JMU_MIDTERM

Joanne Unite

#Background The data set I chose to analyze contains clinical data where patients were examined for cardiovascular and kidney function. This data set was created to help develop prediction models for heart disease, diabetes, and impaired kidney function.

The link to where I downloaded the data is provided: https://www.kaggle.com/datasets/simaanjali/diabetes-classification-dataset/data?select=Diabetes+Classification.csv

```
blood_df <- read.csv("/Users/junite/Desktop/BIFX551/BIFX551_SPRING24/JMU_Week 4/blood_test.csv")
```

##Install all necessary packages

```
library(readr)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggsci)
library(egg)
```

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

###To begin we will take a look at the summary data and the first few columns of the date

```
summary(blood_df)
```

```
##        X             Age           Gender               BMI
##  Min.   :   0   Min.   :20.00   Length:5132        Min.   :15.00
##  1st Qu.:1283   1st Qu.:36.00   Class :character   1st Qu.:22.00
##  Median :2566   Median :49.00   Mode  :character   Median :24.00
##  Mean   :2566   Mean   :48.95                      Mean   :24.61
##  3rd Qu.:3848   3rd Qu.:59.00                      3rd Qu.:27.00
##  Max.   :5131   Max.   :93.00                      Max.   :47.00
```

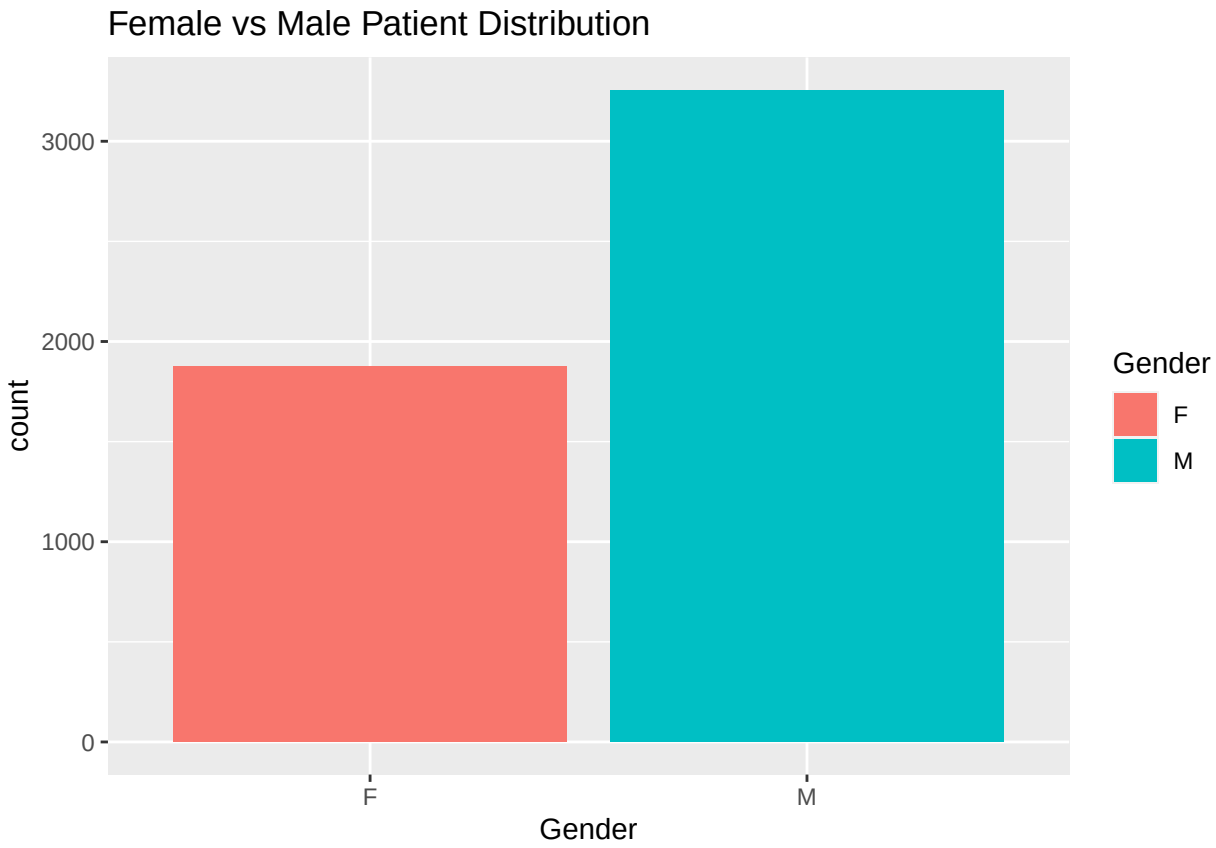
```
##       Chol              TG              HDL             LDL
##  Min.   : 0.000   Min.   : 0.000   Min.   :0.000   Min.   :0.300
##  1st Qu.: 4.190   1st Qu.: 0.910   1st Qu.:1.090   1st Qu.:2.290
##  Median : 4.800   Median : 1.380   Median :1.300   Median :2.790
##  Mean   : 4.867   Mean   : 1.719   Mean   :1.593   Mean   :2.914
##  3rd Qu.: 5.460   3rd Qu.: 2.100   3rd Qu.:1.590   3rd Qu.:3.400
##  Max.   :11.650   Max.   :32.640   Max.   :9.900   Max.   :9.900
##        Cr               BUN           Diagnosis
##  Min.   :  4.861   Min.   : 0.500   Min.   :0.0000
##  1st Qu.: 58.000   1st Qu.: 3.900   1st Qu.:0.0000
##  Median : 70.200   Median : 4.720   Median :0.0000
##  Mean   : 71.145   Mean   : 4.897   Mean   :0.3883
##  3rd Qu.: 81.600   3rd Qu.: 5.600   3rd Qu.:1.0000
##  Max.   :800.000   Max.   :38.900   Max.   :1.0000
```

```
head(blood_df)
```

```
##   X Age Gender BMI Chol  TG HDL LDL Cr BUN Diagnosis
## 1 0  50      F  24  4.2 0.9 2.4 1.4 46 4.7         0
## 2 1  26      M  23  3.7 1.4 1.1 2.1 62 4.5         0
## 3 2  33      M  21  4.9 1.0 0.8 2.0 46 7.1         0
## 4 3  45      F  21  2.9 1.0 1.0 1.5 24 2.3         0
## 5 4  50      F  24  3.6 1.3 0.9 2.1 50 2.0         0
## 6 5  48      M  24  2.9 0.8 0.9 1.6 47 4.7         0
```

