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?**

Ans: After analysis of the data, we can infer the below points about the categorical variables:

- Fall season has the most bookings
- Most of the bookings are done during the month of May, June, July, August, September and October.
- The bookings are less during holidays
- Bookings on clear weather are more.

The above categorical variables have more effect on the dependent variable. i.e. cnt

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

Ans:

- drop_first = True helps in reducing the extra columns created during dummy variable creation
- if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables

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

Ans: Temp variable has the highest correlation

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans: Linear Regression models are validated based on the 5 assumptions:

- Normality of error terms
- Multicollinearity Check
- Linear Relationship Validation
- Homoscedasticity
- Independence of residuals

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

Ans:

- Year, Workingday and Season from the generated model contribute significantly towards the shared bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

2. Explain the Anscombe's quartet in detail.

Ans:

- Anscombe's Quartet is a famous dataset constructed by Francis Anscombe. It is made of 4 different subsets of data. Each subset has very different characteristics, even though common summary statistics such as mean and variance are identical.

3. What is Pearson's R?

Ans:

- The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:
- Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
- The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Ans:
   - This means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points are, like support vector machines (SVM) or k-nearest neighbors (KNN). With these algorithms, a change of "1" in any numeric feature is given the same importance.

Normalization	Standardization
This method scales the model using min imum and maximum values	This method scales the model using the mean a nd standard deviation.
When features are on various scales, it i s functional	When a variable's mean and standard deviation are both set to 0, it is beneficial
Values on the scale fall between [0, 1] a nd [-1, 1]	Values on a scale are not constrained to a parti cular range

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

Ans:
- If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
- When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**

Ans:

- A QQ plot provides a powerful visual assessment, pinpointing deviations between distributions and identifying the data points responsible for them. When comparing a sample to a probability distribution, you'll typically use this graph with a distribution test, such as a normality test, to verify statistical assumptions.
  Importance of Q-Q plot:
- When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests