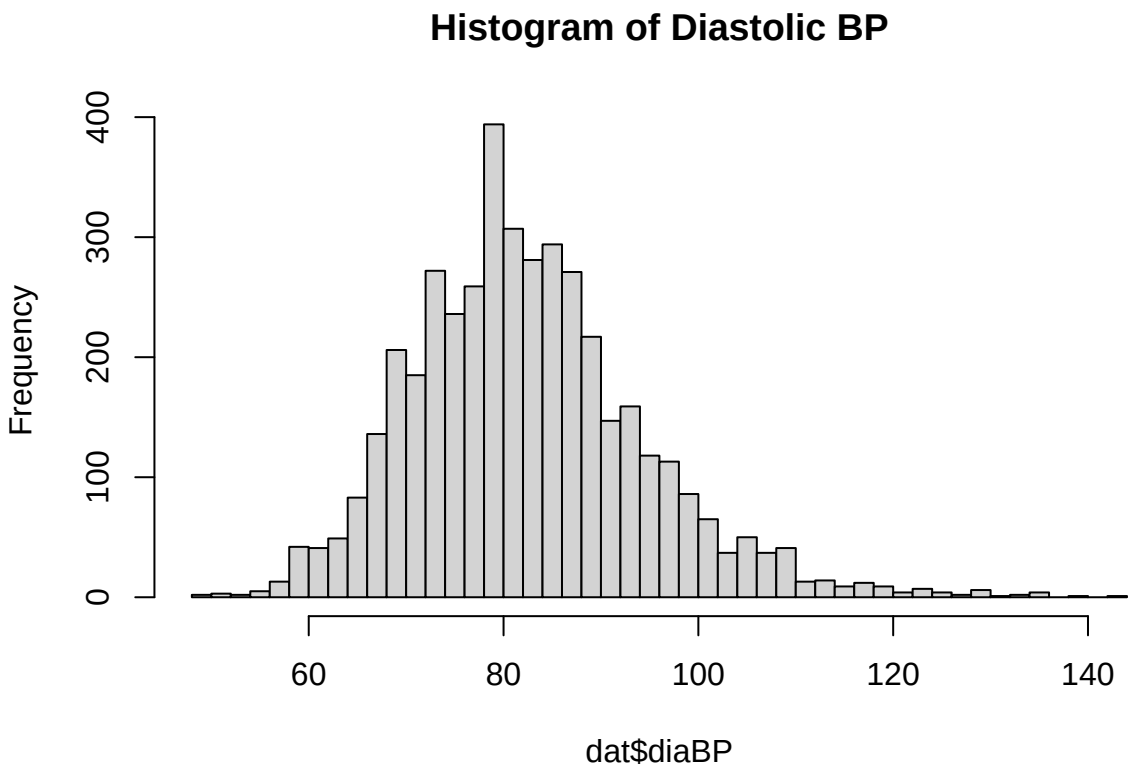


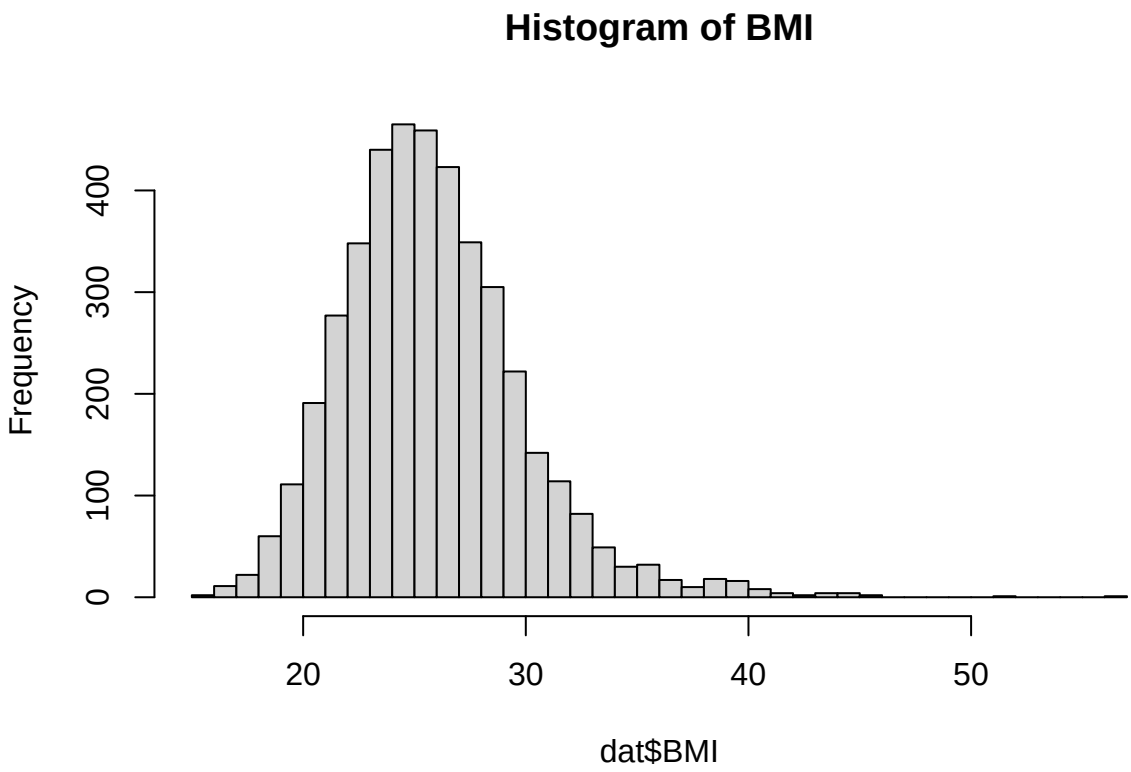
Part (a)

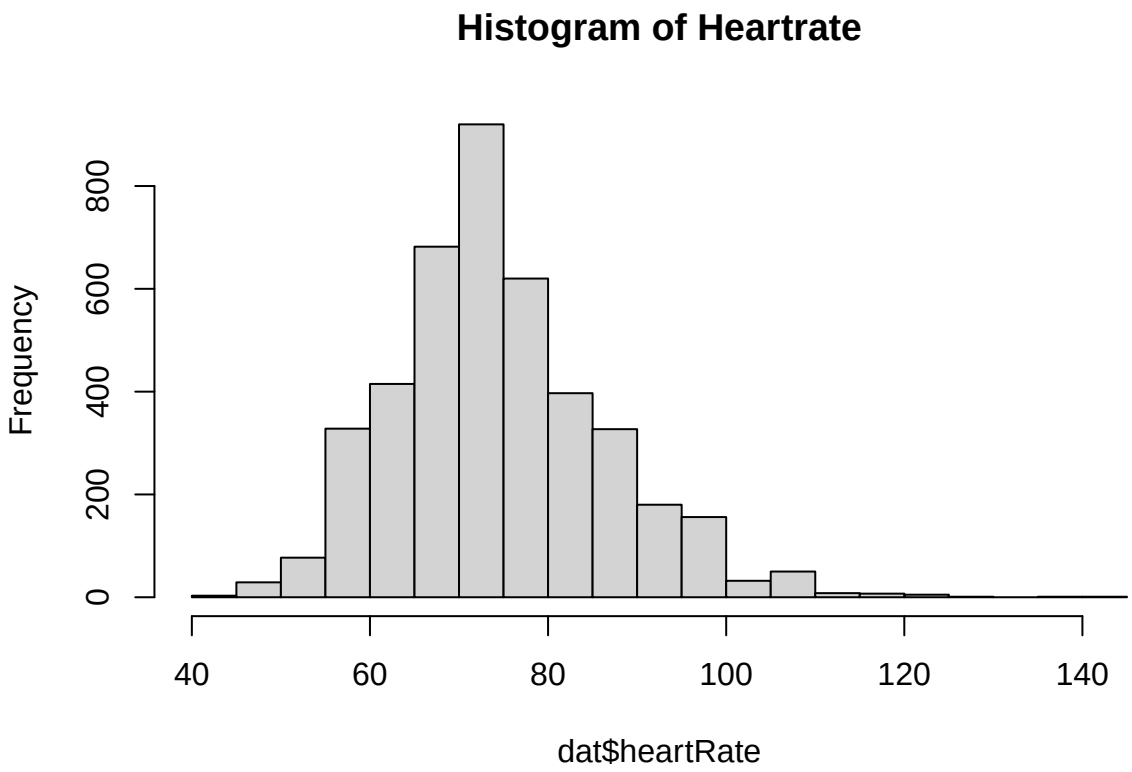


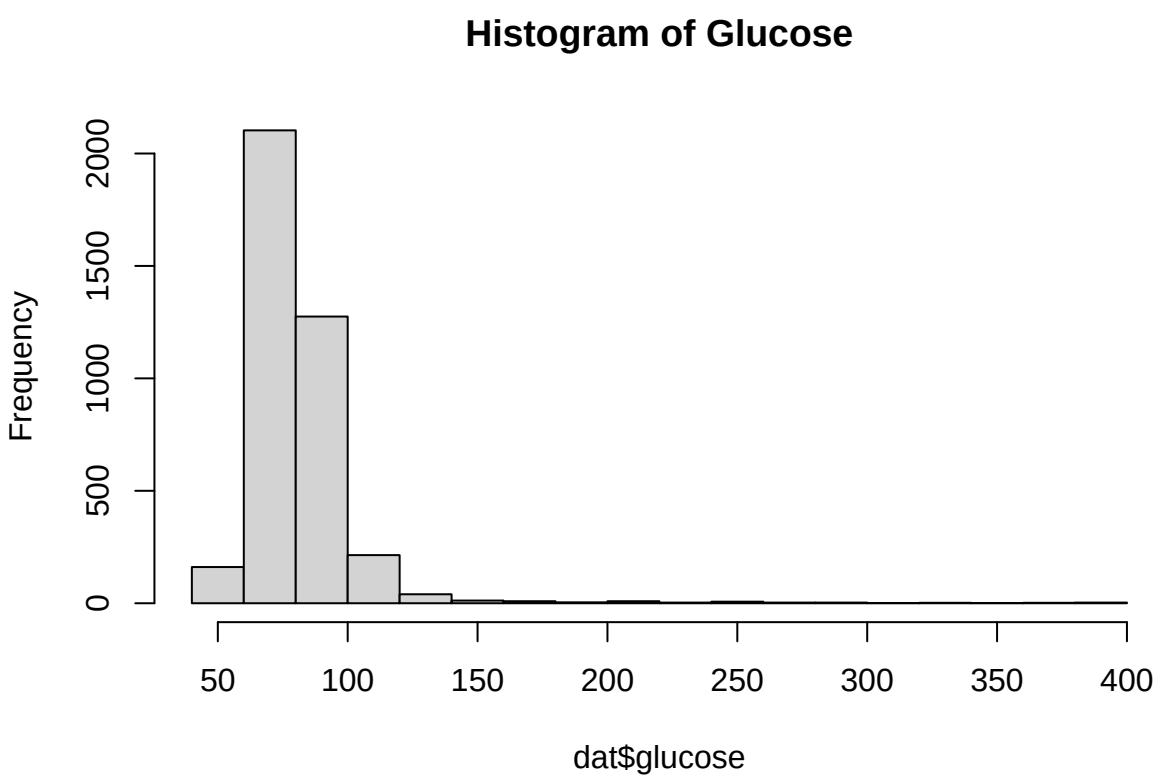

Histogram of Total Cholesterol
Frequency
0
100
200
300
400
100
200
300
400
500
600
700
dat$totChol


Histogram of Systolic BP
Frequency
0
100
200
300
400
100
150
200
250
300
dat$sysBP


Histogram of Diastolic BP
Frequency
0
100
200
300
400
60
80
100
120
140
dat$diaBP


Histogram of BMI
Frequency
0
100
200
300
400
20
30
40
50
dat$BMI


Histogram of Heartrate
Frequency
0
200
400
600
800
40
60
80
100
120
140
dat$heartRate



Part (b)

- **Age:** looks fine with a small amount being very young.
- **Cigarettes per day:** Most people that smoke have around 20 cigarettes per day.
- **Total cholesterol:** It is bell shaped curved, but seems high since typically 200 is ideal
- **Systolic BP and Diastolic BP:** Both look typical.
- **BMI:** Typically want to see between 18 and 25. This data seems a little higher, but not concerning.
- **Heart Rate:** This looks average
- **Glucose:** The typical range is 70-100 so most data falls in the range. However, the data over 150 does raise concerns.

