Ryan Timbrook

Data Science 350 – Homework Assignment 4

Assignment:

1.) Compare and test Normality of the distributions of price and log price - use both a graphical method and a formal test
2.) Test significance of price (log price) stratified by: a) fuel type, b) aspiration, c) rear vs. front wheel drive - use both graphical methods and the format test.
3.) Apply ANOVA to the auto price data to compare the price (or log price if closer to a Normal distribution) of autos stratified by number of doors, and body style - two sets of tests -Graphically explore the differences between the price conditioned by the categories of each variable
-Use standard ANOVA and Tukey ANOVA to test the differences of these groups.

Observations:

- Auto Prices do not have a normal distribution. This is shown in Table 1 and 2 where the plots are shown to not follow a close straight line between the sample sets. This outcome is visualized more clearly in Table 4 where the log normal price sample set 1 and 2 were plotted against each other.
- Price by Fuel Type:
  - At 95% confidence we **cannot reject** the null hypothesis that these means are the same. The p-value is greater than .025 and the confidence interval **overlaps zero**. This is represented in Table 6 below.
- Price by Aspiration:
  - At 95% confidence we **can reject** the null hypothesis that these means are the same. The p-value is significantly less than .025 and the confidence **interval does not overlaps zero**. This is represented in Table 7 below.
- Price by Rear vs Front Wheel Drive:
  - At 95% confidence we **can** reject the null hypothesis that these means are the same. The p-value is significantly less than .025 and the confidence **interval does not overlaps zero**. This is represented in Table 8 below.

- ANOVA testing:
  - Price by Body Style
    - Body Style has a **significant** impact on auto prices. Based on the high F statistic shown below and the very small p-value we **can reject** the null hypothesis that these groups mean values are the same for all body styles. This is represented in Tables 9 and 12 below.
  - Price by Number of Doors
    - Number of Doors does **not have a significant** impact on auto prices. Based on the low F statistic shown below and the greater than .025 p-value we **cannot reject** the null hypothesis that these groups mean values are the same for both door types. This is represented in Tables 10 and 11 below.

Table 1: Normality Test – Q-Q plot Graph of two samples of auto prices

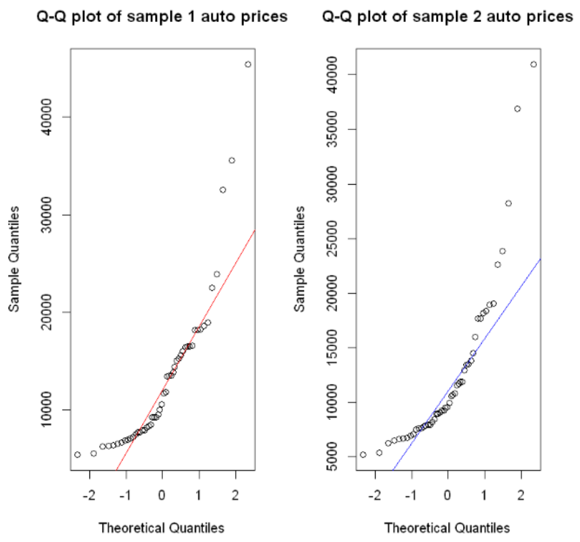


Table 2: Normality Test – Q-Q plot Graph of two samples of log normal auto prices

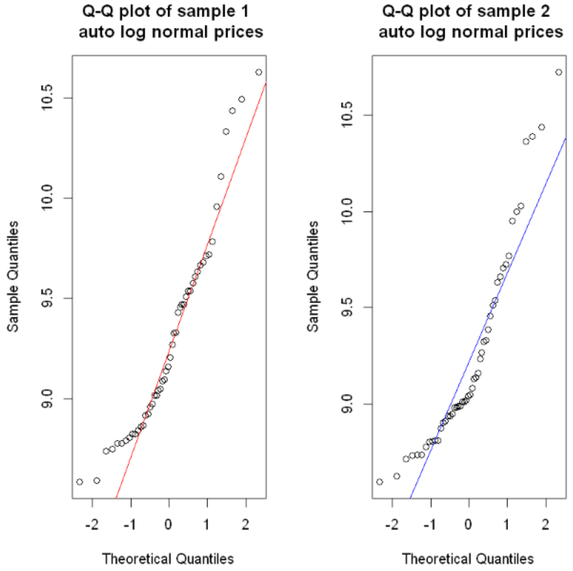


Table 3: Normality Test – Sample 1 price distribution plotted against Sample 2 price distribution

