STOR 455 Homework #8

25 points - Due 4/13 at 5:00pm

Situation (again): Suppose that you are interested in purchasing a used car. How much should you expect to pay? Obviously the price will depend on the type of car you get (the model) and how much it's been used. For this assignment you will investigate how the price might depend on the age, mileage, and the model of car.

Data Source: Your sample of cars will again be taken from the UsedCar CSV file on Sakai. The data was acquired by scraping TrueCar.com for used car listings on 9/24/2017 and contains more than 1.2 million used cars.

For this assignment, you will need to select six new samples, each with *exactly* 50 vehicles, for six different *Model* of used vehicles for sale in North Carolina from the UsedCar dataset. There will likely be more than 50 of your selected models for sale in North Carolina, so you should randomly select those 50 vehicles from the larger number that are available. The six models of vehicles should be selected such that three models of vehicles are selected from Japanese companies, and another three from US companies (i.e. *Make*; It does not matter where the cars were actually manufactured). Within each country, you should select a compact car, a mid-sized car, and a SUV (Note that the country and types of vehicles are not given in the data and are for you to determine). You should add new variables to the dataframes for the country of the company and type of vehicle (compact vs mid-sized vs SUV) and combine these six samples into one dataframe (just as rbind was used in a previous assignment). When selecting these samples make sure to use set.seed(). This will select the same sample each time that you run (and knit) your code.The code below is an example of how you could select a random sample of 50 cars for a given model:

```
library(readr)
UsedCars <- read_csv("UsedCars.csv")

## Rows: 1048575 Columns: 9

## -- Column specification --------------------------------------------
------------
## Delimiter: ","
## chr (5): City, State, Vin, Make, Model
## dbl (4): Id, Price, Year, Mileage

##
## i Use `spec()` to retrieve the full column specification for this da
ta.
## i Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

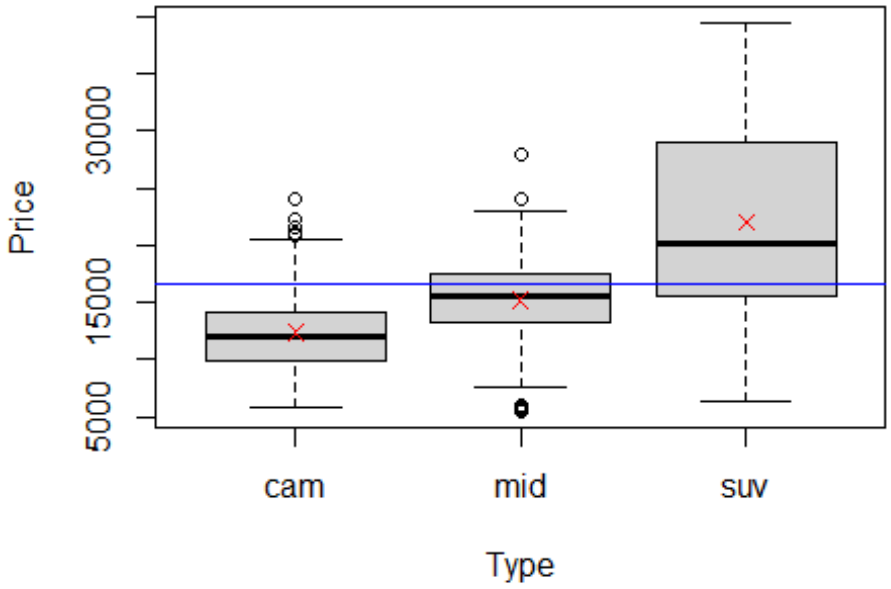
```
StateOfMyChoice = "NC"
MakeOfMyChoice1 = "Honda"
MakeOfMyChoice2 = "Ford"

NCUsedCars_Honda = subset(UsedCars, Make=="Honda" & State=="NC")
NCUsedCars_Ford = subset(UsedCars, Make=="Ford" & State=="NC")

NCUsedCars_Honda$Make[which(NCUsedCars_Honda$Make=='Honda')]<-'JP'
NCUsedCars_Ford$Make[which(NCUsedCars_Ford$Make=='Ford')]<-'USA'

