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?

Answer:

- Light_rainsnow: A coefficient value of '-0.237' indicates that the light snow and rain deters people from renting out bikes.
- Demands increases in the month of Aug, Sep, season of summer, winter, day of Saturday, yr.
- Demands decreases in the month of Jan, Light_rainsnow, Misty.

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

Answer: It drops the redundant dummy variable. It also reduces the correlation between dummy variables.

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

Answer: Variables temp and atemp are high correlated.

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

Answer: The assumptions of Linear Regression are

- Residual analysis of training data: Since, residuals are normally distributed. Hence our assumption for Linear Regression is valid.
- Homoscedasticity: There is no visible pattern in residual values, thus homoscedacity is well preserved.
- Multicolinearity: There is no strong correlation between variables after removing temp variable.
- Linear relationship between X and y: The model is linear between dependent variables and independent variables.

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

Answer: As per the final model, the top 3 predictor variables that influence the bike booking are:

- Feeling Temperature(atemp): A coefficient value of '0.535' indicates that a temperature has significant impact on bike rentals.
- Light_rainsnow: A coefficient value of '-0.237' indicates that the light snow and rain deters people from renting out bikes.
- Year (yr): A coefficient value of '0.231' indicates that a year wise the rental numbers are increasing.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is one of the most fundamental techniques of machine learning, in which a model is trained to predict the behaviour of data based on a set of variables. In the case of linear regression, the x-axis and y-axis variables should be linearly associated.

Types of Linear Regression

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Assumptions
The Linear Regression model makes the following assumptions about the dataset:

Relationship between variables: The linear regression model implies that the relationship between dependent and independent variables should be linear.

Multicollinearity: The linear regression model assumes no multicollinearity in the data. Essentially, multi-collinearity happens when independent variables are dependent on one another.

Normally distributed: The deference between measured output and predicted output (known as error terms) should be normal distributed. They should also be independent of each other.

Constant variance: Variance should not follow any pattern, otherwise model will be unreliable.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is important for data visualization. It demonstrates the significance of plotting data before analysing and building a model. Anscombe's quartet comprises of four datasets that contain statistical observations. These datasets provide the same information (including variance and mean) for each x and y point. However, when these data sets are plotted, they appear significantly differently from one another.

3. What is Pearson's R?

Answer: The Pearson correlation coefficient is calculated by dividing the co-variance of two variables by the product of their standard deviations. It is used to determine a linear relationship between the two variables. The positive value of Pearson's correlation coefficient suggests that changing one of these variables will have a positive effect on the other.

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

Answer: Scaling is a pre-processing step of building a linear model which is used to normalize the independent variables within a particular range.

It can be possible that sometimes independent variables have corrected with different units. Due to this, model will predict wrong values. To solve this issue, we bring all independent variables into same level of magnitudes.

Standardized scaling affects the values of dummy variables, but MinMax scaling does not affect the values of dummy variables.

MinMax scaling lies in the range of 0 and 1, whereas standardized scaling handles its value in Z score with 0 mean and standard deviation 1.

MinMax scaling can lose some information about dataset, mainly about outliers.

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

Answer: An infinite VIF value means that the corresponding variable can be expressed perfect linear combination of other variables.

There is a perfect correlation if VIF is infinity. This shows that two separate variables have a perfect correlation with each other. If there is perfect correlation, R2 = 1, which means that 1/(1-R2) would be infinity, so will be R2 = 1. To overcome this issue, we must eliminate one of the variables so that multicollinearity can be avoided.

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

Answer: A Q-Q plot is a scatter plot made by comparing two sets of quantiles.

Using the Q-Q plot, we can ensure that both sets of data, i.e., training and test, are derived from populations with equal distributions. Q-Q plots are used to determine the distribution of a random variable, such as Gaussian distribution, Uniform distribution, Exponential distribution, Pareto Distribution, etc. Using the Q-Q plot, we may determine the type of distribution just by observing the plot.

Q-Q plots are utilised to visually examine the homoscedasticity and normality assumptions of linear regression.