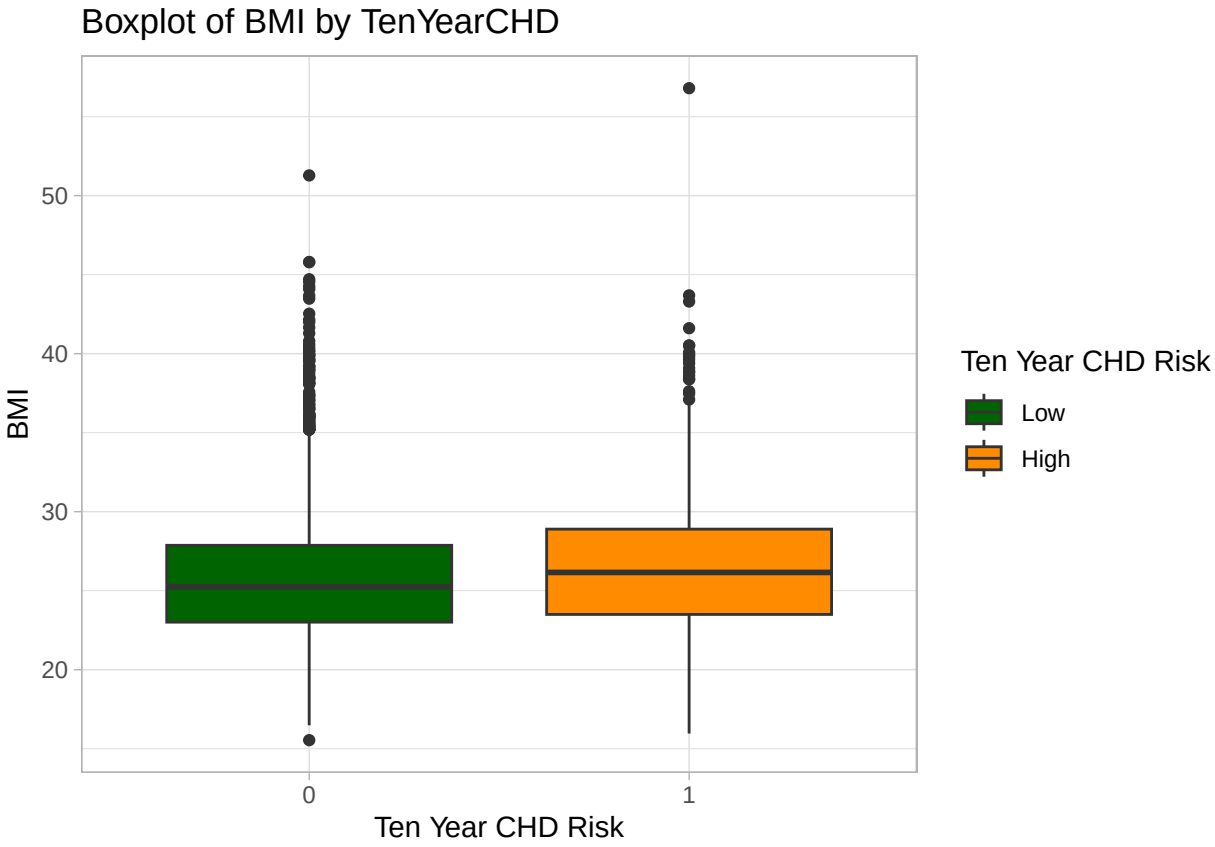
Question 3

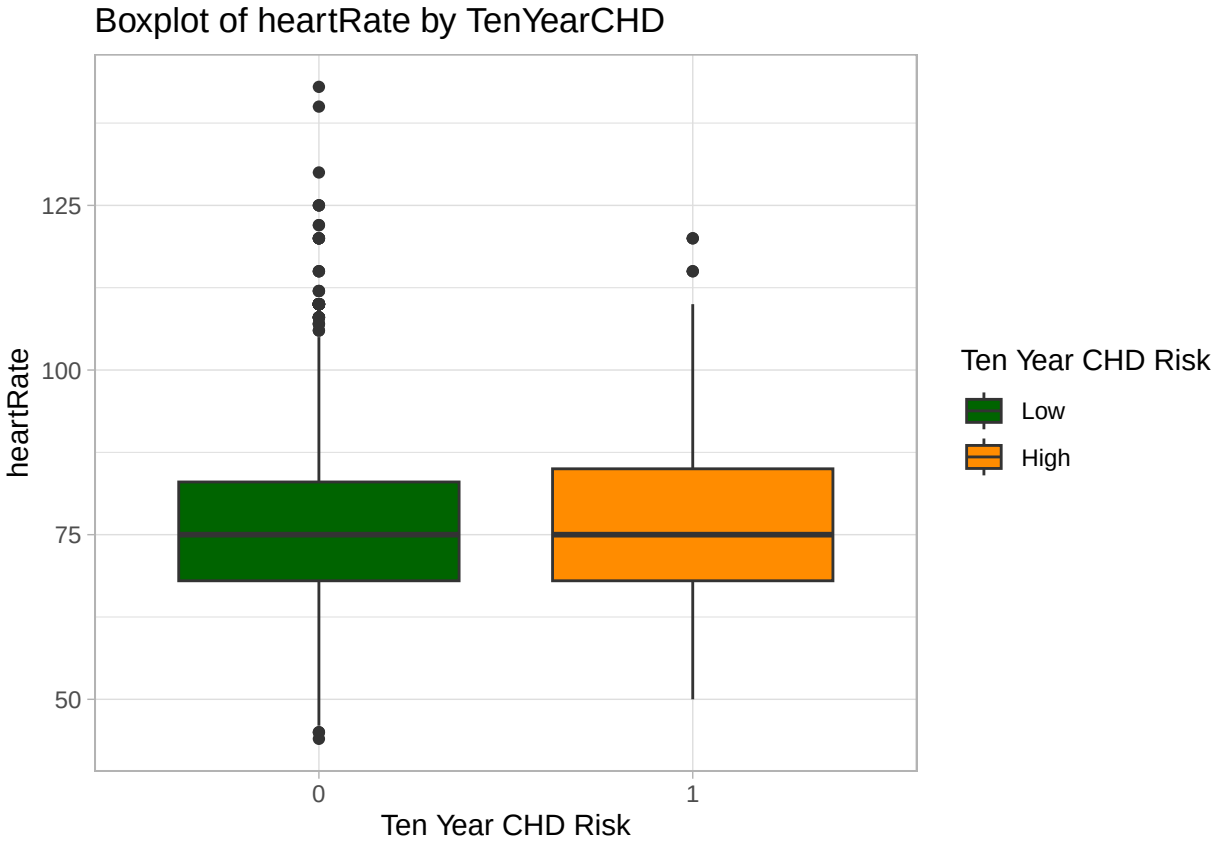
Table 1: Combined Frequency Distribution

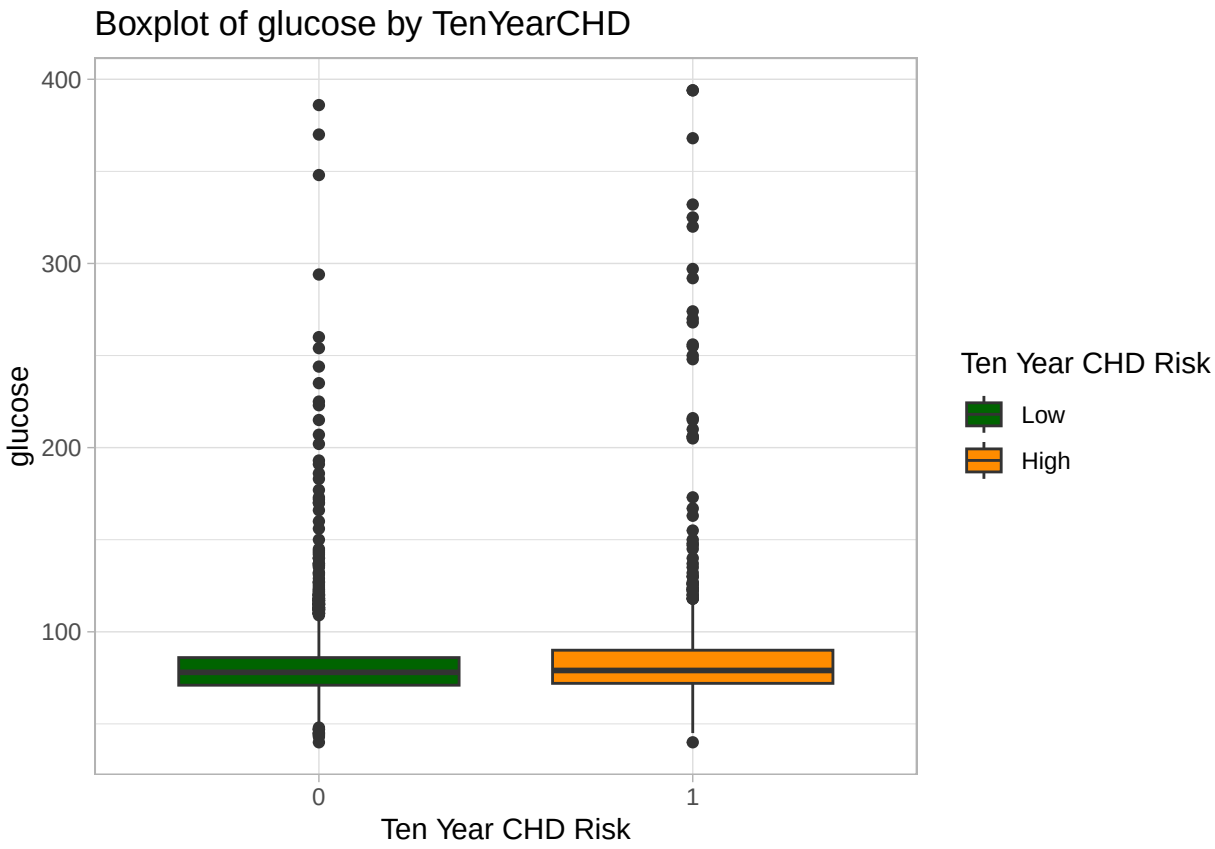
Variable	Value	Count	Percentage
male	0	2420	57.075472
male	1	1820	42.924528
education	1	1720	40.566038
education	2	1253	29.551887
education	3	689	16.250000
education	4	473	11.155660
education	NA	105	2.476415

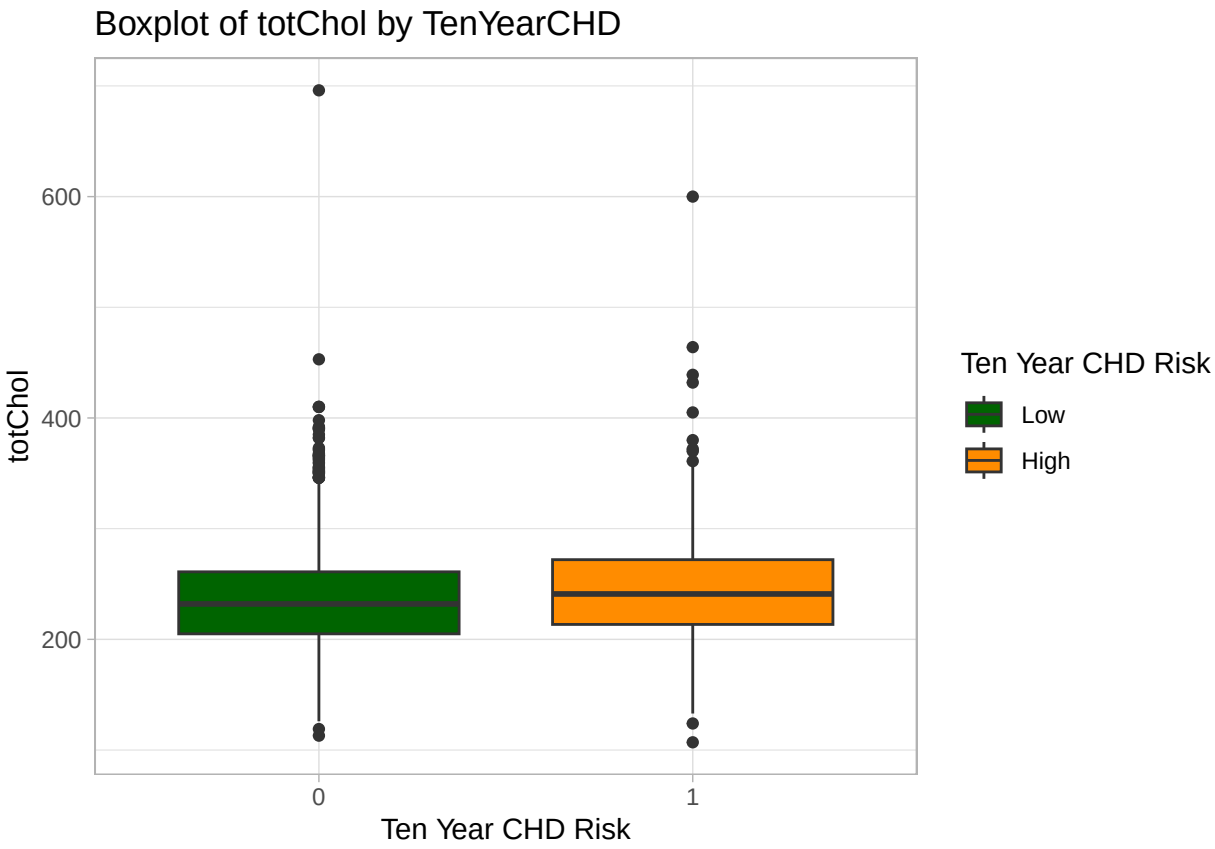
Variable	Value	Count	Percentage
currentSmoker	0	2145	50.589623
currentSmoker	1	2095	49.410377
BPMeds	0	4063	95.825472
BPMeds	1	124	2.924528
BPMeds	NA	53	1.250000
diabetes	0	4131	97.429245
diabetes	1	109	2.570755

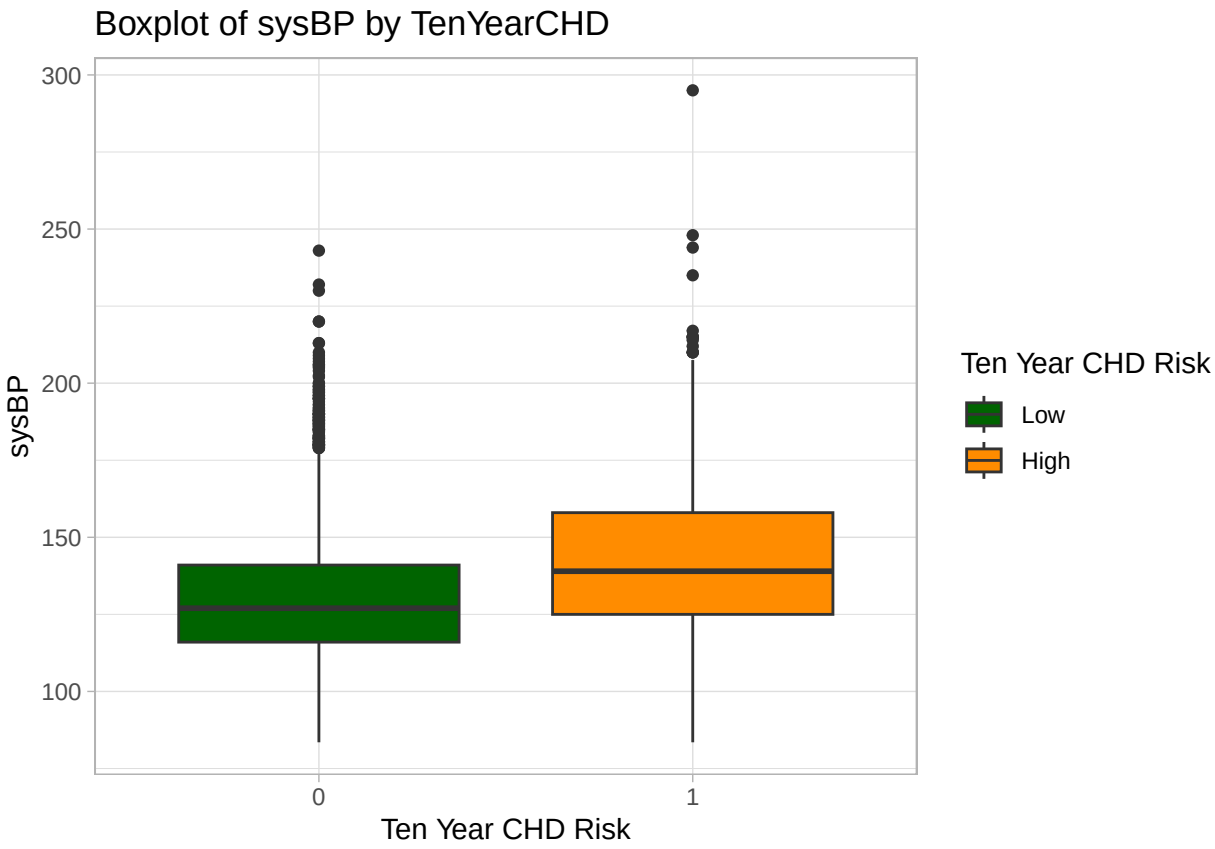
Question 4

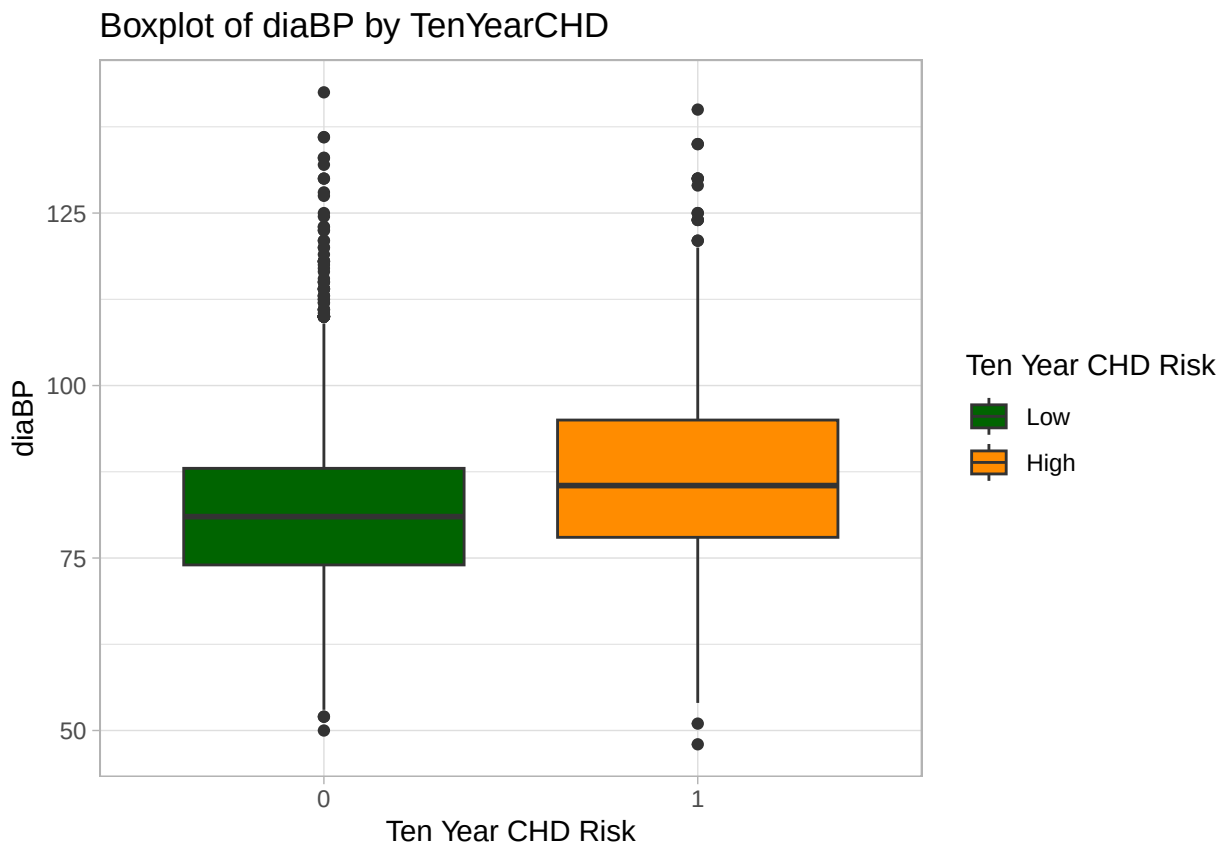

Boxplot of BMI by TenYearCHD
BMI
50
40
30
20
0
1
Ten Year CHD Risk
Ten Year CHD Risk
Low
High


Boxplot of heartRate by TenYearCHD
heartRate
125
100
75
50
0
1
Ten Year CHD Risk
Ten Year CHD Risk
Low
High


Boxplot of glucose by TenYearCHD
400
300
200
100
glucose
0
1
Ten Year CHD Risk
Ten Year CHD Risk
Low
High


Boxplot of totChol by TenYearCHD
totChol
600
400
200
0
1
Ten Year CHD Risk
Ten Year CHD Risk
Low
High