###Next we will create some exploratory plots to better understand the data ###Female vs Male distribution First, I want to see the distribution of how many female to male patients are used in the data. Some of the characters in the "Gender" column were in lowercase which resulted in 3 columns instead of 2 so I mutated the values in the column to be all uppercase. The graph below shows that nearly double the amount of males were documented over females.

```
blood_df <- blood_df %>%
  mutate(Gender = toupper(Gender))
ggplot(blood_df, aes(x =Gender, fill = Gender)) +
  geom_bar() +
  ggtitle("Female vs Male Patient Distribution")
```



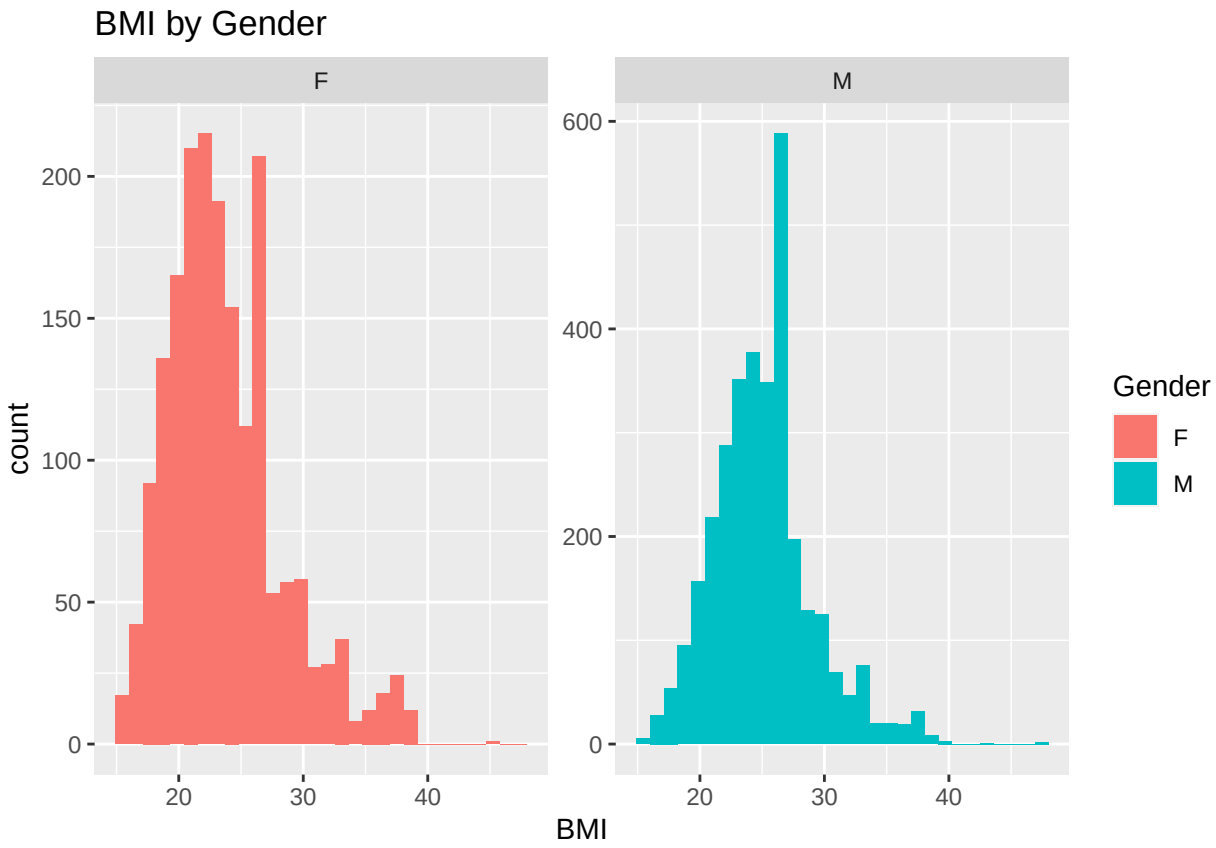
###Next I would like to explore the relationship between Age and Cholesterol I am curious to know if there is a relationship between high cholesterol and age. Based on brief research the normal range for total cholesterol is less than 5.2 mmol/L, at risk is between 5.2 - 6.2 mmol/L, and a dangerously high cholesterol is greater than 6.2mmol/L. We can see that the younger the patients are, the tighter the cluster of dots is. When we look at the older patients we see that the cluster is a bit more dispersed.

```
ggplot(blood_df,aes(x = Age, y = Chol, color = Gender)) +
  geom_point() +
  ggtitle("Cholesterol levels vs Age")
```

####Males vs Female BMI Now let's take a look at the differences between the BMI of males and females in the study. We see that generally females have a lower BMI than men. The average healthy BMI for a woman in the United states is 26.5 and for a man it is 26.6 (according to the CDC).

```
p <- ggplot(blood_df, aes(x = BMI, fill = Gender)) +
  geom_histogram() +
  ggtitle("BMI by Gender")
p + facet_wrap(~Gender, scales = "free_y")
```

