QMB 6304 Analytical Methods for Business | Module 8 Assignment

Pablo Zumba | U54252888

8/5/2022

Contents

Write a simple R script to execute the following data preprocessing and statistical analysis.
Where required show analytical output and interpretations.

Preprocessing

Preprocessing 1

1. Load the file "6304 Module 8 Assignment Data.xlsx" into R. This file contains information on 46,484 vehicles listed for sale on Craig's List in the United States. This will be your master data set.

```
rm(list=ls())
library(rio)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'hms'

## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'tibble'

## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.2
```

```
## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.2
```

```
library(moments)
```

```
## Warning: package 'moments' was built under R version 4.1.2
```

```
library(dplyr)
```

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

## The following object is masked from 'package:car':
##
##     recode

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

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

```
set.seed(54252888)
master_dataset=rio::import("6304 Module 8 Assignment Data.xlsx")
colnames(master_dataset)=tolower(make.names(colnames(master_dataset)))
str(master_dataset)
```

```
## 'data.frame':    46484 obs. of  10 variables:
##  $ region      : chr  "albuquerque" "albuquerque" "albuquerque" "albuquerque" ...
##  $ asking.price: num  15500 17995 18995 8998 22500 ...
##  $ year        : chr  "1965" "2015" "2014" "2012" ...
##  $ make        : chr  "ford" "ford" "ram" "volkswagen" ...
##  $ model       : chr  "mustang" "transit" "promaster 2500" "jetta tdi" ...
##  $ condition   : chr  "excellent" "good" "good" "excellent" ...
##  $ cylinders   : chr  "8" "6" "6" "4" ...
##  $ fuel        : chr  "gas" "gas" "gas" "diesel" ...
##  $ odometer    : num  4800 71181 80483 89000 15700 ...
##  $ paint.color : chr  "blue" "white" "white" "white" ...
```

Filtering and creating a stratified sample

Preprocessing 2

2. Create a single data frame for your analysis which will be your primary data set. The primary data set should have the following characteristics:

- Only includes cars from the regions of Vermont, Appleton, green bay, Indianapolis, and Worcester.

- Only includes cars with 4, 6, or 8-cylinder engines.
- Includes all variables appearing in the master (N=46,484) data set.
- Be a random sample of n=50 cars from each of the five regions listed above. This is referred to as a stratified sample. (Remember to use the numerical portion of your U number as the random number seed.)

```
primary_dataset=subset(master_dataset,region=="vermont"|region=="appleton"|region=="green bay"
                       |region=="indianapolis"|region=="worcester")
unique(primary_dataset$region)
```

```
## [1] "appleton"     "green bay"    "indianapolis" "vermont"      "worcester"
```

```
primary_dataset=subset(primary_dataset,cylinders=="4"|cylinders=="6"|cylinders=="8")
unique(primary_dataset$cylinders)
```

```
## [1] "8" "4" "6"
```

```
stratifiedSample = primary_dataset %>%
  group_by(region) %>%
  sample_n(size=50)
stratifiedSample$region=as.factor(stratifiedSample$region)
stratifiedSample$cylinders=as.factor(stratifiedSample$cylinders)
str(stratifiedSample)
```

```
## grouped_df [250 x 10] (S3: grouped_df/tbl_df/tbl/data.frame)
##  $ region      : Factor w/ 5 levels "appleton","green bay",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ asking.price: num [1:250] 19995 5475 10000 6000 3900 ...
##  $ year        : chr [1:250] "2014" "2013" "2012" "2013" ...
##  $ make        : chr [1:250] "ford" "chevrolet" "volkswagen" "hyundai" ...
##  $ model       : chr [1:250] "e350 superduty" "sonic" "jetta" "elantra" ...
##  $ condition   : chr [1:250] "excellent" "excellent" "excellent" "good" ...
##  $ cylinders   : Factor w/ 3 levels "4","6","8": 3 1 1 1 1 2 2 1 1 3 ...
##  $ fuel        : chr [1:250] "gas" "gas" "diesel" "gas" ...
##  $ odometer    : num [1:250] 109826 132000 99746 153000 184000 ...
##  $ paint.color : chr [1:250] "white" "blue" "black" "black" ...
##  - attr(*, "groups")= tibble [5 x 2] (S3: tbl_df/tbl/data.frame)
##   ..$ region: Factor w/ 5 levels "appleton","green bay",..: 1 2 3 4 5
##   ..$ .rows : list<int> [1:5]
##   .. ..$ : int [1:50] 1 2 3 4 5 6 7 8 9 10 ...
##   .. ..$ : int [1:50] 51 52 53 54 55 56 57 58 59 60 ...
##   .. ..$ : int [1:50] 101 102 103 104 105 106 107 108 109 110 ...
##   .. ..$ : int [1:50] 151 152 153 154 155 156 157 158 159 160 ...
##   .. ..$ : int [1:50] 201 202 203 204 205 206 207 208 209 210 ...
##   .. ..@ ptype: int(0)
##   ..- attr(*, ".drop")= logi TRUE
```

```
attach(stratifiedSample)
```

