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)

- I have plotted the categorical variables with the target variables on boxplot and has inferred following effect on target:
    - Season: 3 - fall has the highest demand for rental bikes.
    - Demand for next year has grown.
    - Demand is continuously growing each month till June.
    - September has the highest demand.
    - After September, demand is decreasing.
    - When there is a holiday, demand has decreased.
    - Weekdays are not giving a clear picture about demand.
    - The clear weathersit has highest demand

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

- The number of dummy variables is p-1 where p signifies the categories of a given field/variable. 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. If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with the dataset as we will have a constant variable(intercept) which will create multicollinearity issues.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- The feature “temp” has the highest correlation. It is linearly related with target “cnt”.

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

- I have checked the following assumptions:
    - Error terms are normally distributed with mean 0.
    - Error Terms do not follow any pattern.
    - Multicollinearity check using VIF(s).
    - Linearity Check.
    - Ensured the overfitting by looking at the R2 value and Adjusted R2.

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

- ❖ Features “holiday”, “temp” and season “hum” are highly related to the target column, therefore, these are the top contributing features in model building.

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. It's a simple but powerful algorithm, often used for tasks like predicting numeric values, understanding relationships between variables, and making forecasts. Let's delve into the details of linear regression:

1. Types of Linear Regression:

Simple Linear Regression: In this form, there is only one independent variable that is used to predict a single dependent variable. The relationship is represented as a straight line.
Multiple Linear Regression: In this form, there are multiple independent variables used to predict a single dependent variable. The relationship is represented as a hyperplane in a multi-dimensional space.

2. The Linear Regression Equation:

In simple linear regression, the relationship between the independent variable (X) and the dependent variable (Y) is represented by the equation:
$Y = \beta0 + \beta1 * X + \varepsilon$
Y: The dependent variable (the one we want to predict).
X: The independent variable (the one used for prediction).
β0: The intercept (the value of Y when X is 0).
β1: The slope (the change in Y for a unit change in X).
ε: The error term (representing the difference between the predicted and actual values, accounting for noise and other factors).

3. Training the Model:

The goal is to find the best values for β0 and β1 that minimize the error (ε) between the predicted Y and the actual Y in the training data.
This is typically done using a method called "least squares" that minimizes the sum of the squared differences between predicted and actual values.

4. Evaluation and Prediction:

Once the model is trained, you can use it to make predictions on new, unseen data.

Common evaluation metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2), which help assess the model's performance and accuracy.

5. Assumptions of Linear Regression:

Linear relationship: The relationship between the independent and dependent variables is linear.
Independence: The residuals (error terms) are independent of each other.
Homoscedasticity: The variance of the residuals is constant across all levels of the independent variable.
Normality: The residuals are normally distributed.
No or little multicollinearity: Independent variables should not be highly correlated.

6. Regularization Techniques:

Sometimes, linear regression may overfit or perform poorly due to multicollinearity. In such cases, regularization techniques like Ridge Regression and Lasso Regression can be applied to prevent overfitting and improve model performance.

7. Use Cases:

Linear regression is commonly used in fields such as economics, finance, social sciences, and various other domains where relationships between variables need to be quantified and analyzed.

8. Limitations:

Linear regression assumes a linear relationship, which might not hold for all real-world data.
It can be sensitive to outliers.
Other more complex models may outperform linear regression for highly complex problems.
In practice, linear regression serves as a foundational algorithm in data analysis and machine learning. It's a valuable tool for understanding and predicting relationships between variables when the underlying assumptions are met.

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

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics but appear very different when graphed. This famous example was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and to show that summary statistics alone can be deceptive. Anscombe's quartet is often used to illustrate the concept that data should be graphically explored and not solely summarized through statistical measures.

Here are the details of Anscombe's quartet:

Dataset 1:

x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
y-values: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

Dataset 2:

x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
y-values: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]

Dataset 3:

x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
y-values: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

Dataset 4:

x-values: [8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 19.0, 8.0, 8.0, 8.0]
y-values: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

The key takeaways and the purpose of Anscombe's quartet:

Similar Descriptive Statistics: When you calculate the summary statistics for each of the four datasets, you'll find that they have nearly identical means, variances, correlations, and linear regression coefficients.

Visual Differences: However, when you plot these datasets, they look vastly different. Dataset 1 appears to have a simple linear relationship, Dataset 2 has a non-linear pattern, Dataset 3 has an outlier that strongly affects the linear regression, and Dataset 4 is heavily influenced by a single point, creating a different linear regression line.

Importance of Data Visualization: Anscombe's quartet underscores the importance of data visualization. Relying solely on summary statistics can lead to incorrect conclusions about the data. Visualization helps reveal the underlying patterns, outliers, and relationships in the data.

