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?

   Season, weather situation, holiday, month, working day and weekday were the categorical variables in the dataset. A boxplot was used to visualize these. These variables influenced our dependent variable.

   We infer from the above information that
   - More bikes go on rent in fall
   - Bad Weather conditions affect the count (Light Snow, Misty)
   - September month has more count
   - Holiday decreases the count
   - Jan and dec has lest count

   (3 marks)

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

   Your dummy variables if you don't remove the first column(redundant). This may have negative impact on some models and the effect is amplified when the cardinality is low. Iterative models, for example, may have difficult convergent, and lists of variable importance may be distorted.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Temp and atemp are two numerical variables which are highly correlated with target variable (cnt)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The distribution of residuals should be normal and centred around 0. The mean is 0. We test these residuals assumptions by producing distplot of residuals to see if they follow a normal distribution or not. The residuals are scattered around mean=0.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top three predictive variables that influence bike booking, according to our final model are temp, weathersit and year

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Linear regression is a type of supervised ML algorithm that is used for the prediction of numeric values. It is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. It is based on the popular equation y=mx+c. It assumes that there is linear relationship between the dependent variable and the predictor variable or independent variable. In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed

when the dependent variable is of continuous data type and predictors or independent variables could be of any datatype like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least errors.

Regression is broadly divided into two types.

Simple Linear Regression: SLR is used when the dependent variable is predicted using multiple independent variables.

Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet was developed by statistician Francis Anscombe. It includes four data sets that have most identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analysing it and the effect of outliers and other observations.

3. What is Pearson's R? (3 marks)

Pearson correlation coefficient, also known as Pearson R, is a statistical test that estimates the strength between the different variables and their relationships. Hence, whenever any statistical test is performed between the two variables, it is always a good idea for the person to estimate the correlation coefficient value to know the strong relationship between them.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

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. It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and

1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

Standardization Scaling:

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

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The value of VIF is calculated by the below formula: VIF =1/(1-Ri^2). VIF = infinity if there is perfect correlation. If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

The quantiles in the first data set are plotted against the quantiles of the second dataset in a q-q graphic. It's a tool for comparing the shapes of different distributions. A scatter plot generated by plotting two sets of quantiles against each other is known as Q-Q plot.