Part 1

Model 1

1. How many observations are in the sample data?

The number of observations in the sample data can be found by looking at the degrees of freedom (DF) for the residuals plus the degrees of freedom for the model.

From the ANOVA table:
- Residuals DF = 67
- Model DF = 4

Adding them together:
67 + 4 = 71

There are 71 observations in the sample data.

2. Write out the null and alternate hypotheses for the t-test for Beta1.

For the t-test for $\beta 1$(which corresponds to the predictor $X1$ in the regression model), the null and alternate hypotheses are:

$H0 \colon \beta 1 \ = \ 0$ (The predictor $X1$ has no effect on the response variable.)

$H1 \colon \beta 1 \ \neq \ 0$ (The predictor $X1$ has an effect on the response variable.)

Where:

- $H0$ is the null hypothesis.
- $H0$ is the alternate hypothesis.
- $\beta 1$ is the regression coefficient for the predictor $X1$.

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

To compute the t-statistic for $\beta 1$, we use the formula:

$$t = \frac{Estimate}{Std.Error}$$

From the coefficients table for $X1$:Estimate = 2.186

Std. Error = 0.4104

$$t = \frac{2.186}{0.4104} = 5.33$$

However, looking at the Coefficients table, the t-value for $X1$ is already given as 5.68.

Now, the hypothesis test:
$H0: \beta1 \; = \; 0$ (The predictor $X1$ has no effect on the response variable.)

$Ha: \beta1 \; \neq \; 0$ (The predictor $X1$ has an effect on the response variable.)

Given the t-value of 5.68 and the corresponding p-value of <0.0001 (from the coefficients table), the result is statistically significant.

Interpretation: Since the p-value is less than the typical significance level (0.05), we reject the null hypothesis. This indicate that X1 has a statistically significant effect on the response variable.

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

The R^2 (R-Squared) value for a regression model represents the proportion of variance in the dependent variable that can be explained by the predictors (independent variables) in the model. It is computed using the following formula:

$$R^2 = \frac{ModelSS}{TotalSS}$$

From the ANOVA table:

Model SS (Sum of Squares) = 2126
Total SS = 2756.37

Plugging in the values:

$$R^2 = \frac{2126}{2756.37} = 0.7713$$

This matches the Multiple R-squared value provided in the output, which is 0.7713.

Interpretation:
The R^2 value of 0.7713 means that approximately 77.13% of the variance in the dependent variable can be explained by the predictors $X1, X2, X3, and\ X4$ in Model 1. This indicates a relatively strong fit of the model to the data.

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

The Adjusted $R^2$is a modified version of R^2 that takes into account the number of predictors in the model. It is computed using the following formula:

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

Where:

- R^2 is the R-squared value.
- n is the number of observations.
- k is the number of predictors.
-

From the provided information:

- R^2 = 0.7713
- n (number of observations) = 71
- k (number of predictors) = 4

Plugging in the values:

Adjusted $R^2 \ = \ 1 \ - (\frac{(1-0.7713)(71-1)}{71-4-1})$
Adjusted $R^2 \ = \ 1 \ - (\frac{0.2287*70}{66})$
Adjusted $R^2 \ = \ 0.7574$

Discussion:

The key difference between R^2 and Adjusted R^2lies in their adjustment for the number of predictors. While R^2simply quantifies the proportion of variance explained by the model (without considering how many predictors are in the model), the Adjusted R^2 takes into account the number of predictors and adjusts the statistic accordingly. This adjustment is crucial when comparing models of different complexities or when trying to avoid overfitting.

In essence, while R^2provides a raw measure of model fit, the Adjusted R^2 offers a more nuanced perspective that takes model complexity into account.

To determine if the addition of $X4$ has reduced the accuracy of Model 1, we need to compare the Adjusted $R^2$of Model 1 with the Adjusted $R^2$of a model without $X4$) (let's call it Model 1.1). Unfortunately, we don't have the statistics for Model 1.1, so we can't make this direct comparison.

