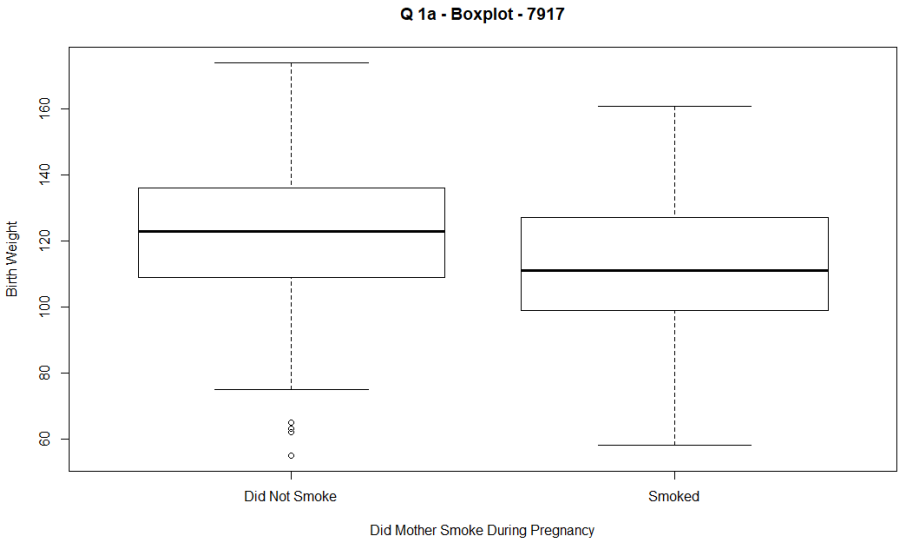
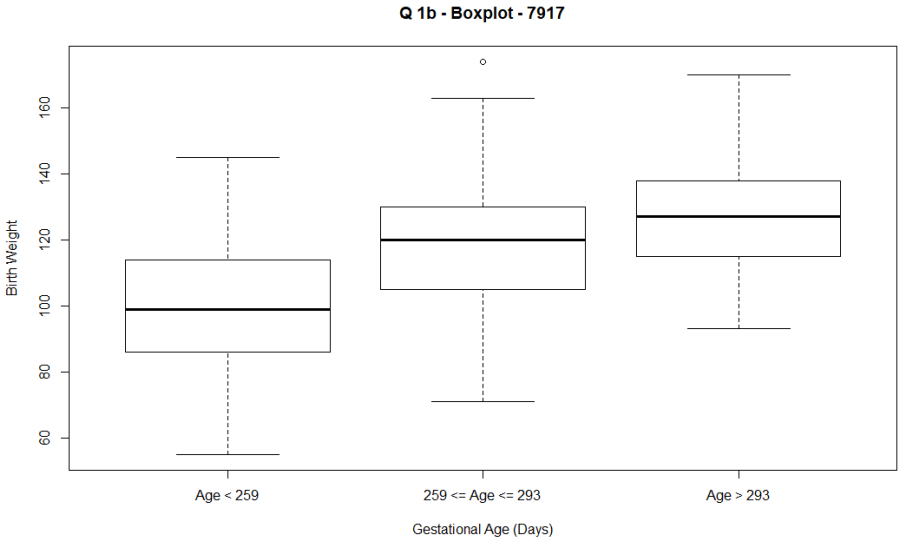
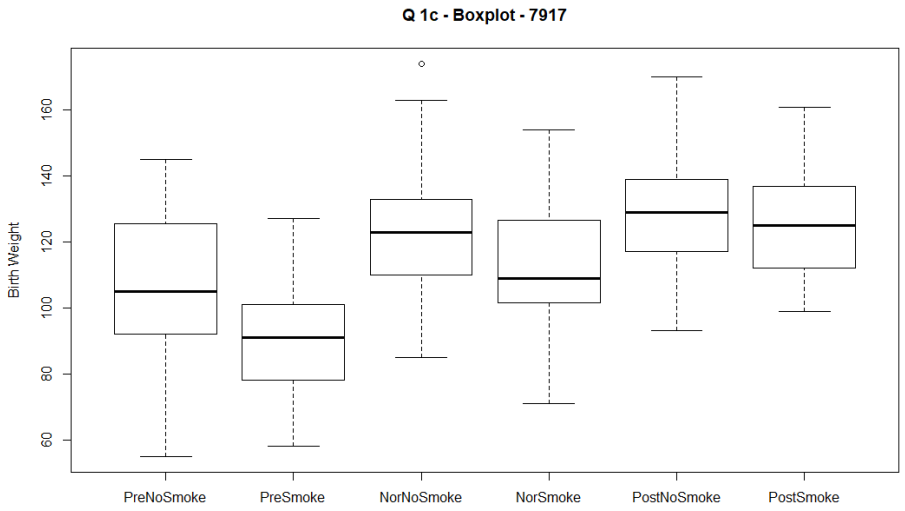
1.

