Concept

* : This represents the total variation in your data. It's like asking, "How much do the values of my dependent variable (the one I'm trying to predict) vary around their mean?"
* : This measures how much of that total variation your model can explain. It's like saying, "How much of this variation can be attributed to the relationship I’ve modeled with the independent variables (the predictors)?"
* : This is the part of the total variation that your model cannot explain. It's like admitting, "Despite my model, there's still some variation in the data that I can’t account for."
* : All the variation in my data (TSS) can be split into two parts: what my model explains (ESS) and what it doesn’t (RSS).
* : While gives a proportionate measure of fit quality, it doesn't always tell the whole story, especially in nonlinear models or models with high leverage points.
* : Mean Squared Error, the MSE gives us a single number representing the error of our model. The smaller the MSE, the closer our predictions are to the actual values, indicating a better model fit.
* They both are important in assessing a model's performance, but they should be used together to get a comprehensive picture. For instance, a high with a high indicates a model that explains a large proportion of variance but still has large errors, possibly due to outliers.
* Q: What’s the expected impact on when adding more features to the model?

Subset Selection

* Reduce the complexity of model
* Forward Stepwise Method: Starts with an empty model and adds variables step by step, testing each addition. Pros & Cons: More computationally efficient than best subset selection but might not find the best model if interactions between predictors are important.
* Backward Method: Begins with the full model and removes variables step by step, testing each removal. Pros & Cons: Useful when the number of predictors is not too large. It can find better models than forward selection if interactions are present, but it still may not find the best possible model.

Linear Regression Assumptions and Pitfalls:

* Heteroscedasticity: It occurs when the variance of the residuals is not constant across all levels of the independent variables. In other words, the spread of the residuals is uneven across the range of predicted values. X increases, Residuals increase
* Normality: Residuals are normally distributed. QQ-Plot to check
* Outliers: Cook’s distances, estimate the influence of any data givens. Remove points above a certain threshold
* Multicollinearity: it occurs when two or more independent variables in a regression model are highly correlated. This can make it difficult to determine the individual effect of each independent variable and can inflate the variance of the coefficient estimates. Use VIF to detect, and use PCA/PLS to reduce the dimensionality, or linear combine the variables