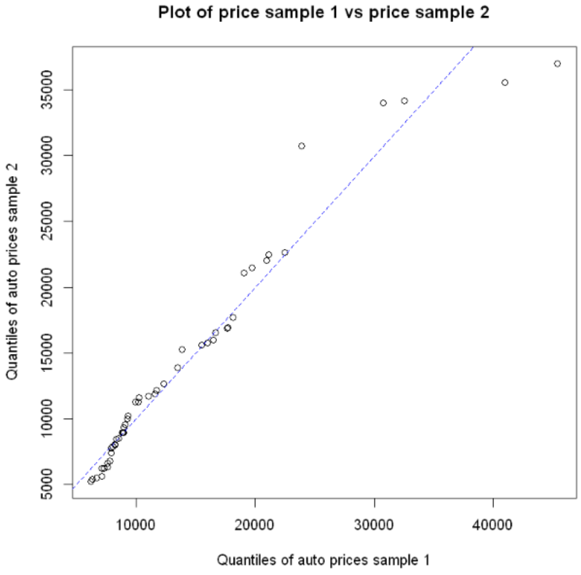


Table 4: Normality Test – Sample 1 log normal price distribution plotted against Sample 2 log normal price distribution

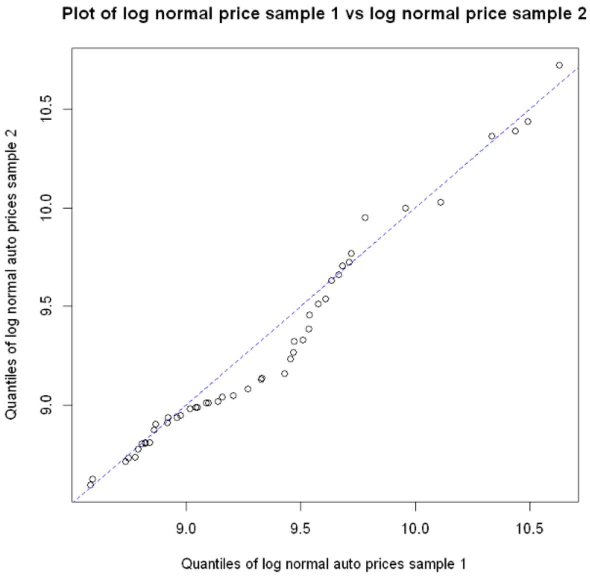


Table 5: Kolmogorov-Smirnov test for distributions of two samples of the log normal price

