Northwestern University
Master of Science in Data Science

By: Ali Gowani

Table of Contents

Part I: MECHANICS AND COMPUTATIONS

Model 1:

ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1974.53	1974.53	209.8340	< 0.0001
X2	1	118.8642568	118.8642568	12.6339	0.0007
X3	1	32.47012585	32.47012585	3.4512	0.0676
X4	1	0.435606985	0.435606985	0.0463	0.8303
Residuals	67	630.36	9.41		
Note: You can make the following calculations from the ANOVA table above to get Overall F statistic					
Model (adding 4 rows)	4	2126	531.50		<0.0001
Total (adding all rows)	71	2756.37			

Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	11.3303	1.9941	5.68	<.0001
X1	2.186	0.4104		<.0001
X2	8.2743	2.3391	3.54	0.0007
X3	0.49182	0.2647	1.86	0.0676
X4	-0.49356	2.2943	-0.22	0.8303
Residual standard error: 3.06730 on 67 degrees of freedom				
Multiple R-sqaured: 0.7713, Adjusted R-squared: 0.7577				
F-statistic: on 4 and 67 DF, p-value < 0.0001				

Number of predictors	C(p)	R-square	AIC	BIC	Variables in the model
4	5	0.7713	166.2129	168.9481	X1 X2 X3 X4

Question 1: How many observations are in the sample data?

The total number of observations in the sample data is derived from summing the Df column (67 + 1 + 1 + 1 + 1 = 71) and then adding 1 (71 + 1 = 72). There are 72 total observations in the sample data.

Question 2: Write out the null and alternative hypotheses for the t-test for Beta1?

- Null Hypothesis (H_o): $ß_1 = 0$
- Alternative Hypothesis (H_a): $ß_1 \neq 0$

Question 3: Compute the t-statistic for Beta1. Conduct the hypothesis test and interpret the result.

T-statistic (t)is derived by taking the slope of $ß_1$ over the Standard Error (S) for $ß_1$. the

$$T = \frac{\left(\beta_{\{1\}} - \beta_{\{1\}}^{\{(0)\}}\right)}{S_{(\beta_1)}}$$

$$T = \frac{(2.186 - 0)}{0.4104} = 5.3256$$

If we use the alpha of 0.05 then we can calculate the t-statistics for a two-sided t-test as:

$$t_{\left(n-1,1-\frac{\alpha}{2}\right)} = t_{(70,0.975)} = 1.9944$$

The t-statistic that we measured for $ß_1$ is 5.3256, which is greater than the critical t-statistics of 1.9944. This would mean that we should reject the null hypothesis (Ho) as a model with X_1 is more meaningful than a model without it in predicting Y.

Question 4: Compute the R-Squared value for Model 1, using information from the ANOVA table. Interpret this statistic.

R-squared is calculated by subtracting Sum of Square Errors (SSE) from Sum of Squares Predicted (SSY) and then dividing it by SSY.

$$R^2 = \frac{SSY - SSE}{SSY}$$

$$R^2 = \frac{2756.37 - \ 630.36}{2756.37} = 0.7713$$

The R-squared value of 0.7713 indicates that the independent variables in this model accounted for approximately 77% of variance in Y. The higher the r-squared value, the better. However, an extremely high r-squared value (e.g.: close to 1) could mean that the r-squared may have a biased estimate or the model is overfitting. Also, as more variables are added, the r-squared value will continue to rise, even if it may not provide significant new information to explain the variance in Y.

Question 5: Compute the Adjusted R-Squared value for Model 1. Discuss why Adjusted R-squared and the R-squared values are different.

R-squared we can determine the linear relationship of a model, so in that sense it is useful. However, Adjusted R-squared accounts for degrees of freedom (n – 1), as well as, the residual degrees of freedom (n – k – 1). This essentially penalizes the model for adding variables that do not had relevant information in predicting the response variable. The formula for this is as follows along with the calculation for Model 1.

$$Adjusted\ R^2 = 1 - (1 - R^2) * \left(\frac{n-1}{n-k-1}\right)$$

$$Adjusted\ R^2 = 1 - (1 - 0.7713) * \left(\frac{72 - 1}{72 - 4 - 1}\right) = 0.7577$$

The Adjusted R-squared value of 0.7577 is lower than R-squared value of 0.7713 as Adjusted R-squared penalized the model for irrelevant information to predict Y.

