Final Lab

AUTHOR
Jessica Tran

Part 1

I will now load my packages.

```
library(ggplot2)
wcgs <- read.csv("wcgs.csv")
```

We will now find a few summary statistics.

```
mean(wcgs$sbp)
```

```
[1] 128.6328
```

```
mean(wcgs$weight)
```

```
[1] 169.9537
```

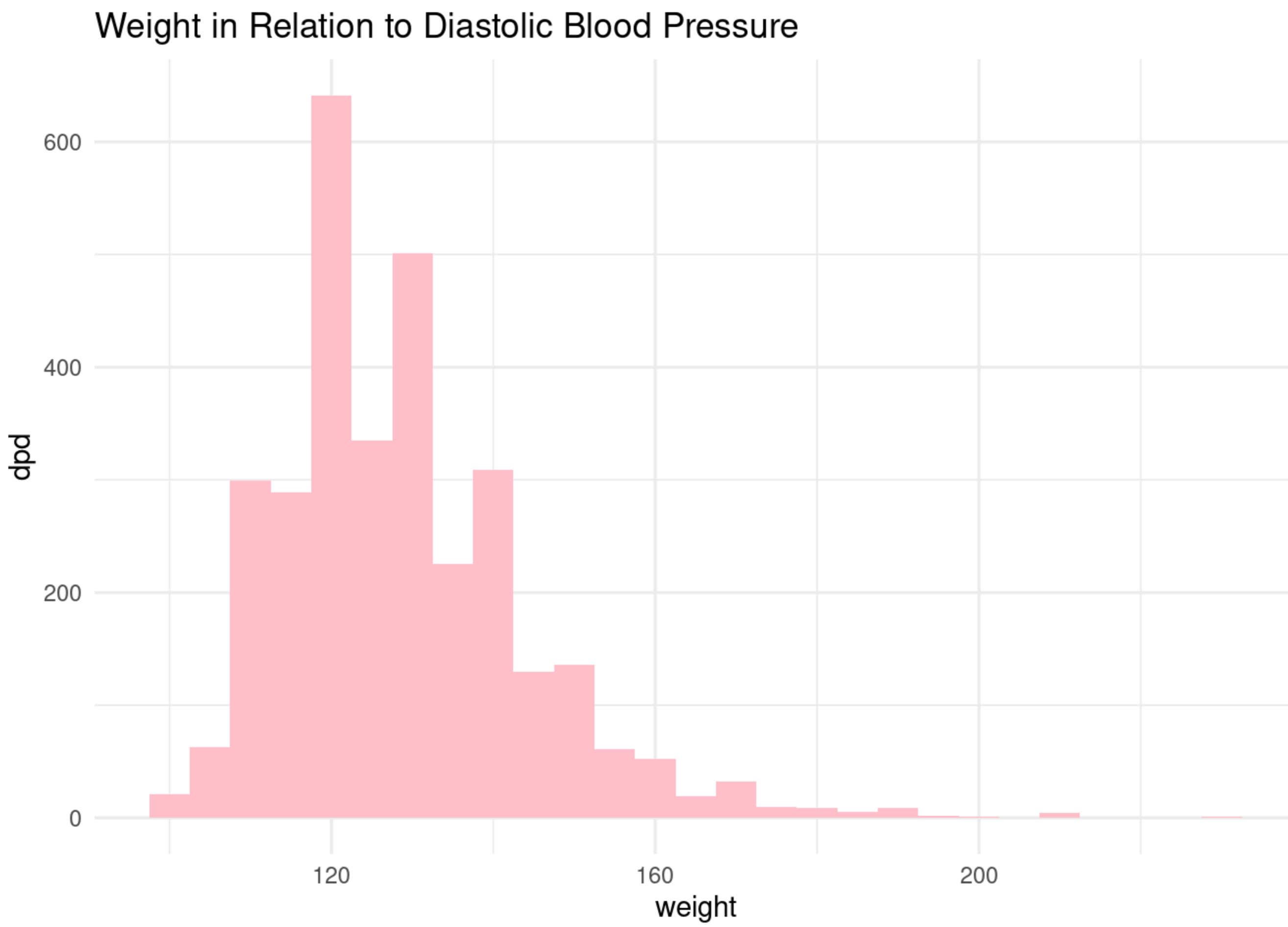
```
mean(wcgs$dbp)
```

```
[1] 82.01554
```

I found the mean of systolic blood pressure, weight, and diastolic blood pressure.