From the boxplots there do appear to be differences. In boxplot 1a it appears babies from mothers who did not smoke have higher birthweights. In boxplot 1b it appears the higher the gestational age in days the higher the birthweight. 1c seems to agree with 1a and 1b; as gestational age increases so does birthweight, and for a given gestational age maternal smoking appears to decrease the birthweight.







2.
The t-test from R would indicate there is a difference between the weights of babies born from smokers vs non-smokers. The t-stat of 4.6409 (333.49 df) from the Welch two-sample t-test, and associated p-value of 4.994e-06, suggests there is strong evidence of these weights not being equal.

3.
The small p-value (2e-16) in the one-way ANOVA would suggest there is a difference in mean birth weight among babies classified by gestational maturity.

To see which levels of maturity differ we can use a pairwise t-test using a Bonferroni correction. The low p-values reported in the pairwise t-test would suggest that all of the means differ from each other.

4.
The small p-value (2e-16) in the one-way ANOVA would suggest there is a difference in mean birth weight among babies classified in one of six categories related to maturity level and mother's smoking status.

Looking at the pairwise t-test results it appears that of the 15 (${}_6C_2$) comparisons, all groups differ from one another except: PostNoSmoke-NorNoSmoke, PostSmoke-NorNoSmoke, PostSmoke-PostNoSmoke, PreNoSmoke-NorSmoke

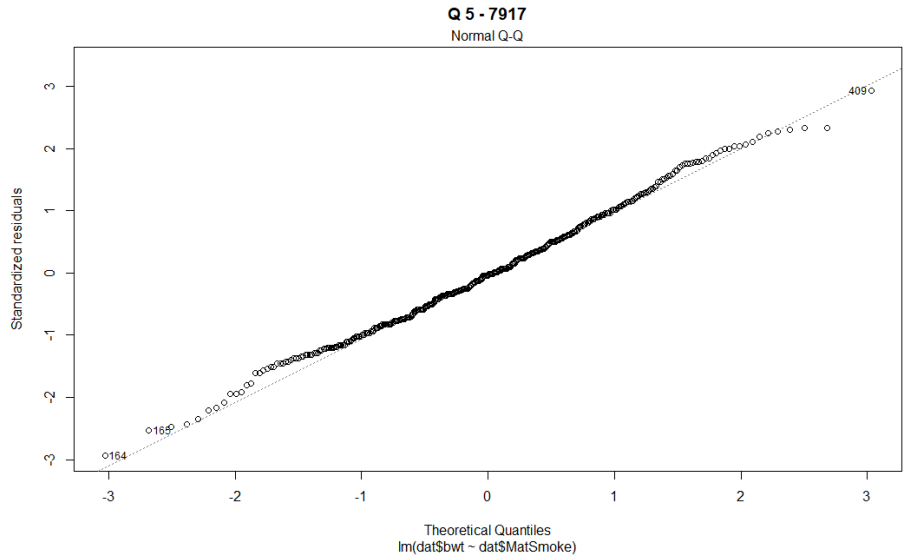
5.
At first glance, I trust the statistical test carried out in part 4 given it aligns with the visual representation of the data in the box plot graph.