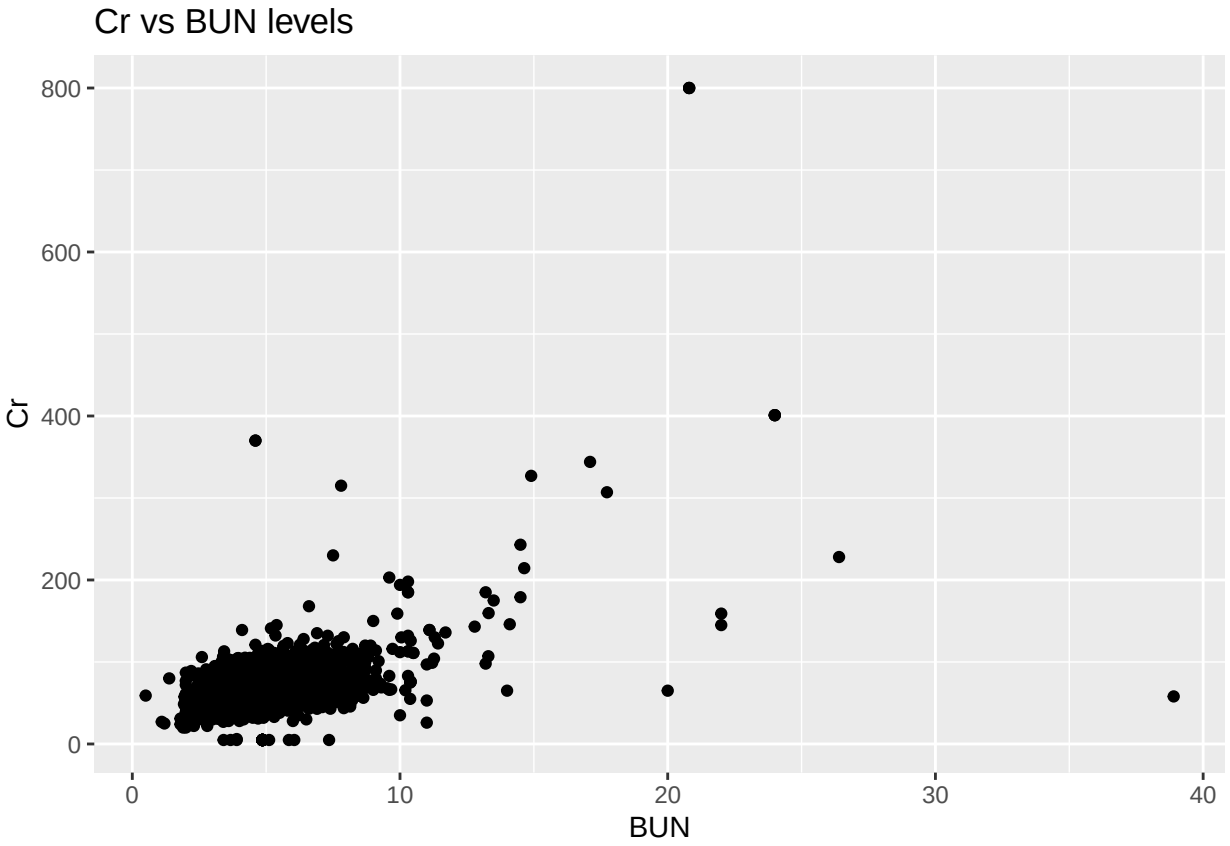
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



###Analysis Goal I aim to analyze the relationship between the indicators of kidney disease and heart disease/diabetes.I am curious to analyze if there is a pattern between the indicators Cr and BUN, if both tend to be high at the same time or otherwise. I would like to explore the same for TG and LDL. Additionally, I would like to explore any major differences between the men and women who are diagnosed with diabetes.

###Analysis of Kidney disease indicators The indicators for Kidney disease in the blood test include: Cr (Creatinine) and BUN (Blood Urea Nitrogen). Creatinine is a waste product of muscle metabolism and high levels of creatinine can be indicative poor kidney function. Similarly, blood urea nitrogen is indicative of poor kidney and liver function.

```
ggplot(blood_df, aes(x = BUN, y = Cr)) +
  geom_point() +
  ggtitle("Cr vs BUN levels")
```



>The graphs above gives us some insight into the reltionship between Cr and BUN levels. The normal levels for BUN is 2.1 to 8.5 mmol/L. We can see several outliers of both Cr and BUN but we cannot confidently say there is indeed a significant correlation between the two. To truly test if there is a pattern, we will run a t-test on the data. The results show a p-value of <2.2e-16, therefore we can reject the null hypothesis. This confirms that there is a significant relationship between Cr and BUN levels.

```
t.test(blood_df$Cr, blood_df$BUN)
```