However, we can look at the coefficients table for Model 1 to glean some insight:

From the provided data for $X4$:

- Estimate (coefficient) for $X4$ = -0.49356
- t-value for $X4$ = -0.22
- p-value for $X4$ = 0.8303

The t-value for $X4$ is small and the p-value is quite large (0.8303), indicating that $X4$ is not statistically significant at common significance levels (like 0.05). This suggests that $X4$ might not be a meaningful predictor in the context of this model.

Given that $X4$ is not statistically significant and its inclusion doesn't seem to contribute valuable information, it's possible that a model without $X4$ could have a higher Adjusted R^2. If that's the case, then the inclusion of $X4$could have reduced the accuracy of Model 1.

However, to definitively conclude that the addition of $X4$ has reduced the accuracy of the model, we would need to compare Model 1's Adjusted $X4$ to that of Model 1.1 (without $X4$). Without this comparison, we can only speculate based on the given data for X4.

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

From Model 1:

- F-statistic = 531.50
- p-value for the F-statistic = <0.0001

Null Hypothesis $H0$: All regression coefficients (except the intercept) in Model 1 are equal to zero, meaning none of the predictors $(X1, X2, X3, X4)$ have an effect on the dependent variable.

Alternate Hypothesis Ha: At least one regression coefficient in Model 1 is different from zero, indicating that at least one predictor is significant in explaining variation in the dependent variable.

Since the p-value is much less than typical significance levels (e.g., 0.05), we reject the null hypothesis. This means that, in the context of Model 1, at least one of the predictors (X1, X2, X3, X4) is significant in explaining the variation in the dependent variable.

Given this result, the Overall F-test indicates that the model with the predictors is statistically significant and does a better job of explaining the variation in the dependent variable than a model with no predictors.

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

To compute the F-statistic for the Overall F-test in the context of multiple regression, the formula is:

$$F = \frac{ModelMS}{ResidualMS}$$

Where:

- Model MS is the Mean Square for the Model (or Regression)
- Residual MS is the Mean Square for the Residuals (or Error)

From the provided ANOVA table for Model 1:

- Model MS = 531.50
- Residual MS = 9.41

Plugging in the values:

$$F = \frac{531.50}{9.41} = 56.48$$

However, the provided F-statistic in the output is 531.50. So, we'll use the given value for our hypothesis test.

Null Hypothesis $H0$: All regression coefficients (except the intercept) in Model 1 are equal to zero, meaning none of the predictors (X1, X2, X3, X4) have an effect on the dependent variable.

Alternate Hypothesis Ha: At least one regression coefficient in Model 1 is different from zero, indicating that at least one predictor is significant in explaining variation in the dependent variable.

Considering the provided F-statistic and its p-value: p-value = <0.0001

Since the p-value is much less than typical significance levels (e.g., 0.05), we reject the null hypothesis.

Interpretation: The Overall F-test with an F-statistic of 531.50 and a p-value of <0.0001 indicates that the regression model is statistically significant. At least one of the predictors (X1, X2, X3, X4) is relevant in explaining the variation in the dependent variable. The model with these predictors does a better job of explaining the variation in the dependent variable than a model with no predictors.

Model 2

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.

   R^2 for Model 1 = 0.7713
   R^2 for Model 2 = 0.7923

The R^2 value for Model 2 is higher than that for Model 1. This makes sense because Model 2 includes all the predictors from Model 1 plus two additional predictors ($X5\ and\ X6$). When more predictors are added to a model, the R^2 value will either increase or stay the same, because the model is now accounting for more variance in the dependent variable, even if the added predictors are not statistically significant.

To definitively prove the nested relationship, we observe:

- Model 1's predictors ($X1, X2, X3, X4$) are a subset of Model 2's predictors ($X1, X2, X3, X4, X5, X6$).
- The R^2 value for Model 1 is less than that for Model 2.