Question 6: Write out the null and alternate hypotheses for the Overall F-test.

- Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = ß_4 = 0$
- Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2, 3 or 4)

Question 7: Compute the F-statistic for the Overall F-test. Conduct the hypothesis test and interpret the result.

The following is the equation for F-statistic:

$$F = \frac{(Mean\ Squared\ Regression)}{(Mean\ Squared\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n - k - 1}\right)}$$

$$F = \frac{(Mean\ Squared\ Regression)}{(Mean\ Squared\ Residual)} = \frac{\left(\frac{2756.37 - 630.36}{4}\right)}{\left(\frac{630.36}{72 - 4 - 1}\right)} = 56.4926$$

Now that we have calculated the F-statistic for the model (56.4926), we need to determine the critical F-statistic in order to compare our findings. We will need Degrees of Freedom 1, Degrees of Freedom 2 and Probability Level for us to calculate the critical F-statistic.

$$F_{k,n-k-1,1-a} = F_{4,67,0.95} = 2.5087$$

We must reject the null hypothesis (alpha = 0.05) as the F-statistic of 56.4926 is greater than the critical F-statistic of 2.5087. This shows that our independent variables play a significant role in predicting the target response variable of Y.

Model 2:

ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1928.27000	1928.27000	218.8890	<.0001
X2	1	136.92075	136.92075	15.5426	0.0002
X3	1	40.75872	40.75872	4.6267	0.0352
X4	1	0.16736	0.16736	0.0190	0.8908
X5	1	54.77667	54.77667	6.2180	0.0152
X6	1	22.86647	22.86647	2.5957	0.112
Residuals	65	572.60910	8.80937		
Note: You can make the following calculations from the ANOVA table above to get Overall F statistic					
Model (adding 6 rows)	6	2183.75946	363.96	41.3200	<0.0001
Total (adding all rows)	71	2756.37			

Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	14.3902	2.89157	4.98	<.0001
X1	1.97132	0.43653	4.52	<.0001
X2	9.13895	2.30071	3.97	0.0002
X3	0.56485	0.26266	2.15	0.0352
X4	0.33371	2.42131	0.14	0.8908
X5	1.90698	0.76459	2.49	0.0152
X6	-1.0433	0.64759	-1.61	0.112
Residual standard error: 2.968 on 65 degrees of freedom				
Multiple R-sqaured: 0.7923, Adjusted R-squared: 0.7731				
F-statistic: 41.32 on 6 and 65 DF, p-value < 0.0001				

Number of predictors	C(p)	R-square	AIC	BIC	Variables in the model
6	7	0.7923	163.2947	166.7792	X1 X2 X3 X4 X5 X6

Question 8: Now let's consider Model 1 and Model 2 as a pair of models. Does Model 1 nest Model 2 or does Model 2 nest Model 1? Explain.

In our example, Model 1 is nested under Model 2. This means that Model 2 has all the explanatory (X_i) variables in addition to other explanatory variables that are not part of Model 1.

Question 9: Write out the null and alternate hypotheses for a nested F-test using Model 1 and Model 2.

- Null Hypothesis (H_o): $ß_5 = ß_6 = 0$
- Alternative Hypothesis (H_a): $ß_5 \neq 0$ or $ß_6 \neq 0$

Question 10: Compute the F-statistic for a nested F-test using Model 1 and Model 2. Conduct the hypothesis test and interpret the results.

- SSE_R is the Sum of Square Errors for the reduced model (Model 1).
- SSE_C is the Sumer of Square Errors for the complete model (Model 2).
- i is the number of additional ß

$$F = \frac{\left(\frac{(SSE_R - SSE_C)}{i}\right)}{\left(\frac{SSE_C}{n-(k+p+1)}\right)}$$

$$F = \frac{\left(\frac{(630.36 - 572.609)}{2}\right)}{\left(\frac{572.609}{65}\right)} = 3.278$$

We need to determine the critical F-statistic with 95% confidence (1 – alpha) in order to compare our findings.

$$F_{i,n-k-p-1,1-a} = F_{2,65,0.95} = 3.1381$$

We must reject the null hypothesis (alpha = 0.05) as the F-statistic of 3.278 is greater than the critical F-statistic of 3.1381. This shows that the 2 additional independent variables play a significant role in predicting the target response variable of Y.