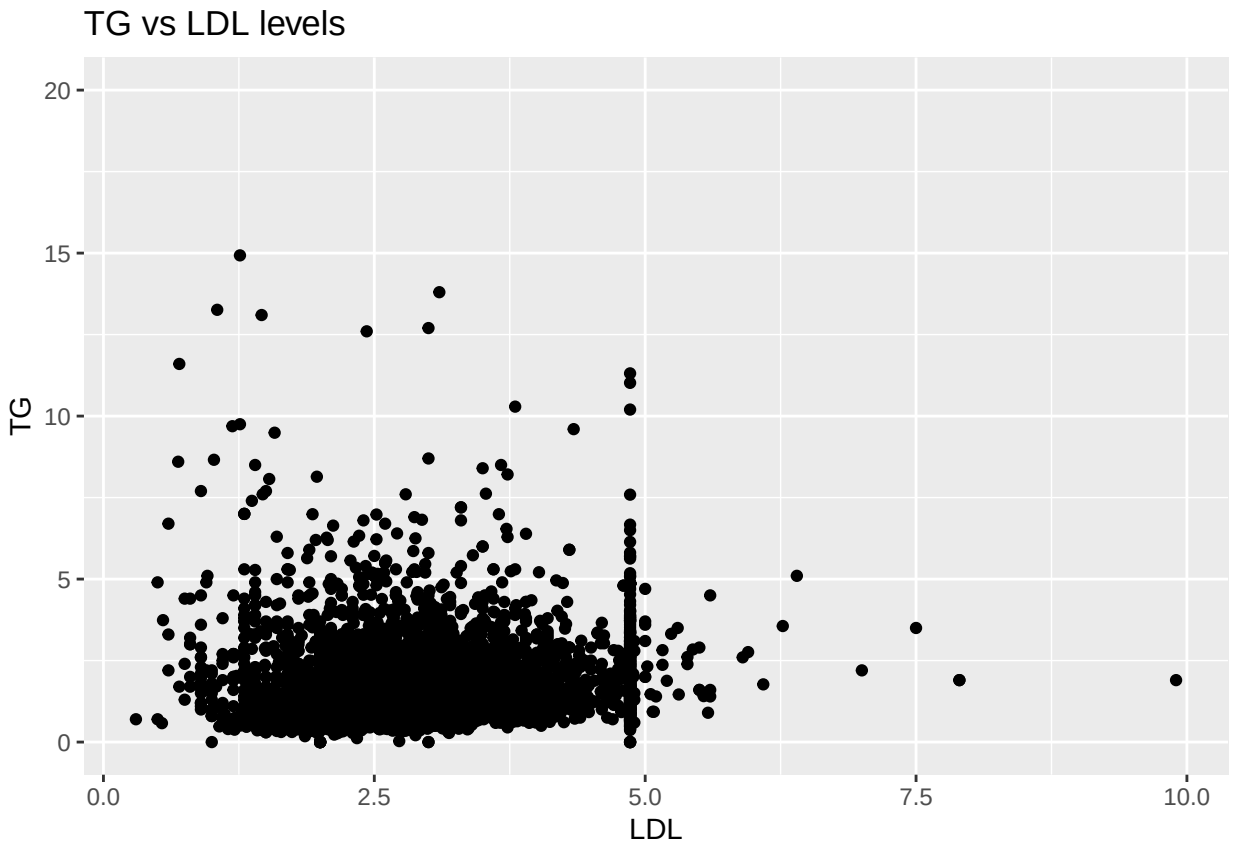
```
##
##  Welch Two Sample t-test
##
## data:  blood_df$Cr and blood_df$BUN
## t = 166.26, df = 5167.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  65.46669 67.02897
## sample estimates:
## mean of x mean of y
## 71.144800  4.896969
```

