Midterm Exam 1 - Open Book Section (R) - Part 2

Instructions

This R Markdown file includes the questions, the empty code chunk sections for your code, and the text blocks for your responses. Answer the questions below by completing this R Markdown file. You must answer the questions using this file. You can change the format from pdf to Word or html and make other slight adjustments to get the file to knit but otherwise keep the formatting the same. Once you've finished answering the questions, submit your responses in a single knitted file (just like the homework peer assessments).

There are 13 questions total, each worth between 2-6 points. Partial credit may be given if your code is correct but your conclusion is incorrect or vice versa.

Next Steps:

1. Save this .Rmd file in your R working directory - the same directory where you will download the "real estate.csv" data file into. Having both files in the same directory will help in reading the .csv file.
2. Read the question and create the R code necessary within the code chunk section immediately below each question. Knitting this file will generate the output and insert it into the section below the code chunk.
3. Type your answer to the questions in the text block provided immediately after the response prompt.
4. Once you've finished answering all questions, knit this file and submit the knitted file on Canvas.

Example Question 12 - 4pts

This will be the exam question - each question is already copied from Canvas and inserted into individual text blocks below, *you do not need to copy/paste the questions from the online Canvas exam.*

```
# Example code chunk area. Enter your code below the comment‘
# Read the data set
house <- read.csv('Real estate.csv', header=TRUE)
# Set Purchase Year as categorical
#house$Purchase.Year<-as.factor(house$Purchase.Year)
```

Response to Question 12: This is the section where you type your written answers to the question. Depending on the question asked, your typed response may be a number, a list of variables, a few sentences, or a combination of these elements.

Ready? Let's begin.

Real Estate Data Analysis

For this exam, you will be building a model to predict the Price per Area in Taipei, Taiwan.

The "real estate.csv" data set consists of the following variables:

- *Purchase Year*: year the property was purchased
- *Purchase Month*: month the property was purchased
- *Age*: age of the property, years since property was built
- *MTR Distance*: distance to the nearest Metro Rapid Transit Station (meters)
- *Number of Stores*: number of market stores within 1 km
- *Latitude*: latitude coordinates
- *Longitude*: longitude coordinates
- *Price per Area*: sell price per unit area (dollar per square meter)

Read the data and answer the questions below. Assume a significance threshold of 0.05 for hypothesis tests unless stated otherwise.

```
# Read the data set
house = read.csv('C:/Users/mjpearl/Desktop/omsa/ISYE-6414-OAN/midterm1/Real estate.csv', header=TRUE)
#Set Purchase Year as categorical
house$Purchase.Year<-as.factor(house$Purchase.Year)
head(house)
```

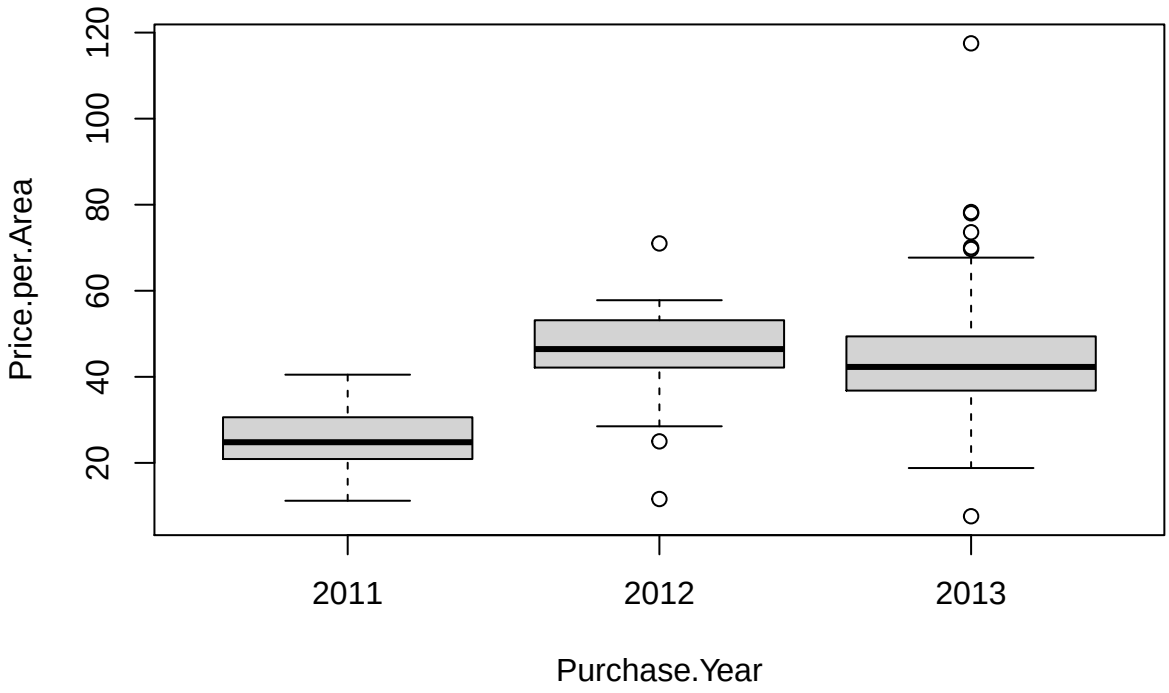
```
##   Purchase.Year Purchase.Month  Age  MRT.Dist Num.of.Stores Latitude Longitude
## 1          2013              4 14.8  393.2606             6 24.96172  121.5381
## 2          2011              7 17.4 6488.0210             1 24.95719  121.4735
## 3          2012              1 16.0 4066.5870             0 24.94297  121.5034
## 4          2011              0 30.9 6396.2830             1 24.94375  121.4788
## 5          2011              0 16.5 4082.0150             0 24.94155  121.5038
## 6          2011              1 32.0 1156.7770             0 24.94935  121.5305
##   Price.per.Area
## 1            7.6
## 2           11.2
## 3           11.6
## 4           12.2
## 5           12.8
## 6           12.8
```

Note: For all of the following questions, treat all variables as quantitative variables except for *Purchase.Year*.

Question 1 - 4pts

Create a box plot comparing the response variable *Price per Area* across the *Purchase Year* categories. Based on this box plot, does *Purchase Year* appear useful in predicting the house price? Include your reasoning.

```
# Code to create boxplot...
boxplot(Price.per.Area ~ Purchase.Year,data=house)
```



Response to Question 1: Based on our results, we can see that the purchase year seems to have an impact on price year to year as the average seems to increase. Hence, there does seem to be a relationship between the price and purchase year of the car.

Question 2 - 4pts

Create an ANOVA model to compare the mean *Price per Area* among the three years. Display the corresponding ANOVA table.

A) How many Degrees of Freedom are there for the residuals?

B) Provide the formula for this calculation.

```
# Code to create an ANOVA model and display the ANOVA table...
anova_model = aov(Price.per.Area ~ Purchase.Year,data=house)
summary(anova_model)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Purchase.Year   2  31283   15641   142.3 <2e-16 ***
## Residuals     411  45179     110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Response to Question 2A: There are 411 DF

Response to Question 2B: Formula for DF Residuals = N - k N: Number of observations k: Number of classes for the categorical variable (i.e in our case the 3 years)

Question 3 - 4pts

A) What can you conclude from the ANOVA table with respect to the test of equal means at a significance level of 0.05 (accept or reject null hypothesis)?
B) Provide conclusions in the context of the problem.

Response to Question 3A: Reject the null hypothesis because our p-value $<$ the alpha level 0.05

Response to Question 3B: We reject the null hypothesis in that the mean Price per area of the 3 different car years is equal. Therefore we can conclude that the mean Price per area is different for at least one of the Purchase years.

Question 4 - 4pts

Conduct a Tukey pairwise comparison of the mean prices per unit area for the three years.

A) Which pairs of years have significantly different means at the 0.05 significance level?

B) Also provide conclusions in the context of the average purchase price per area. Use the difference in average purchase price in your explanation.

