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)

A - The final predictors seem to have fairly low correlations. Thus, the final model consists of the 6 variables mentioned above. One can go ahead with this model and use it for predicting count of daily bike rentals

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

A - drop_first=**True is important** to use, as it helps in reducing the extra column created during dummy variable creation. 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)

A- temp

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

A- The final predictors seem to have fairly low correlations. Thus, the final model consists of the 6 variables mentioned above. One can go ahead with this model and use it for predicting count of daily bike rentals

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

A – Temp , holiday and Casual

General Subjective Questions

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

A- Linear Regression is **a supervised machine learning algorithm where the predicted output is continuous and has a constant slope**. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cup and plates).

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

A- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven points.

3. What is Pearson’s R? (3 marks)

A- In statistics, the Pearson correlation coefficient (PCC, pronounced /ˈpɪərsən/) — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is **a measure of linear correlation between two sets of data**

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    - Scaling - Feature scaling is **a method used to normalize the range of independent variables or features of data**. In data processing, it is also known as data normalization and is generally performed during the data pre processing step.
    - 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.
    - In the business world, "normalization" typically means that the range of values are "normalized to be **from 0.0 to 1.0"**. "Standardization" typically means that the range of values are "standardized" to measure how many standard deviations the value is from its mean.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A- If all the independent variables are orthogonal to each other, then VIF = 1.0. **If there is perfect correlation, then VIF = infinity**. A large value of VIF indicates that there is a correlation between the variables.

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

A - Quantile-Quantile (Q-Q) plot, is **a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution**. Also, it helps to determine if two data sets come from populations with a common distribution.

If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.