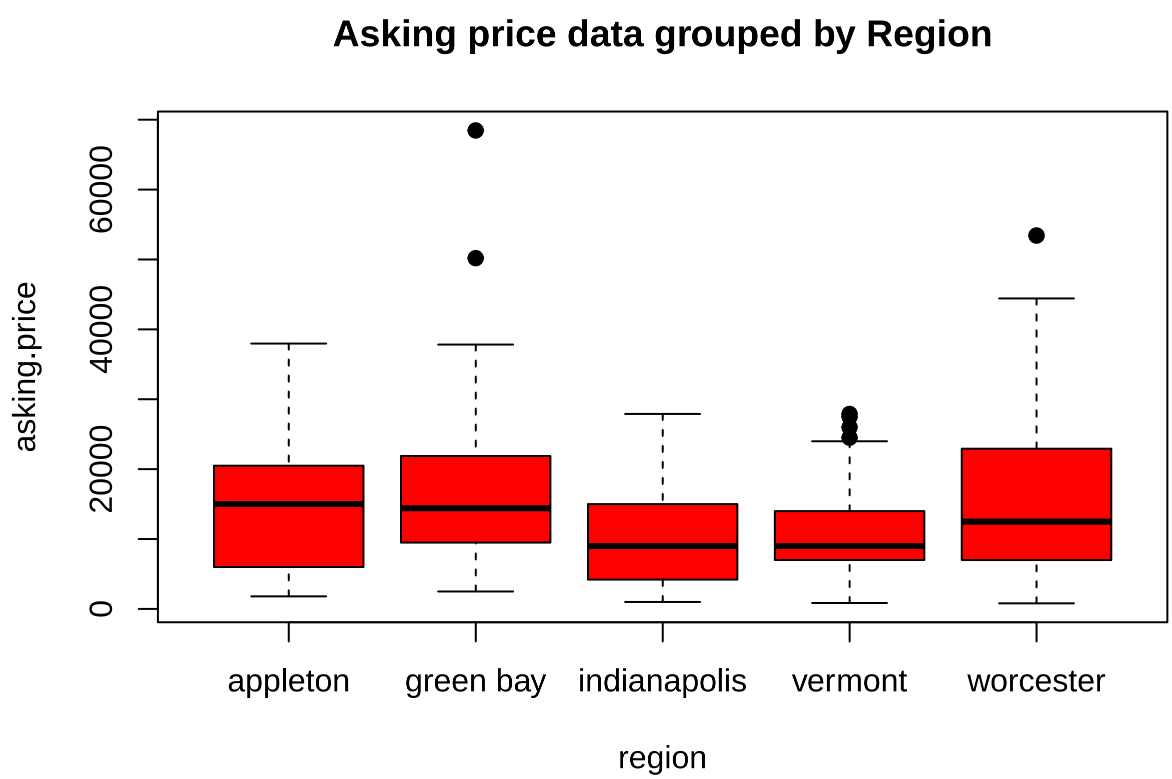
Analysis 1 | Determining equality of variance