We will now create a visual graph showing the relation between variables weight and diastolic blood pressure.

```
ggplot(wcgs, aes(x=sbp)) +
  geom_histogram(binwidth = 5, fill="pink")+
  xlab("weight")+
  ylab("dpd")+
  ggtitle("Weight in Relation to Diastolic Blood Pressure") + theme_minimal()
```



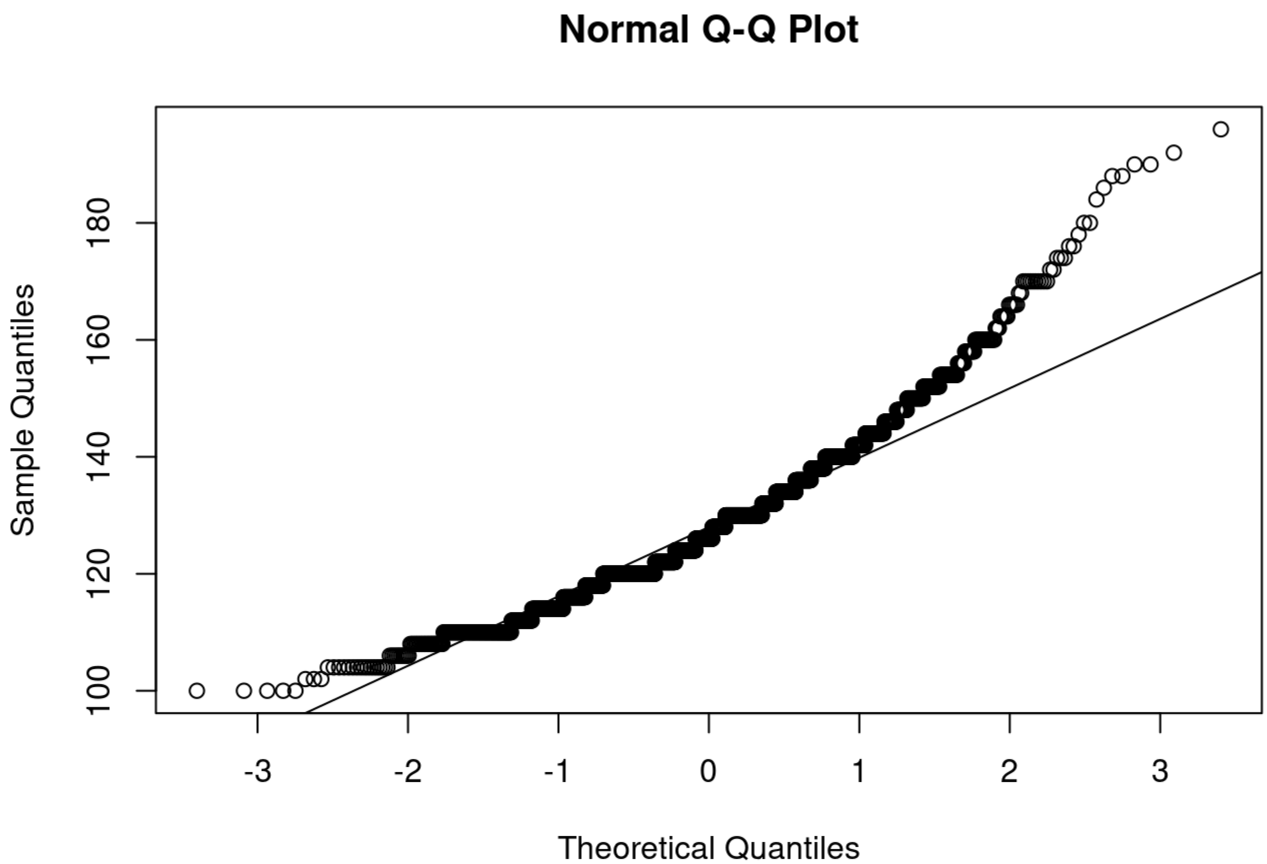
This histogram shows the relation between weight and diastolic blood pressure.