###Analysis of Heart Disease/Diabetes indicators The blood test indicators for Heart Diseases are TG (Triglycerides) and LDL (Low-Density Lipoprotein). Triglycerides are a type of fat found in the blood; high levels of Triglycerides increase the risk of heart disease and diabetes. Low-Density Lipoprotein is referred to as “bad cholesterol”, which can build up in the arteries. This increases the risk of heart attack, stroke, and diabetes.

```
ggplot(blood_df, aes(x = LDL, y = TG)) +
  geom_point() +
  ylim(0,20) +
```

```
  ggtitle("TG vs LDL levels")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



>The graph above shows us a large cluster in the normal ranges, however there does not seem to be a correlation between the two indicators. Below we can see a t-test performed on the data results in a P-value of < 2.2e-16, which confirms a significant relationship between TG and LDL.

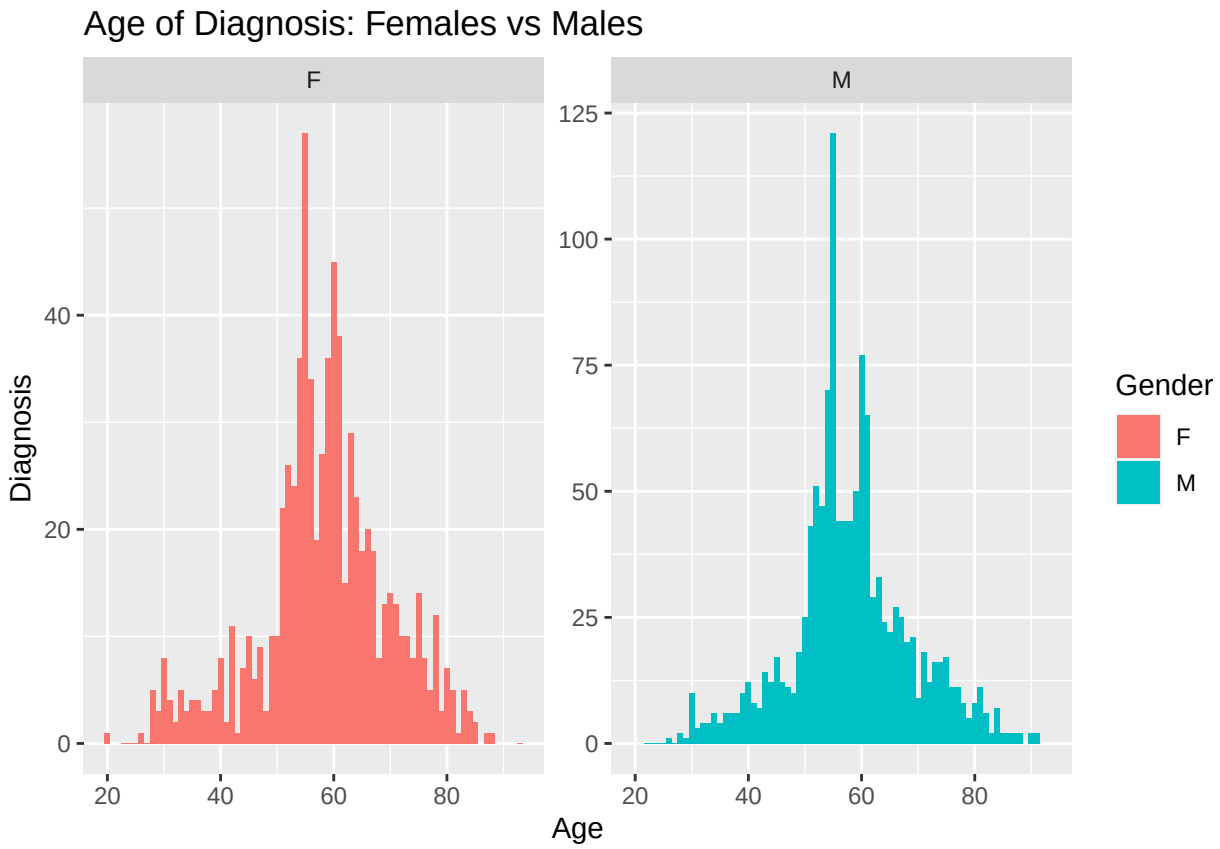
```
t.test(blood_df$TG, blood_df$LDL)
```

```
##
##  Welch Two Sample t-test
##
## data:  blood_df$TG and blood_df$LDL
## t = -52.53, df = 9272.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.239377 -1.150208
## sample estimates:
## mean of x mean of y
##  1.719328  2.914121
```

###Males vs Female age of diagnosis Next we will explore the differences is the age of diagnosis for Males vs. Females. In the two graphs below we see that in both the Male and Female plots, the highest amounts of diagnoses were given to people aged between 50-70 years old. We observe some higher peaks between the ages of 25-45 in women compared to the men.

