1. The number of observations in the sample data refers to the total count of individual entries or data points collected. We have (x1,x2,x3,x4) for our variables, the variables count of 67, with the addition of the intercept makes the total amount **72**.
2. Null Hypothesis (H0): The coefficient for the independent variable x1 (Beta1) is equal to zero, indicating that x1 has no significant effect on the dependent variable.

   H0:β1=0

   Alternative Hypothesis (Ha): The coefficient for the independent variable x1 (Beta1) is not equal to zero, indicating that x1 has a significant effect on the dependent variable.

   Ha:β1!=0

3. To compute the t-statistic for the coefficient Beta1 (β1), we divide the estimated coefficient by its standard error. If the computed t-statistic for Beta1 is 5.327 as given, then the calculation would have been:

   $t=SE(\beta 1)\beta 1=5.327$

   The p-value associated with this t- is less than 1.258 x 10^-6. This is much lower than the target 0.05.

   Therefore, we reject the null hypothesis H0: β1 = 0. The small p-value indicates that the result is statistically significant, meaning there is strong evidence against the null hypothesis. As a result, we conclude that it has a statistically significant effect on the dependent variable in our regression model.

4. The formula to compute the R-Squared value for Model 1 is

   $R2=SSR/SST$

   Inputting our variables, we get R2=2216/2756.37 resulting in 0.8039. The interpretation of this R-squared value is that approximately 80.39% of the variation in the dependent variable

5. The adjusted R-squared statistic is used to account for the number of predictors in the model relative to the number of observations, providing a more accurate picture of how well the model predicts the dependent variable.
   The formula for adjusted R-squared (adj. R^2) is:
   Ad.R^2 = 1 – (1-R^2)(n-1/n-p-1)
   Where
   R^2 is the R-squared value

n represents the number of observations in our dataset.

p is the number of predictors, or independent variables, in our model.

Swapping the values, we get the adjusted value of 0.7577. The adjusted R-squared is different from the R-squared because it takes the number of variables into account. The adjusted R-squared is lower than the R-squared because adding more variables to the model will always increase the R-squared value, but it may not necessarily.
increase the accuracy of the model.

6. Null Hypothesis ($H0$): $H0:\beta1=\beta2=\beta3=\beta4=0$
   This states that all coefficients of the independent variables are equal to zero, implying that none of the independent variables have a statistically significant effect on the dependent variable.
   Alternative Hypothesis (Ha): At least one $\beta i \neq 0$ for i =1,2,3,4
   This states that at least one coefficient of the independent variables is not equal to zero, suggesting that at least one of the independent variables significantly contributes to the model's prediction of the dependent variable.

7. Given the mean square of the model (MSR) is 531.50 and the mean square of the residuals (MSE) is 9.41, the F-statistic can be calculated as:

   F = 531.50/9.41 = 56.48

   We can consult the F-distribution to find the corresponding p-value.

   Given that the p-value is indeed well below the 0.05 threshold, we reject the null hypothesis. This indicates that there is a statistically significant relationship between the independent variables and the dependent variable. In other words, the regression model has a very high significance.

8. Model 1 nests model two because model 2 includes all the variables of model 1 plus any additional variables. Model 1 is a more restricted or simplified version of Model 2.

9. Assuming that Model 2 includes all the predictors of Model 1 plus two additional predictors (let's denote them as x5 and x6), the null hypothesis (H0) for the nested F-test would be:
   Null Hypothesis ($H0$):$\beta5=\beta6=0$ This hypothesis states that the additional variables in Model 2 ($\beta5$ and $\beta6$) do not significantly improve the fit of the model, implying that both coefficients are equal to zero.
   Alternative Hypothesis (Ha): At least one of the additional variables in Model 2 significantly improves the fit of the model. At least one of these coefficients is not equal to zero:
   Ha:At least one $\beta i=0$ for $i=5,6$

10. F = (630.36-572.61)/(6-4)/572.61/(72-6-1) = 3.28

The null hypothesis is

H0 = $\beta 5 = \beta 6 = 0$

This hypothesis states that the coefficients for the additional predictors in Model 2 (assuming these are $\beta 5$,$\beta 6$, etc.) are all equal to zero, suggesting that these additional predictors do not significantly improve the model fit.

The alternative hypothesis is

Ha:At least one $\beta i = 0$ for $i = 5,6$

This hypothesis asserts that at least one coefficient for the additional predictors in Model 2 is not equal to zero, implying that the additional predictors do significantly improve the model fit.

There is at least one of the additional predictors included in Model 2 that provides a statistically significant improvement to the model's fit over Model 1. Therefore, Model 2 is considered statistically better than Model 1.

