565_Quiz3

Question 1

Beta0 is the intercept for group 1 and is negative Beta1 is the slope for group 1 and is positive Beta2 is the difference in intercept between group 1 and 2 and is positive Beta3 is the difference in slope between group 1 and group two and is positive.

Question 2

a. Confounding.

If higher lactic acid concentration results in both better tasting cheese and a higher acetic acid concentration, then these two variables are confounded in their effect on the final taste score.

b. Interaction

It is possible that the impact of lactic acid on tasete varies depending on the acetic acid level. For example, at a high level of acetic acid, high lactic acid may have a negative effect while at a low level of acetic acid, the lactic acid may have a positive effect.

#Question 3

```r
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
```

```r
library(emmeans)
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

```r
library(ggplot2)

diesel1 <- read.table("Data/diesel1.txt", header=TRUE)
str(diesel1)
```

```
## 'data.frame':    80 obs. of  3 variables:
##  $ DPF        : chr  "A" "A" "B" "A" ...
##  $ engine_size: num  16.5 17.1 11.4 8.7 16 9.4 15.2 13.7 10.6 8.9 ...
##  $ PM_emission: int  110 140 115 79 142 91 121 129 110 95 ...
```

a

```r
fit1 <- aov(PM_emission ~ DPF, data=diesel1)
summary(fit1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## DPF          2   3599  1799.6   5.564 0.00553 **
## Residuals   77  24904   323.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(fit1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = PM_emission ~ DPF, data = diesel1)
##
## $DPF
##          diff        lwr       upr     p adj
## B-A -2.641723 -14.487613  9.204166 0.8553979
## C-A 17.238619   3.829794 30.647443 0.0081646
## C-B 19.880342   4.236714 35.523970 0.0090437
```

There are significant differences between Brands. Specifically, C emits significantly more than brand A and B. Brands A and B are not significantly different than each other

b

```r
fit2 <- aov(PM_emission ~ engine_size + DPF, data=diesel1)

emmeans(fit2, "DPF") %>% pairs()
```

```
##  contrast estimate   SE df t.ratio p.value
##  A - B        8.36 2.98 76   2.805  0.0173
##  A - C      -12.22 3.36 76  -3.639  0.0014
##  B - C      -20.58 3.89 76  -5.296  <.0001
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

```r
fit3 <- aov(PM_emission ~ engine_size * DPF, data=diesel1)
summary(fit3)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)    
## engine_size      1  16647   16647 168.247  < 2e-16 ***
## DPF              2   3197    1599  16.156 1.51e-06 ***
## engine_size:DPF  2   1337     669   6.756  0.00202 ** 
## Residuals       74   7322      99                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit2, fit3)
```