These two points confirm that Model 1 nests inside Model 2. Model 1 can be viewed as a simpler, more restricted version of Model 2, making Model 2 the more general or comprehensive model of the two.

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

For a nested F-test comparing a simpler model (Model 1) to a more complex model (Model 2), the hypotheses are set up to test whether the additional predictors in the more complex model significantly increase the proportion of variance explained.

In this context:

- Model 1 (simpler model) includes predictors: $X1, X2, X3, X4$
- Model 2 (more complex model) includes predictors: $X1, X2, X3, X4, X5, X6$

The null and alternate hypotheses for the nested F-test are as follows:

Null Hypothesis $H0$: The additional predictors in Model 2 (in this case, X5 and X6) do not significantly improve the model fit compared to Model 1. In more technical terms, the regression coefficients for the additional predictors (X5 and X6) are equal to zero in the context of the larger model. $\beta 5 = \beta 6 = 0$ when considering them in Model 2.

Alternate Hypothesis Ha: At least one of the additional predictors in Model 2 provides a significant improvement in the model fit compared to Model 1. This means at least one of the coefficients for the additional predictors is different from zero. At least one $\beta i \neq 0$ for $i = 5, 6$ when considering them in Model 2.

Interpretation:

- If the nested F-test is significant (typically, if the p-value is less than a chosen significance level, e.g., 0.05), then we would reject the null hypothesis. This suggests that the more complex model (Model 2) provides a significantly better fit to the data compared to the simpler model (Model 1).

- If the F-test is not significant, then we would fail to reject the null hypothesis, suggesting that the addition of the predictors X5 and X6 in Model 2 does not provide a significant improvement over Model 1.

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

To conduct a nested F-test comparing Model 1 (simpler model) to Model 2 (more complex model), we use the following formula:

$$F = \left(\frac{\frac{RSSModel1 - RSSModel2}{dfModel1 - dfModel2}}{\frac{RSSModel2}{RSSModel2}} \right)$$

Where:

- RSS is the Residual Sum of Squares for a model.
- df is the degrees of freedom for the residuals of a model.

Using the information from the ANOVA tables:
For Model 1:

- $RSSModel1$ = 630.36 (Residual Sum Sq)
- $dfModel1$ = 67 (Residual DF)

For Model 2:

- $RSSModel2$ = 572.60910 (Residual Sum Sq)
- $dfModel2$ = 65 (Residual DF)

Plugging in the values:

$$F = \frac{\frac{630.36 - 572.60910}{67 - 65}}{\frac{572.60910}{65}}$$

$$F = 3.2776$$

To interpret the result, we would compare this F-statistic with a critical F-value based on a chosen significance level (e.g., 0.05) and the respective degrees of freedom (2 for the numerator and 65 for the denominator). However, we don't have a specific p-value for this computed F-statistic from the provided information.

Interpretation:

If the p-value associated with the computed F-statistic of 3.2776 is less than 0.05 (or another chosen significance level), we would reject the null hypothesis, suggesting that the inclusion of X5 and X6 in Model 2 provides a significant improvement over Model 1. If the p-value is greater than 0.05, we would not reject the null hypothesis, indicating that X5 and X6 might not add significant explanatory power to the model beyond the predictors in Model 1.

Part2

Model 3

11. 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.

Set 1: Size-related Variables - House Structure: Variables that give us insights into the size of the main parts of the house structure.

a. `FirstFlrSF`: First floor square feet.
b. `SecondFlrSF`: Second floor square feet.
c. `GrLivArea`: Above-grade (ground) living area square feet.
d. `TotalBsmtSF`: Total square feet of basement area.

Set 2: Size-related Variables - Lot and Ancillary Areas: Variables related to the size of the lot and other ancillary areas in the house.

a. `LotFrontage`: Linear feet of street connected to the property.
b. `LotArea`: Lot size in square feet.
c. `BsmtFinSF1`: Finished square feet of basement area type 1.
d. `BsmtFinSF2`: Finished square feet of basement area type 2.
e. `BsmtUnfSF`: Unfinished square feet of basement area.
f. `MasVnrArea`: Masonry veneer area in square feet.