Boxplot of sysBP by TenYearCHD
300
250
200
150
100
sysBP
0
1
Ten Year CHD Risk
Ten Year CHD Risk
Low
High



Question 5

Part (a)

Table 2: Summary Statistics for Continuous Variables

Variable	Mean	SD
age	49.580189	8.572942
cigsPerDay	9.005937	11.922462
BMI	25.800801	4.079840
heartRate	75.878981	12.025348
glucose	81.963655	23.954335
totChol	236.699523	44.591284
sysBP	132.354599	22.033300
diaBP	82.897759	11.910395

Part (b)

Table 3: Frequency Distribution for Categorical Variables

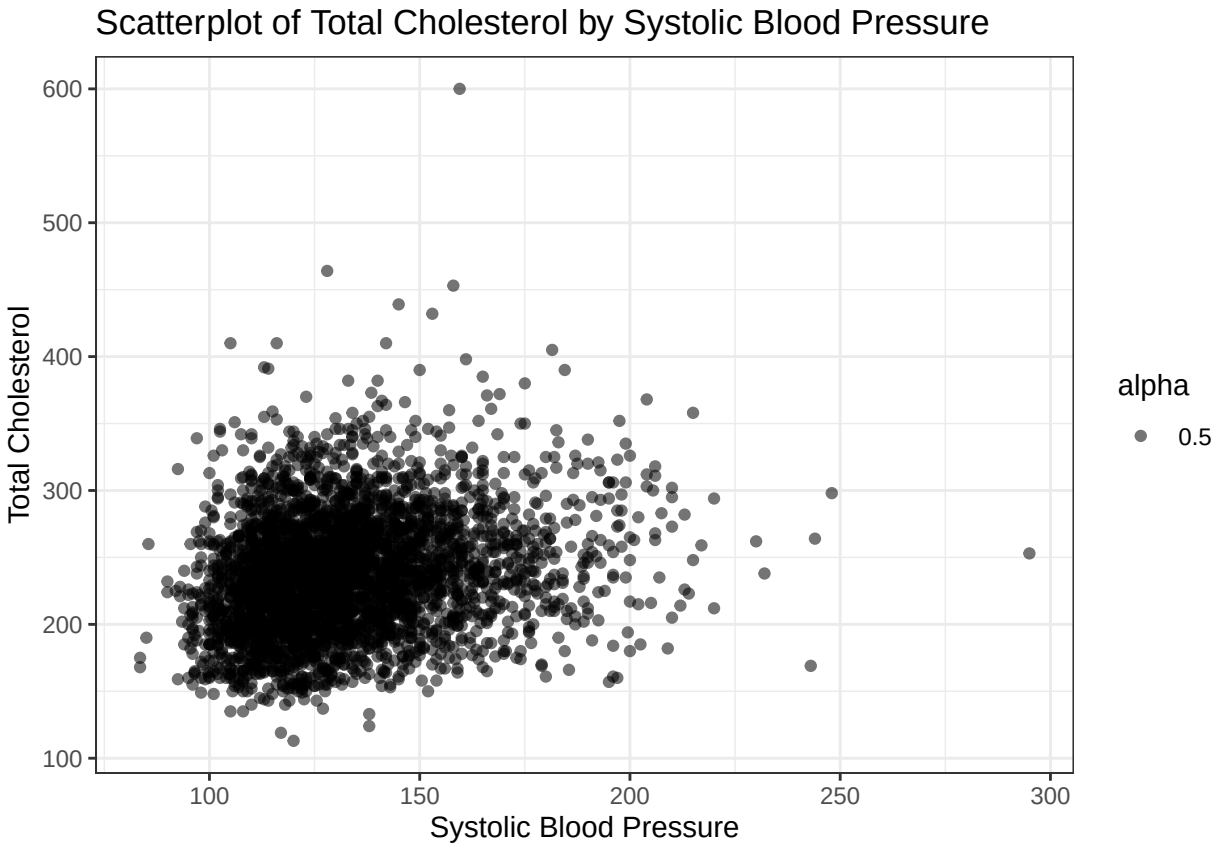
Variable	Count	Percentage
male 0	2420	57.075
male 1	1820	42.925

education 1	1720	40.566
education 2	1253	29.552
education 3	689	16.250
education 4	473	11.156
education NA	105	2.476
currentSmoker 0	2145	50.590
currentSmoker 1	2095	49.410
BPMeds 0	4063	95.825
BPMeds 1	124	2.925
BPMeds NA	53	1.250
prevalentStroke 0	4215	99.410
prevalentStroke 1	25	0.590
prevalentHyp 0	2923	68.939
prevalentHyp 1	1317	31.061
diabetes 0	4131	97.429
diabetes 1	109	2.571
TenYearCHD 0	3596	84.811
TenYearCHD 1	644	15.189

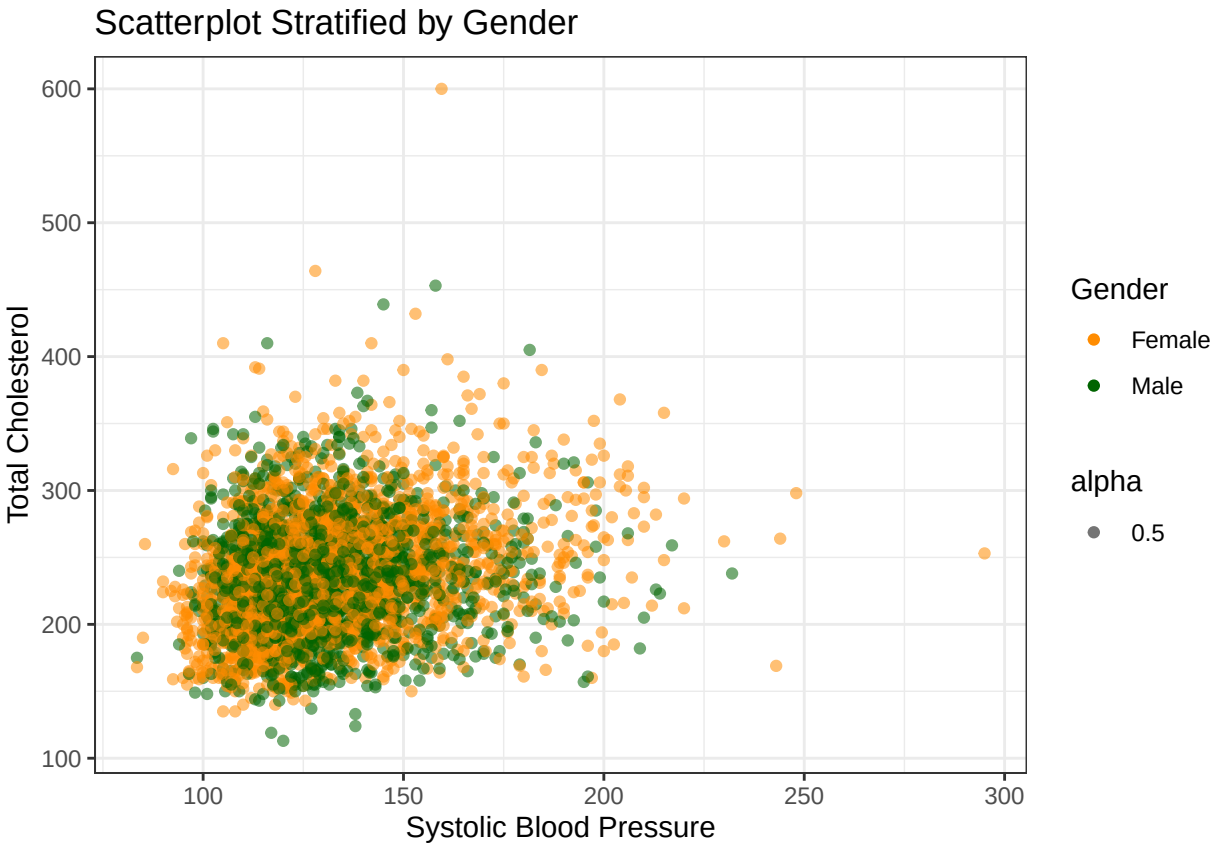
Question 6

- Original dataset rows: 4240
- Rows after removal: 3658

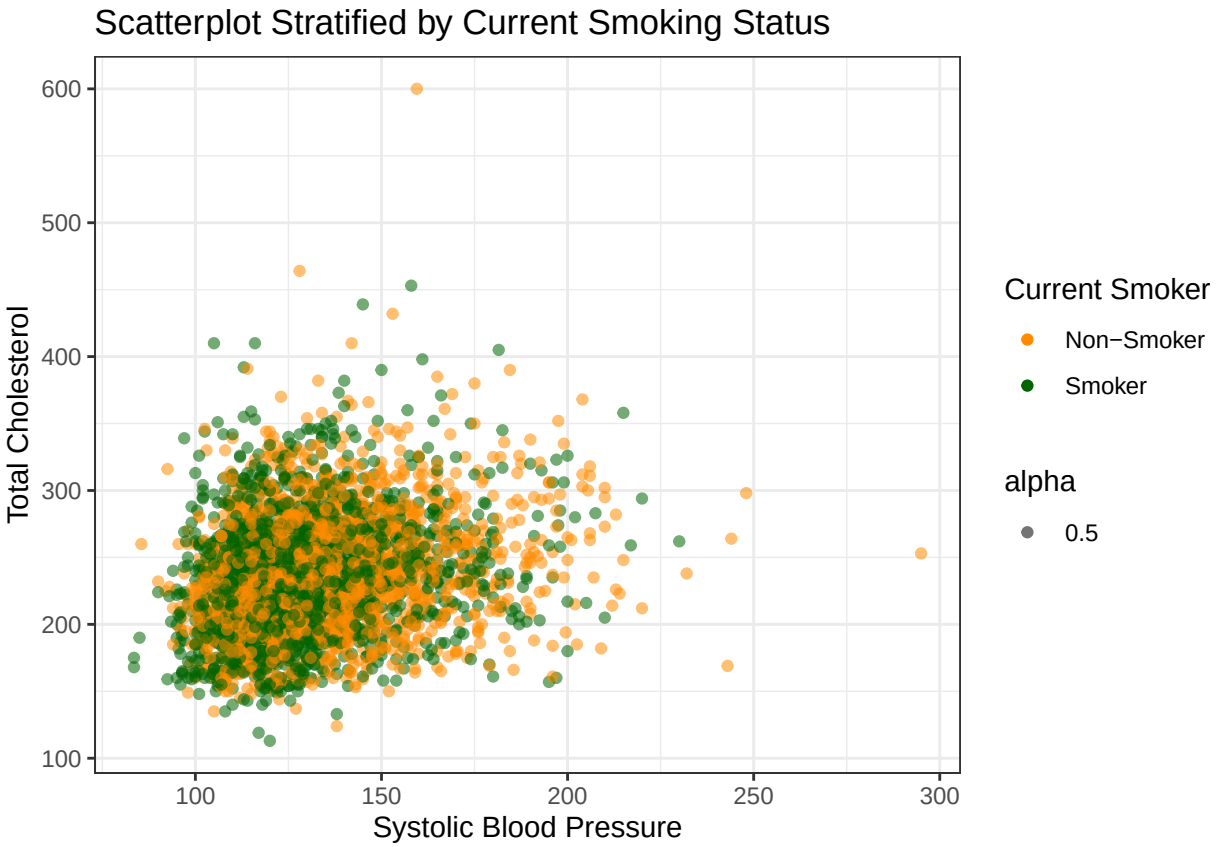
Question 7


Scatterplot of Total Cholesterol by Systolic Blood Pressure
Total Cholesterol
600
500
400
300
200
100
100
150
200
250
300
Systolic Blood Pressure
alpha
0.5

