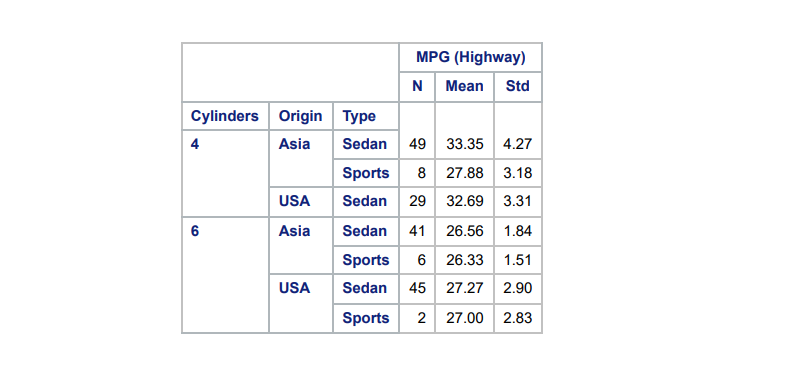
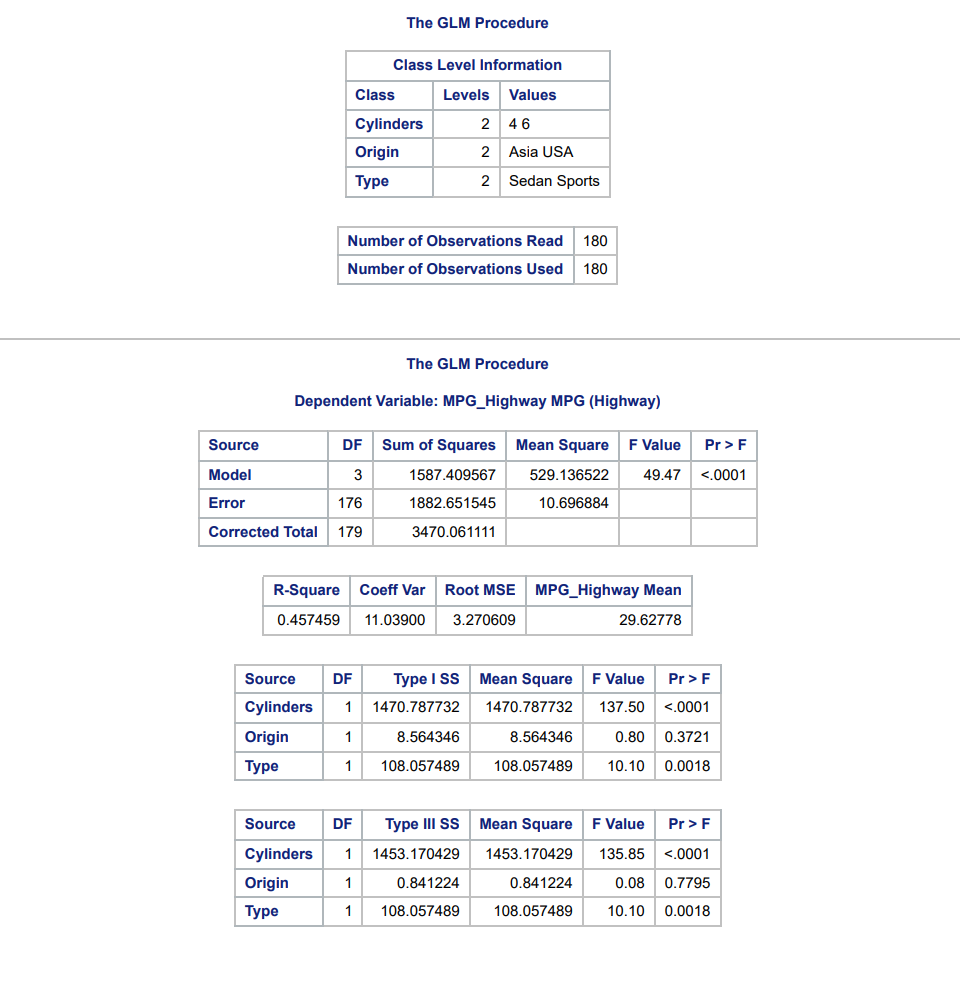