Delving deeper, the necessary assumptions for the one-way ANOVA to hold are: the six samples drawn from six specific populations with unknown means, each population has the same variance, and each population is normally distributed.

On the first point, the six samples drawn from six specific populations with unknown means, this would appear to hold given the dataset provided, but we would need more information about how the data was collected to be sure.

Checking whether each population has the same variance we can use Bartlett's test for the homogeneity of variances. The Bartlett test returns a p-value of 0.09627, above the benchmark significance level of 5%. Therefore we cannot reject the null hypothesis that the variances of the populations are unequal.

To check if the populations are normally distributed we can observe the QQ-plot of the linear model. The QQ-plot is relatively well behaved suggesting the populations are normally distributed.



In summary it appears the necessary assumptions of the model hold.

6.

a)

In question 4 we fit a model with five predictor variables (with n groups need only n-1 indicator variables). The model had five predictor variables and took the form:

$$Y_i = \beta_0 + \beta_{NorSmoke}*X_{NorSmoke} + \beta_{PostNoSmoke}*X_{PostNoSmoke} + \beta_{PostSmoke}*X_{PostSmoke} + \beta_{PreNoSmoke}*X_{PreNoSmoke} + \beta_{PreSmoke}*X_{PreSmoke}$$

If we used a model of the form maternal smoking status, maturity level, and their interaction, the model would have had the following five variables (the same amount as in question 4):

Smoking Terms:

$\mathbf{1}_{Smoke}$ = 1 if mother smoked, 0 if not

Maturity Terms:

$\mathbf{1}_{Maturity1}$ = 1 if maturity level = 1, 0 if not

$\mathbf{1}_{Maturity2}$ = 1 if maturity level = 2, 0 if not

Interaction Terms:

$\mathbf{1}_{Smoke} \times \mathbf{1}_{Maturity1}$ = 1 if mother smoked and baby had maturity level 1, 0 if not

$\mathbf{1}_{Smoke} \times \mathbf{1}_{Maturity2}$ = 1 if mother smoked and baby had maturity level 2, 0 if not

b)
Yes the F-Test for the presence of interaction would be significant. On one hand, you can observe this from the boxplot where the effect of smoking does not appear constant for different maturity levels.

More specifically, the pairwise t-test would confirm this result. For example, the mean weights of NorNoSmoke and PostNoSmoke are not statistically significantly different from each other (pairwise t-test p-value of 0.1625). If there was no interaction term one would expect NorSmoke and PostSmoke to not be statistically significantly different as well. However the p-value in the pairwise t-test for NorSmoke and PostSmoke is 0.0033, below our significance level suggesting that the effect of smoking was different for these groups (suggesting the interaction term $\underline{\mathbf{1}}_{\text{Smoke}} \times \underline{\mathbf{1}}_{\text{Maturity2}}$ is statistically significant).

7.
The fact there is different numbers of babies in the three maturity levels risks that the design is unbalanced. Given the somewhat small size for some groups (all groups <100 observations), and the disparity between the largest and smallest groups (PreSmoke has 41 observations vs. PostNoSmoke has 97 observations), there is some cause for concern. The potential risk from unequal groups is that the constant variance assumption is violated. The Bartlett test from question 5 tested for unequal variances. This test gives us a p-value of 0.09627 suggesting we cannot reject the assumption that the variances are equal. Given this we do not need to be too concerned about the fact the group sizes are different.

8.
Using gestation as a factor variable as opposed to a quantitative variable has a couple of advantages. First, boxplot 1b suggests that Post and Nor maturity levels (baby spent over 259 days in the womb) have very similar birthweights, it is only Pre maturity levels (baby spent less than 259 days in the womb) that have much different birthweights. Had gestation been a quantitative variable it would have suggested that as maturity level increases so does birthweight, but this does not capture the apparent nuance that after a certain point increasing the maturity level does not seem to affect birthweight.