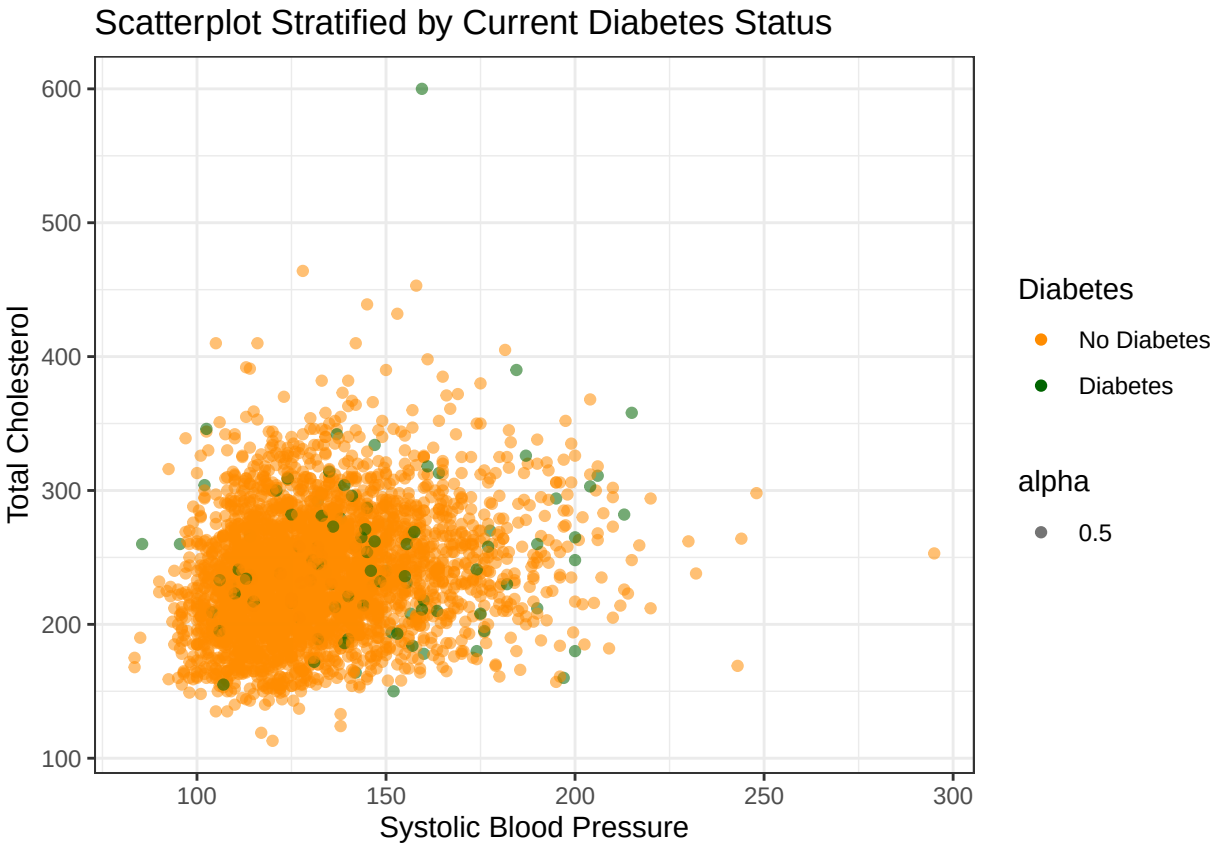
Question 8


Scatterplot Stratified by Gender
Total Cholesterol
600
500
400
300
200
100
100
150
200
250
300
Systolic Blood Pressure
Gender
Female
Male
alpha
0.5

Question 9


Scatterplot Stratified by Current Smoking Status
Total Cholesterol
600
500
400
300
200
100
100
150
200
250
300
Systolic Blood Pressure
Current Smoker
Non−Smoker
Smoker
alpha
0.5

Question 10



Question 11

- After conducting the exploratory analysis we can see that the data set is fairly average. The data includes smokers that smoke about 20 cigarettes a day and have a slightly above average **BMI**. Most of the patients also have an elevated amount of **total cholesterol**. This is typical since we are studying the affect of **total cholesterol** on **systolic BP**. There are no outliers of concern at this time.

Question 12

Term	Estimate	Std..Error	P.value
Intercept	107.900	1.798	< 0.001
Total Cholesterol	0.103	0.007	< 0.001

Question 13

- When comparing the the simple model with the new model we see that the p-value still shows that **total cholesterol** is significant. However, there is a decrease in the estimate of **total cholesterol** that means the introduction of **BMI** and **current smoker status** to the model improves the model.

Table 5: Linear Regression Model Coefficients

Term	Estimate	Std..Error	t.value	P.value
Intercept	72.769	2.579	28.211	< 0.001
Total Cholesterol	0.085	0.007	11.931	< 0.001
BMI	1.592	0.079	20.203	< 0.001
Current Smoker	-3.267	0.639	-5.113	< 0.001

Question 14

Table 6: Linear Regression Model Coefficients

Term	Estimate	Std_Error	P_value
Intercept	-17.065	2.183	< 0.001
Total Cholesterol	0.011	0.005	0.0240
Male	-3.889	0.436	< 0.001
Age	0.591	0.025	< 0.001
Diabetes	1.440	1.576	0.3609
DiaBP	1.347	0.019	< 0.001
BMI	0.113	0.055	0.0382
Current Smoker	0.351	0.650	0.5890
Cig Per Day	0.040	0.028	0.1564
Glucose	0.051	0.011	< 0.001

Question 15

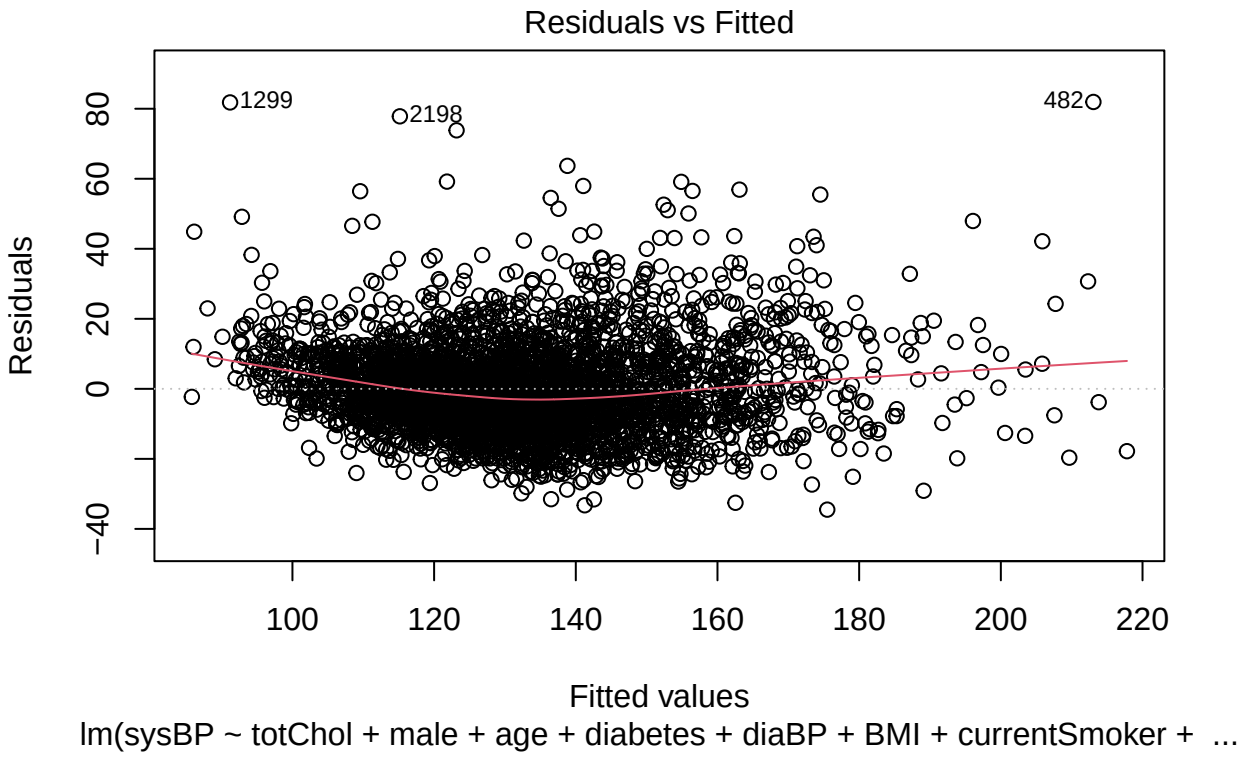
- The goodness of fit test shows that when comparing my model with the simple model, my model provides as significantly better fit for predicting **systolic BP**. This can be interpreted by the reduction in RSS. This indicates that my model explains a much larger portion of the variance in **systolic BP**. The significant F-statistic from the summary also suggests that the predictors added to my model improve the overall model.

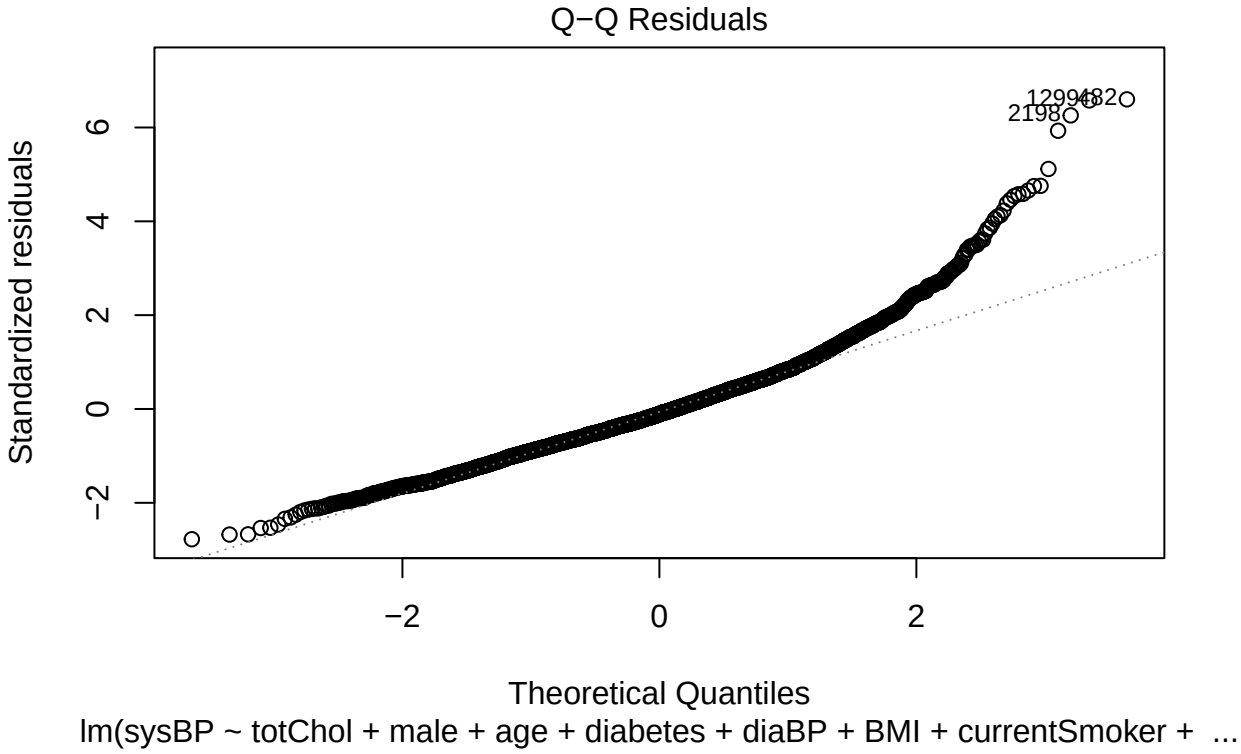
Table 7: ANOVA Results for Comparing Two Models