Part II: APPLICATION

For this part of the assignment, you are to use the AMES Housing Data you worked with during Modeling Assignment #1.

Question 11: Model 3

Based on your EDA from Modeling Assignment #1, focus on 10 of the continuous quantitative variables that you though/think might be good explanatory variables for SALESPRICE. Is there a way to logically group those variables into 2 or more sets of explanatory variables? For example, some variables might be strictly about size while others might be about quality. Separate the 10 explanatory variables into at least 2 sets of variables. Describe why you created this separation. A set must contain at least 2 variables.

Variable_Name	Variable_Type	Variable_Group
SalePrice	Continuous	Target
FirstFlrSF	Continuous	Interior
SecondFlrSF	Continuous	Interior
GrLivArea	Continuous	Interior
TotalBsmtSF	Continuous	Interior
EnclosedPorch	Continuous	Interior

GarageArea	Continuous	Interior
LotFrontage	Continuous	Exterior
LotArea	Continuous	Exterior
MasVnrArea	Continuous	Exterior
WoodDeckSF	Continuous	Exterior

I grouped the 10 independent variables into two groups: interior and exterior. Out of the 10 continuous variables that I selected, 6 explanatory variables were focused on the area within the house and the house itself (e.g.: FirstFlrSF, SecondFlrSF, GrLivArea, TotalBsmtSF, EnclosedPorch and GarageArea), while the remaining 4 explanatory variables focused on the exterior of the house (e.g.: LotFrontage, LotArea, MasVnrArea and WoodDeckSF).

Question 12: Individual Model 3 Coefficients & Omnibus Overall F-test

Pick one of the sets of explanatory variables. Run a multiple regression model using the explanatory variables from this set to predict SALEPRICE(Y). Call this Model 3. Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:

a) *all model coefficients individually*
b) *the Omnibus Overall F-test*

Model 3: lm(formula = SalePrice ~ FirstFlrSF + SecondFlrSF + GrLivArea + TotalBsmtSF + EnclosedPorch + GarageArea, data = model3_df)

Model 3	**Target Variable: SalePrice**			
Predictors	***Estimates***	***std. Error***	***t-statistic***	***p-value***
(Intercept)	-26157.455	2941.021	-8.894	0.000
FirstFlrSF	90.536	18.375	4.927	0.000
SecondFlrSF	93.679	18.097	5.177	0.000
GrLivArea	-21.721	17.918	-1.212	0.226
TotalBsmtSF	55.134	3.186	17.303	0.000
EnclosedPorch	-89.897	13.043	-6.892	0.000
GarageArea	99.898	4.727	21.132	0.000
Observations	**2928**			
R^2 / R^2adjusted	**0.688 / 0.687**			

Residual standard error: 44670 on 2921 degrees of freedom (2 observations deleted due to missingness)
Multiple R-squared: 0.688
Adjusted R-squared: 0.6874
F-statistic: 1074 on 6 and 2921 DF
p-value: < 0.00000000000000022

Let's define our critical T-statistic to compare our variables:

$$t_{n-1,\left(1-\frac{a}{2}\right)} = t_{2927,0.975} = 1.961$$

T-statistic (t) for each of the variables is derived by taking the slope of $ß_i$ over the Standard Error (S_i) for $ß_i$.

$$T = \frac{\left(\beta_{\{1\}} - \beta_{\{1\}}^{\{(0)\}}\right)}{S_{(\beta_1)}}$$

1. **Intercept:**
    a. Null Hypothesis (H_o): $ß_0 = 0$
    b. Alternative Hypothesis (H_a): $ß_0 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(-26157.455-0)}{2941.021} = -8.894$
    d. This value indicates that we should reject the Null Hypothesis. However, it does not make sense to have a SalePrice of a home with a negative value.

2. **FirstFlrSF:**
    a. Null Hypothesis (H_o): $ß_1 = 0$
    b. Alternative Hypothesis (H_a): $ß_1 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(90.536-0)}{18.375} = 4.927$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the FirstFlrSF variable provides significant information for predicting SalePrice.

3. **SecondFlrSF:**
    a. Null Hypothesis (H_o): $ß_2 = 0$
    b. Alternative Hypothesis (H_a): $ß_2 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(93.679-0)}{18.097} = 5.177$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the SecondFlrSF variable provides significant information for predicting SalePrice.