```
# Code to create pairwise-comparison...
TukeyHSD(anova_model,"Purchase.Year",conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Price.per.Area ~ Purchase.Year, data = house)
##
## $Purchase.Year
##                diff       lwr       upr     p adj
## 2012-2011 20.395979 16.492590 24.299368 0.0000000
## 2013-2011 17.785680 15.111209 20.460152 0.0000000
## 2013-2012 -2.610298 -6.303272  1.082676 0.2209272
```

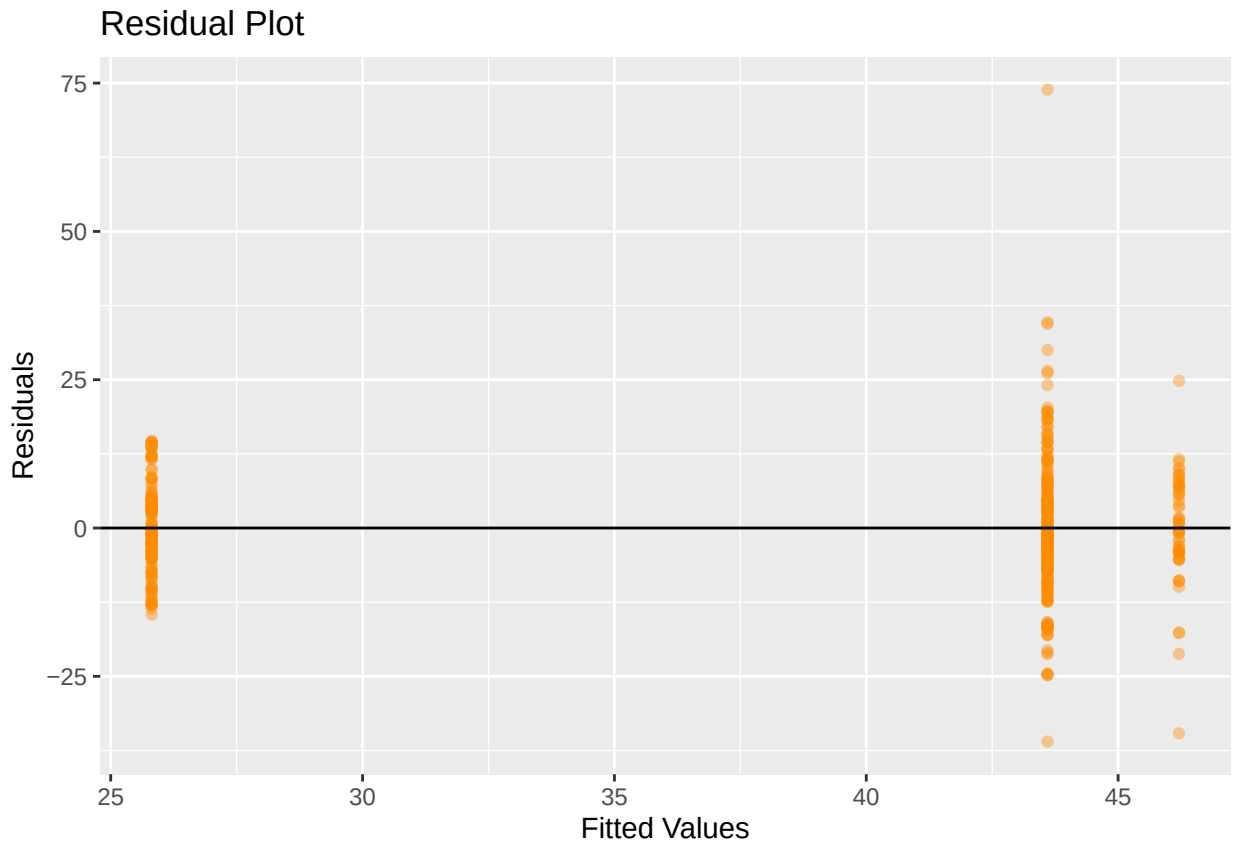
Response to Question 4A: 2012-2011 and 2013-2011 are the two pairs of purchase years that have statistically significantly different means at the signifcance level of 0.05 since the p-values of the pairwise comparisons are small the significance level of 0.05.