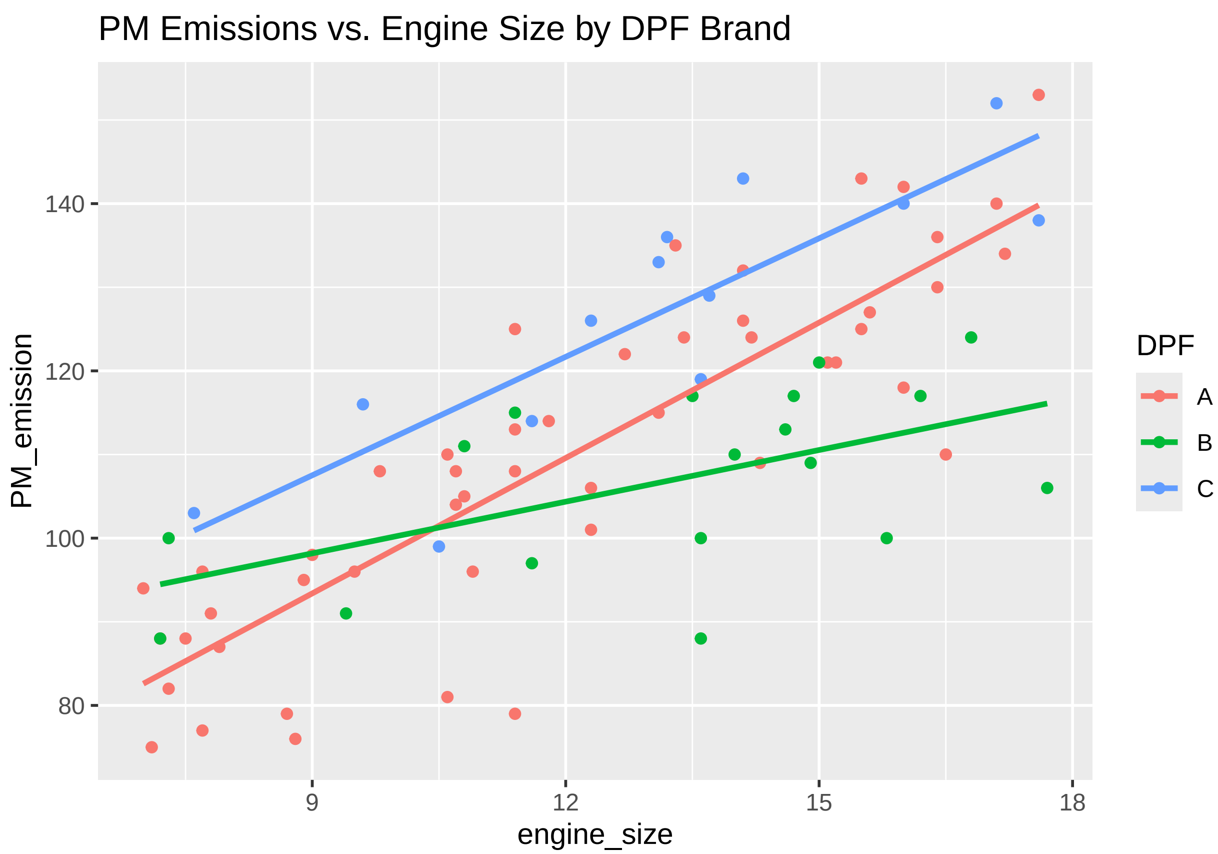
```
## Analysis of Variance Table
## 
## Model 1: PM_emission ~ engine_size + DPF
## Model 2: PM_emission ~ engine_size * DPF
##   Res.Df  RSS Df Sum of Sq      F   Pr(>F)   
## 1     76 8659                                
## 2     74 7322  2      1337 6.7563 0.002018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After including the covariate, engine size and brand are both significant.The adjusted pairwise comparisons now suggest that all 3 groups differ. Specifically, A emits more than B, and C emits more than both A and B. That said, there is a significant interaction between the engine size and the brand. To determine which is best, we must see where the lines intersect in the plot below.

c

```r
ggplot(diesel1, aes(x=engine_size, y=PM_emission, color=DPF)) +
  geom_point() +
  geom_smooth(method="lm", aes(fill=DPF), se=FALSE) +
  labs(title="PM Emissions vs. Engine Size by DPF Brand")
```

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



d

C has consistently the highest emissions. For engines less than size 10, A has the lowest emissions, and for engines greater than size 10, B has the lowest emissions.

e

In this case, the engine size interacts with the efficiency of each manufacturer meaning that at different sizes, different engines are better for reducing emissions. As such we need to account for it in our model.

#Question 4

```
diesel2 <- read.table("Data/diesel2.txt", header=TRUE)
str(diesel2)
```

```
## 'data.frame':    80 obs. of  3 variables:
##  $ DPF        : chr  "A" "B" "B" "A" ...
##  $ engine_size: num  7.4 13.6 15.4 9.9 8.2 12.2 11.8 12.8 7.1 10.9 ...
##  $ PM_emission: int  90 115 128 106 109 122 105 103 84 113 ...
```

a

```r
fit1 <- aov(PM_emission ~ DPF, data=diesel2)
summary(fit1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)    
## DPF          2   8073    4037   14.05 6.3e-06 ***
## Residuals   77  22129     287                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(fit1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = PM_emission ~ DPF, data = diesel2)
## 
## $DPF
##          diff        lwr       upr     p adj
## B-A  22.73073  12.450025 33.011439 0.0000034
## C-A  10.67073  -1.870447 23.211910 0.1110880
## C-B -12.06000 -25.584189  1.464189 0.0902300
```

There are significant differences between Brands at a .05 level. Specifically, B emits significantly more than brand A. All other brands are not significantly different.

b