1. Within your n=250 stratified sample, determine if asking.price has an equal variance across the five regions. Briefly interpret your results

```
leveneTest(asking.price~region,data=stratifiedSample)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value   Pr(>F)
## group    4  4.1369 0.002917 **
##        245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(asking.price~region,pch=19,col="red",
        main="Asking price data grouped by Region")
```




```
list_variance1=aggregate(asking.price~region,stratifiedSample,var)
list_variance1[order(-list_variance1$asking.price),]
```

```
##         region asking.price
## 5    worcester    162684253
## 2    green bay    149899736
## 1     appleton    100099380
```

```
## 4      vermont     43684628
## 3 indianapolis     43267309
```

Interpretation: "asking.price" does not satisfy the equality of variances. Using the "asking.price" variable, we find the Lavene test produces a p-value of 0.2917% (less than 5%), which means we can reject the Null hypothesis in favor of the alternate hypothesis, namely that there is at least one variance in a region that is different from the other ones. According to the descending order variance table above, Vermont and Indianapolis regions have significantly different variances from the rest regions (this can also be seen in the boxplot).

Analysis 2 | One Way ANOVA: asking.price ~ region

2. Conduct a one-way analysis of variance on your sample data with asking.price as the dependent variable and region as the independent variable. Plot the results of a Tukey HSD test to show whether/where differences in asking.price among the regions exist. Briefly explain the results shown in the plot, stating which pairs of regions do and do not appear to show significant mean differences in asking.price. Make sure region names can be clearly and completely read on the appropriate axis of your plot.

```
analysis2.out=aov(asking.price~region,data=stratifiedSample)
summary(analysis2.out)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## region        4 1.769e+09 442150190   4.425 0.0018 **
## Residuals   245 2.448e+10  99927061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
analysis2.out$coefficients
```

```
##        (Intercept)    regiongreen bay regionindianapolis      regionvermont
##           15444.12            1230.94           -5079.88           -4269.30
##    regionworcester
##             776.94
```

```
list_means2=aggregate(asking.price~region,stratifiedSample,mean)
list_means2[order(-list_means2$asking.price),]
```