# Suppose that NCUsedCars_Honda is a subset of
# only Hondas sold in NC from the Usedcars dataset.
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

set.seed(8675309)
FocusSE = sample_n(subset(NCUsedCars_Ford, Model=='FocusSE'), 50)
FusionSE = sample_n(subset(NCUsedCars_Ford, Model=='FusionSE'), 50)
EdgeSEL = sample_n(subset(NCUsedCars_Ford, Model=='EdgeSEL'), 50)

Civic = sample_n(subset(NCUsedCars_Honda, Model=='Civic'), 50)
Accord = sample_n(subset(NCUsedCars_Honda, Model=='Accord'), 50)
Pilot4WD = sample_n(subset(NCUsedCars_Honda, Model=='Pilot4WD'),50)

data= rbind(FocusSE,Civic,FusionSE,Accord,EdgeSEL,Pilot4WD)
data$Model[which(data$Model=='FocusSE')]<-'cam'
data$Model[which(data$Model=='Civic')]<-'cam'
data$Model[which(data$Model=='FusionSE')]<-'mid'
data$Model[which(data$Model=='Accord')]<-'mid'
data$Model[which(data$Model=='EdgeSEL')]<-'suv'
data$Model[which(data$Model=='Pilot4WD')]<-'suv'

names(data)[8] = "Country"
names(data)[9] = "Type"
```

One Way ANOVA

1. Produce a set of side-by-side boxplots to compare the price distributions of your three types of vehicles (not the models). Comment on any obvious differences in the distributions.

Compared with cam and mid size, suv size has more variance and is much more spread out of price.

```
boxplot(Price ~ Type, data = data)
means = tapply(data$Price, data$Type, mean)
points(means, col="red", pch=4)
abline(h = mean(data$Price), col = "blue")
```



2. Produce summary statistics (mean and standard deviation) for each of the groups (vehicle types) AND the entire sample of vehicle prices.

```
tapply(data$Price, data$Type, mean)

##      cam      mid      suv
## 12371.78 15216.29 22063.37

tapply(data$Price, data$Type, sd)

##      cam      mid      suv
## 3757.270 4031.583 8629.474

mean(data$Price)

## [1] 16550.48

sd(data$Price)

