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)

Observations from the categorical variables:

- Booking count is highest in fall season.
- Booking count in 2019 is higher than 2018.
- Highest booking were made in June and September month
- 97.12 % of bookings were made when there is no holiday.
- Bookings were highest on Thursday and Sunday.
- 68.23 % of bookings were made on working day.
- Maximum bookings were made when the weather is clear.

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

It helps in reducing the extra column created during dummy variable creation and hence it reduces the correlations created among dummy variables.

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

'registered' column has the maximum correlation with the target variable 'count'.

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

I have validate the result by :

1. Getting R2 score of y_validate (y_test) dataset and y_predict.
2. Observing the linear relationship between the y_validate and y_predict by plotting sns.regplot.
3. By observing the distribution plot of difference between the y_validate and y_predict (also known as residual plot).
4. By getting the accurate error by:

$$\mu\left(\sqrt{}(original - predicted)^2\right) \div \mu(orginal) * 100$$

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

- year
- weathersit_light_rain_snow
- season_spring

General Subjective Questions

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

Linear Regression is a statistical model based on supervised learning that establishes the existence of linear relationship between the dependant variable and independent variables. The dependant variable is a continuous number.
It typically based on the linear equation i.e. y = mx + c + e. Where m is the slope and c is the interceptor and e is the random error. It aims at finding the best values od m and c which minimises the squared loss (error) between the actual value and the predicted value. The cost function helps to figure out the best possible values for m (ß1 ….. ßN) and c (ß0), which provides the best fit line for the data points. To update m and c values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random m and c values and then iteratively updating the values, reaching minimum cost.
Multiple Linear Regression is given by: y = ß0 + ß1.x1 + ß2.x2 + …… + ßN.xN + e

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

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set. There are some peculiarities in the dataset that fools the regression model if built.
It was constructed in 1973 by statistician Francis Anscombe to illustrate
the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

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

The Pearson correlation coefficient **(*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- Between 0 and 1 (Positive Correlation): When one value of variable changes, the value of other variable changes in the same direction (Direct Proportional).
- Equal to 0 (No Correlation): There is no relationship between the variables.
- Between -1 and 0 (Negative Correlation): When one value of variable changes, the value of other variable changes in the opposite direction (Inversely Proportional).

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

Scaling changes the range of independent variables that helps in speeding up the calculation of the 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.

- Normalized Scaling (Min Max Scaling): changes the range of dataset between 0 and 1.
  Given by: x = (x – min(x)) / (max(x) – min(x)).
- Standard Scaling replaces the value with their z-score and brings all the data to standard normal distribution which has mean 0 and standard deviation 1.
  Given by: x = (x – mean(x))/(std(x)).

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q Plot stands for Quantile-Quantile Plot. It is graphical method of comparing two probability distributions by plotting their quantiles against each other. Common quantiles have special names, such as quartiles (four groups), deciles (ten groups), and percentiles (100 groups).