Response to Question 4B: When looking further at the results of the two pairs, we can see that both pairs are statistically signifcant and they don't include 0 in the lower and upper bounds. From this we can conlclude that the mean performance of the year value = 2011 is significantly higher than the mean purchase year of the other two values 2012 and 2013.

Question 5 - 6pts

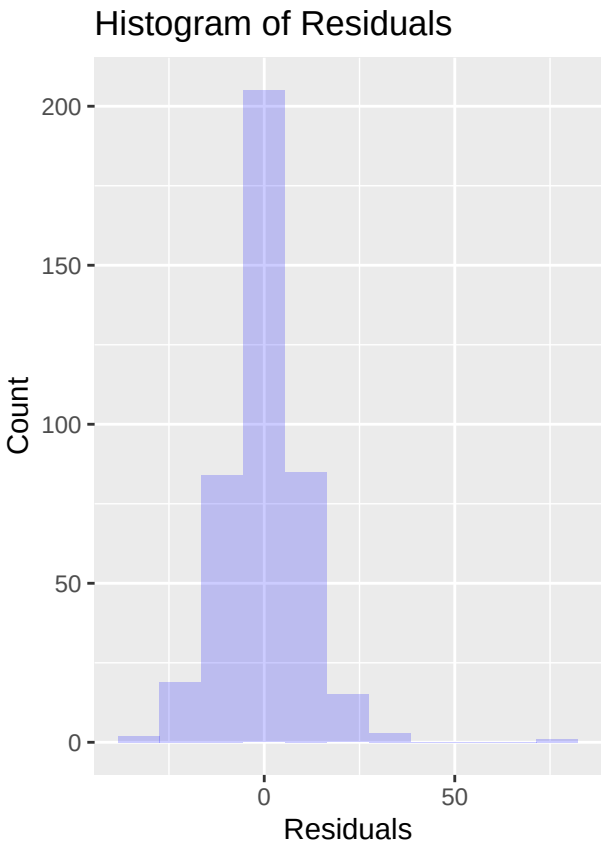
Perform a residual analysis on the ANOVA model. Comment on **each** model assumption and whether the assumption holds.

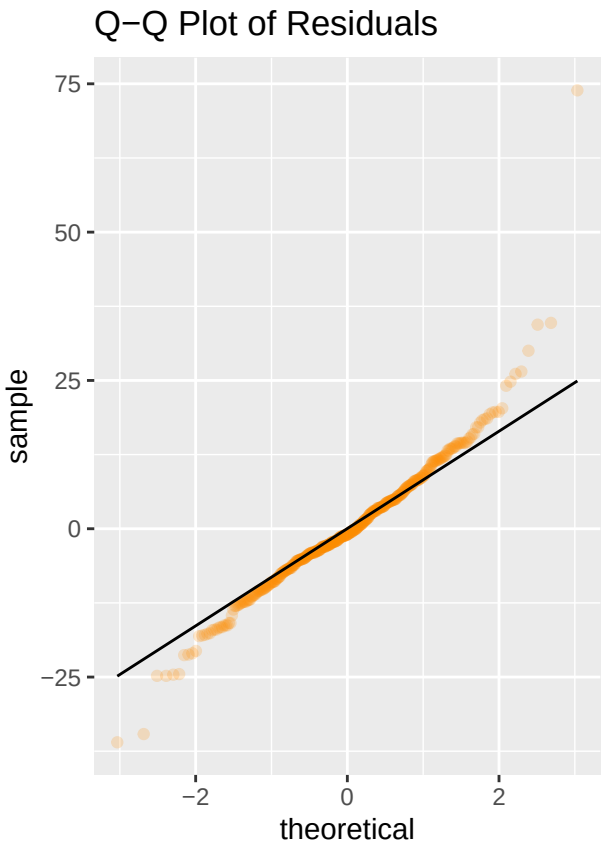
```
# Code to create residual anaylsis
library(ggplot2)
library(ggpubr)
ggplot(data=house, aes(x=anova_model$fitted.values, y=anova_model$residuals)) +
  geom_point(alpha=I(0.4),color='darkorange') +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Residual Plot') +
  geom_hline(yintercept=0)
```