Part 2

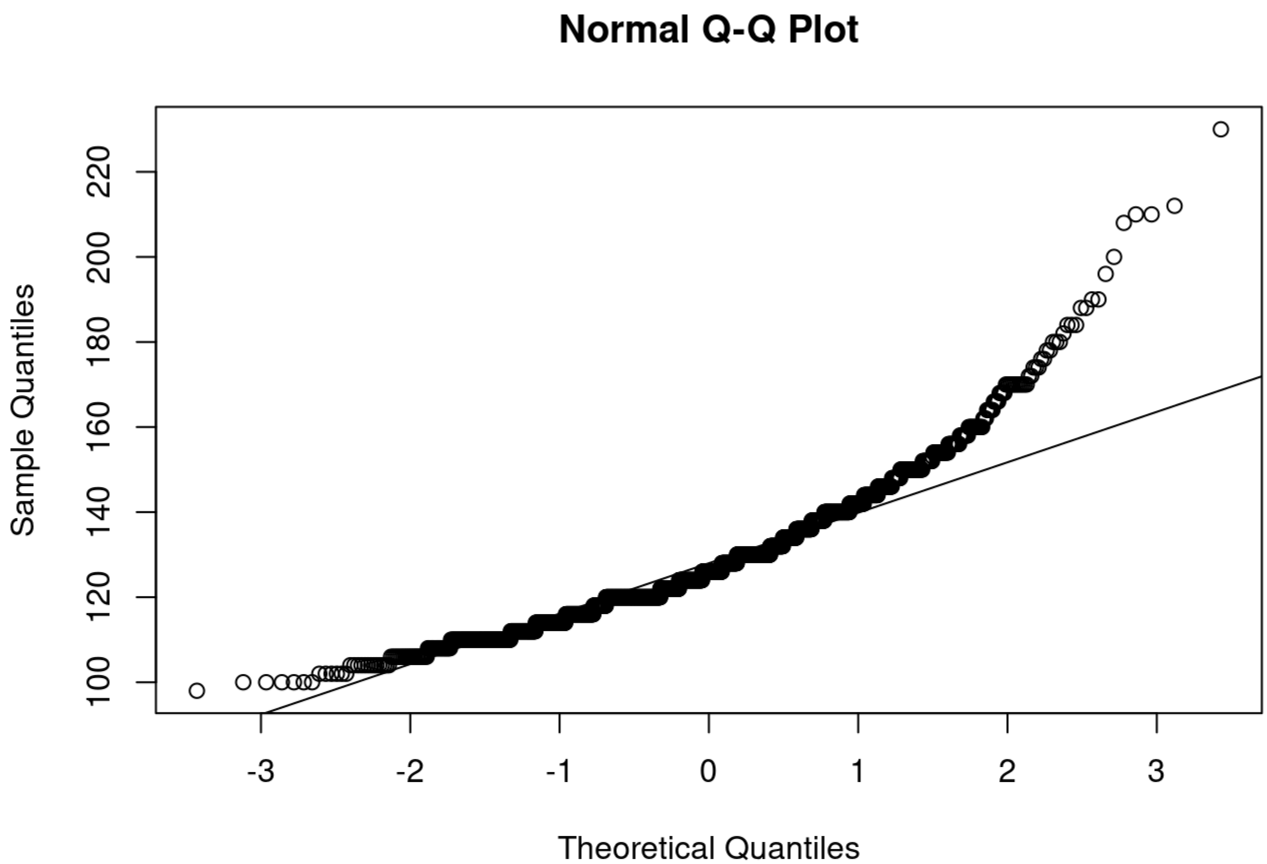
We will now conduct a hypothesis test for independent samples that considers the effects of smoking. Group 1 are the smokers while group 2 are the non-smokers.

```
group1<-with(wcgs,sbp[smoking == "Smoker"])
group2<-with(wcgs,sbp[smoking == "Non-smoker"])

qqnorm(group1)
qqline(group1)
```




```
qqnorm(group2)
qqline(group2)
```




```
var.test(group1,group2)
```

```
	F test to compare two variances

data:  group1 and group2
F = 0.88494, num df = 1501, denom df = 1651, p-value = 0.01555
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8017114 0.9770401
sample estimates:
ratio of variances
         0.8849404
```

We will now conduct a t.test.

```
t.test(sbp ~ smoking, data = wcgs, alternative = "two.sided", conf.level = 0.95,
```

```
	Two Sample t-test

data:  sbp by smoking
t = -0.14729, df = 3152, p-value = 0.8829
```

```
alternative hypothesis: true difference in means between group Non-smoker and group
Smoker is not equal to 0
95 percent confidence interval:
 -1.1363606  0.9775651
sample estimates:
mean in group Non-smoker     mean in group Smoker
                128.5950                 128.6744
```

Since the p-value is not less than 0.05, we fail to reject the null hypothesis. There is not enough evidence to suggest group 1 and group 2 had different systolic blood pressures.

Part 3

We can check the assumptions by running `chisq.test()`. We can check the expected counts with the following code:

```
output <- chisq.test(table(wcgs$personality))
output$expected
```

```
   A1    A2    B3    B4 
788.5 788.5 788.5 788.5 
```

We can find the p-value with the following code:

```
output$p.value
```

```
[1] 7.725469e-258
```

There were different proportions of personality types discovered.