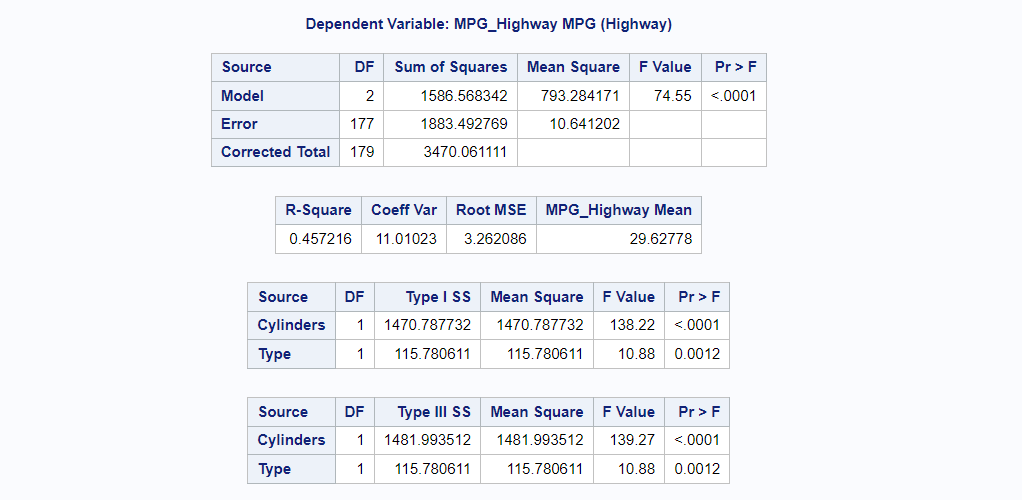
HW 3