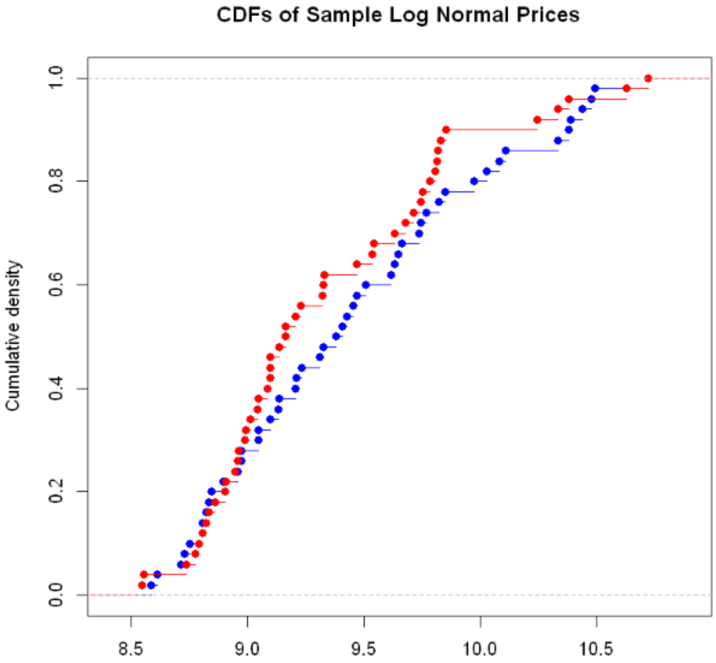
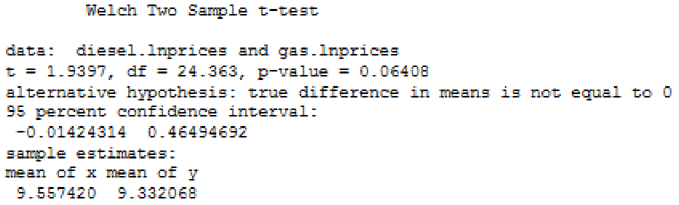


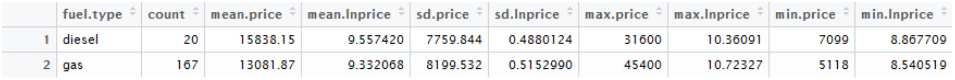
Table 6: Significance test, Price comparison by Diesel vs. Gas Fueled Cars

At 95% confidence we **cannot** reject the null hypothesis that these means are the same. The p-value is greater than .025 and the confidence interval overlaps zero.

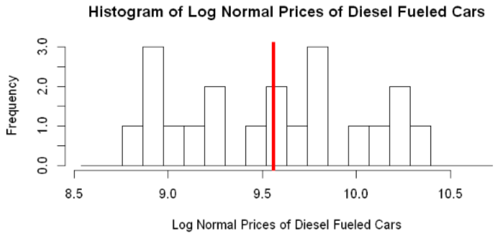


```
        Welch Two Sample t-test

data:  diesel.lnprices and gas.lnprices
t = 1.9397, df = 24.363, p-value = 0.06408
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01424314  0.46494692
sample estimates:
mean of x mean of y 
 9.557420  9.332068 
```



	fuel.type	count	mean.price	mean.lnprice	sd.price	sd.lnprice	max.price	max.lnprice	min.price	min.lnprice
1	diesel	20	15838.15	9.557420	7759.844	0.4880124	31600	10.36091	7099	8.867709
2	gas	167	13081.87	9.332068	8199.532	0.5152990	45400	10.72327	5118	8.540519



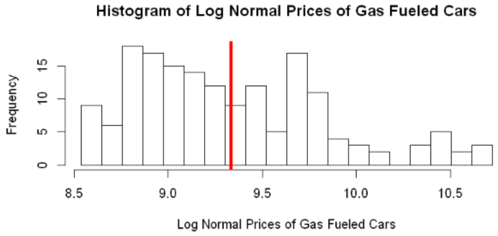
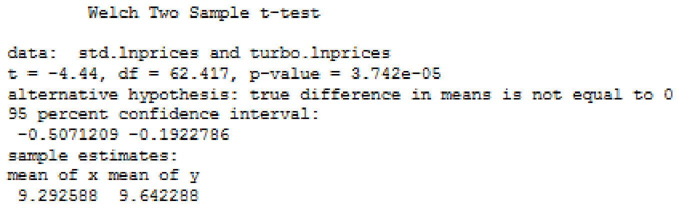


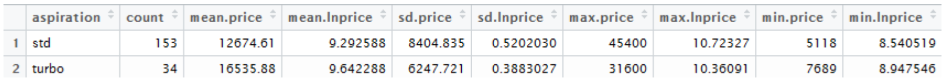
Table 7: Significance test, Price comparison by Turbo vs. Standard Cars

At 95% confidence we **can** reject the null hypothesis that these means are the same. The p-value is significantly less than .025 and the confidence interval does not overlaps zero.