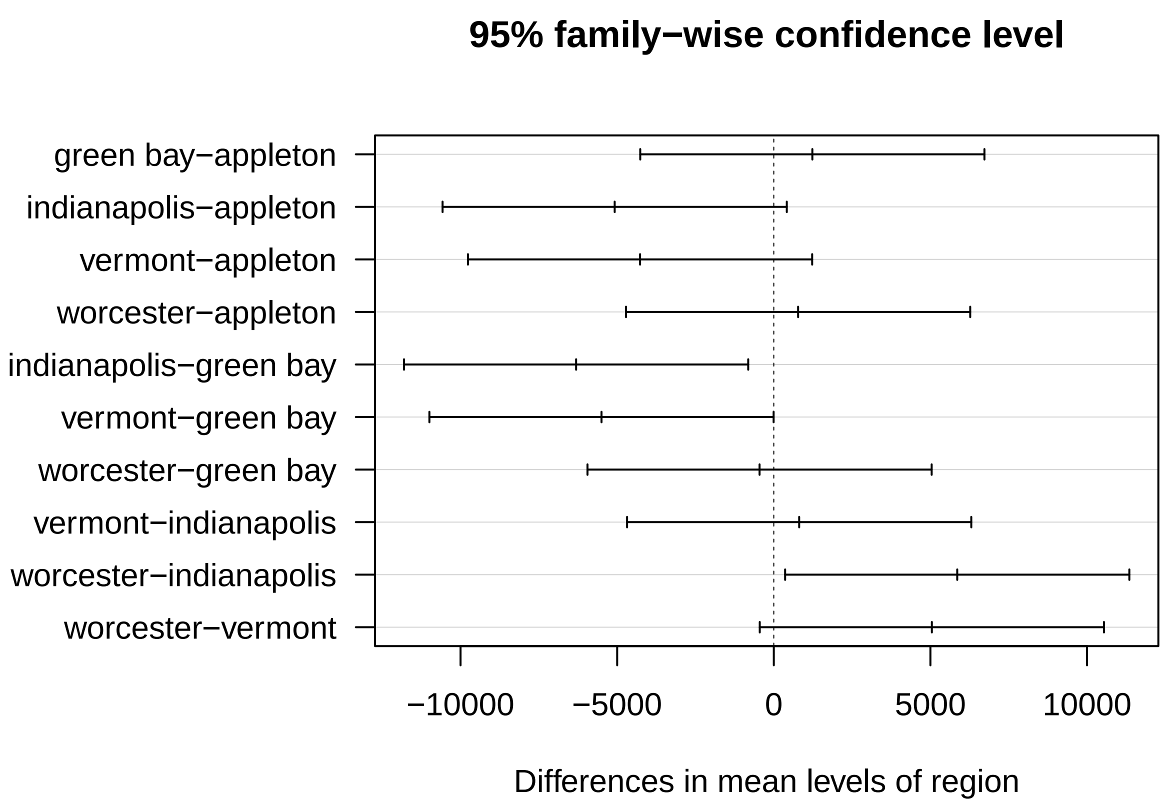
```
##         region asking.price
## 2    green bay     16675.06
## 5    worcester     16221.06
## 1     appleton     15444.12
## 4      vermont     11174.82
## 3 indianapolis     10364.24
```

```
tukey1=TukeyHSD(analysis2.out)
tukey1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = asking.price ~ region, data = stratifiedSample)
##
## $region
##                               diff         lwr          upr     p adj
## green bay-appleton         1230.94  -4263.4577  6725.337724 0.9725046
## indianapolis-appleton     -5079.88 -10574.2777   414.517724 0.0850873
## vermont-appleton          -4269.30  -9763.6977  1225.097724 0.2084941
## worcester-appleton          776.94  -4717.4577  6271.337724 0.9951487
## indianapolis-green bay    -6310.82 -11805.2177  -816.422276 0.0153221
## vermont-green bay         -5500.24 -10994.6377    -5.842276 0.0496098
## worcester-green bay        -454.00  -5948.3977  5040.397724 0.9994073
## vermont-indianapolis        810.58  -4683.8177  6304.977724 0.9942882
## worcester-indianapolis     5856.82    362.4223 11351.217724 0.0302234
## worcester-vermont          5046.24   -448.1577 10540.637724 0.0886505
```

```
par(mar=c(5.1,10,4.1,2.1))
plot(tukey1,las=1)
```




```
par(mar=c(5.1,4.1,4.1,2.1))
```

Interpretation: Indianapolis seems to be the region where car-salesman ask less price on average compared to the other regions. The plot shows only two significant differences in the mean between Indianapolis-Green Bay and Worcester-Indianapolis. If the asking price variable is in USD, then the asking price in Green Bay is 6,310.82 USD higher than the asking price in Indianapolis. Worcester's asking price is 5,856.82 USD higher than Indianapolis'. It appears that Vermont-Green Bay may have a significant difference in mean, but we won't know for sure until we have more observations to make a narrow confidence interval. The other combinations seem to have no difference in mean.

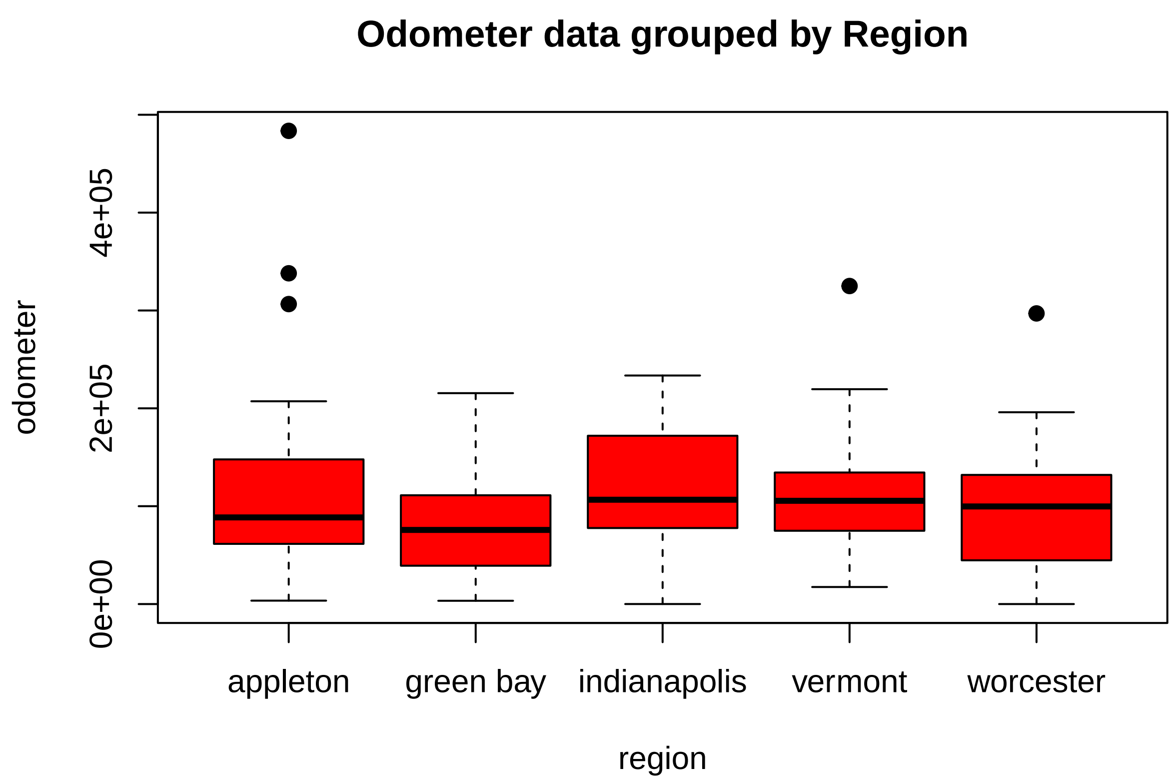
Analysis 3 | One Way ANOVA: odometer ~ region

3. Repeat Steps 1 and 2 above using the odometer as the dependent variable and the region as the independent variable. Again, briefly explain your analysis results and make sure region names can be clearly and completely read on the appropriate axis of your plot.

```
leveneTest(odometer~region,data=stratifiedSample)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  1.8313 0.1234
##       245
```

```
boxplot(odometer~region,pch=19,col="red",
        main="Odometer data grouped by Region")
```




```
list_variance2=aggregate(odometer~region,stratifiedSample,var)
list_variance2[order(-list_variance2$odometer),]
```

```
##         region   odometer
## 1     appleton 7784319868
## 3 indianapolis 4015769406
## 5    worcester 3393090230
## 4      vermont 2965265560
## 2    green bay 2717930264
```

```
analysis3.out=aov(odometer~region,data=stratifiedSample)
summary(analysis3.out)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## region        4 3.592e+10 8.979e+09   2.151 0.0752 .
## Residuals   245 1.023e+12 4.175e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
analysis3.out$coefficients
```

```
##        (Intercept)    regiongreen bay regionindianapolis      regionvermont
##          109756.80          -27268.38            5993.88            -180.74
##    regionworcester
##          -13570.12
```