```r
fit2 <- aov(PM_emission ~ engine_size + DPF, data=diesel2)
summary(fit2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)    
## engine_size  1  21448   21448 188.157 <2e-16 ***
## DPF          2     92      46   0.403   0.67    
## Residuals   76   8663     114                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
emmeans(fit2, "DPF") %>% pairs()
```

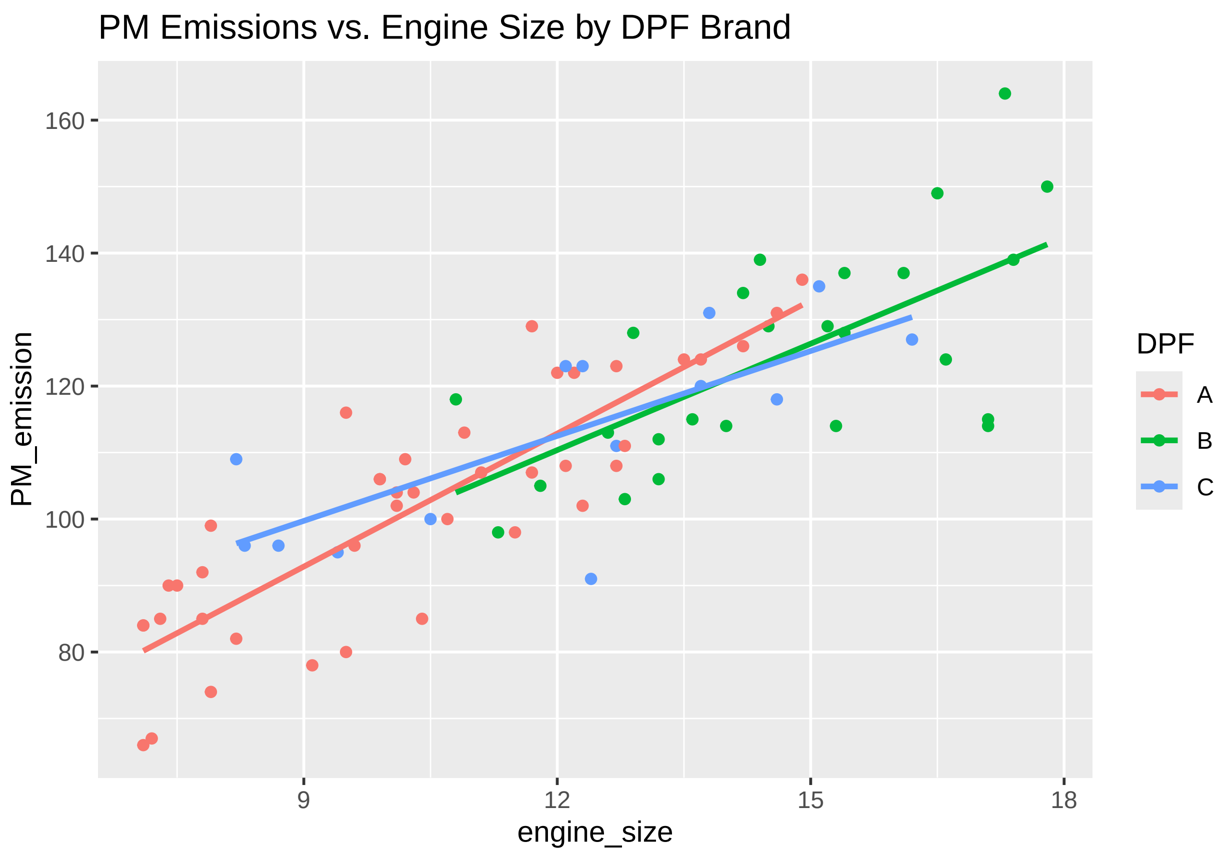
```
##  contrast estimate   SE df t.ratio p.value
##  A - B        2.31 3.56 76   0.649  0.7937
##  A - C       -1.07 3.42 76  -0.314  0.9473
##  B - C       -3.38 3.84 76  -0.881  0.6540
## 
## P value adjustment: tukey method for comparing a family of 3 estimates
```

After controlling for Engine size, the brand becomes no longer significant. This implies that the changes we see in emissions are due to engine size not brand. It also follows tha the signifficance seen before was due to certain brands primarily making engines of specific sizes.

c

```r
ggplot(diesel2, aes(x=engine_size, y=PM_emission, color=DPF)) +
  geom_point() +
  geom_smooth(method="lm", aes(fill=DPF), se=FALSE) +
  labs(title="PM Emissions vs. Engine Size by DPF Brand")
```

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



d

No brand differs significantly once engine size is accounted for; none can be recommended over the others on the basis of emissions.

e

Engine size is a strong confounder as larger engine sizes mean more emissions. Without adjustment, apparent brand effects simply reflect differences in the engine-size mix.