4. **GrLivArea:**
    a. Null Hypothesis (H_o): $ß_3 = 0$
    b. Alternative Hypothesis (H_a): $ß_3 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(-21.721-0)}{17.918} = -1.212$
    d. We fail to reject the Null Hypothesis as the T-statistic value for this variable is less than the critical T-Statistic value of 1.961. This means that the GrLivArea variable provides insignificant information for predicting SalePrice or the relationship between the variables is not linear.

5. **TotalBsmtSF:**
    a. Null Hypothesis (H_o): $ß_4 = 0$
    b. Alternative Hypothesis (H_a): $ß_4 \neq 0$

c. T-Statistic:
    i. $T = \frac{(55.134-0)}{3.186} = 17.303$
d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the 5. TotalBsmtSF variable provides significant information for predicting SalePrice.

6. **EnclosedPorch:**
    a. Null Hypothesis (H_o): $ß_5 = 0$
    b. Alternative Hypothesis (H_a): $ß_5 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(-89.897-0)}{13.043} = -6.892$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the EnclosedPorch variable provides significant information for predicting SalePrice.

7. **GarageArea:**
    a. Null Hypothesis (H_o): $ß_6 = 0$
    b. Alternative Hypothesis (H_a): $ß_6 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(90.898-0)}{4.727} = 21.132$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the GarageArea variable provides significant information for predicting SalePrice.

The Omnibus Overall F-statistic for Model 3:

a. Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = ß_4 = ß_5 = ß_6 = 0$
b. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2, 3, 4, 5 or 6)

$$F = \frac{(Mean\ Sqrd\ Regression)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{18681275513864 - 5827979153622}{6}\right)}{\left(\frac{5827979153622}{2928-6-1}\right)} = 1075$$

The critical F-statistic for Model 3 is:

$$F_{i,n-k-p-1,1-a} = F_{6,2928-6-1,0.95} = 2.10$$

Since the F-statistic for Model 3 is 1075, which is greater than the critical F-statistic for Model 3 at 2.10 and p-value of less than 0.00001 then we can reject the Null Hypothesis. This means that our model contains significant relationship between the explanatory variables and the response variable of SalePrice.

Question 13: Individual Model 4 Coefficients & Omnibus Overall F-test

Pick the other set (or one of the other sets) of explanatory variables. Add this set of variables to those in Model 3. In other words, Model 3 should be nested within Model 4. Run a multiple regression model using the explanatory variables from this set to predict SALEPRICE(Y). Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:

a) all model coefficients individually
b) the Omnibus Overall F-test

Model 4: lm(formula = SalePrice ~ FirstFlrSF + SecondFlrSF + GrLivArea + TotalBsmtSF + EnclosedPorch + GarageArea + LotFrontage + LotArea + MasVnrArea + WoodDeckSF, data = model3_df)

Model 4	**Target Variable: SalePrice**			
Predictors	***Estimates***	***std. Error***	***t-statistic***	***p***
(Intercept)	-14837.506	3662.961	-4.051	1.000
FirstFlrSF	71.680	19.526	3.671	0.000
SecondFlrSF	73.450	19.142	3.837	0.000
GrLivArea	-9.185	18.868	-0.487	0.626
TotalBsmtSF	47.249	3.541	13.342	0.000
EnclosedPorch	-72.082	14.491	-4.974	0.000
GarageArea	95.388	5.191	18.377	0.000
LotFrontage	-93.135	48.319	-1.928	0.054
LotArea	0.395	0.167	2.360	0.018
MasVnrArea	64.961	5.933	10.950	0.000
WoodDeckSF	54.323	7.971	6.815	0.000
Observations	**2421**			
R^2 / R^2adjusted	**0.713 / 0.712**			

Residual standard error: 44710 on 2410 degrees of freedom (509 observations deleted due to missingness)
Multiple R-squared: 0.7134
Adjusted R-squared: 0.7122
F-statistic: 599.9 on 10 and 2410 DF
p-value: < 0.00000000000000022

Let's define our critical T-statistic to compare our variables:

$$t_{n-1,\left(1-\frac{a}{2}\right)} = t_{2421,0.975} = 1.961$$

T-statistic (t) for each of the variables is derived by taking the slope of $ß_i$ over the Standard Error (S_i) for $ß_i$.