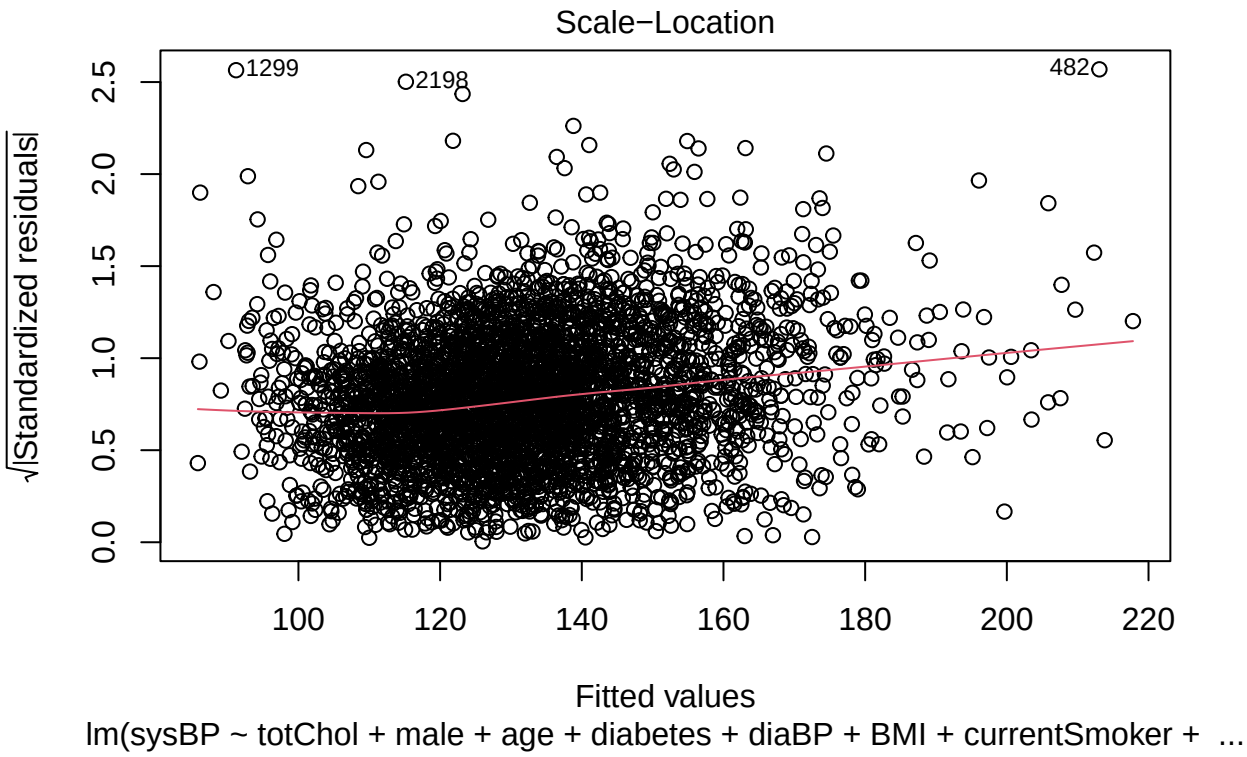
Model	Res.Df	RSS	Df	Sum of Sq	F	P-value
Model 1: sysBP ~ totChol	3656	1697707	-	-	-	-
Model 2: sysBP ~ totChol + male + age + diabetes + diaBP + BMI + currentSmoker + cigsPerDay + glucose	3648	566305	8	1131402	911.03	< 0.001

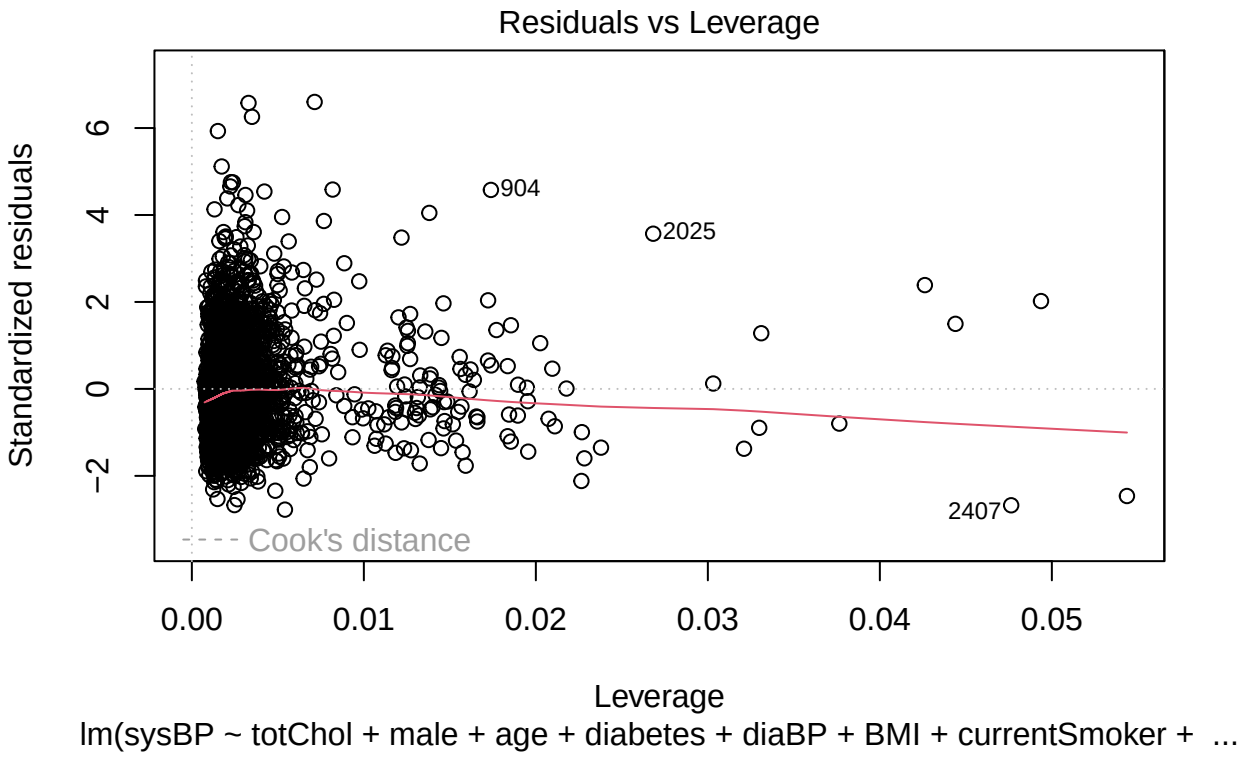
Question 16

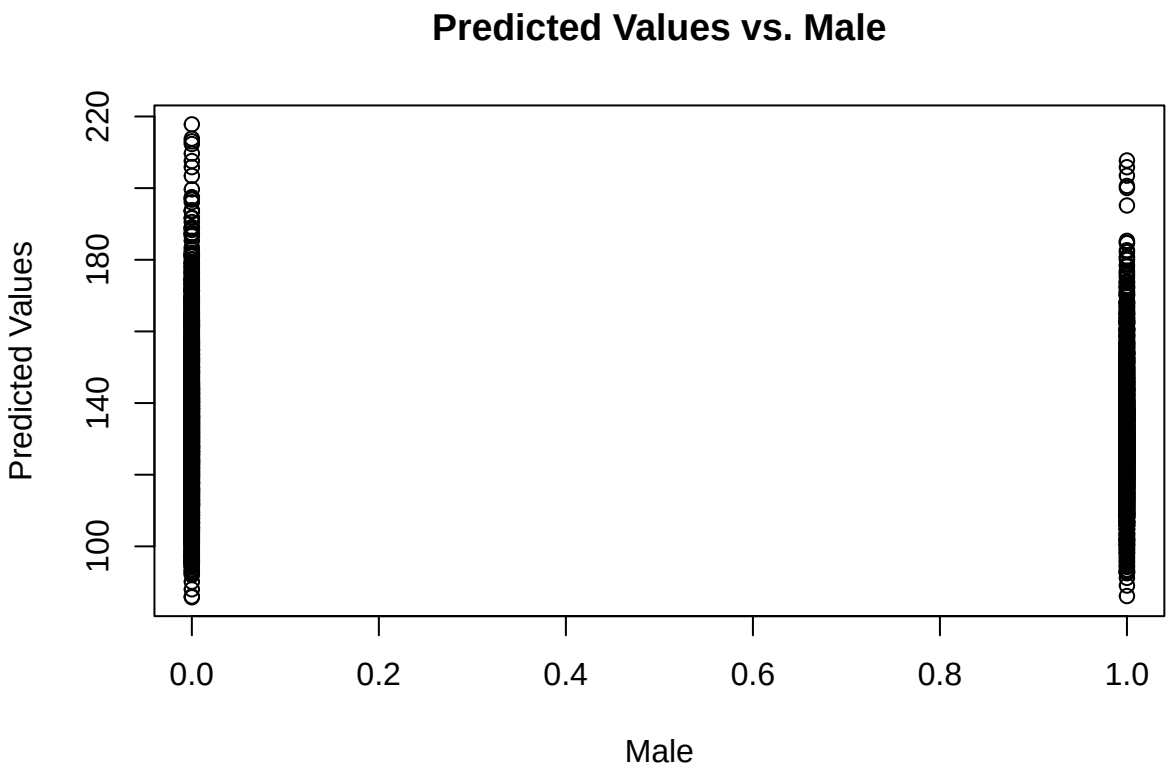
- The Residuals vs. Fitted plot might show a slight funnel shape. This plot indicates the equal variance, but our data does not seem to be concerning.
- The Q-Q plot is the only one that indicates a linear regression assumption may be violated. The plot shows the the tail ends starting to turn upwards. This could indicate that the normaility assumption is violated.
- No comments on the other two plots as they look normal.