```
hp = qplot(anova_model$residuals,
geom="histogram",
bins=11,
main = "Histogram of Residuals",
xlab = "Residuals",
ylab = "Count",
fill=I("blue"),
alpha=I(0.2))
qqp = ggplot(house, aes(sample=anova_model$residuals)) +

stat_qq(alpha=I(0.2),color='darkorange') +
stat_qq_line() +
ggtitle("Q-Q Plot of Residuals")
ggarrange(hp, qqp, ncol=2, nrow=1)
```





Response to Question 5: Based on the output, we can see that we have 3 bands in the residual analysis, which makes sense in our case because the Purchase.Year variable is taking on 3 distinct values. Therefore the independence assumption does hold.

However, when looking at the histogram and qq-plot, we do see that the normality assumption holds as our histogram plot is exhibiting no skewness and a clear normal distribution. In addition the qq-plot does not have any significant tails when looking at the plot.

Question 6 - 2 pts

Based on your assessment of the assumptions (Question 5), is the ANOVA model a good fit? If not, comment on how you would improve the fit. *Do not apply the recommendation.*

Response to Question 6
Based on our results we can see that all assumptions for the model hold, and therefore we can conclude that the Anova model provides a good fit. This also aligns with our previous answers, in that there is a relationship between price and purchase year, and that the purchase year does explain variability in the target variable for price per area.

Question 7 - 4pts

Now consider the quantitative variables.

Compute the correlation between each quantitative variable and the response, *Price.per.Area*.

A) Which predicting variables have a correlation above 0.50 or below -0.50 with the response?

B) Interpret the value of the strongest correlation coefficient in the context of the problem.

```
# Code to calculate correlation...
cat("cor(Price.per.Area, Age):", cor(house$Price.per.Area, house$Age)[1], end="\n")
```

```
## cor(Price.per.Area, Age): -0.210567
```

```
cat("cor(Price.per.Area, MRT.Dist):", cor(house$Price.per.Area, house$MRT.Dist)[1], end="\n")
```

```
## cor(Price.per.Area, MRT.Dist): -0.6736129
```

```
cat("cor(Price.per.Area, Num.of.Stores):", cor(house$Price.per.Area, house$Num.of.Stores)[1], end="\n")
```

```
## cor(Price.per.Area, Num.of.Stores): 0.5710049
```

```
cat("cor(Price.per.Area, Latitude):", cor(house$Price.per.Area, house$Latitude)[1], end="\n")
```

```
## cor(Price.per.Area, Latitude): 0.5463067
```

```
cat("cor(Price.per.Area, Longitude):", cor(house$Price.per.Area, house$Longitude)[1], end="\n")
```