Also using factor variables helps with the interpretability of the analysis. From our analysis one could interpret that "babies born prematurely have lower birthweights by Y ounces". If gestation was a quantitative variable you would have to interpret the results of the linear model and perform calculations to determine how much lower a premature baby's birthweight is. On the other-hand, the results of the analysis using gestation as a quantitative variable better helps us answer questions such as "if a baby spends an additional X days in the womb its birthweight should be expected to change by Y".

The models would look like this:

Additive Model – Gestation as Factor

$$Y_i = \beta_0 + \beta_{\text{NorSmoke}} * X_{\text{NorSmoke}} + \beta_{\text{PostNoSmoke}} * X_{\text{PostNoSmoke}} + \beta_{\text{PostSmoke}} * X_{\text{PostSmoke}} + \beta_{\text{PreNoSmoke}} * X_{\text{PreNoSmoke}} + \beta_{\text{PreSmoke}} * X_{\text{PreSmoke}}$$

Additive Model – Gestation as Quantitative Variable

$$Y_i = \beta_0 + \beta_{\text{GestationDays}} * X_{\text{GestationDays}} + \beta_{\text{Smoke}} * X_{\text{Smoke}}$$

9.

An study out of India from Metgud, Naik, and Mallapur[1] (suggests some additional factors affecting the birth weight of a newborn.

Factor 1 – The mother previously had a baby with a low birthweight

Levels
0 – mother has not previously had a child with low birthweight
1 – mother has previously had a child with low birthweight

Factor 2 – Education Level

Levels
0 – mother has not been awarded a university degree
1 – mother has been awarded a university degree

[1] Metgud CS, Naik VA, Mallapur MD. Factors Affecting Birth Weight of a Newborn – A Community Based Study in Rural Karnataka, India. Szecsi PB, ed. *PLoS ONE*. 2012;7(7):e40040. doi:10.1371/journal.pone.0040040.