Reason for Separation:

i. House Structure: The first set focuses on the main living areas of the house, which is often the primary consideration for potential buyers. The total living area (`GrLivArea`), whether on the first floor, second floor, or the basement, directly impacts the functionality and desirability of a house.

ii. Lot and Ancillary Areas: The second set encompasses the overall lot size and other ancillary areas that, while not being primary living spaces, can still influence the house's appeal and functionality. For example, having a large `LotFrontage` might be desirable for those who want more privacy or a bigger yard. Similarly, the finished or unfinished sections of the basement can be useful for recreational areas or storage, while a larger `MasVnrArea` might add aesthetic appeal.

This grouping ensures that each set contains at least two variables, and it allows us to understand the impact of the main structural size versus additional features and land size on the sale price.

12. 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(SalePrice ~ FirstFlrSF + SecondFlrSF + GrLivArea + Tot
alBsmtSF, data=mydata)
> summary(model_3)

Call:
lm(formula = SalePrice ~ FirstFlrSF + SecondFlrSF + GrLivArea + 
    TotalBsmtSF, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-732026  -21634     209   20924  265225 

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)    
(Intercept) -21719.191   3136.111  -6.926    0.000000000005319 *
FirstFlrSF     142.294     19.782   7.193    0.000000000000802 *
SecondFlrSF    139.417     19.509   7.146    0.000000000001120 *
GrLivArea      -54.473     19.356  -2.814              0.00492  
TotalBsmtSF     68.902      3.396  20.288 < 0.0000000000000002 *
---
Signif. codes:  0 ‘*’ 0.001 ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48490 on 2924 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.632,     Adjusted R-squared:  0.6315 
F-statistic:  1255 on 4 and 2924 DF,  p-value: < 0.00000000000000022
```

Hypothesis Testing

For each coefficient in the model, we can test the null and alternative hypotheses:

$H0: \beta i = 0$

$Ha: \beta i \neq 0$

- For `FirstFlrSF`:
  - $H0$: The coefficient of `FirstFlrSF` is zero (it has no effect).
  - Ha:The coefficient of `FirstFlrSF` is not zero (it has a significant effect).
  - p-value: < 0.000000000000802 -> Reject $H0$.`FirstFlrSF` has a significant effect on `SalePrice`.
- For `SecondFlrSF`:
  - p-value: < 0.000000000001120 -> $H0$.`SecondFlrSF` has a significant effect on `SalePrice`.
- For `GrLivArea`:
  - p-value: 0.00492 -> Reject $H0$. `GrLivArea` has a significant effect on `SalePrice`, though the negative coefficient suggests that as `GrLivArea` increases, the predicted `SalePrice` may decrease, which is unusual and might need further investigation.
- For `TotalBsmtSF`:
  - p-value: < 0.0000000000000002 -> Reject $H0$. `TotalBsmtSF` has a significant effect on `SalePrice`.

All predictors in `model_3` are statistically significant based on their very low p-values.

The Omnibus Overall F-test:

$H0: \beta 1 = \ \beta 2 = \ \beta 3 = \ \beta 4 = 0$

$Ha: At\ least\ one\ \beta i \neq 0$

- F-statistic from the summary: 1255
- p-value for the F-test: < 0.00000000000000022

Given this extremely low p-value, we Reject $H0$ for the overall F-test, indicating that at least one predictor variable is significantly related to the dependent variable `SalePrice`. The model as a whole explains a significant amount of the variance in `SalePrice`.

Conclusion: Both the individual coefficient tests and the overall F-test confirm the statistical significance of the model. Each predictor (`FirstFlrSF`, `SecondFlrSF`, `GrLivArea`, and `TotalBsmtSF`) significantly affects `SalePrice`. The model as a whole (with all these predictors combined) also significantly explains the variation in `SalePrice`. However, the negative coefficient for `GrLivArea` is worth investigating further, as it might suggest multicollinearity or other data-related issues.

Model 4

13. Pick the other set (or one of the other sets) of explanatory variables. Add this set of variables to those in Model 3. You are preparing to fit a multiple regression model with this combined set of explanatory variables – call this Model 4. You should note that Model 3 is nested within Model 4. Fit the multiple regression model using the explanatory variables from the combined set of explanatory variables to predict SALEPRICE(Y). In other words, fit Model 4. 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