## [1] 7163.259
```

3. Based on just what you see in the boxplots and summary statistics comment on whether you think there are significant differences in the mean prices among your three vehicle types. Also comment on any concerns you see about the conditions for the ANOVA for means model.

   I think there is a significant difference in the mean price among 3 types of cars. From seeing sd data we know that constant variance might be a issue because the one of suv is more than double two of the others.(8629>37572 8629>40312)

4. Construct an ANOVA model for the mean price by vehicle type. Include the output showing the ANOVA table; state hypotheses, and provide a conclusion in the context of your data.

   Ho: μ1 = μ2=…=μi Ha:some μi≠μj

   H0: There is no difference in means between Types of cars. Ha:there is at least one type of cars that the mean is different.

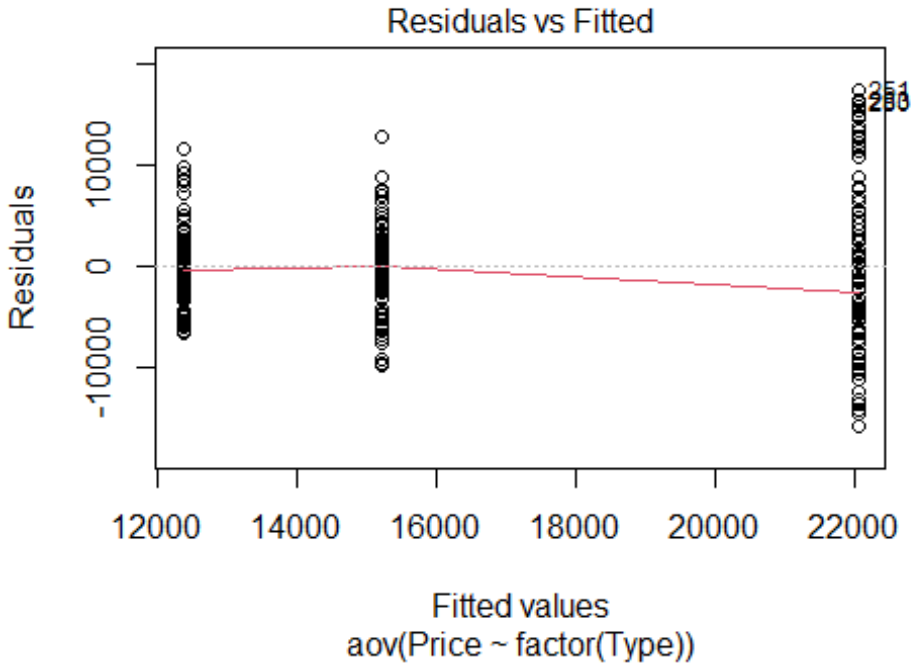
   Since P value is small, there is a significant difference in mean price between different types of cars.

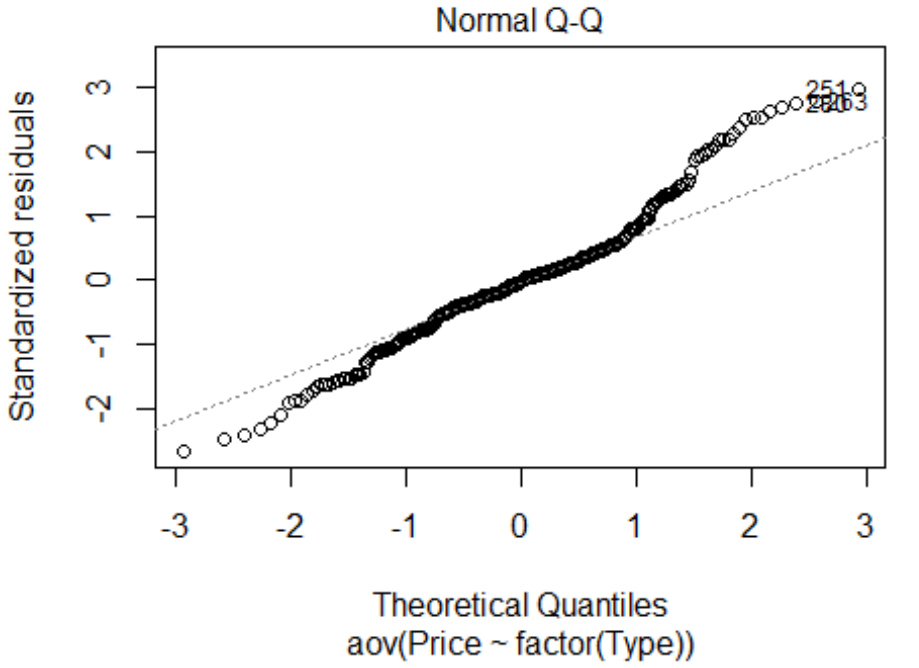
```
PT=aov(Price~factor(Type),data=data)
PT

## Call:
##    aov(formula = Price ~ factor(Type), data = data)
##
## Terms:
##                 factor(Type)   Residuals
## Sum of Squares    4963355280 10379017483
## Deg. of Freedom            2         297
##
## Residual standard error: 5911.53
## Estimated effects may be unbalanced

summary(PT)

##               Df    Sum Sq   Mean Sq F value Pr(>F)
## factor(Type)   2 4.963e+09 2.482e+09   71.01 <2e-16 ***
## Residuals    297 1.038e+10 3.495e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Produce plots and/or summary statistics to comment on the appropriateness of the following conditions for your data: normality of the residuals, and equality of the variances. If you find that the conditions are *not* met, You can still continue with analysis of your data for this homework. We will soon discuss how to deal with violations of these conditions.

equality: There is a little bit falling trend on the right side. normality: There is a little bit questionable because both of the tails fall off the line.

```
plot(PT, 1:2)
```