$$T = \frac{\left(\beta_{\{1\}} - \beta_{\{1\}}^{\{(0)\}}\right)}{S_{(\beta_1)}}$$

1. **Intercept:**
   a. Null Hypothesis (H_o): $ß_0 = 0$
   b. Alternative Hypothesis (H_a): $ß_0 \neq 0$
   c. T-Statistic:
      i. $T = \frac{(-14837.506-0)}{3662.961} = -4.051$
   d. This value indicates that we should reject the Null Hypothesis. However, it does not make sense to have a SalePrice of a home with a negative value.

2. **FirstFlrSF:**
   a. Null Hypothesis (H_o): $ß_1 = 0$
   b. Alternative Hypothesis (H_a): $ß_1 \neq 0$
   c. T-Statistic:
      i. $T = \frac{(71.680-0)}{19.526} = 3.671$
   d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the FirstFlrSF variable provides significant information for predicting SalePrice.

3. **SecondFlrSF:**
   a. Null Hypothesis (H_o): $ß_2 = 0$
   b. Alternative Hypothesis (H_a): $ß_2 \neq 0$
   c. T-Statistic:
      i. $T = \frac{(73.450-0)}{19.142} = 3.837$
   d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the SecondFlrSF variable provides significant information for predicting SalePrice.

4. **GrLivArea:**
   a. Null Hypothesis (H_o): $ß_3 = 0$
   b. Alternative Hypothesis (H_a): $ß_3 \neq 0$
   c. T-Statistic:
      i. $T = \frac{(-9.185-0)}{18.868} = -0.487$
   d. We fail to reject the Null Hypothesis as the T-statistic value for this variable is less than the critical T-Statistic value of 1.961. This means that the GrLivArea variable provides insignificant information for predicting SalePrice or the relationship between the variables is not linear.

5. **TotalBsmtSF:**
   a. Null Hypothesis (H_o): $ß_4 = 0$
   b. Alternative Hypothesis (H_a): $ß_4 \neq 0$
   c. T-Statistic:
      i. $T = \frac{(47.249-0)}{3.541} = 13.342$
   d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the 5. TotalBsmtSF variable provides significant information for predicting SalePrice.

6. **EnclosedPorch:**
    a. Null Hypothesis (H_o): $ß_5 = 0$
    b. Alternative Hypothesis (H_a): $ß_5 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(-72.082-0)}{14.491} = -4.974$
    d. We fail to reject the Null Hypothesis as the T-statistic value for this variable is less than the critical T-Statistic value of 1.961. This means that the EnclosedPorch variable provides insignificant information for predicting SalePrice or the relationship between the variables is not linear.

7. **GarageArea:**
    a. Null Hypothesis (H_o): $ß_6 = 0$
    b. Alternative Hypothesis (H_a): $ß_6 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(95.388-0)}{5.191} = 18.377$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the GarageArea variable provides significant information for predicting SalePrice.

8. **LotFrontage:**
    a. Null Hypothesis (H_o): $ß_7 = 0$
    b. Alternative Hypothesis (H_a): $ß_7 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(-93.135-0)}{48.319} = -1.928$
    d. We fail to reject the Null Hypothesis as the T-statistic value for this variable is less than the critical T-Statistic value of 1.961. This means that the LotFrontage variable provides insignificant information for predicting SalePrice or the relationship between the variables is not linear.

9. **LotArea:**
    a. Null Hypothesis (H_o): $ß_8 = 0$
    b. Alternative Hypothesis (H_a): $ß_8 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(0.395-0)}{0.167} = 2.360$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the LotArea variable provides significant information for predicting SalePrice.

10. **MasVnrArea:**
    a. Null Hypothesis (H_o): $ß_9 = 0$
    b. Alternative Hypothesis (H_a): $ß_9 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(64.961-0)}{5.933} = 10.950$

d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the MasVnrArea variable provides significant information for predicting SalePrice.

11. **WoodDeckSF:**
    a. Null Hypothesis (H_o): $ß_{10} = 0$
    b. Alternative Hypothesis (H_a): $ß_{10} \neq 0$
    c. T-Statistic:
        i. $T = \frac{(54.323-0)}{7.971} = 6.815$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the WoodDeckSF variable provides significant information for predicting SalePrice.