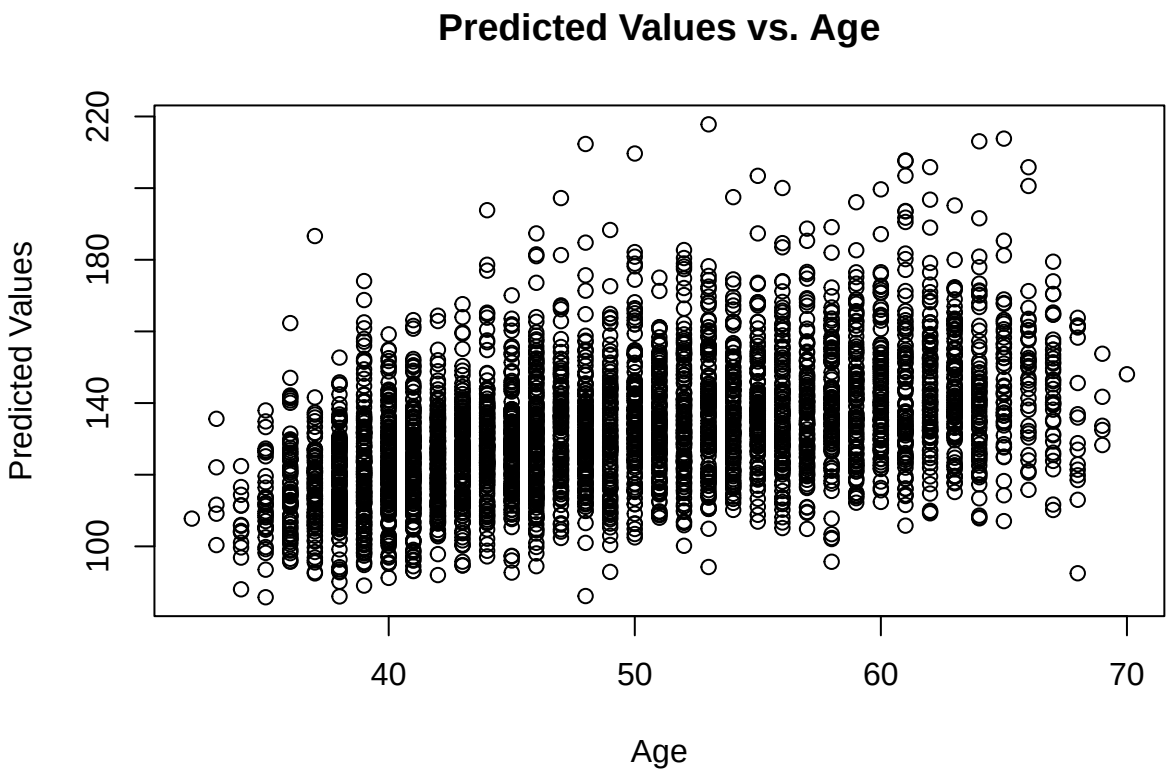

Residuals vs Fitted
1299
2198
482
Residuals
80
60
40
20
0
−40
100
120
140
160
180
200
220
Fitted values
lm(sysBP ~ totChol + male + age + diabetes + diaBP + BMI + currentSmoker + ...

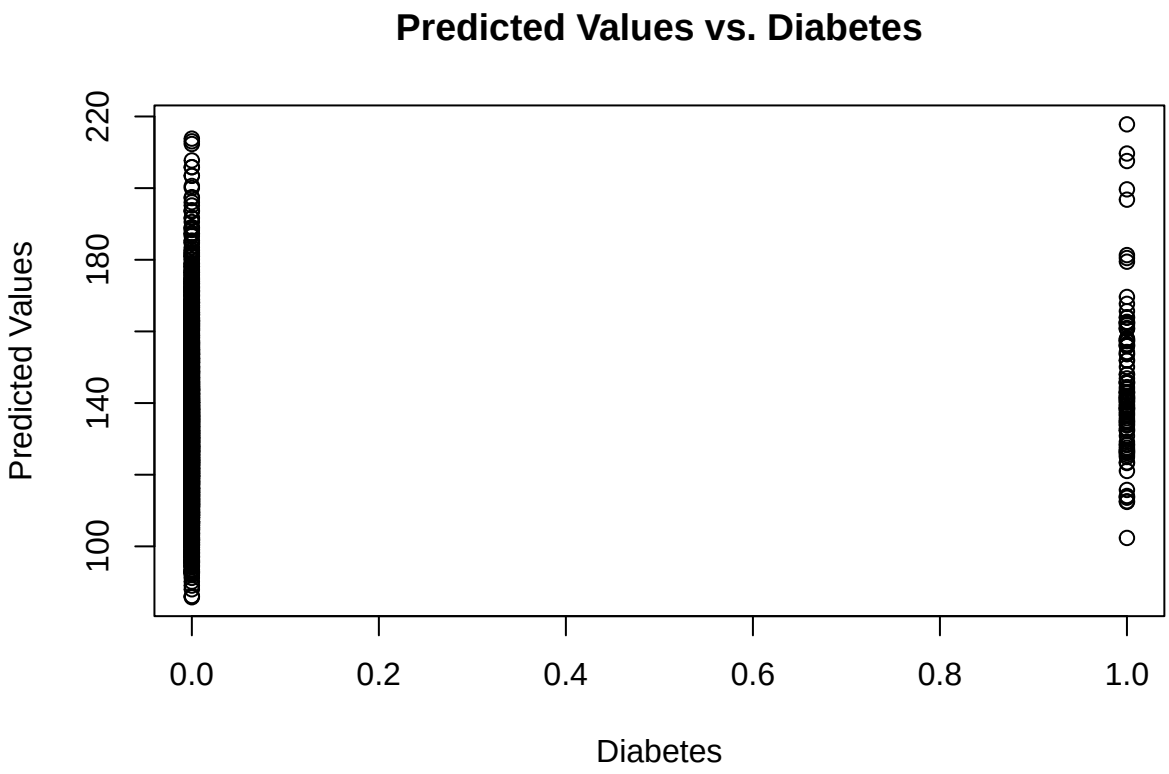

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(sysBP ~ totChol + male + age + diabetes + diaBP + BMI + currentSmoker + ...
2198
−2
0
2
4
6

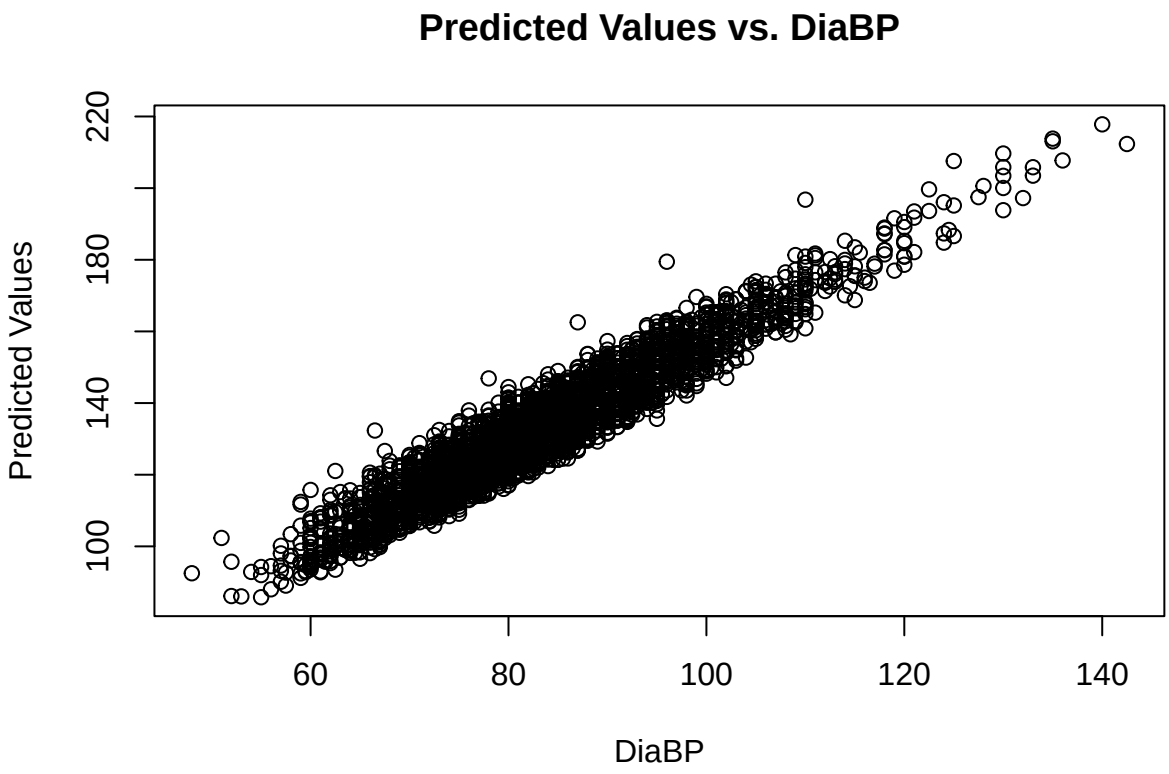

Scale−Location
1299
2198
482
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
100
120
140
160
180
200
220
Fitted values
lm(sysBP ~ totChol + male + age + diabetes + diaBP + BMI + currentSmoker + ...

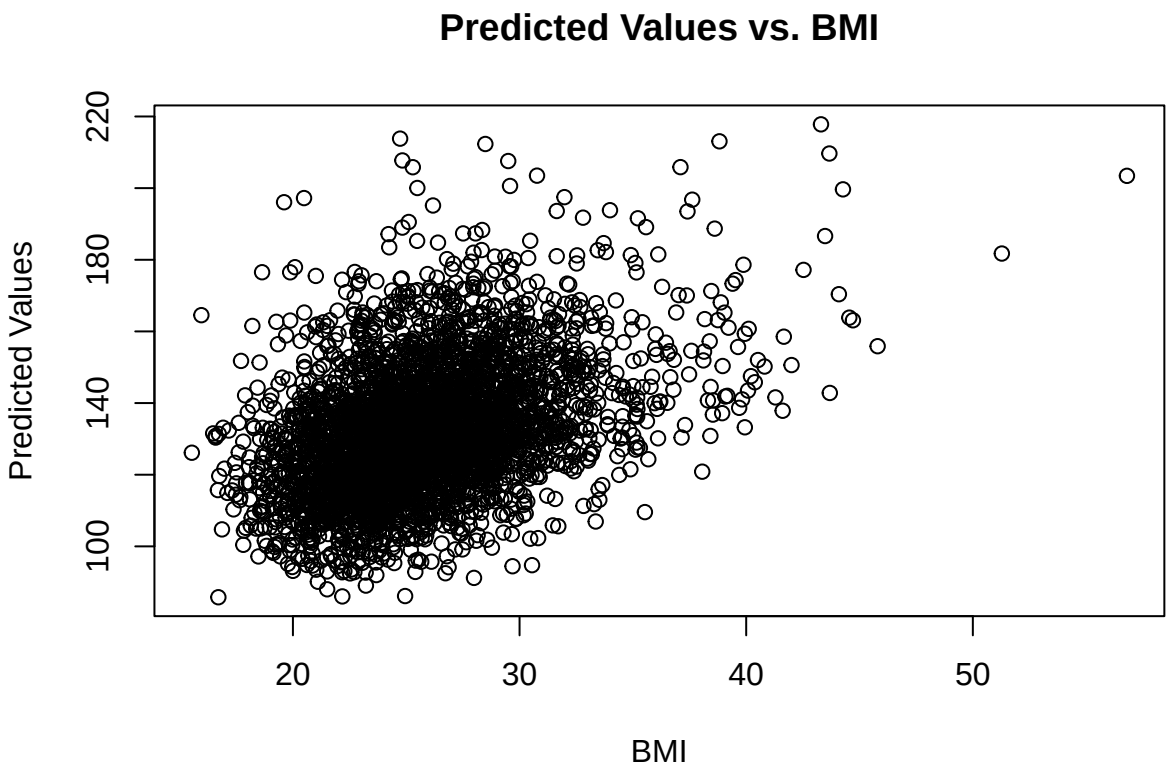

Residuals vs Leverage
Standardized residuals
6
4
2
0
−2
904
2025
2407
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
lm(sysBP ~ totChol + male + age + diabetes + diaBP + BMI + currentSmoker + ...