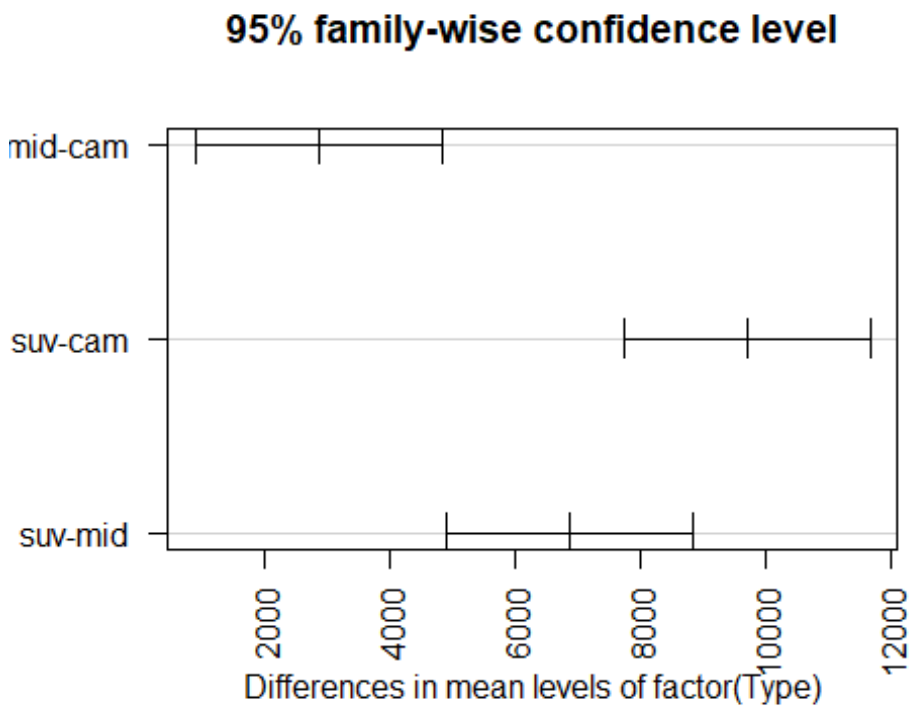




6. If your ANOVA model indicates that there are significant differences among the vehicle type price means, discuss where the significant differences occur using Tukey HSD methods. If your ANOVA indicates there are not significant differences among the vehicle type price means, determine how different your means prices would need to be in order to find a significant difference using the Tukey HSD methods.

   From the graph and small P value we know that significant difference occur in each pair of them, especially in suv and the others.

```
TukeyHSD(PT)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Price ~ factor(Type), data = data)
##
## $`factor(Type)`
##             diff       lwr       upr     p adj
## mid-cam 2844.51  875.2522  4813.768 0.0021864
## suv-cam 9691.59 7722.3322 11660.848 0.0000000
## suv-mid 6847.08 4877.8222  8816.338 0.0000000

hsd = TukeyHSD(PT)
plot(hsd, las=2)
```



Two Way ANOVA

7. Construct an ANOVA model for the mean price using the country of the company and the type of vehicle as predictors (without an interaction). Include the output showing the ANOVA table; state hypotheses and provide a conclusion in the context of your data. If your ANOVA model indicates there are significant differences among the vehicle price means: Discuss where the significant differences occur using Tukey HSD methods.

   Ho: $\alpha 1 = \alpha 2 = \ldots = \alpha K = 0$ Ha:some $\alpha k \neq 0$ Ho: $\beta 1 = \beta 2 = \ldots = \beta j = 0$

   Ha: some $\beta j \neq 0$

   H0: There is no difference in means between Types of cars and There is no difference in means between countries. Ha:there is at least one different type or country where the mean is different.

   Since P value is small, we have 95% confidence to say that there is a significant difference in mean price between different subgroups(counties and types of cars).