The Omnibus Overall F-statistic for Model 4:

a. Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = ß_4 = ß_5 = ß_6 = ß_7 = ß_8 = ß_9 = ß_{10} = 0$
b. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10)

$$F = \frac{(Mean\ Sqrd\ Regression)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{16811844355807 - 4818005605587}{10}\right)}{\left(\frac{4818005605587}{2421 - 10 - 1}\right)} = 599.90$$

The critical F-statistic for Model 4 is:

$$F_{i,n-k-p-1,1-a} = F_{10,2421-10-1,0.95} = 1.83$$

Since the F-statistic for Model 4 is 599.90, which is greater than the critical F-statistic for Model 4 at 1.83 and p-value of less than 0.00001 then we can reject the Null Hypothesis. This means that our model contains significant relationship between the explanatory variables and the response variable of SalePrice.

Question 14: Nested Model

Write out the null and alternate hypotheses for a nested F-test using Model 3 and Model 4, to determine if the Model 4 variables, as a set, are useful for predicting SALEPRICE or not. Compute the F-statistic for this nested F-test and interpret the results.

For a nested F-test, we use two models (Model 3 and Model 4), these models are considered nested if they both have the same variables and one of the models (Model 4) has at least one additional variable. In our case, Model 3 is nested within Model 4. Model 3 is considered reduced and Model 4 is considered complete. By conducting a nested F-test between Model 3 and Model 4, we will determine whether the additional explanatory variables in Model 4 are more robust than the reduced model.

The values for i represent the additional variables added to our model.

a. Null Hypothesis (H_o): $ß_7 = ß_8 = ß_9 = ß_{10} = 0$
b. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 7, 8, 9 or 10)

We can calculate the F-test of the nested model by using the following formula:

$$F = \frac{(SSE_R - SSE_C)}{\left(\frac{SSE_C}{n-k-p-1}\right)} = \frac{(5827979153622 - 4818005605587)}{\left(\frac{4818005605587}{2421-10-1}\right)} = 505.20$$

The critical F-statistic value is:

$$F_{i,n-k-p-1,1-a} = F_{10,2421-10-1,0.95} = 1.83$$

Since the F-statistic value of 505.20 is greater than the critical value of 1.83 at a confidence of 95%, then we would reject the null hypothesis that the complete Model 4 is no more robust than the reduced model (Model 3). This means that the additional variables add significant information in predicting SalePrice.

CONCLUSION & REFLECTION:

This was another grueling assignment where I learned to manually compute various formula and values. I got more comfortable in reading and understanding the Summary and ANOVA tables for multivariate linear regression models and how to make statistical inferences based on coefficients and residual variances. In addition, we were asked to formulate a hypothesis about the overall fit of the various models using both R-squared and Adjusted R-square metrics.

We learned more about Adjusted R-squared and how it penalizes the model for adding variables that do not had relevant information in predicting the response variable. However, with R-squared we can determine the linear relationship of a model, so in that sense it is useful but Adjusted R-squared is better as it provides the best estimate for strength of the relationship.

Formulating hypothesis for validating individual components, such as, beta coefficients, performing t-tests on individual variables, formulating an overall F-statistic, calculating how to generate statistics for nested models, etc. were really beneficial for me as it allowed me to get a bit more in the weeds to understand how to assess models and variables within them.

The Application portion of the computation assignment was quite beneficial for me as it reinforced calculating statistics of variables and models, so it becomes engrained in our minds. The mathematics behind each metric and being able to compare it between models as we did in the nested portion of the assignment was something that I enjoyed as it took several hours for me to understand and become more comfortable with. Repetition is key!

Comparing the models using statistics so we can measure them and evaluate them is something that is already way above my level of understanding and experience, so it takes me much longer. However, I still believe these things are fundamental to Data Science, no matter how cool and catchy AI sounds. We need a grounding in statistics in order to have a meaningful business discussion with our clients.