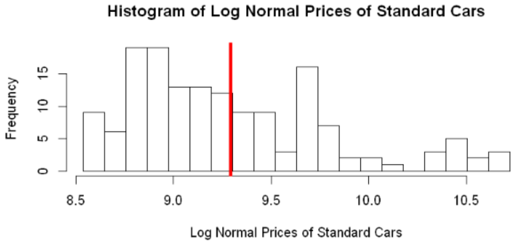


```
        Welch Two Sample t-test

data:  std.lnprices and turbo.lnprices
t = -4.44, df = 62.417, p-value = 3.742e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5071209 -0.1922786
sample estimates:
mean of x mean of y 
 9.292588  9.642288 
```



	aspiration	count	mean.price	mean.lnprice	sd.price	sd.lnprice	max.price	max.lnprice	min.price	min.lnprice
1	std	153	12674.61	9.292588	8404.835	0.5202030	45400	10.72327	5118	8.540519
2	turbo	34	16535.88	9.642288	6247.721	0.3883027	31600	10.36091	7689	8.947546



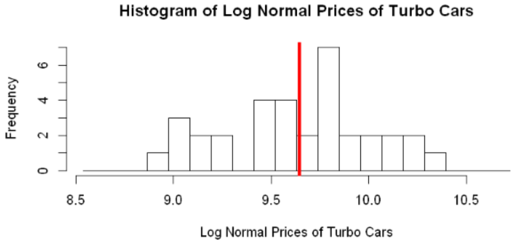
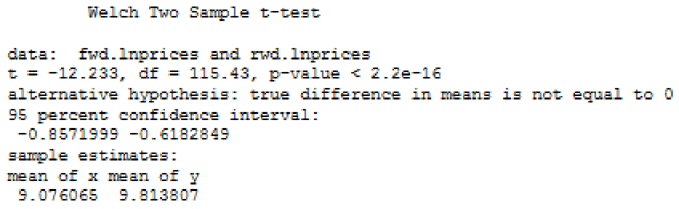


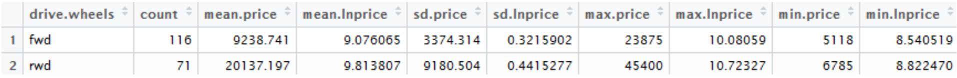
Table 8: Significance test, Price comparison by Front Wheel Drive vs. Rear Wheel Drive Cars

At 95% confidence we **can** reject the null hypothesis that these means are the same. The p-value is significantly less than .025 and the confidence interval does not overlaps zero.

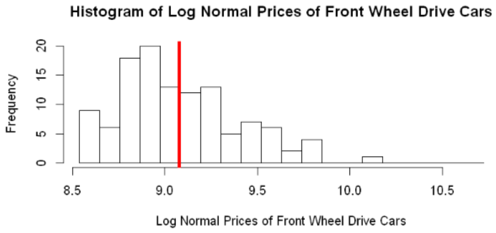


```
        Welch Two Sample t-test

data:  fwd.lnprices and rwd.lnprices
t = -12.233, df = 115.43, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8571999 -0.6182849
sample estimates:
mean of x mean of y 
 9.076065  9.813807 
```



	drive.wheels	count	mean.price	mean.lnprice	sd.price	sd.lnprice	max.price	max.lnprice	min.price	min.lnprice
1	fwd	116	9238.741	9.076065	3374.314	0.3215902	23875	10.08059	5118	8.540519
2	rwd	71	20137.197	9.813807	9180.504	0.4415277	45400	10.72327	6785	8.822470