   From the small P value we know that significant difference occur in each subgroups, especially in suv and the others.

```
amod = aov(Price~factor(Type)+Country, data=data)
summary(amod)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## factor(Type)    2 4.963e+09 2.482e+09  71.779 <2e-16 ***
## Country         1 1.452e+08 1.452e+08   4.199 0.0413 *
## Residuals     296 1.023e+10 3.457e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

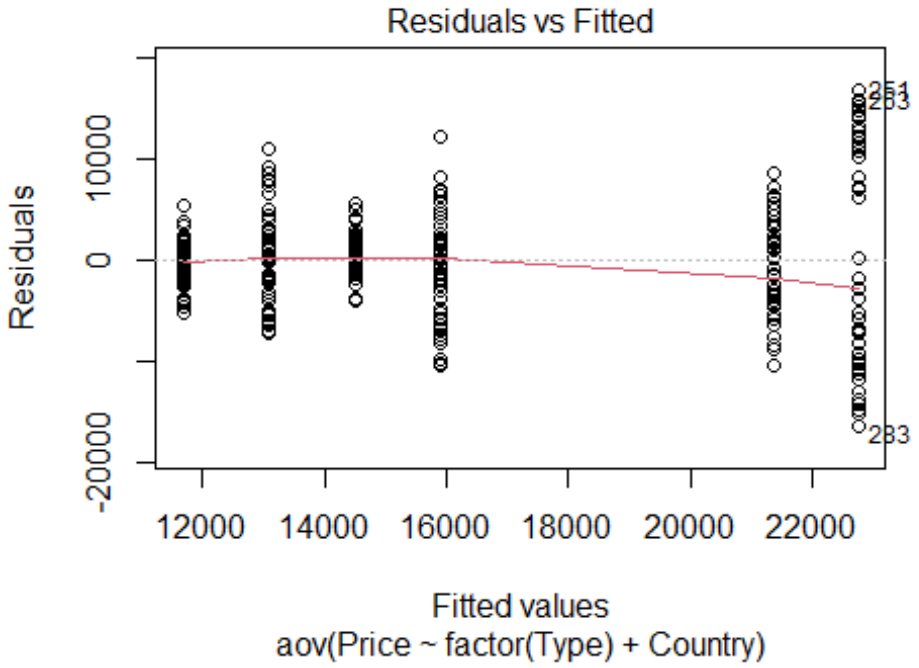
```
TukeyHSD(amod)
```

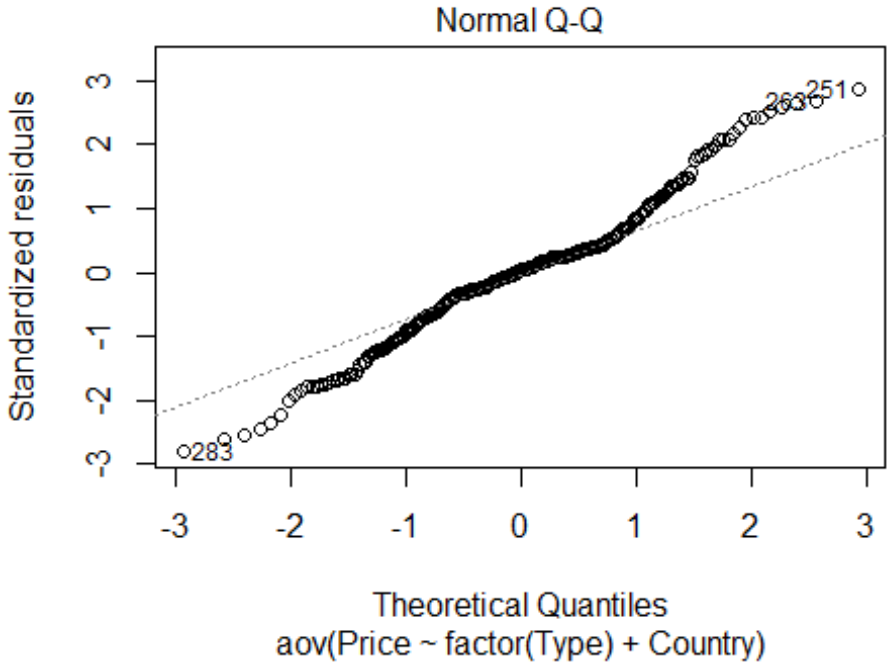
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Price ~ factor(Type) + Country, data = data)
##
## $`factor(Type)`
##             diff       lwr      upr     p adj
## mid-cam 2844.51  885.7404  4803.28 0.0020526
## suv-cam 9691.59 7732.8204 11650.36 0.0000000
## suv-mid 6847.08 4888.3104  8805.85 0.0000000
##
## $Country
##              diff       lwr      upr     p adj
## USA-JP -1391.333 -2727.529 -55.1375 0.0413219
```