Appendix

A: Model 3 Summary

```
Residuals:
    Min      1Q  Median      3Q     Max
-683608  -19169    -604   19093  264757

Coefficients:
                Estimate Std. Error t value            Pr(>|t|)
(Intercept)   -26157.455   2941.021  -8.894 < 0.0000000000000002 ***
FirstFlrSF        90.536     18.375   4.927       0.0000008805838 ***
SecondFlrSF       93.679     18.097   5.177       0.0000002413212 ***
GrLivArea        -21.721     17.918  -1.212                 0.226
TotalBsmtSF       55.134      3.186  17.303 < 0.0000000000000002 ***
EnclosedPorch    -89.897     13.043  -6.892       0.0000000000067 ***
GarageArea        99.898      4.727  21.132 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44670 on 2921 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.688,     Adjusted R-squared:  0.6874
F-statistic:  1074 on 6 and 2921 DF,  p-value: < 0.00000000000000022
```

B: Model 3 ANOVA

```
Analysis of Variance Table

Response: SalePrice
                Df        Sum Sq       Mean Sq  F value                Pr(>F)
FirstFlrSF       1 7217737055505 7217737055505 3617.551 < 0.00000000000000022
SecondFlrSF      1 3597013158116 3597013158116 1802.833 < 0.00000000000000022
GrLivArea        1   24966479892   24966479892   12.513             0.0004104
TotalBsmtSF      1  967909116599  967909116599  485.119 < 0.00000000000000022
EnclosedPorch    1  154726641995  154726641995   77.549 < 0.00000000000000022
GarageArea       1  890943908134  890943908134  446.544 < 0.00000000000000022
Residuals     2921 5827979153622    1995199984

FirstFlrSF    ***
SecondFlrSF   ***
GrLivArea     ***
TotalBsmtSF   ***
EnclosedPorch ***
GarageArea    ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

C: Model 4 Summary

```
Residuals:
    Min      1Q  Median      3Q     Max 
-662795  -20297     205   19376  285142 

Coefficients:
                 Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)   -14837.5063   3662.9613  -4.051      0.0000526821785 ***
FirstFlrSF        71.6796     19.5258   3.671             0.000247 ***
SecondFlrSF       73.4496     19.1424   3.837             0.000128 ***
GrLivArea         -9.1850     18.8683  -0.487             0.626447    
TotalBsmtSF       47.2493      3.5413  13.342 < 0.0000000000000002 ***
EnclosedPorch    -72.0820     14.4907  -4.974      0.0000007008028 ***
GarageArea        95.3879      5.1905  18.377 < 0.0000000000000002 ***
LotFrontage      -93.1345     48.3187  -1.928             0.054034 .  
LotArea            0.3948      0.1673   2.360             0.018367 *  
MasVnrArea        64.9607      5.9325  10.950 < 0.0000000000000002 ***
WoodDeckSF        54.3230      7.9710   6.815      0.0000000000119 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44710 on 2410 degrees of freedom
  (509 observations deleted due to missingness)
Multiple R-squared:  0.7134,	Adjusted R-squared:  0.7122 
F-statistic: 599.9 on 10 and 2410 DF,  p-value: < 0.00000000000000022
```

D: Model 4 ANOVA

```
Analysis of Variance Table

Response: SalePrice
                Df        Sum Sq       Mean Sq   F value               Pr(>F)
FirstFlrSF       1 6859258172816 6859258172816 3431.0488 < 0.00000000000000022
SecondFlrSF      1 2965954403779 2965954403779 1483.5911 < 0.00000000000000022
GrLivArea        1   27978475852   27978475852   13.9950            0.0001876
TotalBsmtSF      1  804506806349  804506806349  402.4199 < 0.00000000000000022
EnclosedPorch    1  133457215408  133457215408   66.7562 0.0000000000000004904
GarageArea       1  846737212766  846737212766  423.5439 < 0.00000000000000022
LotFrontage      1    4049277402    4049277402    2.0255            0.1548098
LotArea          1    8838378006    8838378006    4.4210            0.0356023
MasVnrArea       1  250206867354  250206867354  125.1552 < 0.00000000000000022
WoodDeckSF       1   92851940487   92851940487   46.4452 0.0000000000118628699
Residuals     2410 4818005605587    1999172450                               
                 
FirstFlrSF    ***
SecondFlrSF   ***
GrLivArea     ***
TotalBsmtSF   ***
EnclosedPorch ***
GarageArea    ***
LotFrontage      
LotArea       *  
MasVnrArea    ***
WoodDeckSF    ***
Residuals        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```