```
## cor(Price.per.Area, Longitude): 0.5232865
```

Response to Question 7A: cor(Price.per.Area, MRT.Dist): -0.6736129 cor(Price.per.Area, Num.of.Stores): 0.5710049 cor(Price.per.Area, Latitude): 0.5463067 cor(Price.per.Area, Longitude): 0.5232865

Based on our output, the Num.of.Stores, Latitude and Longitude features.

Response to Question 7B:

MRT.Dist which I'm going to assume is the number of miles driven by the vehicle, has a negative correlation with the Price.Per.Area target variable. Which means that higher values for MRT.Dist result in a decrease of the Price.Per.Area of the car.

Question 8 - 4pts

Create a full model with all variables (quantitative and qualitative) called **lm.full** with *Price.per.Area* as the response variable. Include an intercept. Display the summary. Which coefficients are significant at the 0.05 significance level?

```
# Code to create model and find significant coefficients...
model1 <- lm(Price.per.Area ~ Purchase.Year + Purchase.Month + Age + MRT.Dist + Num.of.Stores + Latitud
summary(model1)
```

```
##
## Call:
## lm(formula = Price.per.Area ~ Purchase.Year + Purchase.Month +
##     Age + MRT.Dist + Num.of.Stores + Latitude + Longitude, data = house)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -36.661  -4.795  -0.514   3.551  73.963
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.359e+03  5.799e+03  -0.579    0.563
## Purchase.Year2012  8.553e+00  1.535e+00   5.571 4.62e-08 ***
## Purchase.Year2013  8.886e+00  1.196e+00   7.432 6.42e-13 ***
## Purchase.Month     2.346e-02  1.246e-01   0.188    0.851
## Age               -2.403e-01  3.711e-02  -6.476 2.73e-10 ***
## MRT.Dist          -3.188e-03  6.893e-04  -4.625 5.05e-06 ***
## Num.of.Stores      9.050e-01  1.806e-01   5.010 8.15e-07 ***
## Latitude           1.802e+02  4.282e+01   4.210 3.15e-05 ***
## Longitude         -9.093e+00  4.586e+01  -0.198    0.843
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.333 on 405 degrees of freedom
## Multiple R-squared:  0.6322, Adjusted R-squared:  0.625
## F-statistic: 87.03 on 8 and 405 DF,  p-value: < 2.2e-16
```

Response to Question 8:

Based on the result, the Purchase year variables for 2012 and 2013, the Age, MRT.Dist, Num.of.Stores, and Latitude columns are statistically significant.

Question 9 - 2pts

State and interpret the estimated coefficient for *Age*.

Response to Question 9:

The age coefficient means that for ever 1 unit increase in Age, results in a -2.403e-01 decline in the Price.per.Area.

Question 10 - 4pts

What are the bounds for a 90% confidence interval on the coefficient for *Age*? Is the coefficient for *Age* plausibly equal to zero at this confidence level?

```
# Code to calculate 90% CI...
confint(model1,level=0.90)['Age',]
```

```
##        5 %       95 %
## -0.3014857 -0.1791334
```

Response to Question 10:

-0.3014857 -0.1791334 are the confidence interval values for the Age variable.

Since they do not include 0 in the bounds, we can conclude that they are not plausibly equal to 0 at the given confidence level.

Question 11 - 3pts

Create a third model called *lm.full2* by adding a second order *Age* variable (Age^2) to *lm.full*. Display the summary. Comment on the addition of this predicting variable by comparing to the model without it. Is there any significant change in the direction and/or statistical significance of the regression coefficients?

```
# Code to create lm.full2'
house2 = house
house2$Age_Squared = house2$Age^2
model2 <- lm(Price.per.Area ~ Purchase.Year + Purchase.Month + Age + Age_Squared + MRT.Dist + Num.of.St
summary(model2)
```