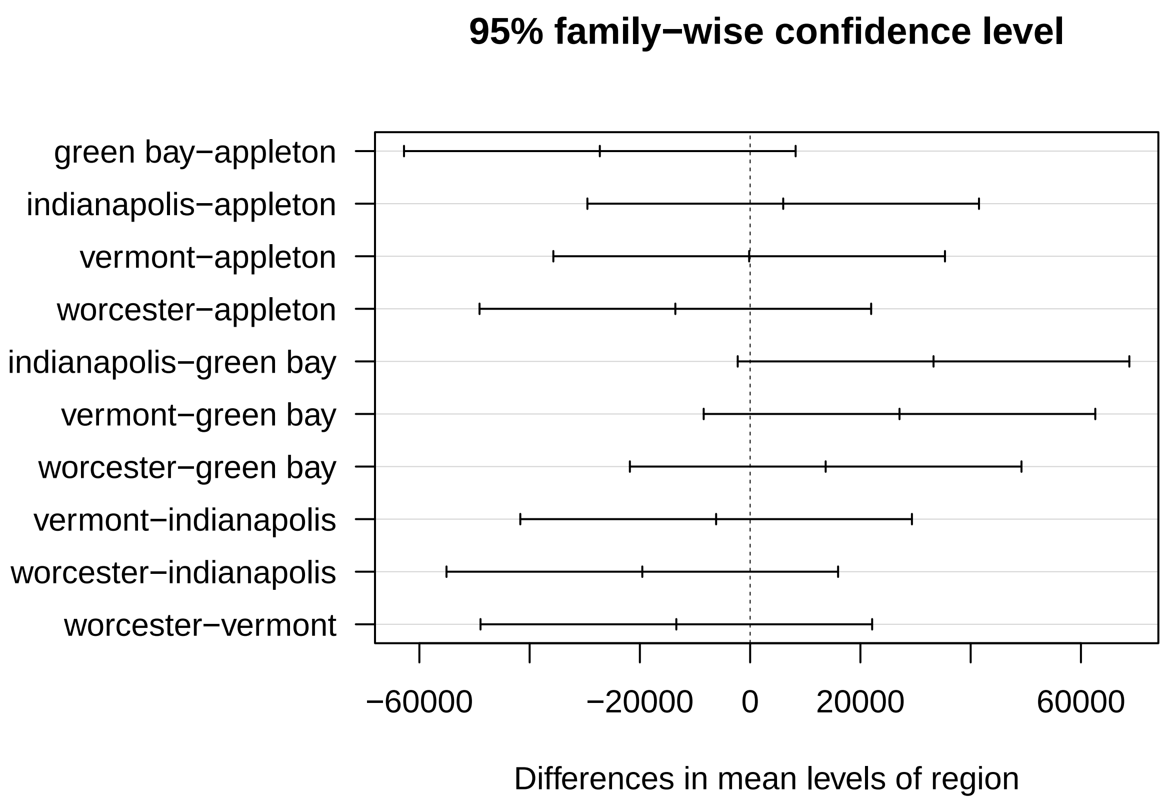
```
list_means3=aggregate(odometer~region,stratifiedSample,mean)
list_means3[order(-list_means3$odometer),]
```

```
##         region  odometer
## 3 indianapolis 115750.68
## 1     appleton 109756.80
## 4      vermont 109576.06
## 5    worcester  96186.68
## 2    green bay  82488.42
```

```
tukey3=TukeyHSD(analysis3.out)
tukey3
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = odometer ~ region, data = stratifiedSample)
##
## $region
##                              diff        lwr       upr     p adj
## green bay-appleton     -27268.38 -62784.137  8247.377 0.2190928
## indianapolis-appleton    5993.88 -29521.877 41509.637 0.9904473
## vermont-appleton         -180.74 -35696.497 35335.017 1.0000000
## worcester-appleton     -13570.12 -49085.877 21945.637 0.8316361
## indianapolis-green bay  33262.26  -2253.497 68778.017 0.0784440
## vermont-green bay       27087.64  -8428.117 62603.397 0.2250785
## worcester-green bay     13698.26 -21817.497 49214.017 0.8267632
## vermont-indianapolis    -6174.62 -41690.377 29341.137 0.9893078
## worcester-indianapolis -19564.00 -55079.757 15951.757 0.5544622
## worcester-vermont      -13389.38 -48905.137 22126.377 0.8383946
```

```r
par(mar=c(5.1,10,4.1,2.1))
plot(tukey3,las=1)
```




```r
par(mar=c(5.1,4.1,4.1,2.1))
```

Interpretation: Odometer readings in the five regions do not differ significantly (p-value 12.3% > 5% | fail to reject the Null Hypothesis) according to the Lavene test. According to the plot "Odometer data grouped by Region", there are no significant differences in mean among the five regions. An Indianapolis-Green Bay combination has a very close difference in mean, but it needs to be analyzed using more observations to make a narrow confidence interval.