```
p <- ggplot(blood_df, aes(x = Age, y = Diagnosis, fill = Gender)) +
  geom_col() +
```

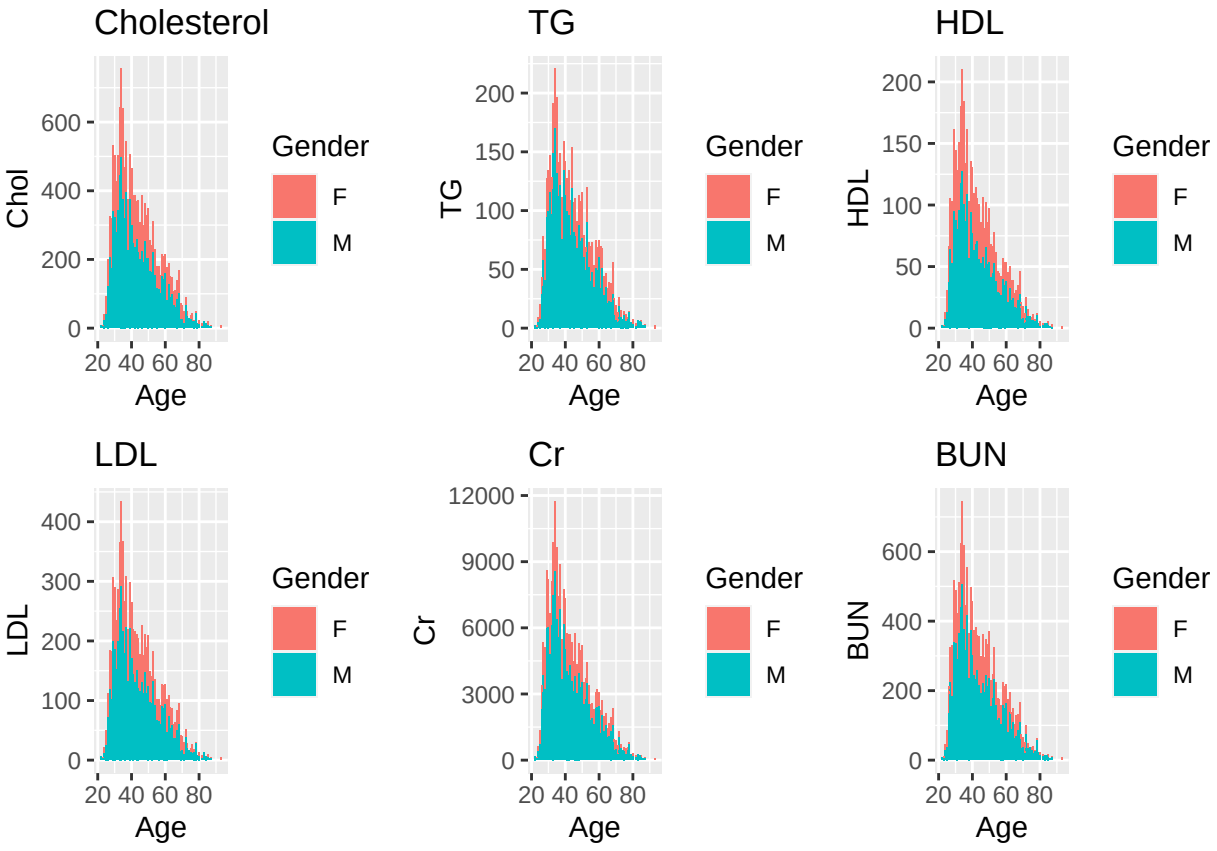
```
  ggtitle("Age of Diagnosis: Females vs Males")
p + facet_wrap(~Gender, scales = "free_y")
```



###Comparing blood test results of Diagnosed Males vs Females Let's look at the difference of indicator levels between diagnosed males and females. First let's look at the diagnosed levels of men