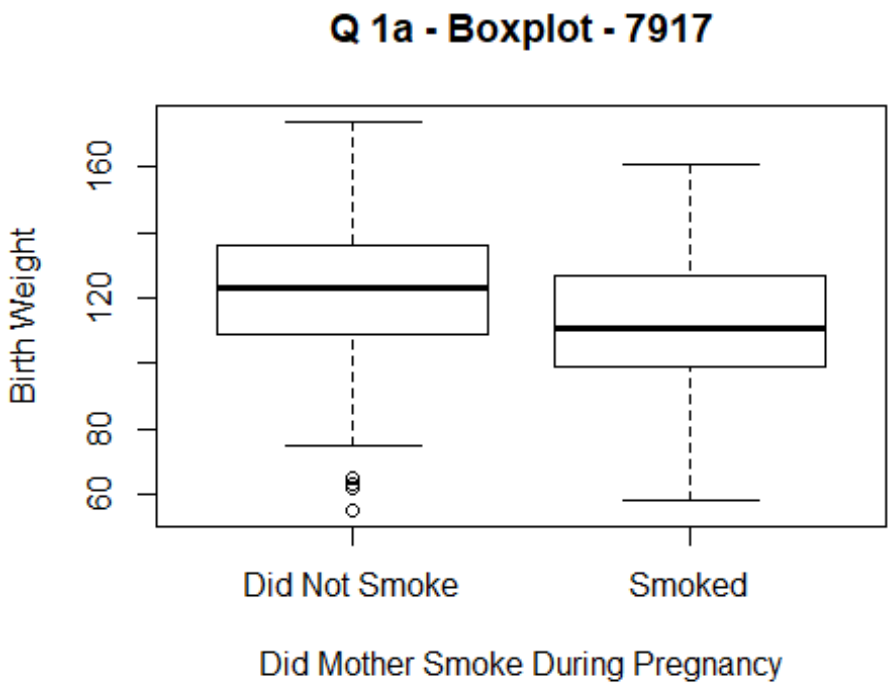
APPENDIX - R Code and Output

David

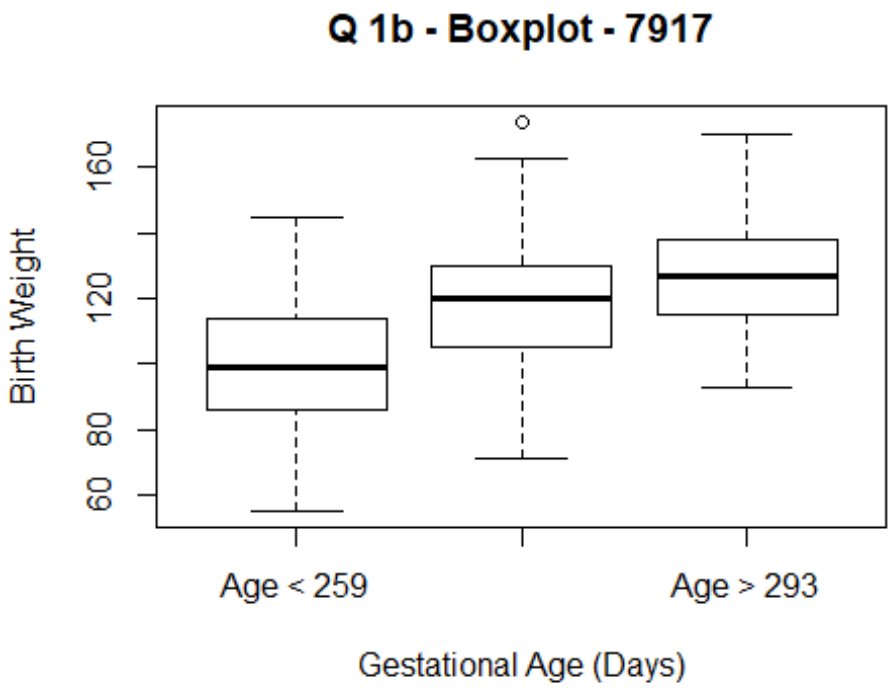
Tue Feb 13 09:42:30 2018

```
## DAVID VEITCH ASSIGNMENT 2 - 1004657917

#Import CSV to variable dat
setwd("C:/Users/David/Google Drive/Documents/UofT/NonDegree/303/Assignment 2"
)
dat <- read.csv("bbw.csv")

# Create variables maturity & MatSmoke with Factor levels
maturity=array(0,length(dat$gestation))
MatSmoke=array(0,length(dat$smoke))
for (i in 1:length(dat$gestation))
{
  if (dat$gestation[i]<259)
  {maturity[i]=1}
  else if (dat$gestation[i]>293)
  {maturity[i]=3}
  else {maturity[i]=2}
}
for (i in 1:length(dat$smoke))
{
  if (maturity[i]==1 & dat$smoke[i]==1)
  {MatSmoke[i]="PreSmoke"}
  else if (maturity[i]==1 & dat$smoke[i]==0)
  {MatSmoke[i]="PreNoSmoke"}
  else if (maturity[i]==2 & dat$smoke[i]==1)
  {MatSmoke[i]="NorSmoke"}
  else if (maturity[i]==2 & dat$smoke[i]==0)
  {MatSmoke[i]="NorNoSmoke"}
  else if (maturity[i]==3 & dat$smoke[i]==1)
  {MatSmoke[i]="PostSmoke"}
  else {MatSmoke[i]="PostNoSmoke"}
}

# Append maturity & MatSmoke to dat
dat$MatSmoke <- MatSmoke
dat$maturity <- maturity

# QUESTION 1 - Create Boxplots for 3 factors

# BOXPLOT - Mothers Smoked/not smoked
boxplot(dat[dat$smoke==0,]$bwt, dat[dat$smoke==1,]$bwt,
        xlab="Did Mother Smoke During Pregnancy",ylab="Birth Weight",
```

