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)

- season: Maximum bike booking happened during ‘fall’ season with a median of around 5000. This number was followed by ‘summer’ and ‘winter’ season.
- mnth: Maximum number of bike booking happened during May to October with a median of above 4000.
- holiday: Maximum number of bike booking happened when it is a holiday.
- workingday: Median for both working and non-working day is almost same, which is around 4000.
- weekday: Weekday shows a very similar trend throughout the week having medians in the range of 4000 to 5000.
- weathersit: Highest number of bookings were happening in ‘clear’ weathersit with a median of around 5000. It is followed by ‘mist_cloudy’ weather.

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

The key idea behind creating dummy variables is that for a categorical variable with ‘n’ levels, we need to create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. So **drop_first=True** helps in deleting the first column. Therefore, if the column has 3 levels, we need 2 new columns.

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

‘temp’ has the highest correlation with the target variable ‘cnt’.

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

- Performed the residual analysis, then plotted the density plot of the error terms, Check whether the error terms are normally distributed with mean = 0.
- Check if multicollinearity exists or not by calculating the VIF. In this case for all the predictor variables VIF is less than 5. It explains that multicollinearity does not exist.
- Check whether there is a linear relationship between variables X and Y by creating a pair plot.

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

- temp: It has a coefficient of 0.5499
- weathersit_lightsnow_lightrain : It has a coefficient of -0.2880
- yr: It has a coefficient 0.2332

These three predictor variables has the highest coefficients (excluding positive and negative sign). So an unit change in these variables, will affect the target variable ‘cnt’ to the maximum extent. Therefore, these three variables will be given utmost importance, while planning to achieve maximum bike booking

General Subjective Questions

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

Linear regression is a form of predictive modelling technique, which tells us the relationship between the dependent (target variable) and independent variables (predictors).

There are two types of linear regression.

- Simple Linear Regression - The most elementary type of regression model is the simple linear regression, which explains the relationship between a dependent variable and one independent variable using a straight line.

- Multiple Linear Regression - Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Linear regression algorithm tries to find the coefficients by minimizing the sum of squared error.

It checks how adding a variable helps in explain the variance in the data in a better way.

It tries to eliminate multicollinearity and overfitting in the data and come up with a significant model with high R- square and adjusted R-square.

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

Anscombe’s quartet is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. In this case, it is important to visualize the data to get a clear picture of what’s going on.

3. What is Pearson’s R? (3 marks)

Pearson’s R measures the linear correlation between two variables X and Y. It has a value between 1 and -1.

- 1 means total positive correlation.
- 0 means no linear correlation
- -1 means total negative correlation

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

When there are many independent variables in a model, a lot of them might be on very different scales, which will lead to a model with very weird coefficients that might be difficult to interpret. So scaling is the process of converting the variables to a comparable scale.
Scaling is performed because of two reasons:

- Ease of interpretation
- Faster convergence for gradient descent methods

Normalized Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

Standardized Scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

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

A high value of VIF happens if there is a perfect correlation between the variables.
Also, if the value of VIF is infinite, then R-squared will be 1. It means that the variance in the data is completely explained by the model.

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

A Q-Q plot or quantile-quantile plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line, which is almost straight.