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: Year, season, month, working day, weekday are categorical variables in dataset.

- Clear weather results in high booking, which appears to be evident.
- Working & non-Working day will not impact the booking.
- most of the booking done during May-Oct.
- More bookings are observed during the fall season.
- booking increasing 2019 as compared to last year 2018.

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

Ans: -

- This helps to reducing the additional column created during dummy variable creation, to achieve k-1 dummy variable.
- It also used to reduce the collinearity between dummy variables.

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

Ans: - atemp & temp both has 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:-

- By checking the normality of error terms or error terms should correspond to a normal curve.
- There is no visible pattern to be observed. *Line number [103,104]*

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

Ans: - Top 3 features are.

Year & Month Sep (+ve influencing) and Light Snow_Rain (-ve influencing)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is useful for finding the linear relation between the target variable and one or more predictors.

There are two types of LR

Simple Linear Regression: -

one dependent & one independent variable

$$\boldsymbol{y = \beta_0 + \beta_1 X}$$

Multiple Leaner Regression.

one dependent & multiple independent variable

$$\boldsymbol{y = \beta_0 + \beta_1 X + \beta_1 X + \cdots + \beta_n X}$$

The Steps will be as follows:

a) Reading, understanding and visualising the data.

   After looking into the data and cleaning it with exploratory data analysis and analysis the univariate & Bivariate analysis.

b) Preparing the data for modelling (train-test, split, rescaling etc..)

   split the dataset into training set (which would be used to train a model) and the testing set (which would be used to check how close is our model to the actual output).

c) Training the model

   After checking the collinearity of variables and using the requisite variables to train the model and checking the R-value of the model and the p-values of dependent variables, after dealing/dropping the necessary columns and reiterating the steps (feature elimination), we come to a final model.

d) Residual analysis

   linear regression which states that the error curve must be a normal one.

e) Predictions and evaluation on the test set

   proceed to testing the model with the test dataset. The conclusion hence drawn on the model would be used to provide valuable insights.

2. Explain the Anscombe’s quartet in detail.

Ans:- It comprise a set of four dataset, having identical descriptive statistical properties in terms of means, variance , R-squared, correlation and linear regression lines but having different representations when we scatter plot on graph.

It used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on the summary statistics. "It highlights the significance of leveraging data visualization to identify trends, outliers, and other critical details that may not be apparent solely from summary statistics.

3. What is Pearson’s R?

Ans:- Pearson’s correlation coefficient, also known as Pearson’s is a measure of strength of correlation between two variables. It commonly used in linear regression (LR). The values lies between -1 to +1 that measures the strength and direction of the relationship between two variables.

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

Ans:- It’s a step of data pre-Processing which is applied to independent variables to normalize the data within a particular range

Most of the time the data set contains high varying in magnitudes, units and a range. If scaling is not done, then algorithm only takes magnitude in account and not units this makes the model incorrect. To solve this, make all the variable in common scaling or level magnitude.

Scaling just affects only coefficients and not on any other parameters like t-statistic, F-statistic, P-value, R-squared etc..

- **Normalization/Min-Max Scaling**: - It brings all of the data in the range of 0 and 1.

$$MinMax\ Scaling\ x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **Standardization Scaling: -** It replaces the values by their Z score. It brings all of data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ)

$$MinMax\ Scaling\ x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

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

Ans:- The VIF Calculated by

$$VIF_i = \frac{1}{1 - R_i^2}$$

If R-squared values is equal to 1 then the denominator of the above formula becomes 0 and the overall value become infinite. It denotes perfect correlation between variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Ans:- Q-Q plot is a graphical tool to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both the data comes from the same background, to maintain the sanity of the model.