```
summary(model_4)

Call:
lm(formula = SalePrice ~ FirstFlrSF + SecondFlrSF + GrLivArea +
    TotalBsmtSF + LotFrontage + LotArea + BsmtFinSF1 + BsmtFinSF2 +
    BsmtUnfSF + MasVnrArea, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-784522  -22152     711   20262  248947

Coefficients: (1 not defined because of singularities)
                Estimate   Std. Error t value               Pr(>|t|)
(Intercept) -10338.27874   3595.38489  -2.875                0.00406
FirstFlrSF     104.73938     19.04840   5.499           0.00000004159 *
SecondFlrSF    108.33554     18.72676   5.785           0.00000000802 *
GrLivArea      -29.46386     18.55287  -1.588                0.11237
TotalBsmtSF     52.11047      3.45551  15.080 < 0.0000000000000002 *
LotFrontage     79.63446     46.31582   1.719                0.08565 .
LotArea          0.06186      0.12059   0.513                0.60801
BsmtFinSF1      22.40277      2.28026   9.825 < 0.0000000000000002 *
BsmtFinSF2      -0.92130      5.15805  -0.179                0.85825
BsmtUnfSF             NA           NA      NA                     NA
MasVnrArea      66.61511      5.49304  12.127 < 0.0000000000000002 *
---
Signif. codes:  0 ‘*’ 0.001 ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46320 on 2919 degrees of freedom
Multiple R-squared:  0.6648,    Adjusted R-squared:  0.6638
F-statistic: 643.3 on 9 and 2919 DF,  p-value: < 0.00000000000000022
```

```
> anova(model_3, model_4)
Analysis of Variance Table

Model 1: SalePrice ~ FirstFlrSF + SecondFlrSF + GrLivArea + TotalBsm
tSF
Model 2: SalePrice ~ FirstFlrSF + SecondFlrSF + GrLivArea + TotalBsm
tSF +
    LotFrontage + LotArea + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
    MasVnrArea
  Res.Df           RSS Df    Sum of Sq      F                Pr(>F)
1   2924 6875635981651
2   2919 6261734940281  5 613901041370 57.236 < 0.00000000000000022
*
---
Signif. codes:  0 ‘*’ 0.001 ‘’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The output of the `anova(model_3, model_4)` function is an Analysis of Variance (ANOVA) table that compares the two models: `model_3` (Model 1 in the output) and `model_4` (Model 2 in the output).

Model 1 (model_3): The formula represents the predictors used in Model 1 (`FirstFlrSF`, `SecondFlrSF`, `GrLivArea`, and `TotalBsmtSF`).

- `Res.Df`: Residual degrees of freedom, which is the difference between the number of observations and the number of parameters estimated (including the intercept). For Model 1, this value is 2924.
- `RSS`: Residual sum of squares for Model 1. It is a measure of the discrepancy between the data and the estimation model. A smaller RSS indicates a better fit to the data. For Model 1, RSS is 6875635981651.

Model 2 (model_4): The formula represents the predictors used in Model 2, which includes all predictors from Model 1 plus additional predictors (`LotFrontage`, `LotArea`, `BsmtFinSF1`, `BsmtFinSF2`, `BsmtUnfSF`, and `MasVnrArea`).

- `Res.Df`: For Model 2, the residual degrees of freedom is 2919, which is less than Model 1 due to the inclusion of more predictors.
- RSS`: For Model 2, RSS is 6261734940281, which is smaller than the RSS for Model 1, indicating a better fit.