and women. Then let's create the same plots for diagnosed patients. The plots below show that across all the indicators, women show higher levels than men. Additionally, the diagnosed patients have similar values as the diagnosed patients. This suggests that the diagnosed patients likely have the same diseases.

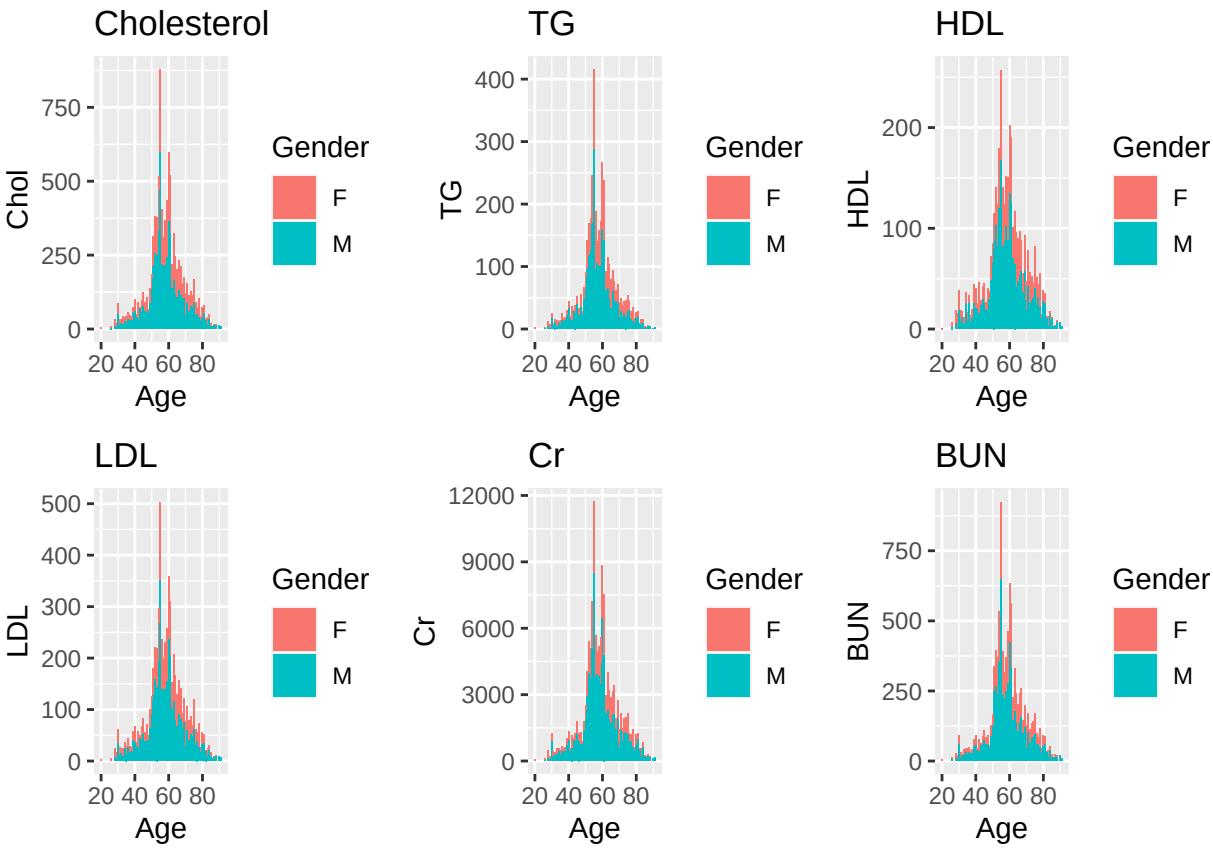
```
blood_ctrl <- blood_df %>% group_by(Gender) %>% filter(Diagnosis == 0)

