Assignment-based Subjective Questions:

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- fall has the highest average rentals followed by summer.

- 2019 has had a median 2000 increase in rentals compared to 2018.

- September has the highest rentals, followed by the October & August months.

- no significant difference between rentals vs weekdays, except that Wednesday and Saturdays have a higher variation in rentals than others.

- clear weathers has highest rentals folowed by cloudy days

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- we can predict the Kth item using K-1 item. drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.

Q3. 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 target variable('cnt)

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- By ploting the graph of 'Actual vs Predicted No of rentals'
- Compared R-squared score of train & test data set.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- temp
- yr
- workingday

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

Q2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Q3. What is Pearson’s R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

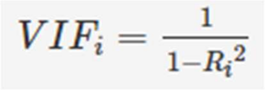
Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is the process of transforming the numerical features of a dataset to a standard range. It is performed to ensure that no variable dominates the others due to differences in their scales. Normalized scaling brings the values within a specific range, typically [0, 1].

Standardized scaling (z-score normalization) transforms the data to have a mean of 0 and a standard deviation of 1, making it suitable for algorithms sensitive to the scale of variables.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The value of VIF is calculated by the below formula:



$$VIF_i = \frac{1}{1-{R_i}^2}$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

he Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q 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 that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.