```
##
## Call:
## lm(formula = Price.per.Area ~ Purchase.Year + Purchase.Month +
##     Age + Age_Squared + MRT.Dist + Num.of.Stores + Latitude +
##     Longitude, data = house2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.095  -4.414  -0.584   3.622  75.247
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4.801e+03  5.618e+03  -0.855    0.393
## Purchase.Year2012  7.334e+00  1.503e+00   4.879 1.54e-06 ***
## Purchase.Year2013  8.220e+00  1.164e+00   7.063 7.19e-12 ***
## Purchase.Month     1.612e-02  1.206e-01   0.134    0.894
## Age               -9.536e-01  1.386e-01  -6.882 2.27e-11 ***
## Age_Squared        1.777e-02  3.334e-03   5.330 1.64e-07 ***
## MRT.Dist          -2.652e-03  6.747e-04  -3.930 9.99e-05 ***
## Num.of.Stores      8.673e-01  1.750e-01   4.957 1.06e-06 ***
## Latitude           1.959e+02  4.154e+01   4.717 3.30e-06 ***
## Longitude         -4.176e-01  4.442e+01  -0.009    0.993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.064 on 404 degrees of freedom
## Multiple R-squared:  0.6564, Adjusted R-squared:  0.6487
## F-statistic: 85.75 on 9 and 404 DF,  p-value: < 2.2e-16
```

Response to Question 11:

Yes when including the variable we can see that we've included a variable of Age Squared, which has a higher statistical significance compared to the original Age variable. We can also see that the overall R^2 value improves from 0.6322 to 0.6564.

However, although it does improve the coefficient of determination. When taking out the Age variable out of the equation, we can see that for the other coefficients there is either a slight increase or decrease in value, but not a significant change.

Question 12 - 4pts

Perform a partial F-test on the new model (lm.full2) vs the previous model (lm.full), using $\alpha = 0.05$

A) State the null hypothesis of the partial f-test.

B) Do you *reject* or *fail to reject* the null hypothesis?

C) Based on these results, what conclusion do you make in context of the problem?

```
# Code for Partial F-Test...
anova(model2,model1)
```

```
## Analysis of Variance Table
##
## Model 1: Price.per.Area ~ Purchase.Year + Purchase.Month + Age + Age_Squared +
##     MRT.Dist + Num.of.Stores + Latitude + Longitude
## Model 2: Price.per.Area ~ Purchase.Year + Purchase.Month + Age + MRT.Dist +
##     Num.of.Stores + Latitude + Longitude
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    404 26273
## 2    405 28120 -1   -1847.2 28.404 1.641e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Response to Question 12A:
H0: a = 0 Meaning that the additional variable does not provide any additional explanatory power HA: a != 0 Meaning that the additional variable does provide additional explanatory power

Response to Question 12B:
We reject the null hypothesis that the regression coefficients for Age squared are 0 at alpha level 0.05

Response to Question 12C: Both indicate the Age squared is a statistically significant predictor for explaining the response variable. However, we can observe that the f-value is higher in the case of the marginal relationship as the other predictors were not fixed. That's because Age Squared is obviously correlated with the other predictors.

Question 13 - 5pts

Using ***lm.full*** model, what is the predicted *Price.per.Area* and corresponding *90% prediction interval* for an 9 year old house purchased in December 2012 that is 104 meters from the MRT station, located at 24.966 (latitude) and 121.54 (longitude) with 5 stores within a kilometer? Provide an interpretation of your results.

Note: (Ensure you are using ***lm.full*** not **lm.full2**)

```
# Code to predict point and prediction interval...
data1 <- house[1,]
data1['Purchase.Year']='2012'
data1['Purchase.Month']=12
data1['Age']=9
data1['MRT.Dist']=104
data1['Latitude']=24.966
data1['Longitude']=121.54
data1['Num.of.Stores']=12

predict(model1, data1, interval="prediction")
```

```
##        fit      lwr      upr
## 1 52.61636 35.90411 69.32861
```

Response to Question 13:
From our result, we can see that with a .95 alpha level for our prediction interval, we've calculated a Price.per.Area of 52.616 for this observation, which lower and upper bounds of 35.90411 and 69.32861, respectively.

The End.