Critical Thinking: It encourages data analysts and statisticians to critically examine their data through graphical exploration before drawing conclusions or making decisions based on statistical measures alone.

In practice, Anscombe's quartet serves as a reminder that statistics should complement data visualization and not replace it. The quartet highlights the limitations of summary statistics in capturing the richness of real-world data and the potential for different datasets to exhibit distinct patterns despite having similar statistical properties.

3. What is Pearson’s R? (3 marks)

Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the linear relationship between two continuous variables. It assesses the strength and direction of the linear association between two variables, indicating how well they are correlated.

Key characteristics of Pearson's correlation coefficient:

Range: Pearson's r values range from -1 to 1, inclusive.

A value of 1 indicates a perfect positive linear relationship. As one variable increases, the other also increases in a perfectly straight line.
A value of -1 indicates a perfect negative linear relationship. As one variable increases, the other decreases in a perfectly straight line.
A value of 0 indicates no linear relationship. The variables are not correlated in a linear fashion.
Strength: The absolute value of Pearson's r measures the strength of the linear relationship. The closer the absolute value is to 1, the stronger the relationship. A value of 0 indicates no linear relationship.

Direction: The sign (positive or negative) of Pearson's r indicates the direction of the linear relationship:

Positive r: As one variable increases, the other tends to increase. There's a positive linear correlation.
Negative r: As one variable increases, the other tends to decrease. There's a negative linear correlation.
Assumptions: Pearson's correlation coefficient assumes that the relationship between the variables is linear and that the data is roughly normally distributed. It also assumes that there are no outliers that could unduly influence the correlation.

Mathematically, Pearson's correlation coefficient (r) can be calculated as:

$$r = \frac{\sum(Xi - \overline{X})(Yi - \overline{Y})}{\sqrt{\sum(Xi - \overline{X})2 \ \sum(Yi - \overline{Y})2}}$$

Where
Xi and Yi are data points in two variables.
$\overline{X}$ and $\overline{Y}$ are the means of the respective variables.

Pearson's r is widely used in statistics, data analysis, and scientific research to assess the relationships between variables. It is valuable for understanding how variables are associated

and for making predictions based on these relationships. However, it's important to note that Pearson's correlation coefficient only measures linear associations and may not capture more complex or non-linear relationships between variables. For those situations, other correlation measures or modeling techniques might be more appropriate.

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

Scaling is a data preprocessing technique used in machine learning and statistics to transform the features (variables) of a dataset so that they all have the same scale or range. The primary goals of scaling are to ensure that no single feature dominates the learning algorithm, to improve the convergence of certain algorithms, and to make the interpretation of model coefficients or feature importance more meaningful.

Here's why scaling is performed and the key differences between normalized scaling and standardized scaling:

1. Why Scaling is Performed:

Avoid Dominance: In many machine learning algorithms, the magnitude of features can influence the algorithm's performance. Features with larger scales can dominate features with smaller scales, potentially leading to biased or inefficient models.
Algorithm Convergence: Some algorithms, such as gradient-based optimization methods (e.g., gradient descent), converge faster and more reliably when features are on similar scales.
Interpretability: Scaling makes it easier to interpret the importance of each feature in a model, such as coefficients in linear regression or feature importance scores in decision trees.

2. Normalized Scaling (Min-Max Scaling):

In normalized scaling, each feature is scaled to a specific range, usually between 0 and 1. It's also known as Min-Max scaling.
The formula for normalized scaling is: $X_{new} = (X - X_{min})/(X_{max} - X_{min})$
In this approach, the minimum value in the dataset is mapped to 0, and the maximum value is mapped to 1. All other values are linearly scaled in between.
Normalized scaling is sensitive to outliers because it depends on the range of the data.

3. Standardized Scaling (Z-Score Standardization):

In standardized scaling, each feature is transformed to have a mean (average) of 0 and a standard deviation of 1. This is also known as Z-score standardization.
The formula for standardized scaling is: $X_{new} = (X - mean(X))/std(X)$

In this approach, the mean of the data is subtracted from each data point, and then the result is divided by the standard deviation. This centers the data around 0 and scales it based on its spread.
Standardized scaling is less sensitive to outliers compared to normalized scaling. Outliers have less influence on the scaled values because they are centered around the mean.

Key Differences:

Normalized scaling maps data to a specific range (e.g., 0 to 1), while standardized scaling centers the data around 0 with a standard deviation of 1.
Normalized scaling is sensitive to the range of the data and is useful when the data distribution is relatively uniform.
Standardized scaling is less sensitive to outliers and is appropriate when the data distribution is not necessarily uniform and has a more complex shape.
Standardized scaling is commonly used in statistics and machine learning when the distribution of the data is unknown or when the data might contain outliers.
The choice between normalized scaling and standardized scaling depends on the specific requirements of the algorithm and the characteristics of the data. It's important to consider the context and the potential impact on the model's performance.