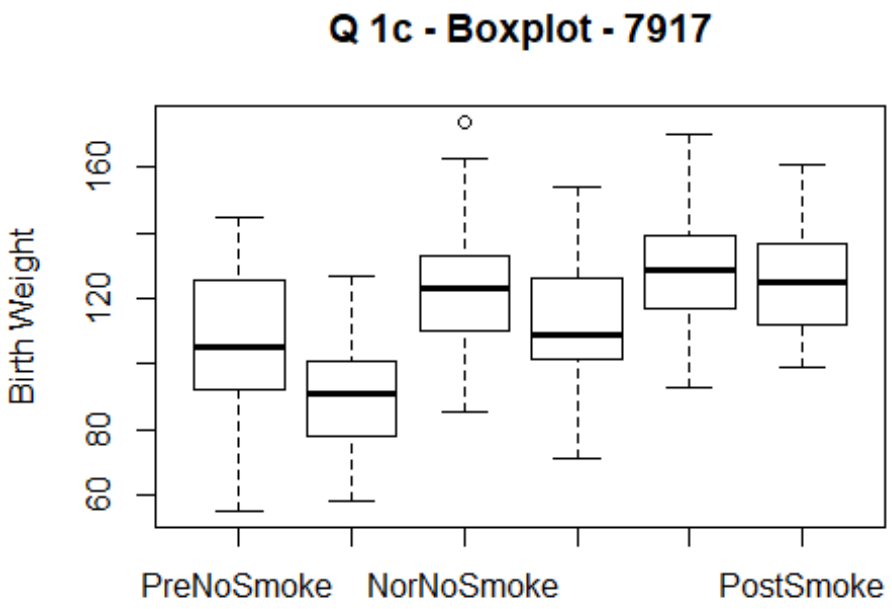
```
        names=c("Did Not Smoke","Smoked"),
        main="Q 1a - Boxplot - 7917")
```




```
# BOXPLOT - Maturity Level
boxplot(dat[maturity==1,]$bwt, dat[dat$maturity==2,]$bwt,
        dat[dat$maturity==3,]$bwt, xlab="Gestational Age (Days)",ylab="Birth
Weight",
        names=c("Age < 259","259 <= Age <= 293","Age > 293"),
        main="Q 1b - Boxplot - 7917")
```



