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?

A).From the summary of the linear regression model, it appears that most of the coefficients for the categorical variables (season, year, month, holiday, weekday, workingday, and weathersit) are close to zero. This suggests that these categorical variables might not have a significant effect on the dependent variable (cnt).

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

A) drop_first=True is used to avoid multicollinearity . This will drop original column before creating other dummy columns

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

A) registered

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

A. Perform residual analysis
B. Check how model is performing on unseen data

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

A)As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (atemp) - A coefficient value of ‘0.443822’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.443822 units.
- Year (yr) - A coefficient value of ‘0.225736’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.225736 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.211022’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.225736 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A. Linear regression is machine learning model. This is supervised learning and used to predicate the continues values. Linear regression is widely used for prediction, inference, and understanding the relationships between variables in various fields such as statistics, economics, finance, and machine learning.
Linear regression models the relationship between the independent variables X and the dependent variable
y using a linear equation: y= b0+b1x1+b2x2+....bnxn
Here objective is to predict the coefficients which minimises the predicted values and actual value

2.Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a famous dataset in statistics and data visualisation that consists of four datasets that have nearly identical statistical properties but vastly different graphical representations..
- Dataset I: This dataset exhibits a linear relationship between x and y, where the points form a fairly straight line.
- Dataset II: Similar to Dataset I, this dataset also shows a linear relationship between x and y, but with an outlier that significantly affects the regression line.
- Dataset III: This dataset contains a non-linear relationship between x and y, where the points form a curved pattern. However, the summary statistics are still similar to the previous datasets.
- Dataset IV: This dataset looks similar to Dataset I, but with one outlier that causes a different regression line from the other datasets.

Anscombe's quartet highlights the limitations of relying solely on summary statistics like means, variances, and correlations. Despite having identical summary statistics, the datasets can have drastically different relationships between variables when visualised. This emphasises the importance of data visualisation in understanding the underlying patterns and relationships in the data, and it serves as a cautionary tale against drawing conclusions based solely on summary statistics without visual inspection of the data.

3. What is Pearson's R?

A) Pearson's correlation coefficient, often denoted as r, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between two variables. Pearson's correlation coefficient can range from -1 to 1, where:

- $r=1$: indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- $r=-1$: indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- $r=0$: indicates no linear relationship between the variables.

Pearson's correlation coefficient measures only the strength of the linear relationship between variables and assumes that the relationship is linear. It does not capture non-linear relationships. Additionally, Pearson's

r is sensitive to outliers, meaning that extreme values can influence its value. Therefore, it's important to consider the context of the data and inspect the scatter plot before interpreting Pearson's correlation coefficient.

4 What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

A . Scaling is method used in ml to make data set standard or normalize.It involves transforming the values of the features so that they all fall within a similar scale. Scaling is typically performed on numerical features, although categorical features can also be scaled in certain cases.

The primary reasons for performing scaling are:

Feature Scaling: Many machine learning algorithms use distance-based calculations to determine the similarity between data points or to optimize the model parameters. If the features have different scales, features with larger scales can dominate the calculations and overshadow the influence of features with smaller scales. Scaling helps to equalize the influence of all features on the model, making it more fair and effective.
Convergence: Some optimization algorithms, like gradient descent, converge faster when features are on similar scales. Scaling ensures that the

optimization process converges more efficiently, leading to faster training times.
Interpretability: Scaling can also improve the interpretability of the model coefficients. When features are on the same scale, it becomes easier to compare their coefficients and understand their relative importance in the model.

There are two common types of scaling techniques:

Normalised Scaling (Min-Max Scaling):

- Normalised scaling rescales the features to a fixed range, usually between 0 and 1.
- It is calculated using the following formula for each feature
- Normalised scaling preserves the relative relationships between the values of the features, maintaining the distribution shape.

Standardised Scaling (Z-score Normalisation):

- Standardised scaling transforms the features so that they have a mean of 0 and a standard deviation of 1.
- It is calculated using the following formula for each feature
- Standardised scaling centres the data around 0 and adjusts the spread of the data to have a standard deviation of 1.
- It is more robust to outliers compared to normalised scaling because it uses the mean and standard deviation.

The key difference between normalised scaling and standardised scaling lies in how they transform the features: normalised scaling rescales the features to a fixed range, while standardised scaling standardises the features to have a mean of 0 and a standard deviation of 1. Both techniques have their advantages and can be used depending on the specific requirements of the problem and the characteristics of the data.

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

- When calculating the Variance Inflation Factor (VIF) for a variable in multiple linear regression, the VIF can become infinite if there is perfect multicollinearity between that variable and one or more other variables in the model.
- Perfect multicollinearity occurs when one or more independent variables in a regression model are perfectly linearly dependent on other independent variables. In other words, one or more independent variables can be expressed as a perfect linear combination of other independent variables.

This means that the coefficient estimates for these variables become unstable or indeterminate, resulting in infinite VIF values.

- When VIF becomes infinite, it indicates that the corresponding variable is perfectly collinear with one or more other variables in the model. In such cases, it is essential to identify and address the multicollinearity issue. This can be done by removing one of the correlated variables, transforming the variables, or using techniques like principal component analysis (PCA) to reduce multicollinearity. Resolving multicollinearity helps in stabilizing the coefficient estimates and improving the interpretability of the regression model.

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

A. Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically the normal distribution. If the points on the Q-Q plot lie approximately along a straight line, it suggests that the data follows the distribution being compared to.

Here's how a Q-Q plot is created:

- Sort the data in ascending order.
- Calculate the quantiles of the sorted data.
- Calculate the quantiles of the theoretical distribution being compared to (e.g., normal distribution).
- Plot the quantiles of the observed data against the quantiles of the theoretical distribution.

If the data points fall approximately along the 45-degree line (the line of equality), it indicates that the data closely follows the distribution being compared to. Deviations from the 45-degree line suggest departures from the assumed distribution.

In linear regression, Q-Q plots are commonly used to assess the normality of residuals, which is one of the assumptions of linear regression. Here's how Q-Q plots are used in linear regression:

Assessing Normality of Residuals: After fitting a linear regression model, the residuals (the differences between observed and predicted values) are

calculated. A Q-Q plot of these residuals is then created to assess whether they are approximately normally distributed. A straight line on the Q-Q plot suggests that the residuals follow a normal distribution, while deviations from the line indicate departures from normality.

Identifying Outliers: Q-Q plots can also help in identifying outliers in the data. Outliers are data points that deviate significantly from the expected pattern. In a Q-Q plot, outliers appear as points that deviate from the expected straight line pattern.

The importance of Q-Q plots in linear regression lies in their ability to visually assess the normality of residuals, which is a crucial assumption for valid inference in regression analysis. Departures from normality can affect the validity of hypothesis tests, confidence intervals, and predictions made by the regression model. Therefore, Q-Q plots provide a useful diagnostic tool for checking the validity of the linear regression model and identifying any issues that need to be addressed.