## Problem 1

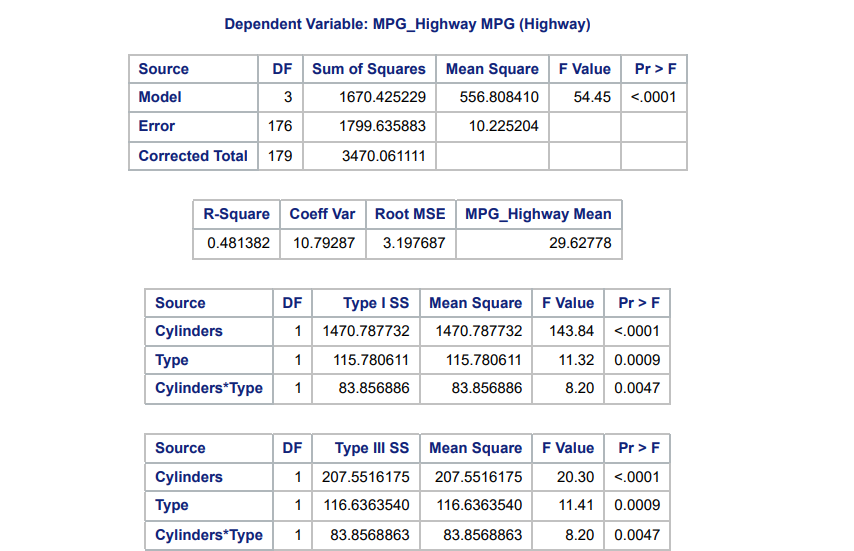
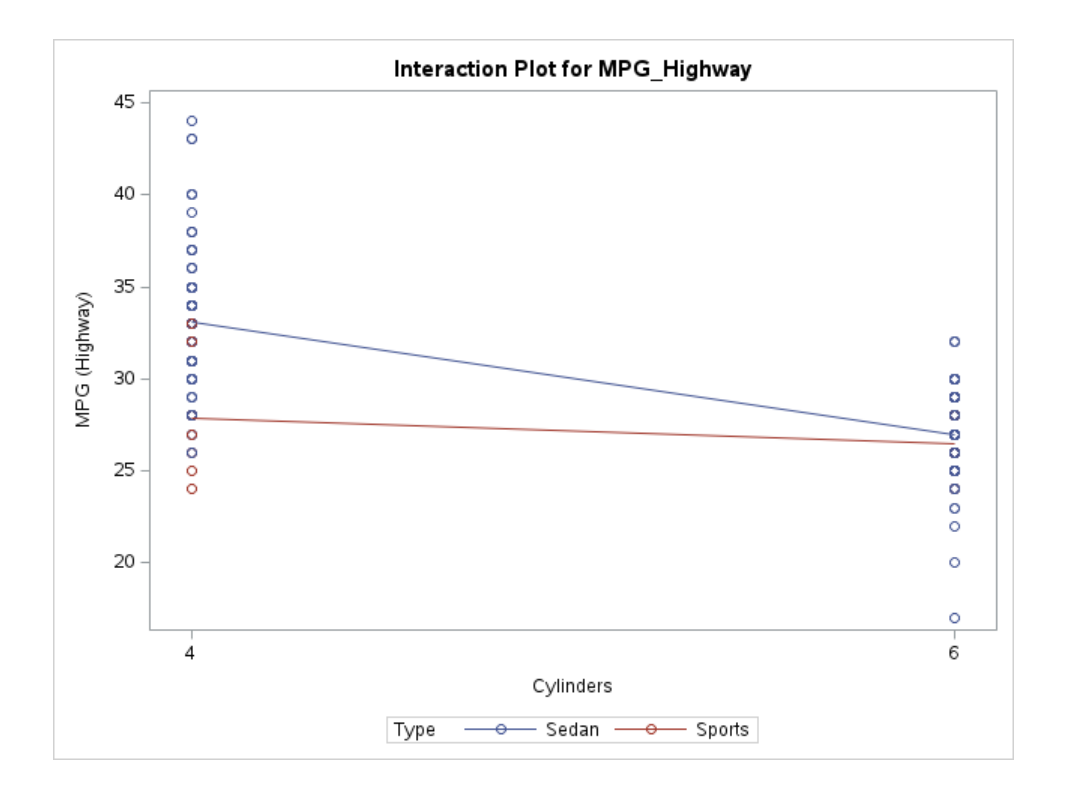
