Assignment-based Subjective Answers:

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans:

- The demand of bikes rental  will be more on during fall i.e. during June, July, August, September
- Bike rental demand is more especially during Friday and Saturday weekdays.
- Bike rental demand is more during holidays than workingdays compartively(by proportion)
- Bike rental demand is more when weather is clear

2. Why is it important to use drop_first=True during dummy variable creation?
Ans: To represent n categorical values in the categorical feature data, n-1 features are sufficient, hence we drop the first categorical value in this feature.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: Temprature has a higher correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans: - By checkiong VIF values for multicollinearity

- scatter plot among errors to verify, there should not be any pattern.
- Residual distribution plot to verify errors are distributted normally.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans: Temprature, season, weekdays

General Subjective Answers:
1. Explain the linear regression algorithm in detail.
Ans:

- Linear regression algorithm is a statistical technique to find the relationship between among independent variable(predictor varibale)

and dependent variables. It is assumed that these variables have a linear relationship, error terms are normally distributed and independent to each other.

Errors terms have a constant variance. Linear regression is an interpolated model and it can not find extroplated relationship among variables.

Feature selections:

- Feature selections will be done using a balanced approach of automated (Coarse Tuning) and manual (Fine Tuning)

Automated: Using RFE method, Regualrization and based on the AIC, BIC values

Build model:

- To find relationship between these variables, The linear equation is used to fit the line and it is called as model.
- To find significance of coefficients which are derived from model, hypothesis testing using P-value is done. If P-value of coefficient is above the threshold value,

these coefficients are ignored.

- if independent variables are multicollinear each other, then one of the feature is used to build the model. This multicollinearity is found using

Variance information factor(VIF). It is good to have less than 5 to a feature.

- Adjusted R-square will be used to penalise the model having number of variables. It will increase or decrease when a new feature is added but R-square will be remain same or increase.
- If model has a higher adjusted R-square, the model explains the variance of the independent.

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe’s quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. Therefore, it is a good practice, visualize the data prior to the analysis.

3. What is Pearson’s R?

Ans: It provides relationship between variables how much are correlated to each other and it's value is range from -1 to 1

if the value is 0: it means, the varibales do not have any relationship.
if the value is greter than zero: it means, the varibales does have the positive linear relationship.
if the value is less than zero: it means, the varibales does have the negative relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: In general, the features data will be in different units and different ranges and this will be a problematic to find cost function of model in less time. For example, Gradient descent algorithm takes more iterations to find minimum cost. If features are not scaled, it does not affect the R-square values.

normalized scaling scales the features data in the range of 0 to 1
standardized scaling scales the features data with unity stadard deviation and zero mean.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF will be infinite, if their R-square value is one. It means, these are highly correlated to each other explaining the same variance of output variable.