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

The Variance Inflation Factor (VIF) is a statistical measure used to assess multicollinearity in a multiple regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to issues in interpreting the model and can affect the stability and reliability of coefficient estimates. A high VIF suggests that a variable is highly correlated with other variables in the model.

VIF is calculated for each independent variable as follows:

$$VIF = \frac{1}{(1-R^2)}$$

Where R^2 is the coefficient of determination obtained when the variable of interest is regressed against all other independent variables in the model.
A VIF value of 1 indicates no multicollinearity, while higher values indicate increasing levels of multicollinearity. However, the VIF can sometimes become infinite or extremely large. This happens for a specific independent variable when there is a perfect linear relationship between that variable and other independent variables in the model. In other words, when one variable can be perfectly predicted by a linear combination of the other variables, the R^2 value in the formula becomes equal to 1, leading to an infinite VIF.

The reason VIF becomes infinite in such cases is because it suggests that the variance of the estimated coefficient for the affected variable is also infinite. In practical terms, this means that

the coefficient estimate for the variable is unreliable and cannot be interpreted, as it is perfectly predictable from the other variables in the model.

When you encounter a VIF of infinity, it's a clear indication of severe multicollinearity, and you should take action to address the issue. Common strategies to deal with multicollinearity include:

Variable Removal: You may decide to remove one or more of the highly correlated variables from the model. This decision should be based on domain knowledge and the research question.

Feature Engineering: You can create new variables that are combinations of the highly correlated variables, reducing the multicollinearity.

Regularization: Techniques like Ridge Regression and Lasso Regression can help mitigate multicollinearity by adding penalty terms to the model, which encourages coefficient shrinkage.

Principal Component Analysis (PCA): PCA can be used to reduce multicollinearity by transforming the original variables into a new set of uncorrelated variables (principal components).

Addressing multicollinearity is important for obtaining stable and interpretable regression models. However, it's crucial to carefully evaluate the consequences of any chosen strategy on the overall performance and interpretability of the model.

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 used to assess whether a dataset follows a specific theoretical distribution, typically the normal distribution. It is a useful tool for visually comparing the distribution of observed data to an expected distribution. In the context of linear regression and statistics, Q-Q plots are particularly valuable for examining the assumptions of normality.

Here's how a Q-Q plot works and its use in linear regression:

How Q-Q Plot Works:

A Q-Q plot is constructed by plotting the quantiles of the observed data against the quantiles of the theoretical distribution (usually the normal distribution). The steps to create a Q-Q plot are as follows:

Sorting: First, you sort the observed data in ascending order.

Calculating Quantiles: For each data point, you calculate its quantile or percentile rank within the dataset. This quantile can be expressed as a fraction or a percentage, indicating how the data point compares to other points in the dataset.

Theoretical Quantiles: You then determine the expected quantiles for the corresponding percentiles based on the chosen theoretical distribution (e.g., normal distribution).

Plotting: Finally, you plot the observed quantiles against the expected quantiles. If the data follows the theoretical distribution closely, the points on the Q-Q plot will fall along a straight line (the 45-degree line), indicating a good fit.

Use and Importance in Linear Regression:

Q-Q plots are particularly important in linear regression for the following reasons:

Assumption of Normality: Linear regression assumes that the residuals (the differences between observed and predicted values) follow a normal distribution. Checking this assumption is critical to ensure the validity of regression results.

Identifying Departures from Normality: A Q-Q plot helps you visually assess whether the residuals are normally distributed. If the points on the Q-Q plot deviate significantly from the straight line, it suggests that the residuals do not follow a normal distribution.

Detecting Outliers and Skewness: Q-Q plots can reveal the presence of outliers, skewness, or other departures from normality. These issues can affect the reliability of regression results.

Model Validity: A well-behaved Q-Q plot is a sign that the linear regression model's assumptions are met. If the plot suggests non-normality, further investigation and potentially model adjustments are necessary.

Transformation Selection: If you find that the residuals are not normally distributed, you might consider data transformations (e.g., log transformation) or different modeling techniques to address the non-normality.

In summary, a Q-Q plot is a valuable diagnostic tool in linear regression, helping to assess the assumption of normality for the residuals. Ensuring that the residuals follow a normal distribution is crucial for the validity and reliability of regression analysis. If departures from normality are detected, it may be necessary to take corrective actions to improve the model or the interpretation of results.