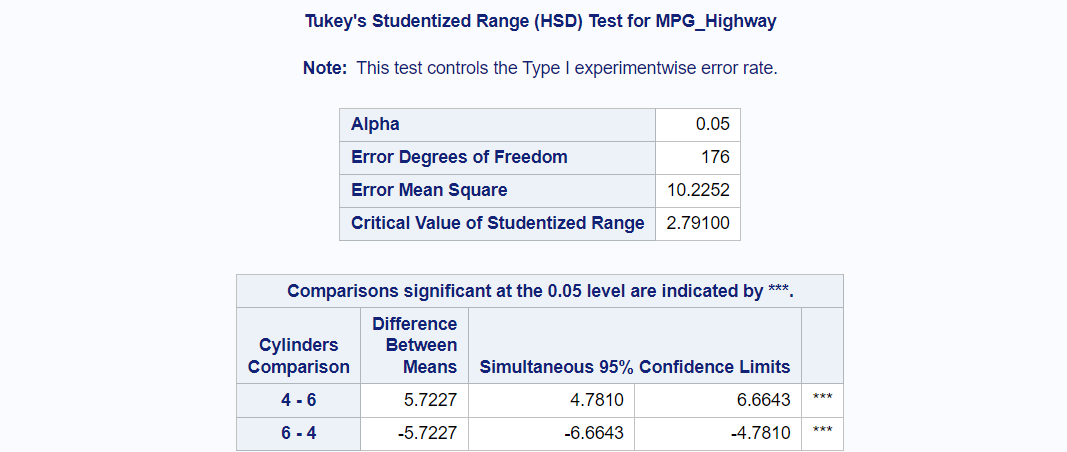
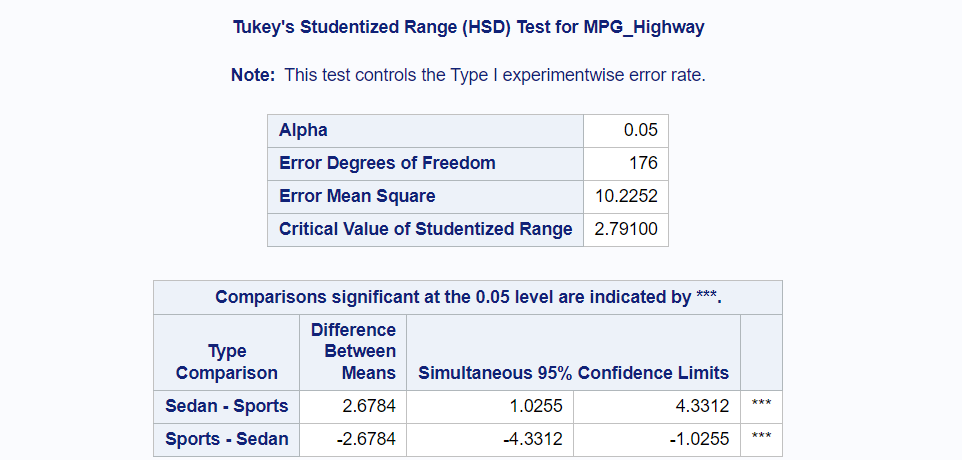
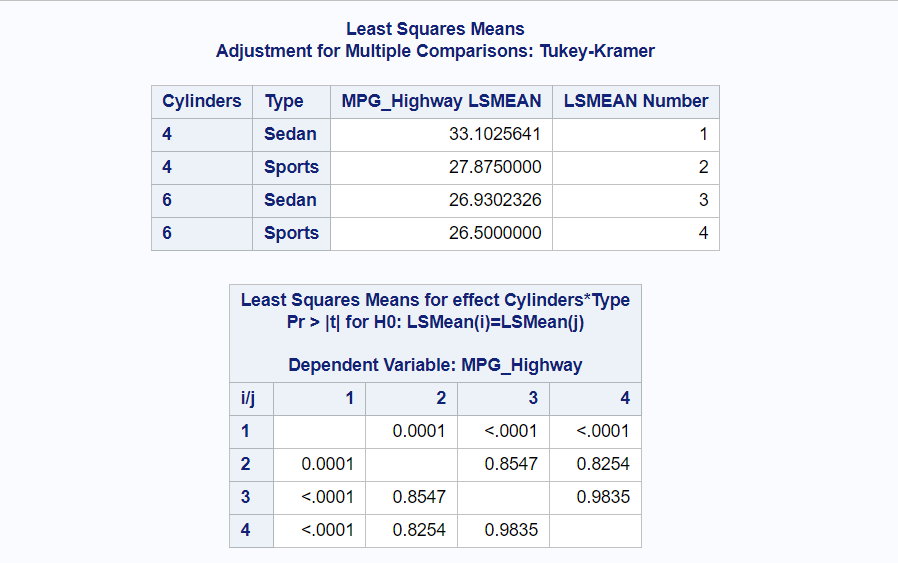
### 1a.)