8. Produce plots and/or summary statistics to comment on the appropriateness of the following conditions for your data: normality of the residuals, and equality of the variances.

equality: There is a little bit falling trend on the right side. normality: There is a little bit questionable because both of the tails fall off the line.

```
plot(amod, 1:2)
```





9. Construct an ANOVA model for the mean price using the country of the company and the type of vehicle as predictors with the interaction. Include the output showing the ANOVA table; state hypotheses and provide a conclusion in the context of your data. If your ANOVA indicates that there are significant differences among the car price means: Discuss where the significant differences occur using Tukey HSD methods.

   Ho: all $\alpha k=0$ Ha:some $\alpha k\neq 0$

   Ho: all $\beta j=0$ Ha:some $\beta j\neq 0$

   Ho: all $\gamma kj=0$ Ha:some $\gamma kj\neq 0$

   From P values (<2e-16, 0.0413, 0.3491), we have 95% confidence to say that there is main effects, but there is no intersection effects.

   Here we are using different criterias in these two test, TukeyHSD have some significant difference is because that it considers the Type one error and adjust the P values. It is also related with the sample size.

```
modint = aov(Price ~Type+Country+ Type*Country, data=data)
summary(modint)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## Type            2 4.963e+09 2.482e+09  71.806 <2e-16 ***
## Country         1 1.452e+08 1.452e+08   4.201 0.0413 *
## Type:Country    2 7.300e+07 3.650e+07   1.056 0.3491
## Residuals     294 1.016e+10 3.456e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(modint)
```

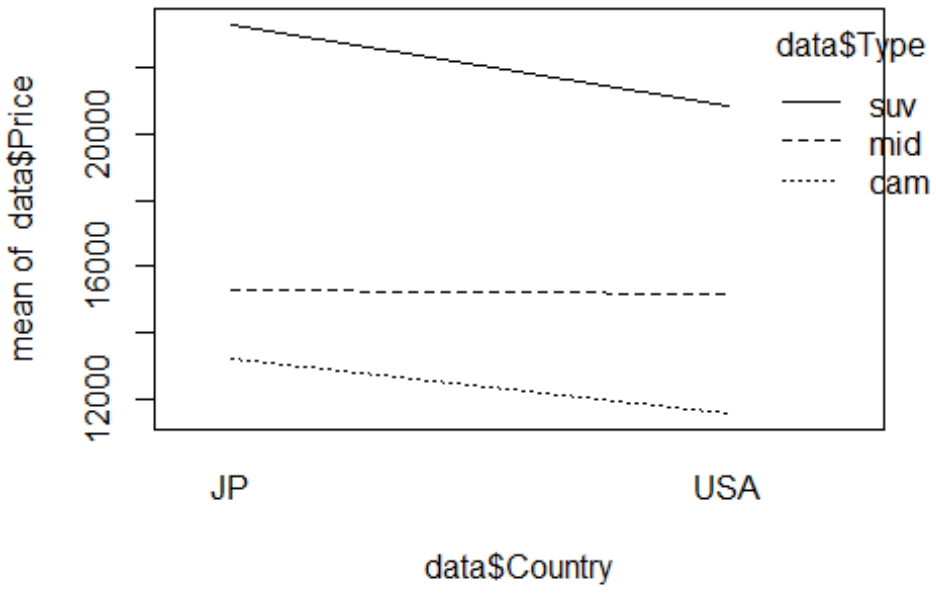
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Price ~ Type + Country + Type * Country, data = d
ata)
##
## $Type
##             diff       lwr       upr     p adj
## mid-cam 2844.51  886.0445  4802.975 0.0020497
## suv-cam 9691.59 7733.1245 11650.055 0.0000000
## suv-mid 6847.08 4888.6145  8805.545 0.0000000
##
## $Country
##              diff       lwr       upr     p adj
## USA-JP -1391.333 -2727.313 -55.35363 0.0412897
##
## $`Type:Country`
##                        diff         lwr        upr     p adj
## mid:JP-cam:JP      2074.44  -1298.3802  5447.2602 0.4905085
## suv:JP-cam:JP     10112.96   6740.1398 13485.7802 0.0000000
## cam:USA-cam:JP    -1623.80  -4996.6202  1749.0202 0.7384206
```