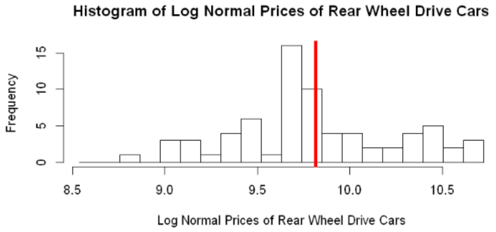
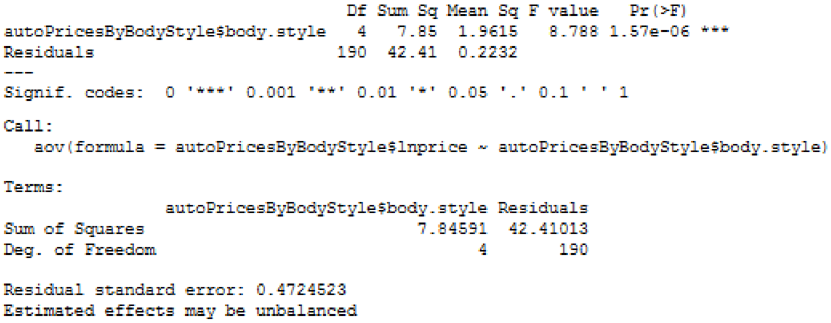


Table 9: Boxplot Graph of LN Auto Prices by Body Style

Body Style **has a significant** impact on auto prices. Based on the high F statistic shown below and the very small p-value we can reject the null hypothesis that these groups mean values are the same for all body styles

ANOVA Summary Data:



```
                             Df Sum Sq Mean Sq F value   Pr(>F)    
autoPricesByBodyStyle$body.style   4   7.85  1.9615   8.788 1.57e-06 ***
Residuals                      190  42.41  0.2232                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
   aov(formula = autoPricesByBodyStyle$lnprice ~ autoPricesByBodyStyle$body.style)

Terms:
                autoPricesByBodyStyle$body.style Residuals
Sum of Squares                           7.84591  42.41013
Deg. of Freedom                                4       190

Residual standard error: 0.4724523
Estimated effects may be unbalanced
```

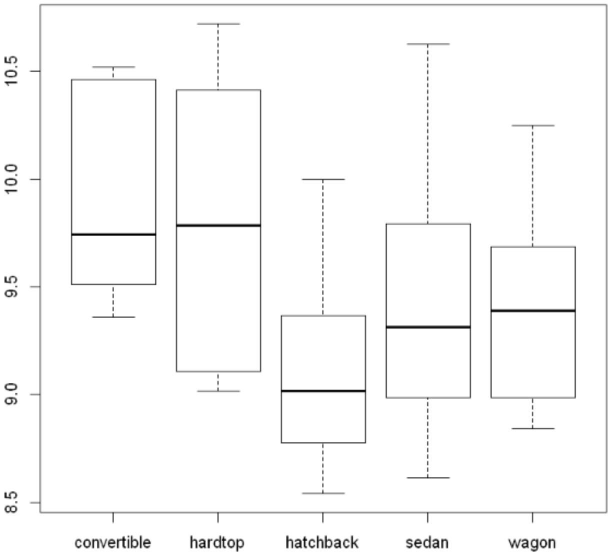
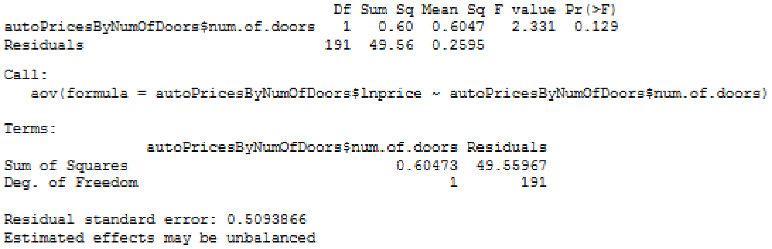


Table 10: Boxplot Graph of LN Auto Prices by Number of Doors

Number of Doors does **not have a significant** impact on auto prices. Based on the low F statistic shown below and the greater than .025 p-value we **cannot reject** the null hypothesis that these groups mean values are the same for both door types