Analysis 4 | One Way ANOVA: asking.price ~ cylinders

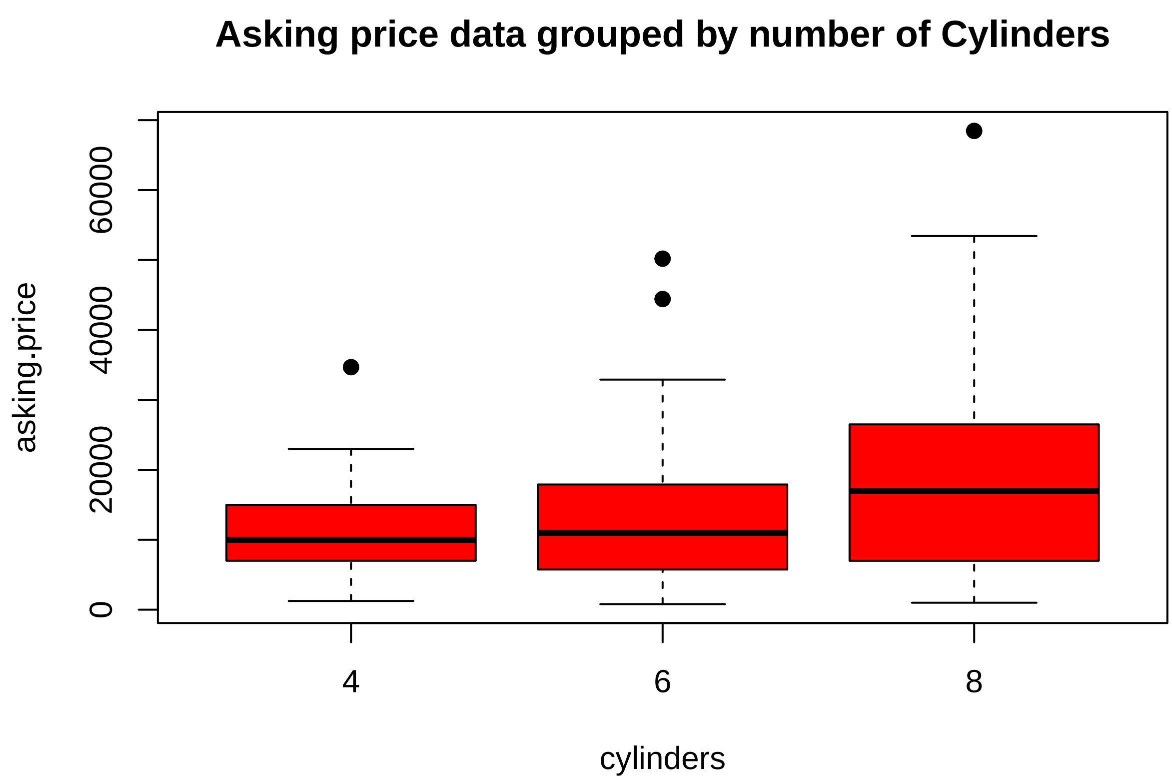
4. Referring to Steps 1 and 2 again, conduct a one-way analysis of variance using asking.price as the dependent variable and cylinders as the independent. Show model output and explain your results as you did in Step 3.

```r
leveneTest(asking.price~cylinders,data=stratifiedSample)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  16.762 1.491e-07 ***
##       247
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(asking.price~cylinders,pch=19,col="red",
        main="Asking price data grouped by number of Cylinders")
```




```
list_variance3=aggregate(asking.price~cylinders,stratifiedSample,var)
list_variance3[order(-list_variance3$asking.price),]
```

```
##   cylinders asking.price
## 3         8    180269319
## 2         6     93615299
## 1         4     32050507
```

```
analysis4.out=aov(asking.price~cylinders,data=stratifiedSample)
summary(analysis4.out)
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## cylinders     2 2.306e+09 1.153e+09    11.9 1.17e-05 ***
## Residuals   247 2.394e+10 9.694e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
analysis4.out$coefficients
```

```
## (Intercept)  cylinders6  cylinders8
##   11170.584    1850.063    7376.895
```