```
## mid:USA-cam:JP    1990.78  -1382.0402  5363.6002 0.5373921
## suv:USA-cam:JP    7646.42   4273.5998 11019.2402 0.0000000
## suv:JP-mid:JP     8038.52   4665.6998 11411.3402 0.0000000
## cam:USA-mid:JP   -3698.24  -7071.0602  -325.4198 0.0223348
## mid:USA-mid:JP     -83.66  -3456.4802  3289.1602 0.9999997
## suv:USA-mid:JP    5571.98   2199.1598  8944.8002 0.0000490
## cam:USA-suv:JP  -11736.76 -15109.5802 -8363.9398 0.0000000
## mid:USA-suv:JP   -8122.18 -11495.0002 -4749.3598 0.0000000
## suv:USA-suv:JP   -2466.54  -5839.3602   906.2802 0.2912596
## mid:USA-cam:USA   3614.58    241.7598  6987.4002 0.0277004
## suv:USA-cam:USA   9270.22   5897.3998 12643.0402 0.0000000
## suv:USA-mid:USA   5655.64   2282.8198  9028.4602 0.0000354
```

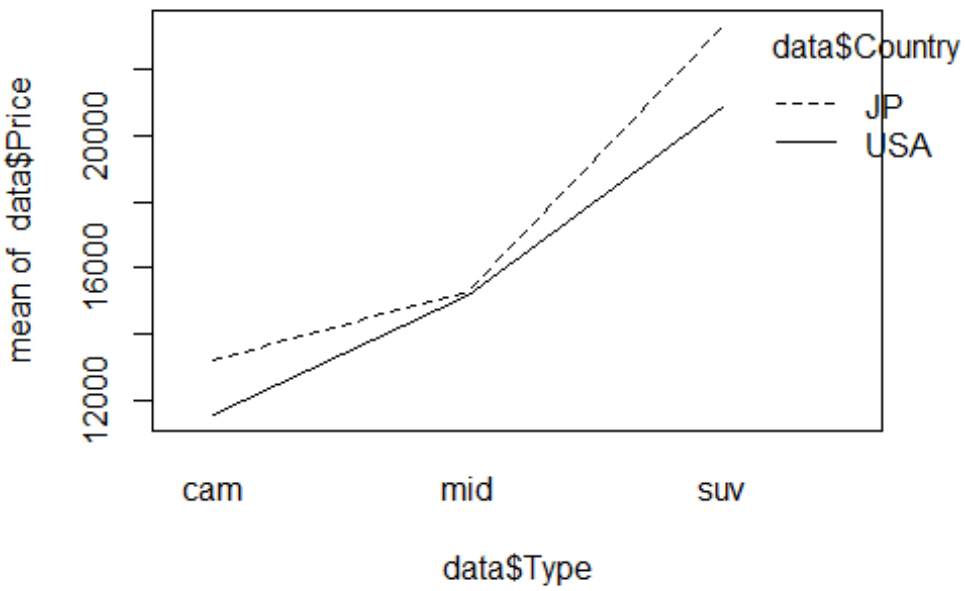
10. Produce two interaction plots for the previous model. If you found significant interactions in your hypothesis test, comment on how these interactions are shown in the plot. If you did not find significant interactions in your hypothesis test, comment on how the (lack of) interactions are shown in the plot.

    There is no significant intersections because the slope of these lines are similar, which means that there is similar change from going to JP to USA for any level of the size of cars, and there is similar change from going to cam to mid to suv for either given countries.