Comparison between the models:

- `Df`: Degrees of freedom for the comparison between the models. It's calculated as the difference in residual degrees of freedom between the two models. Here, it's 5, which is the number of additional predictors added in Model 2.
- `Sum of Sq`: This represents the difference in RSS between the two models, indicating how much more variation is explained by Model 2 compared to Model 1. The value is 613901041370.
- `F`: The F-statistic is used to compare the fits of different models. A larger F-statistic suggests that the additional variables in Model 2 significantly improve the fit over Model 1. Here, the value is 57.236, which is quite large.

- `Pr(>F)`: This is the p-value for the F-statistic. A small p-value (here, it's almost zero) indicates that the improvement in fit from Model 1 to Model 2 is statistically significant.

Conclusion: Given the extremely small p-value, there's strong evidence to reject the null hypothesis that Model 1 fits the data as well as Model 2. The inclusion of the additional predictors in Model 2 (`LotFrontage`, `LotArea`, `BsmtFinSF1`, `BsmtFinSF2`, `BsmtUnfSF`, and `MasVnrArea`) significantly improves the model's fit. Thus, Model 2 (model_4) provides a significantly better explanation of the variation in `SalePrice` compared to Model 1 (model_3).

a. Hypothesis Tests for All Model Coefficients Individually:

For each coefficient, we are testing:

Null Hypothesis ($H0$): The coefficient is equal to zero (i.e., it has no effect).

Alternative Hypothesis (Ha): The coefficient is not equal to zero (i.e., it has some effect).

From the `summary(model_4)` output:

Hypothesis Testing

1. FirstFlrSF:

- Ho: The coefficient for `FirstFlrSF` is equal to zero.
- Ha: The coefficient for `FirstFlrSF` is not equal to zero.
- Interpretation: The p-value is extremely small (<0.00000004159), suggesting we reject the null hypothesis. This means that `FirstFlrSF` has a statistically significant effect on `SalePrice`.

2. SecondFlrSF:

- Ho: The coefficient for `SecondFlrSF` is equal to zero.
- Ha: The coefficient for `SecondFlrSF` is not equal to zero.
- Interpretation: The p-value is extremely small (<0.00000000802), so we reject the null hypothesis. `SecondFlrSF` significantly impacts `SalePrice`.

3. GrLivArea:

- Ho: The coefficient for `GrLivArea` is equal to zero.
- Ha: The coefficient for `GrLivArea` is not equal to zero.
- Interpretation: The p-value is 0.11237, which is greater than the common alpha level of 0.05. We fail to reject the null hypothesis, suggesting `GrLivArea` may not have a significant effect on `SalePrice`.

4. TotalBsmtSF:

- Ho: The coefficient for `TotalBsmtSF` is equal to zero.
- Ha: The coefficient for `TotalBsmtSF` is not equal to zero.

- Interpretation: The p-value is extremely small, so we reject the null hypothesis. `TotalBsmtSF` significantly affects `SalePrice`.

5. LotFrontage:

- Ho: The coefficient for `LotFrontage` is equal to zero.
- Ha: The coefficient for `LotFrontage` is not equal to zero.
- Interpretation: The p-value is 0.08565, slightly greater than 0.05. This means that at a 5% significance level, we fail to reject the null hypothesis. But at a 10% level, it's significant.

6. LotArea:

- Ho: The coefficient for `LotArea` is equal to zero.
- Ha: The coefficient for `LotArea` is not equal to zero.
- Interpretation: With a p-value of 0.60801, we fail to reject the null hypothesis, implying `LotArea` might not be significant in predicting `SalePrice`.

7. BsmtFinSF1:

- Ho: The coefficient for `BsmtFinSF1` is equal to zero.
- Ha: The coefficient for `BsmtFinSF1` is not equal to zero.
- Interpretation: The extremely small p-value indicates we reject the null hypothesis. `BsmtFinSF1` has a significant impact on `SalePrice`.

