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?**

- The frequency distribution or contingency table of each variable can be used to figure out how categorical variables affect the dependent variable. Using the chi-squared test statistic, estimating the coefficients of each variable, and fitting a statistical model, one can estimate the size of an effect. A hypothesis test can be used to determine the effect's significance. The effect is deemed statistically significant if the p-value falls below a threshold for significance (typically 0.05). It is essential to keep in mind that just because a variable has a significant effect doesn't mean it is practically significant.

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

- To avoid the problem of multicollinearity in the data, Drop first = True must be used when creating dummy variables. When creating dummy variables, the entire original variable set is included in the final dataset if drop first = True isn't specified. A direct combination of the other dummy variables can prognosticate one of the real variables, but this may also lead to multicollinearity. The model might overfit as a result of this, among other problems. You can reduce the number of variables and avoid multicollinearity by setting drop first = True, which removes the first dummy variable.

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

- **Registered**

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

- To Validate the assumption of Linear Regression after building the model on a training set, there are few methods: Normality of residual, Outlier detection, and Residual plot analysis.

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

- The three features are instant, season, and yr.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

- The statistical method of linear regression is used to create a linear model of a dependent (target) variable and one or more independent (predictor) variables. Fitting a straight line through the data points is the goal (linear model), which enables us to use the line to predict values for the target variable.

- Finding the line of best fit that reduces the sum of the squared differences (residuals) between the observed and predicted values is the fundamental goal of linear regression. Equation:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn$$

- Here y is a dependent variable, x1,x2,x3…,xn is an independent variable. β0 is the intercept and β1,β2,..,βn are coefficients.

- Ordinary least squares (OLS) or gradient descent are two techniques that can be used to estimate the coefficients. Following the estimation of the coefficients, predictions about the dependent variable for new data points can be made using the linear regression model. Metrics like R-squared, mean squared error, and mean absolute error can be used to assess the accuracy of the predictions.

2. **Explain the Anscombe's quartet in detail.**

- A collection of four datasets known as Anscombe's quartet was produced in 1973 by statistician Francis Anscombe. Although the statistical characteristics of the four datasets are relatively similar, when graphed, they appear extremely different. Anscombe's quartet serves as an example of the value of data visualization prior to statistical analysis.

- A total of 11 (x,y) pairings make up each of the quartet's four datasets. A distinct pattern can be noticed when the data is shown in the first dataset, which exhibits a strong linear connection between x and y. While the association is still evident when the data is plotted, it is less obvious in the second dataset, which likewise has a strong linear relationship. The x and y relationships in the third and fourth datasets are not linear and are not obvious when the data is shown, respectively.

3. **What is Pearson’s R?**

- The Pearson's correlation coefficient, sometimes referred to as Pearson's R, is a metric for the linear relationship between two continuous variables. Its values vary from -1 to 1, with 1 denoting a perfect positive correlation, -1 denoting a perfect negative correlation, and 0 denoting no connection. The strength and direction of a linear relationship between two variables are evaluated using Pearson's R.

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

- In machine learning, scaling is the act of transforming the values of a dataset's feature values such that they fall within a predetermined range, usually between 0 and 1. In order to avoid some machine learning algorithms functioning badly owing to the inclusion of features with wildly divergent ranges, scaling is used to guarantee that features are on a comparable scale and have similar ranges.

- By removing the minimum value of each feature and dividing the result by the range (maximum value minus minimum value), normalized scaling, sometimes referred to as Min-Max scaling, converts the values of the features to a range between 0 and 1.

- By deducting the mean and dividing by the feature's standard deviation, standardized scaling, sometimes referred to as Z-score scaling, modifies the values of the features. Standardized scaling produces a normal distribution with a mean of 0 and a standard deviation of 1.

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

- If there is perfect multicollinearity between two or more independent variables in the model, then the value of VIF can go infinity. When two or more independent variables have a perfect ability to predict one another and are highly correlated with one another, perfect multicollinearity is present. The calculated coefficients' variation in this case grows endlessly enormous, making it impossible to identify the coefficients with absolute certainty. As a result, VIF cannot be computed for these variables, and its value becomes infinite. It is suggested that you take out one of the highly correlated variables from the model and run the study again to address this problem.

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

- A graphical tool used to determine if a collection of data is regularly distributed or not is the Q-Q (Quantile-Quantile) plot. The quantiles of the actual data are shown against the equivalent quantiles of a theoretical distribution in a scatterplot (usually the normal distribution). A Q-Q plot is used to determine how closely observed data adheres to a theoretical distribution.

- The normality of the residuals, which stand for the discrepancies between the actual and projected values, is assessed in linear regression using Q-Q plots. If the residuals are regularly distributed, the linear regression model is suitable for the data, it may be said. If the residuals are not normally distributed, it can be a sign that the linear regression model is inapplicable and that we should look at alternative modeling strategies.

- The use of a Q-Q plot is crucial in linear regression because it may be used to spot residual deviations from normality that could undermine the model's reliability and the reliability of any hypothesis tests that were run on the model's input parameters. The linear regression model's presumptions can be met and the results can be understood by using a Q-Q graphic.