  
The first apparent interesting feature is that there is no observations with 4 cylinders originating in the USA which are of Type Sports. Additionally, the data is not balanced at all with a large amount of variation in the number of counts for each combination of groups. It appears that cars with 6 cylinders have lower average mpg than those with 4 cylinders. For 6 cylinder cars, those originating in the US appear to have a larger mean MPG, but it is not much higher. Additionally for 4 cylinder cars, Sedans appear to have a larger mean MPG which is no the case for 6 cylinder cars.

### 1b.)

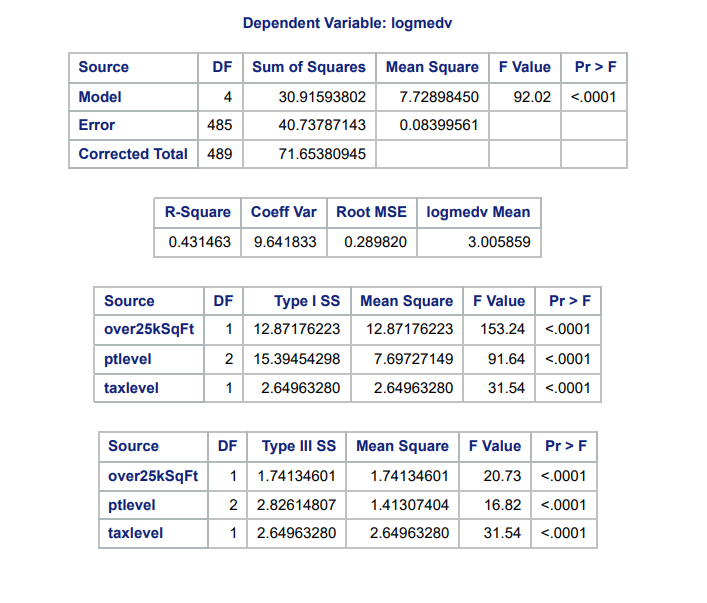
  
Since the data is unbalanced, we use the proc GLM procedure to fit the main effects model using Cylinders, Origin and Type. Since the P-value for Origin is large, we would not keep this term as it does not contribute significantly to predicting MPG. The other two predictors are significant, so we drop the term and fit the main effects model with just Cylinders and Type:  
  
We see that both the terms are still significant and the model itself is significantly different enough from the model with just the global mean. This is because the p-values are all very small, however the model isn't that good as we only explain around 45.7% of the variation in the data with this model as we see from the R-square value.

### 1c.)

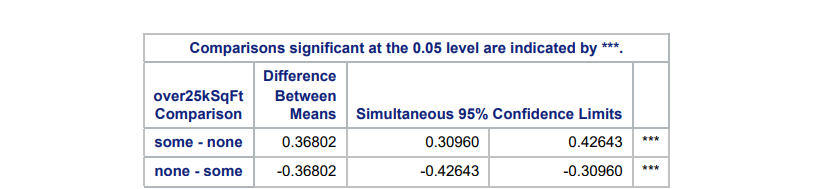
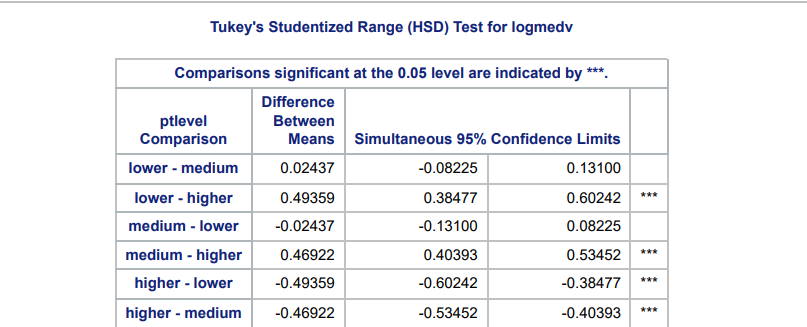
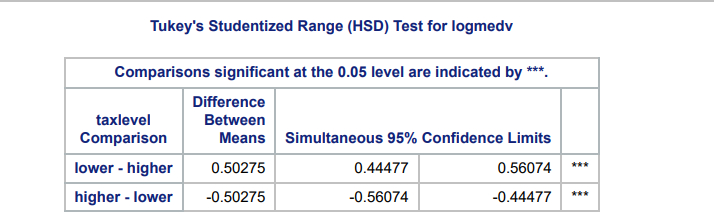
  
Starting with the main effects model of just Cylinders and Type, we add the interaction term between Cylinders and Type. Since the p-value is small for this term, we see that it is also significant so we would decide to keep it in the model. We also see that for the entire model, since the p-value is small, we know this model is significantly different from the model with just the global mean. We also see a small improvement in the amount of variation explained by our model with our R-square number going up to around 0.481 or 48.1% of the variation of the data being explained by our model.  
  
Looking at the interaction plot we see that the lines are not parallel meaning that the mean mpg changes between 4 & 6 Cylinders cars by different amounts when controlling for the type of car.  
  
We see that the difference in mean MPG between is significant between 4 & 6 cylinder cars, with 6 cylinder cars having lower mean mpg then 4 cylinder cars.  
  
We also observe a significant difference in mean mpg between sedans and sports cars with Sedans having higher average mpg when compared to sports cars.  
  
We see that cars that are both Sedans & 4 cylinders have significantly higher LS mean mpg when compared to all of the other group pairings. However, the rest of the group pairings do not have significantly different LS means because their p-values are large.

## Problem 2

### 2a.)

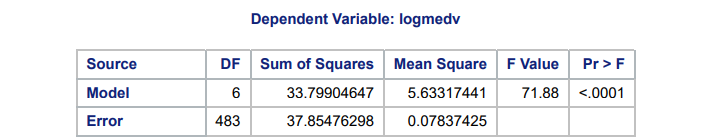
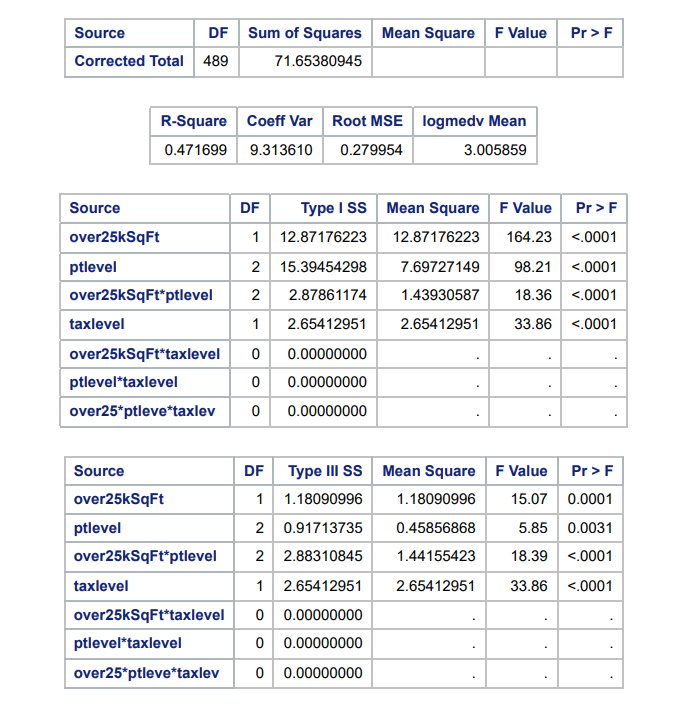
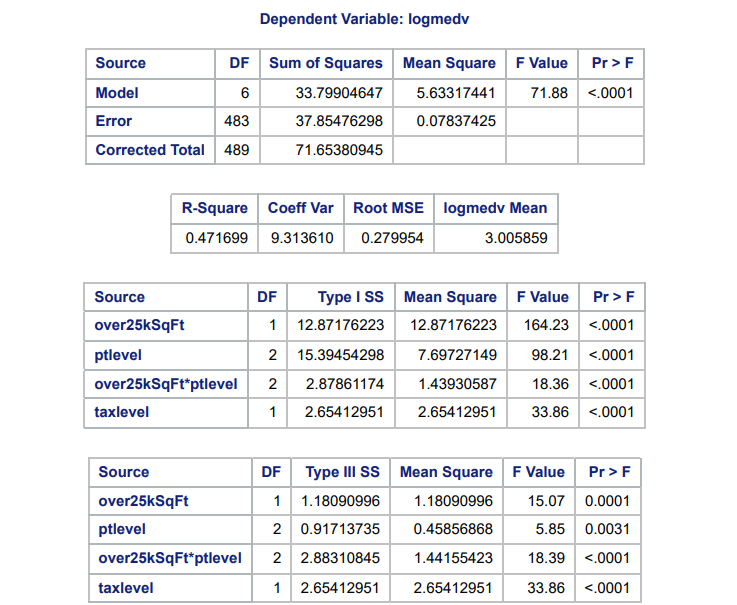
  
I started by fitting the model with all 3 of the terms listed, and since they are all significant due to small p-values, I decided to keep them all in the main effects model. The model itself is also significant since its p-value is small, however we only explain around 43.1% of the variation in the log median home values with our model.

