Assignment-based Subjective Questions

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

Ans: 1--Season: 3--fall has highest demand for rental bikes.

2--Demand for next year has grown.

3--Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing. During September, bike sharing is more.During the year end and beginning, it is less, could be due to extreme weather conditions.

4--When there is a holiday, demand has decreased.

5--Weekday is not giving clear picture about demand.

6--The clear weathersit has highest demand.

2. Why is it important to use drop_first=True during dummy variable creation?

Ans: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans: Both ‘temp’ and ‘atemp’ variables have highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: Residual Analysis—check if error terms are normally distributed by plotting them.

Check if predicted values and values in train data set follow same pattern.

Check if errors are equally distributed about their mean 0 by plotting a scatter plot

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

Ans: --Temperature

--Weather:3(Bad)

--Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a machine learning algorithm based on supervised learning.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

The linear regression model provides a sloped straight line representing the relationship between the variables.

$Y = \beta_1 X + \beta_0$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
β_0 = intercept of the line (Gives an additional degree of freedom)
β_1 = Linear regression coefficient (scale factor to each input value).

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x, y) points.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

Ans: Pearson correlation coefficient (also known as Pearson's *R*) --is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

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

Ans: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Ans: VIF = $1 / (1 - R_i^2)$

VIF will become infinite when R_i^2 value becomes 1. This happens when Linear relationship between two variables perfectly fits a line (i.e., by plotting two variables we get a straight line passing through all points).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.