11. Based on the selection of continuous quantitative variables, we can create a logical separation of these variables into two distinct sets—one focused on the characteristics related to the size of the house and the other on the quality and condition of the house. Here is a rationale for the separation:

Set 1: Size-Related Variables

- TotalBsmtSF: Total square feet of basement area

- FirstFlrSF: First-floor square feet

- GrLivArea: Above-grade (ground) living area square feet

- FullBath: Full bathrooms above grade

- HalfBath: Half baths above grade

The variables in Set 1 are all related to the physical space and size dimensions of the house. TotalBsmtSF, FirstFlrSF, and GrLivArea reflect the total living area in different parts of the house, which can directly influence the sale price because larger homes typically command higher prices. FullBath and HalfBath represent an aspect of size as they quantify the number of bathrooms, which can also be a critical factor in the functionality of a home and, consequently, its value.

Set 2: Quality and Condition Variables

- OverallQual: Rates the overall material and finish of the house

- OverallCond: Rates the overall condition of the house

- YearBuilt: Original construction date

In Set 2, OverallQual and OverallCond reflect the perceived quality and current condition of the property. These are important because they indicate how well the property has been maintained and the level of upgrades that have been put into it. YearBuilt can also be a proxy for quality and condition; newer houses may require less immediate maintenance and offer more modern amenities compared to older homes.

The separation has been performed on the basis that while size is critically important—it often represents the baseline from which buyers begin their house valuation—quality and condition can considerably adjust the price upwards or downwards. Buyers may be willing to pay more for a house that is well-maintained and features high-quality construction or less for a large house that requires significant renovations. By segmenting these variables, we can better understand their unique contributions to the overall sale price of the house.

#Coefficients:

Estimate Std. Error t value Pr(>|t|)

#(Intercept) -35152.145   3241.329 -10.845 < 2e-16 ***

GrLivArea   49.172   3.098 15.872 < 2e-16 ***

TotalBsmtSF   64.290   3.322 19.354 < 2e-16 ***

FirstFlrSF   23.929   4.359   5.489 4.38e-08 ***

FullBath   23937.968   2068.224 11.574 < 2e-16 ***

HalfBath   24674.383   2216.902 11.130 < 2e-16 ***

12.

Model 3 uses a set of size-related variables to predict the SALEPRICE in a multiple regression model. After preparing the dataset by selecting the appropriate variables and dropping any records with missing values, a linear model (Model 3) was fitted using the `lm()` function in R.

a) Hypothesis Tests for Individual Coefficients:

For each explanatory variable in Model 3, we test whether the variable is significantly related to the SALEPRICE. The null and alternative hypotheses are as follows:

Null Hypothesis (H0): The coefficient for the variable is zero, which indicates that the variable does not significantly predict SALEPRICE.

Alternative Hypothesis (Ha): The coefficient for the variable is not zero, which indicates that the variable significantly predicts SALEPRICE.

For example, consider the variable GrLivArea:

- H0: β GrLivArea = 0

- Ha: β GrLivArea ≠ 0

From the model summary output, we see that all the t-values for the predictors are associated with very small p-values (all less than 0.05, many much smaller). Thus, we reject the null hypothesis for each coefficient, concluding that all variables in this set are significant predictors of SALEPRICE.

Each size-related variable has a statistically significant relationship with SALEPRICE. Variables such as GrLivArea, TotalBsmtSF, FirstFlrSF, FullBath, and HalfBath contribute significantly to predicting the sale price of a house.

b) Omnibus Overall F-test:

The overall F-test evaluates whether at least one explanatory variable has a significant relationship with the dependent variable in the model.

Null Hypothesis (H0): All coefficients in the model are equal to zero, which would suggest that none of the explanatory variables significantly predict SALEPRICE.

H0: $\beta1 = \beta2 = \beta3 = \beta4 = \beta5 = 0$

Alternative Hypothesis (Ha): At least one coefficient in the model is not equal to zero, suggesting that at least one explanatory variable significantly predicts SALEPRICE.

Ha: At least one βi for $i = 1,2,3,4,5$

The F-statistic value from summary(model_4) is 1205 and all p-values presented are well below except for YrSold and FirstFlrSF. This shows an extremely high significance for the model, but for those two we cannot reject the null hypothesis.

Nested Model:

Null Hypothesis (H0): The coefficients of the additional variables in Model 4 are all equal to zero, suggesting that these variables do not significantly improve the model's predictive ability for SALEPRICE.

H0: $\beta 7 = \beta 8 = \beta 9 = 0$

Alternative Hypothesis (Ha): At least one of the coefficients of the additional variables in Model 4 is not equal to zero, indicating that these variables provide a significant improvement in the model's predictive ability for SALEPRICE.

Ha:At least one βi: 0 for i = 7, 8, 9

Upon reviewing the summary and ANOVA output for the nested model (Model 4), we focus on the F-statistic and the associated p-value. The computed F-statistic is 1396, and the p-value is virtually zero (< 2.2e-16), indicating an extremely high level of significance.

The very low p-value for the F-test suggests that we reject the null hypothesis. This means that the set of additional variables (OverallQual, OverallCond, and potentially others if Model 4 contains more than these) significantly improve the prediction of SALEPRICE when compared to Model 3, which did not include these variables.

Therefore, Model 4, with the additional quality and condition variables, is statistically better for predicting SALEPRICE than Model 3, which only included size-related variables. The inclusion of variables related to the overall quality and condition of the house has a substantial impact on the model's predictive ability.