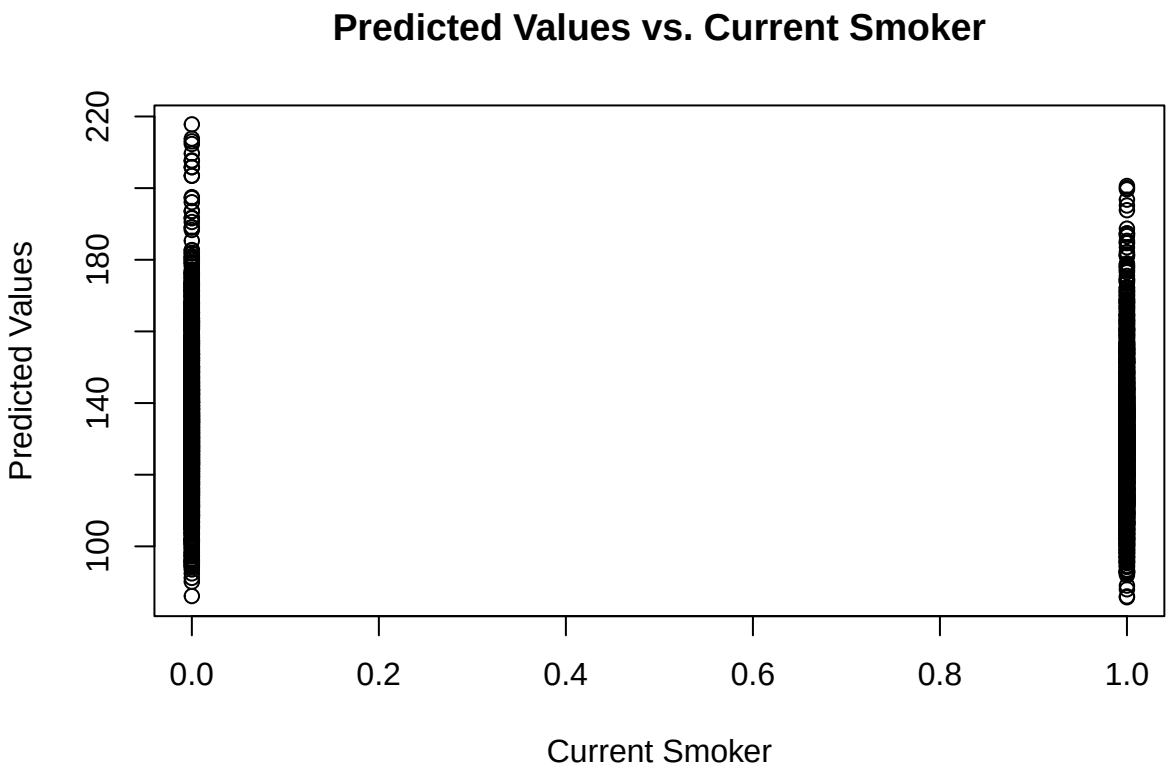

Predicted Values vs. Male
Predicted Values
100
140
180
220
0.0
0.2
0.4
0.6
0.8
1.0
Male

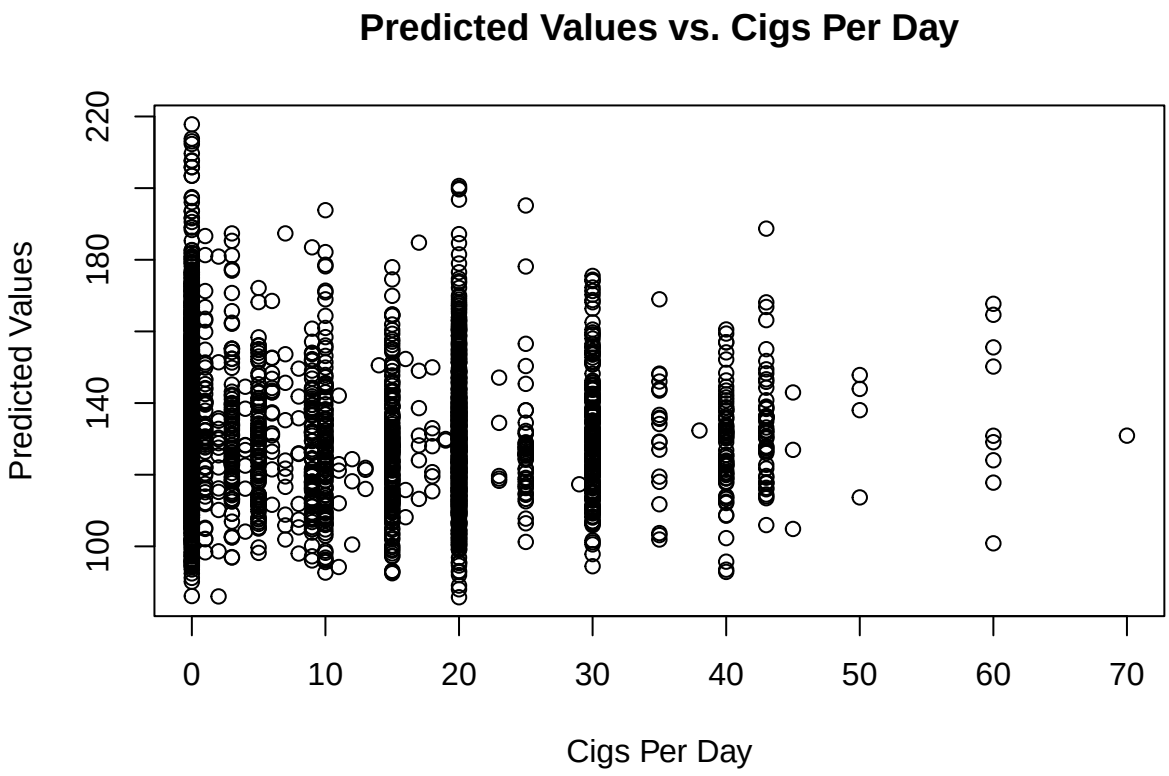

Predicted Values vs. Age
Predicted Values
100
140
180
220
40
50
60
70
Age

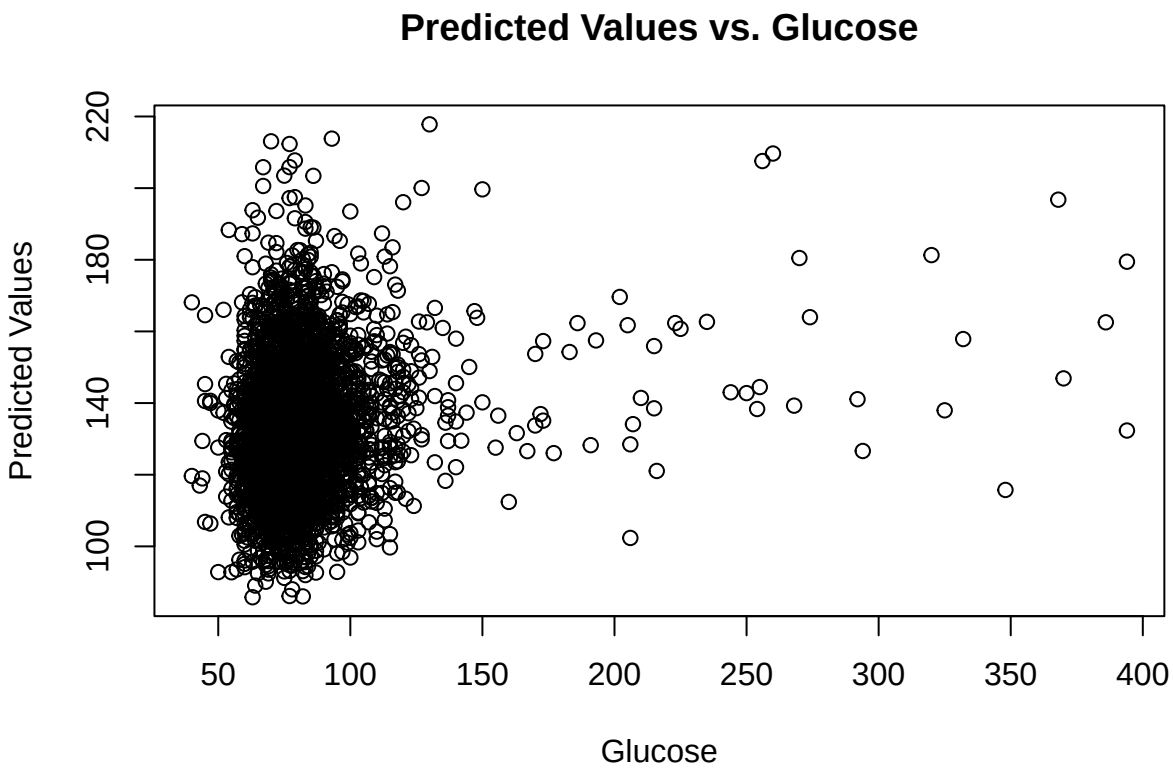

Predicted Values vs. Diabetes
Predicted Values
100
140
180
220
0.0
0.2
0.4
0.6
0.8
1.0
Diabetes


Predicted Values vs. DiaBP
Predicted Values
100
140
180
220
60
80
100
120
140
DiaBP


Predicted Values vs. BMI
Predicted Values
100
140
180
220
20
30
40
50
BMI


Predicted Values vs. Current Smoker
Predicted Values
100
140
180
220
0.0
0.2
0.4
0.6
0.8
1.0
Current Smoker


Predicted Values vs. Cigs Per Day
Predicted Values
100
140
180
220
0
10
20
30
40
50
60
70
Cigs Per Day



Question 17

- The proposed model indicates that there is a significant association between cholesterol and blood pressure. The Linear Regression Model Summary table below highlights the reported significance. The lack of statistical significance for **diabetes** (P-value=0.36) and current **smoking status** (P-value=0.589) was alarming. Typically someone would assume these variables would impact a persons total cholesterol levels.
- I compared an additional model of **sysBP ~ totChol + male + age + diaBP + BMI + glucose** to my original model. While I removed some covariates that where not significant in my original model, this did not make the additional model better than my original model in a goodness of fit test. This indicates to me that while **diabetes, currentSmoker, and cigsPerDay** may not be significant to the model, they are precision variables that increase the overall effectiveness of the model.

Table 8: Linear Regression Model Summary

Term	Estimate	Std_Error	P_value
Intercept	-17.065	2.183	< 0.001
Total Cholesterol	0.011	0.005	0.0240
Male	-3.889	0.436	< 0.001
Age	0.591	0.025	< 0.001
Diabetes	1.440	1.576	0.3609
DiaBP	1.347	0.019	< 0.001
BMI	0.113	0.055	0.0382

Term	Estimate	Std_Error	P_value
Current Smoker	0.351	0.650	0.5890
Cig Per Day	0.040	0.028	0.1564
Glucose	0.051	0.011	< 0.001