8. BsmtFinSF2:

- Ho: The coefficient for `BsmtFinSF2` is equal to zero.
- Ha: The coefficient for `BsmtFinSF2` is not equal to zero.
- Interpretation: With a p-value of 0.85825, we fail to reject the null hypothesis. This suggests `BsmtFinSF2` might not be a significant predictor.

9. BsmtUnfSF:

- This variable has NA for its coefficient, standard error, t-value, and p-value, suggesting a potential multicollinearity issue.

10. MasVnrArea:

- Ho: The coefficient for `MasVnrArea` is equal to zero.
- Ha: The coefficient for `MasVnrArea` is not equal to zero.
- Interpretation: The p-value is extremely small, indicating we reject the null hypothesis. `MasVnrArea` has a significant effect on `SalePrice`.

Overall Model Significance:

The F-statistic tests the hypothesis that all regression coefficients are equal to zero versus at least one is not. Given the extremely small p-value of the F-statistic (<0.00000000000000022), we reject the null hypothesis, indicating that the predictors in the model significantly explain the variability in `SalePrice`.

b. Omnibus Overall F-test:

Null Hypothesis Ho: All regression coefficients in the model are equal to zero. This means that none of the predictors (independent variables) have any effect on the response variable, which in this case is `SalePrice`.

Alternative Hypothesis Ha: At least one regression coefficient in the model is not equal to zero. This implies that at least one of the predictors has a significant effect on `SalePrice`.

Interpretation: The given F-statistic value is 643.3. The extremely small p-value associated with this F-statistic (<0.00000000000000022) is way below the typical significance level of 0.05. This allows us to reject the null hypothesis.

Therefore, we conclude that at least one predictor variable in the model is significant in explaining the variability in `SalePrice`. The model as a whole, with the predictors included, significantly fits the data better than a model with no predictors.

14. Write out the null and alternate hypotheses for a nested F-test using Model 3 and Model 4, to determine if the set of additional variables added to Model 3 to make Model 4 variables are useful for predicting SALEPRICE(Y). Your hypotheses must use symbols. Compute the F-statistic for this nested F-test and interpret the results.

A nested F-test is used to compare two models, one of which (Model 3) is a simpler, "nested" version of the other (Model 4). The goal is to determine if the additional predictors in the more complex model (Model 4) provide a significant increase in explanatory power compared to the simpler model.

Nested F-test:

Null Hypothesis Ho: The additional variables in Model 4 (`LotFrontage`, `LotArea`, `BsmtFinSF1`, `BsmtFinSF2`, `BsmtUnfSF`, `MasVnrArea`) do not improve the fit of the model over Model 3.

Alternative Hypothesis Ha: The additional variables in Model 4 provide a significant improvement in the fit of the model compared to Model 3.

Computing the F-statistic for the nested F-test:

$$F = \frac{\frac{Difference\ in\ RSS\ between\ models}{Difference\ in\ df\ between\ models}}{\frac{RSS\ of\ Model\ 4}{df\ of\ Model4}}$$

From the ANOVA table:

Difference in $RSS\ =\ 6875635981651\ -\ 6261734940281\ =\ 613901041370$

Difference in $df\ = 2924\ -\ 2919\ =\ 5$

RSS of Model 4 = 6261734940281

df of Model 4 = 2919

$$F = \left(\frac{\frac{613901041370}{5}}{\frac{6261734940281}{2919}}\right) = 57.236$$

Interpretation: The computed F-statistic is 57.236. The extremely small p-value (<0.00000000000000022) associated with this F-statistic tells us that this value is highly significant.

Given this, we reject the null hypothesis Ho. This implies that the additional predictors in Model 4 provide a significant improvement in the fit of the model over Model 3. The more complex model, Model 4, explains the variability in `SalePrice` significantly better than the simpler Model 3 when considering the additional variables.