Part 4

We are determining the most correlated variable by:

1. First, we will filter the data to only select numerical values.

```
filtered_data <- Filter(is.numeric, wcgs)
head(filtered_data)
```

```
  X   id age height weight sbp dbp chol ncigs timechd
1 1 2001  49     73    150 110  76  225    25    1664
2 2 2002  42     70    160 154  84  177    20    3071
3 3 2003  42     69    160 110  78  181     0    3071
4 4 2004  41     68    152 124  78  132    20    3064
5 5 2005  59     70    150 144  86  255    20    1885
6 6 2006  44     72    204 150  90  182     0    3102
```

2. Second, we will calculate the correlation matrix.

```r
cor(filtered_data, use = "pairwise.complete.obs")
```

```
                  X           id          age       height       weight
X        1.000000000  0.956757817 -0.039143689 -0.056722904 -0.004811028
id       0.956757817  1.000000000 -0.048160214 -0.052294226 -0.003780646
age     -0.039143689 -0.048160214  1.000000000 -0.095375682 -0.034404537
height  -0.056722904 -0.052294226 -0.095375682  1.000000000  0.532935466
weight  -0.004811028 -0.003780646 -0.034404537  0.532935466  1.000000000
sbp     -0.034136607 -0.044388131  0.165746397  0.018373573  0.253249623
dbp     -0.051566439 -0.052577344  0.139197757  0.010275549  0.295920186
chol     0.050372896  0.057845065  0.089188510 -0.088937779  0.008537442
ncigs    0.013432770  0.011667283 -0.005033852  0.014911292 -0.081747507
timechd  0.048278246  0.041181424 -0.070919630 -0.009895169 -0.065350046
                sbp         dbp         chol        ncigs      timechd
X       -0.03413661 -0.05156644  0.050372896  0.013432770  0.048278246
id      -0.04438813 -0.05257734  0.057845065  0.011667283  0.041181424
age      0.16574640  0.13919776  0.089188510 -0.005033852 -0.070919630
height   0.01837357  0.01027555 -0.088937779  0.014911292 -0.009895169
weight   0.25324962  0.29592019  0.008537442 -0.081747507 -0.065350046
sbp      1.00000000  0.77290641  0.123061297  0.029977529 -0.107884203
dbp      0.77290641  1.00000000  0.129597108 -0.059342317 -0.110693969
chol     0.12306130  0.12959711  1.000000000  0.096031834 -0.095390054
ncigs    0.02997753 -0.05934232  0.096031834  1.000000000 -0.093933141
timechd -0.10788420 -0.11069397 -0.095390054 -0.093933141  1.000000000
```

The variable that is most correlated is dbp, and the secondmost correlated variable is weight. We will use these variables to build this model.

Now, we will build the models.

```r
model <- lm(sbp ~ weight, data = wcgs)
summary(model)
```

```

Call:
lm(formula = sbp ~ weight, data = wcgs)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.549 -10.097  -2.456   7.724  99.544 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 97.78884    2.11473   46.24   <2e-16 ***
weight       0.18148    0.01235   14.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.63 on 3152 degrees of freedom
```

```
Multiple R-squared:  0.06414,   Adjusted R-squared:  0.06384 
F-statistic:   216 on 1 and 3152 DF,  p-value: < 2.2e-16
```

Our model is sbp = 0.1814*weight + 97.7888

```
model <- lm(sbp ~ dbp, data = wcgs)
summary(model)
```

```
Call:
lm(formula = sbp ~ dbp, data = wcgs)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.237  -6.212  -1.394   5.386  62.581 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.11020    1.45075   20.75   <2e-16 ***
dbp          1.20127    0.01757   68.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.594 on 3152 degrees of freedom
Multiple R-squared:  0.5974,    Adjusted R-squared:  0.5973 
F-statistic:  4677 on 1 and 3152 DF,  p-value: < 2.2e-16
```

Our model is sbp = 1.2013*dbp + 30.1102.

The diastolic blood pressure has a stronger correlation to systolic blood pressure than weight.

Part 5

We found that there were not enough evidence to conlude that smokers and non-smokers had different systolic blood pressure. There were different proportions of personalities discovered so we could not complete a Chi-Squared test. The highest correlated variables with systolic blood pressure were weight followed by diastolic blood pressure. From this, we discovered that diastolic blood pressure has a stronger correlation to systolic blood pressure than weight.