### 2b.)

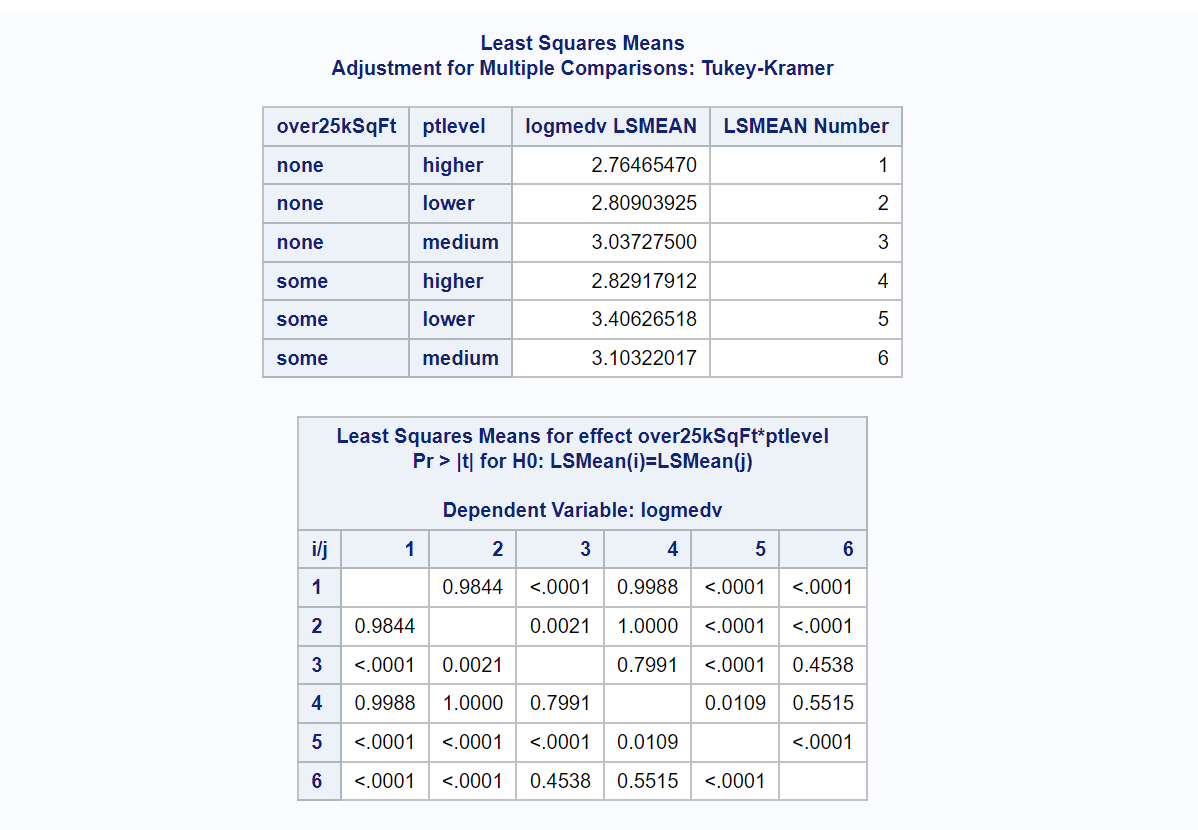
  
We see that suburbs with some houses with over 25k SqFt tend to have significantly higher mean in the log median value of the homes.  
  
We see that there is a significant difference between mean log median home value for suburbs with pupil to teacher ratio that are higher compared to medium and lower. But there was not a significant difference between medium and lower ptlevel suburbs for mean log median home value.  
  
We see that for the tax level, that the mean log median home value differs significantly between higher and lower tax levels, with lower tax levels having higher mean log median home values.

## Problem 3

### 3a.)

  
  
I started by adding the 2nd order interaction term, however from the output we see that it is not significant and in addition to this the interaction terms between tax level and the other two variables in the first order interactions are also not significant. So I removed these 3 terms from the model to get:  
  
This model is significantly different than the null model with just the global mean due to its small p-value and we can see that the individual terms are also all significant. By including the interaction term between over25kSqFt and ptlevel, we gain a slight amount in the R-square value and now the model explains about 47.1% of the variability in the log median home value, which is still not very good.

### 3b.)

  
For the group combinations we see that there is a significant difference between indices 1 an 3, meaning that suburbs with no homes over 25k Sq Ft differ in LS Mean of log Median home value when comparing medium and higher ptlevel.

Additionally there is a significant difference in LS mean of Log median home values between suburbs with no homes over 25k SqFt & higher ptlevel when compared to lower and medium ptlevel but when the suburbs had some homes over 25k SqFt.

Another interesting thing is that when there are some homes over 25k SqFt, there is a significant difference between lower and medium ptlevel in LS mean log median home value which is different then when we don't consider whether there are homes over 25k Sq Ft.  
There also exist more significantly different group pairing as seen by the extremely small p-values in the chart, and the larger p-value group pairs are not statistically significant.