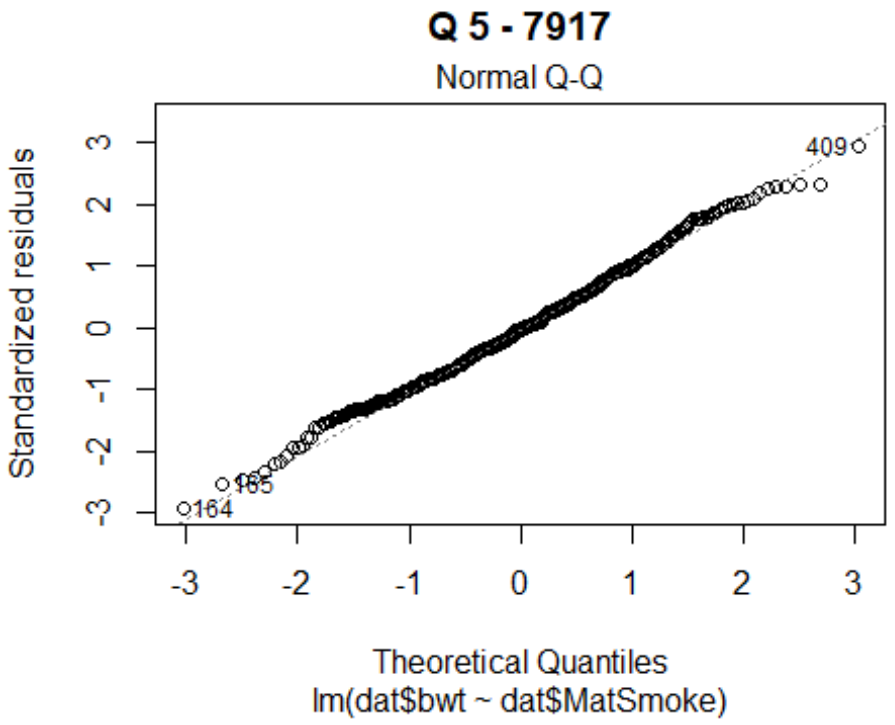

```
# BOXPLOT - Smoking & Maturity
boxplot(dat[dat$MatSmoke=="PreNoSmoke",]$bwt,
        dat[dat$MatSmoke=="PreSmoke",]$bwt,
        dat[dat$MatSmoke=="NorNoSmoke",]$bwt,
        dat[dat$MatSmoke=="NorSmoke",]$bwt,
        dat[dat$MatSmoke=="PostNoSmoke",]$bwt,
        dat[dat$MatSmoke=="PostSmoke",]$bwt,
        ylab="Birth Weight",
        names=c("PreNoSmoke","PreSmoke","NorNoSmoke","NorSmoke",
                "PostNoSmoke","PostSmoke"),
        main="Q 1c - Boxplot - 7917")
```




```
# QUESTION 2 - Use T-Test to investigate if difference in weight to babies
# from mothers who were smokers to mothers who were non-smokers
t.test(dat$bwt~dat$smoke)

##
##  Welch Two Sample t-test
##
## data:  dat$bwt by dat$smoke
## t = 4.6409, df = 333.49, p-value = 4.994e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.582669 13.797101
## sample estimates:
## mean in group 0 mean in group 1
##        121.5488        111.8589

# QUESTION 3 - Investigate whether there is a difference in mean birth weight
# among babies classified by gestational maturity using one way ANOVA
#dat$maturity <- as.factor(dat$maturity)
aov3 = aov(dat$bwt~dat$maturity)
summary(aov3)

##               Df Sum Sq Mean Sq F value Pr(>F)
## dat$maturity   1  44254   44254   133.4 <2e-16 ***
## Residuals    407 135012     332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(dat$bwt, dat$maturity, p.adj="bonf")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  dat$bwt and dat$maturity
##
##   1       2
## 2 1.4e-14 -
## 3 < 2e-16 3.8e-05
##
## P value adjustment method: bonferroni

# QUESTION 4 - Use a one-way analysis of variance to investigate whether ther
e
# is a difference in mean birth weight among six categories based on maturity
# level and mothers smoking status
summary(aov(dat$bwt~dat$MatSmoke))

##               Df Sum Sq Mean Sq F value Pr(>F)
## dat$MatSmoke   5  55448   11090   36.09 <2e-16 ***
## Residuals    403 123818     307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

pairwise.t.test(dat$bwt, dat$MatSmoke, p.adj="bonf")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  dat$bwt and dat$MatSmoke
##
##             NorNoSmoke NorSmoke PostNoSmoke PostSmoke PreNoSmoke
## NorSmoke    0.0114     -        -           -         -
## PostNoSmoke 0.1625     2.4e-07  -           -         -
## PostSmoke   1.0000     0.0033   1.0000      -         -
## PreNoSmoke  3.2e-07    0.2824   2.4e-13     2.1e-07   -
## PreSmoke    < 2e-16    1.7e-08  < 2e-16     < 2e-16   0.0015
##
## P value adjustment method: bonferroni

# QUESTION 5 - Check assumptions of one way ANOVA
# Check equal variances with Bartlett Test
bartlett.test(dat$bwt~dat$MatSmoke)

##
##  Bartlett test of homogeneity of variances
##
## data:  dat$bwt by dat$MatSmoke
## Bartlett's K-squared = 9.3393, df = 5, p-value = 0.09627
```

```r
# Check normarlity with QQ-Plot
plot(lm(dat$bwt~dat$MatSmoke),which=2, main="Q 5 - 7917")
```




```r
# QUESTION 7 - Number of babies per group
table(dat$MatSmoke)
```

```
##
##   NorNoSmoke     NorSmoke PostNoSmoke    PostSmoke  PreNoSmoke     PreSmoke
##           93           68           97           54           56           41
```