ANOVA Summary Data:



```
                              Df Sum Sq Mean Sq F value Pr(>F)
autoPricesByNumOfDoors$num.of.doors   1   0.60  0.6047   2.331  0.129
Residuals                       191  49.56  0.2595               

Call:
   aov(formula = autoPricesByNumOfDoors$lnprice ~ autoPricesByNumOfDoors$num.of.doors)

Terms:
                autoPricesByNumOfDoors$num.of.doors Residuals
Sum of Squares                              0.60473  49.55967
Deg. of Freedom                                   1       191

Residual standard error: 0.5093866
Estimated effects may be unbalanced
```

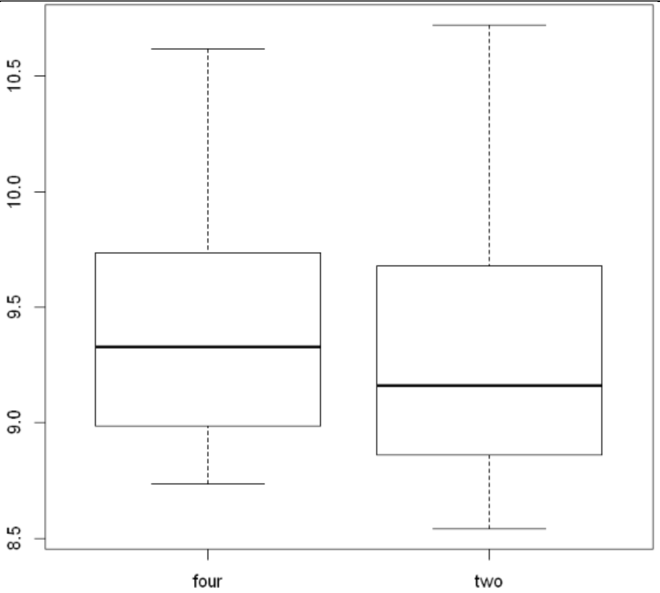
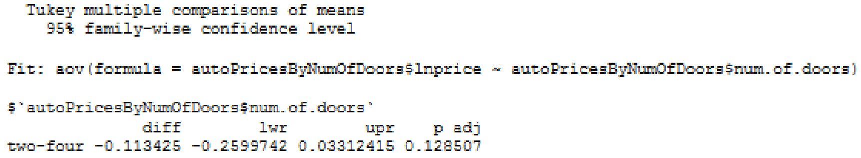


Table 11: Tukey ANOVA – HSD Test – LN Auto Price by Number of Doors

Number of Doors does **not have a significant** impact on auto prices. We **cannot reject** the null hypothesis that these groups mean values are the same for both door types

Summary Data:



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

 Fit: aov(formula = autoPricesByNumOfDoors$lnprice ~ autoPricesByNumOfDoors$num.of.doors)