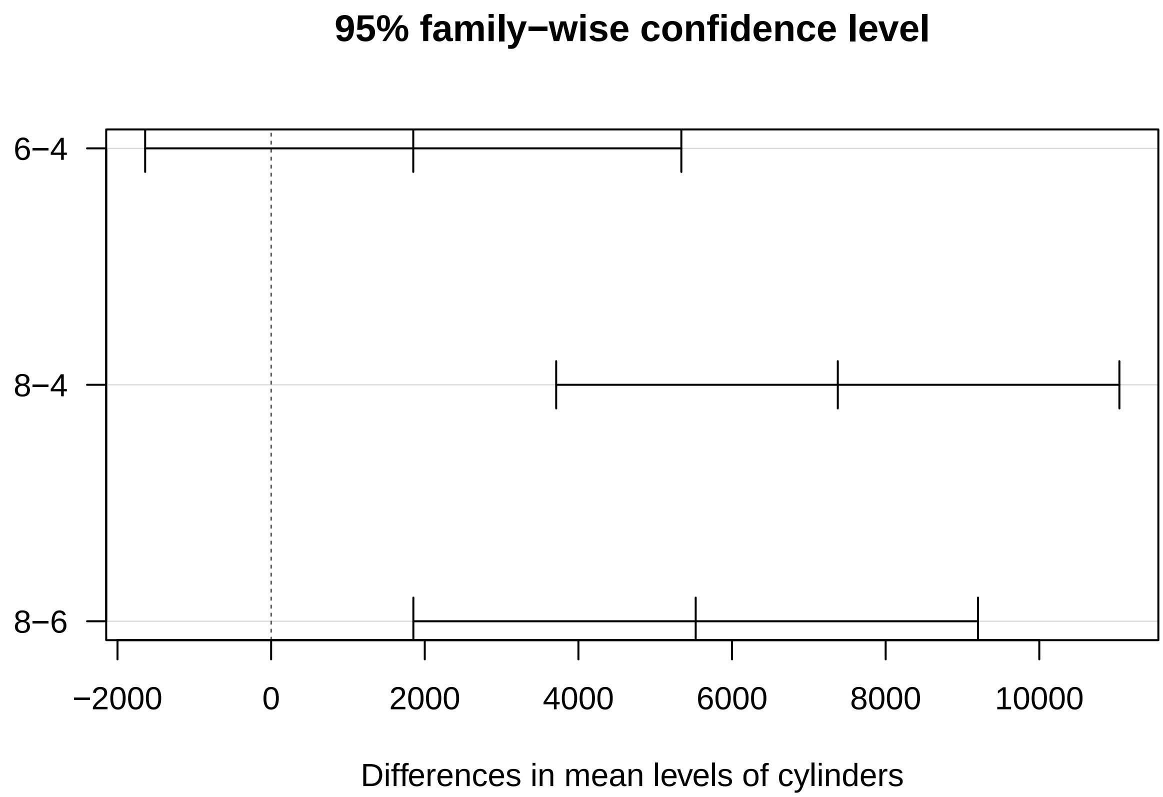
```
list_means4=aggregate(asking.price~cylinders,stratifiedSample,mean)
list_means4[order(-list_means4$asking.price),]
```

```
##   cylinders asking.price
## 3         8     18547.48
## 2         6     13020.65
## 1         4     11170.58
```

```
tukey4=TukeyHSD(analysis4.out)
tukey4
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = asking.price ~ cylinders, data = stratifiedSample)
##
## $cylinders
##         diff       lwr       upr     p adj
## 6-4 1850.063 -1640.009  5340.136 0.4250352
## 8-4 7376.895  3710.956 11042.835 0.0000105
## 8-6 5526.832  1851.518  9202.145 0.0013559
```

```
par(mar=c(5.1,3,4.1,2.1))
plot(tukey4,las=1)
```




```r
par(mar=c(5.1,4.1,4.1,2.1))
```

Interpretation: There is at least one significant variance in the number of cylinders versus the asking price based on the Lavene test (p-value = 1.491e-07). Based on the table and boxplot above, the variance appears to increase as the number of cylinders increases. The Tukey test also shows there is at least one significant difference in the mean by the number of cylinders (p-value = 1.17e-05). On the above plot "Asking price data grouped by the number of Cylinders", we can see that there is a significant difference between 8-4 cylinders and 8-6 cylinders. There is no significant difference in the mean when comparing 6-4 cylinders. The greater the number of cylinders, the greater the variance and the greater the difference in mean when compared to fewer cylinders.