a_1 <- ggplot(blood_ctrl, aes(x = Age, y = Chol, fill = Gender)) +
  geom_col() +
  ggtitle("Cholesterol")

b_1 <- ggplot(blood_ctrl, aes(x = Age, y = TG, fill = Gender)) +
  geom_col() +
  ggtitle("TG")

c_1 <- ggplot(blood_ctrl, aes(x = Age, y = HDL, fill = Gender)) +
  geom_col() +
  ggtitle("HDL")

d_1 <- ggplot(blood_ctrl, aes(x = Age, y = LDL, fill = Gender)) +
  geom_col() +
  ggtitle("LDL")

e_1 <- ggplot(blood_ctrl, aes(x = Age, y = Cr, fill = Gender)) +
```

```
  geom_col() +
  ggtitle("Cr")

f_1 <- ggplot(blood_ctrl, aes(x = Age, y = BUN, fill = Gender)) +
  geom_col() +
  ggtitle("BUN")

ggarrange(a_1,b_1,c_1,d_1,e_1,f_1,ncol = 3, nrow = 2)
```




```
blood_new <- blood_df %>% group_by(Gender) %>% filter(Diagnosis == 1)

a_2 <- ggplot(blood_new, aes(x = Age, y = Chol, fill = Gender)) +
  geom_col() +
  ggtitle("Cholesterol")

b_2 <- ggplot(blood_new, aes(x = Age, y = TG, fill = Gender)) +
  geom_col() +
  ggtitle("TG")

c_2 <- ggplot(blood_new, aes(x = Age, y = HDL, fill = Gender)) +
  geom_col() +
  ggtitle("HDL")

d_2 <- ggplot(blood_new, aes(x = Age, y = LDL, fill = Gender)) +
  geom_col() +
  ggtitle("LDL")
```

```
e_2 <- ggplot(blood_new, aes(x = Age, y = Cr, fill = Gender)) +
  geom_col() +
  ggtitle("Cr")

f_2 <- ggplot(blood_new, aes(x = Age, y = BUN, fill = Gender)) +
  geom_col() +
  ggtitle("BUN")

ggarrange(a_2,b_2,c_2,d_2,e_2,f_2,ncol = 3, nrow = 2)
```



###Conlusion My analysis explored the relationship between indicators for kidney disease and indicators for heart disease/diabetes. I also explore the major differences between the blood test levels of diagnosed males vs females. This analysis revealed that overall women showed higher levels in all 6 indicators than men. Interestingly, the patients who have not been diagnosed show similar levels to that of the diagnosed patients. These blood test levels are indicative of high risks to kidney disease, heart disease, and diabetes. The undiagnosed patients with higher levels of indicators in their blood tests, suggests that younger people may need to be tested if they are at high risk for kidney disease, heart disease, and diabetes