 $`autoPricesByNumOfDoors$num.of.doors`
               diff        lwr        upr    p adj
 two-four -0.113425 -0.2599742 0.03312415 0.128507
```

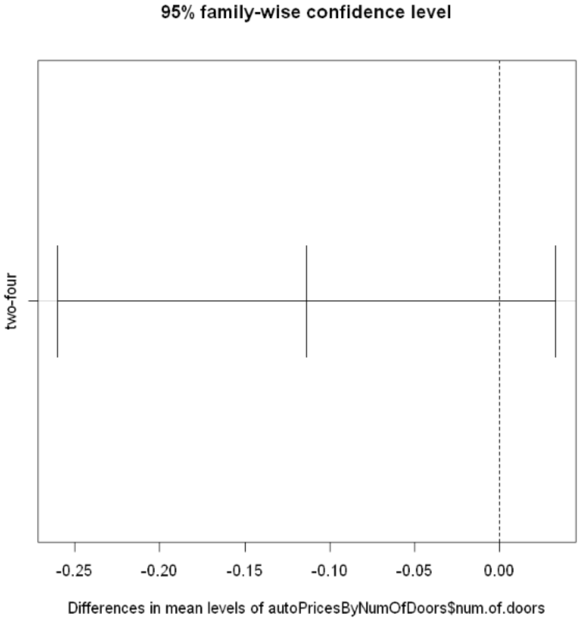
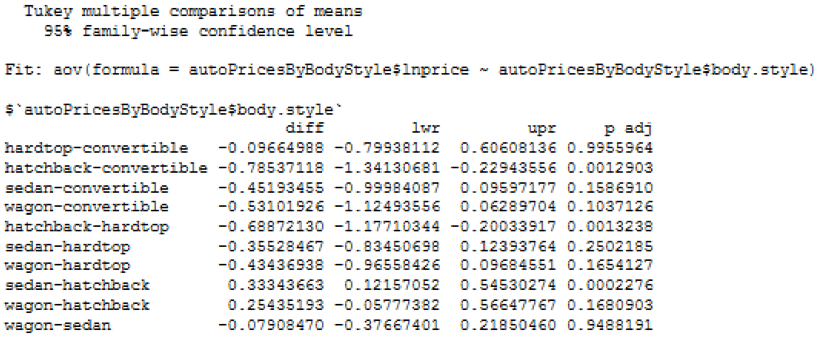


Table 12: Tukey ANOVA – HSD Test

Body Style **has a significant** impact on auto prices. We can reject the null hypothesis that body style mean prices are the same for all groupings. The graph and data summary below shows seven of the groups cross over the zero line representing a significant difference in mean values.

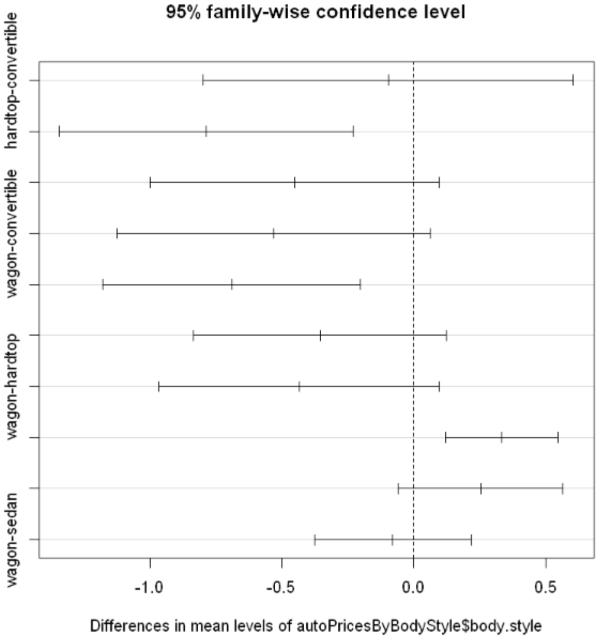
Summary Data:



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

 Fit: aov(formula = autoPricesByBodyStyle$lnprice ~ autoPricesByBodyStyle$body.style)

 $`autoPricesByBodyStyle$body.style`
                             diff         lwr         upr     p adj
 hardtop-convertible   -0.09664988 -0.79938112  0.60608136 0.9955964
 hatchback-convertible -0.78537118 -1.34130681 -0.22943556 0.0012903
 sedan-convertible     -0.45193455 -0.99984087  0.09597177 0.1586910
 wagon-convertible     -0.53101926 -1.12493556  0.06289704 0.1037126
 hatchback-hardtop     -0.68872130 -1.17710344 -0.20033917 0.0013238
 sedan-hardtop         -0.35528467 -0.83450698  0.12393764 0.2502185
 wagon-hardtop         -0.43436938 -0.96558426  0.09684551 0.1654127
 sedan-hatchback        0.33343663  0.12157052  0.54530274 0.0002276
 wagon-hatchback        0.25435193 -0.05777382  0.56647767 0.1680903
 wagon-sedan           -0.07908470 -0.37667401  0.21850460 0.9488191
```


95% family-wise confidence level
hardtop-convertible
wagon-convertible
wagon-hardtop
wagon-sedan
Differences in mean levels of autoPricesByBodyStyle$body.style