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? (3 marks)

- weathersit – Demand is null when there is heavy precipitation. Demand is less if the weather is light precipitation or cloudy. Demand is highest when the weather is clear.
- Holiday **-** Rentals low during holiday. Customers likely to use bikes for office/business related purposes.
- Season versus Cnt: Demand is highest in Fall season, but least in Spring season.
- Year versus Cnt: Demand in year 2019 was high compared with the year 2018
- Month versus Cnt: Demand is high from May to October, However September month has highest demand

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

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

If there are 'n' dummy variables, we would require only 'n-1' columns.

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 the numerical variables which are high correlation with the target variable.

The pair plots are given in the Jupyter notebook for your reference.

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

a. The predicted values have linear relationship with the actual values. - To validate this assumption a scatter plot between actual values and predicted values in the train and test dataset were generated and best fit line were drawn to ensure linear relationship.

b. The error terms are normally distributed. A histogram of error terms was plotted to check whether it follows the normal distribution, and the result was further confirmed using normal Q-Q plot.

c. The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.

d. There is no Multicollinearity between two independent variables. A heatmap of all variables was created to show that none of the variables are highly correlated to each other. Further, VIF of all the variables were also calculated, which indicated that there is no multicollinearity between any variables since all VIF values are well within the acceptable range of less than 5.

e. Homoscedasticity of Residuals. A residual plot was drawn and from that, it was clear that there is no visible trend in the distribution of residuals. Hence, it was confirmed that the residuals are homoscedastic.

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

Light Rain_Snow_mist(Light precipitation), Winter,Year

General Subjective Questions

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation **"y = mx + c".**

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). 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 data type 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 error.
In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple
1. Simple Linear Regression - is used when the dependent variable is predicted using only **one** independent variable.

2. Multiple Linear Regression - is used when the dependent variable is predicted using multiple independent variables.
Some examples of use of linear regression in real life would be:
• For understanding the relationship between budget allocation and the return on investment from each of the invested sectors in an organization.
• To predict the impact of weather change on yield of crops in agricultural industry.

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

Francis Anscombe created Anscombe's Quartet in 1973 to demonstrate the significance of plotting graphs before analysing and modelling, as well as the impact of additional observations on statistical features. It is described as a set of four data sets that are almost equal in terms of simple descriptive statistics but have extremely distinct distributions and look differently on scatter plots. The necessity of visualising data before using various algorithms to develop models is illustrated by Anscombe's Quartet, since data feature plots may help discover abnormalities in the data such as outliers, diversity, and linear separability.

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

Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear r relationship between two sets of data. In simple terms, it tells us can we *draw a line graph to represent the data?* r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association

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

Feature **scaling** is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks

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

When VIF is infinite, it means that there is perfect correlation between the two independent variables under consideration. That’s the R2 will be one with these two variables in the model.
The equation of VIF is: **VIF = 1/(1-R2)**
So, when R2 =1; the equation becomes **VIF = 1/(1-1) = 1/0 = Inf**

In a data when two independent variables are highly correlated, it’s said to have multicollinearity . Thus, to overcome multicollinearity, one of the variables should be removed.

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

The quantile-quantile (Q-Q) plot is a graphical tool for detecting if two data sets are from the same population. A Q-Q plot is a comparison of two dataset’s quantiles. Quantiles are cut points that divide the range of a probability distribution into equal-probability continuous intervals or the observations in a sample in the same way. Q-quantiles are values that divide a finite collection of values into almost equal-sized Q subgroups.

These plots are useful in a linear regression situation where the training and test data sets are obtained separately, and the Q-Q plot is used to validate that both data sets are from populations with similar distributions. In a Q-Q plot, 45-degree reference line is plotted to determine the normality of populations. The points fall roughly along this reference line if the two sets originate from the same population with the same distribution. But if the two data sets are from populations with distinct distributions, then more is the deviation from this reference line.