```
interaction.plot(data$Country, data$Type, data$Price)
```




```
interaction.plot(data$Type, data$Country, data$Price)
```



Additional Topics

11. Recall that we can also handle a categorical predictor with multiple categories using ordinary multiple regression if we create indicator variables for each category and include all but one of the indicators in the model. Run an ordinary multiple regression to predict *Price* using the country of the company, the type of vehicle, and the interaction between the two as predictors. Interpret each of the coefficients in the "dummy" regression by what they mean in the context of mean prices.

    interpretation:

    the mean price of JP_com=13183.7

    the mean price of JP_mid=13183.7+2074.4=15258.1

    the mean price of JP_suv=13183.7+10113.0=23296.7

    the mean price of USA_com=13183.7-1623.8=11559.9

    the mean price of USA_mid=13183.7-1623.8+2074.4+1540.1=15174.4

    the mean price of USA_suv=13183.7-1623.8+10113.0-842.7=20830.2

```
mod=lm(Price~factor(Country)+factor(Type)+factor(Country)*factor(Type),
data=data)
summary(mod)

##
## Call:
## lm(formula = Price ~ factor(Country) + factor(Type) + factor(Country
) *
##     factor(Type), data = data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16896.6  -3080.9     63.3   2532.5  16178.4
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|
t|)
## (Intercept)                         13183.7      831.4  15.857  < 2e
-16 ***
## factor(Country)USA                  -1623.8     1175.8  -1.381   0.1
683
## factor(Type)mid                      2074.4     1175.8   1.764   0.0
787 .
## factor(Type)suv                     10113.0     1175.8   8.601  4.8e
-16 ***
## factor(Country)USA:factor(Type)mid   1540.1     1662.8   0.926   0.3
551
## factor(Country)USA:factor(Type)suv   -842.7     1662.8  -0.507   0.6
127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5879 on 294 degrees of freedom
## Multiple R-squared:  0.3377, Adjusted R-squared:  0.3265
## F-statistic: 29.99 on 5 and 294 DF,  p-value: < 2.2e-16
```

```
13183.7+2074.4
```

```
## [1] 15258.1
```

```
13183.7+10113.0
```

```
## [1] 23296.7
```

```
13183.7-1623.8
```

```
## [1] 11559.9
```

```
13183.7-1623.8+2074.4+1540.1
```

```
## [1] 15174.4
```

```
13183.7-1623.8+10113.0-842.7
```

```
## [1] 20830.2
```

12. One possible drawback of the analysis for this assignment is that different people might have chosen vehicles with quite different mileages when collecting their samples. Thus an apparent "difference" between two countries or vehicle types might be due to one sample having considerably more higher mileage vehicles in it than another. Construct a model that

allows you to check for mean price differences between your vehicles from the model constructed in question 11 after accounting for variability due to the mileage of the vehicles. Explain how you use the output from the model to address this question.

From anova test and small p-values we have confidence to say that country, type, and the interaction between country and type have significant effects to the variability in Price.

Since slopes of pink, green and black lines are similar, we know that the price variability in US_suv, JP_suv, JP_com can be explained by mileage similarly. Since slopes of yellow, red blue line is more smooth relatively, we know that as the mileage increases, the prices of US_mid, US_com, JP_mid drop more slowly.

To sum up, since some subgroups have different slopes between price and mileage, part of the mean differences does affected by mileage.

```
mod1=lm(Price~factor(Country)+factor(Type)+factor(Country)*factor(Type)
+Mileage,data=data)
anova(mod1)

## Analysis of Variance Table
##
## Response: Price
##                               Df     Sum Sq    Mean Sq  F value    P
r(>F)
## factor(Country)                1  145185633  145185633   9.7740  0.0
01948 **
## factor(Type)                   2 4963355280 2481677640 167.0676 < 2.
2e-16 ***
## Mileage                        1 5685484847 5685484847 382.7493 < 2.
2e-16 ***
## factor(Country):factor(Type)   2  196027852   98013926   6.5983  0.0
01574 **
## Residuals                    293 4352319151   14854332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

line1=lm(Price ~ Mileage, data=FocusSE)
line2=lm(Price ~ Mileage, data=Civic)
line3=lm(Price ~ Mileage, data=FusionSE)
line4=lm(Price ~ Mileage, data=Accord)
line5=lm(Price ~ Mileage, data=EdgeSEL)
line6=lm(Price ~ Mileage, data=Pilot4WD)

plot(Price ~ Mileage, data=data)
abline(line1)
abline(line2, col='red')
abline(line3, col='blue')
abline(line4, col='yellow')
```

```
abline(line5, col='green')
abline(line6, col='pink')
```