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)*

   The following categorical variables were present in the dataset: season (spring, summer, fall, winter), year, month, workingday, weathersit (mist_cloudy, clear_few clouds,Light rain_Light snow_Thunderstorm), weekday and holiday. From the data analysis the seasons winter, the year of analysis, seem to affect the demand positively and surprisingly spring season seem to negatively impact the demand. The demand also decreased in the weather situation of Light rain_Light snow_Thunderstorm. Demand was more in May and September.

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

   Setting drop_first=True during dummy variable creation is important because it helps to prevent multicollinearity in regression models. It removes one of the dummy variables, reducing redundancy and preventing perfect prediction of one category based on the others. This can improve the model's stability and interpretability by avoiding the "dummy variable trap."

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

   Temp and atemp had the highest correlation with cnt

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

   a. The residuals should ideally follow a normal distribution: This was checked by plotting a histogram of the residuals. A bell shaped curve with mean about 0 was obtained.
   b. VIF was checked to avoid multicollinearity
   c. The observed value was plotted against the predict the value to ensure linearity between dependent and independent variables

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

   Year, Temperature and humidity

General Subjective Questions

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

Linear regression is a fundamental statistical method used for modelling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

Linear regression aims to find the best-fitting linear relationship between the dependent variable (denoted as 'y') and one or more independent variables (denoted as 'X').

In simple linear regression, which involves one independent variable, the model equation is represented as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where,

y= is the dependent variable.

x is the independent variable

$\beta_0$is the intercept

$\beta_1$is the slope of the line (the change in y for a unit change in x).

ε represents the error term

The algorithm finds the optimal values for the coefficients $\beta_0 \; and \; \beta_1$ \(that minimize the sum of squared differences between the observed and predicted values. This is often done using the method of least squares, minimizing the residual sum of squares (RSS).

Linear regression assumes several things, such as linearity, homoscedasticity (constant variance of errors), independence of errors, and normally distributed errors.

Multiple Linear Regression: Extending to multiple linear regression involves more than one independent variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Where $x_1, x_2, \ldots x_n$ represent multiple dependent variables

Once the model coefficients are estimated, the model can predict the dependent variable 'y' for new or unseen data by plugging in the values of the independent variables into the regression equation.

Linear regression is widely used due to its simplicity, interpretability, and ease of implementation. However, it's important to validate its assumptions and consider its limitations, such as its sensitivity to outliers and the assumption of a linear relationship between variables.

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

Anscombe's quartet is a collection of four datasets, each consisting of 11 (x, y) pairs. Created by the statistician Francis Anscombe in 1973, these datasets were designed to have nearly identical statistical properties despite looking vastly different when graphed. The quartet was constructed to emphasize the importance of visualizing data and not relying solely on summary statistics.

Each dataset within Anscombe's quartet has the following properties:

1. Similar Summary Statistics: All four datasets have nearly identical summary statistics, including means, variances, correlations, and regression coefficients. This means that when examining basic statistical measures like means, variances, and correlations, these datasets appear very similar.
2. Different Relationships: Despite the similar statistical properties, the relationships between the variables in each dataset vary significantly. For instance, one dataset might follow a linear relationship, another might exhibit a quadratic relationship, and so on.
3. Demonstrates the Importance of Visualization: Anscombe's quartet highlights the importance of visualizing data. By plotting the datasets, one can observe the different patterns, outliers, and structures that summary statistics might fail to capture.

The quartet serves as a cautionary example, reminding analysts not to rely solely on summary statistics when exploring data or fitting models. It underscores the significance of visual inspection and exploration of data to gain a comprehensive understanding of its characteristics and relationships.

3. *What is Pearson's R? (3 marks)*

Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges between -1 and 1, where:

$r = 1$ indicates a perfect positive linear relationship (as one variable increases, the other variable increases proportionally).

$r = -1$ indicates a perfect negative linear relationship (as one variable increases, the other variable decreases proportionally).

$r = 0$ indicates no linear relationship between the variables.

Pearson's r is calculated using the following formula:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Where:

- X and Y are the variables' values.
- $\bar{X}$ and $\bar{Y}$ the mean values of X and Y, respectively.

Pearson's r measures the strength and direction of the linear relationship between two variables. However, it's important to note that it only measures linear associations and may not capture other types of relationships (such as non-linear or complex associations) between variables.

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

Scaling refers to the process of altering the range of variables or features within a dataset. It involves transforming the data so that it fits within a specific scale, which could be for instance between 0 and 1 or centred around a mean with a standard deviation of 1. Scaling is performed to ensure that different variables are comparable and to enhance the performance of certain algorithms or models.

Why Scaling is Performed:

1. Algorithm Sensitivity: Many machine learning algorithms are sensitive to the scale of the variables. Scaling helps in ensuring that no variable dominates or has an undue influence on the model.

2. Distance-Based Methods: Algorithms that rely on distances between data points, like k-means clustering or support vector machines (SVM), can be affected by the scale of features. Scaling ensures that features contribute equally to the computation of distances.

Normalized Scaling vs. Standardized Scaling:

The key difference lies in the range and the statistical properties of the scaled data:

- Normalized scaling adjusts values within a specific range (e.g., 0 to 1), preserving the distribution shape.
- Standardized scaling centres the data around the mean, ensuring that the scaled values have a mean of 0 and a standard deviation of 1.

1. Normalized Scaling (Min-Max Scaling):This technique rescales features to a range between 0 and 1. The formula for min-max scaling is:

$$X_{normalised} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, X is the original value, $X_{min}$is the minimum value of the variable, and X_{max} is the maximum value of the variable.

2. Standardized Scaling (Z-score Standardization): Standardization transforms data to have a mean of 0 and a standard deviation of 1. The formula for standardization is:

$$X_{standardised} = \frac{X - \bar{X}}{\sigma}$$

Here, X is the original value, $\bar{X}$ is the mean of the variable, and σ is the standard deviation of the variable.

Choosing between normalized or standardized scaling depends on the specific requirements of the problem, the nature of the data, and the algorithm being used.

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

The Variance Inflation Factor (VIF) measures the extent of multicollinearity in a regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, leading to issues with interpreting the model coefficients and affecting the model's stability and reliability.

The formula for VIF for a predictor variable X_i is:

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

Where R_i^2 represents the R_2 value obtained by regressing the X_i against all other predictor variables.

When the VIF value is infinite, it typically indicates an extremely high correlation between the predictor variable X_i and other predictors in the model. This high correlation makes the

calculation of R_i^2 approach 1, resulting in the denominator becoming very close to zero (approaching 1 - 1 = 0), hence leading to an infinite VIF value.

An infinite VIF signifies an extremely problematic scenario of multicollinearity, suggesting that one or more predictors can be almost perfectly predicted by a linear combination of the other predictors in the model. This situation makes it challenging to separate the individual effects of the predictors, and it can significantly affect the stability and reliability of the regression coefficients and predictions.

Addressing multicollinearity might involve strategies such as removing one of the highly correlated variables, combining correlated variables, or using regularization techniques to mitigate the impact of multicollinearity in 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 plot or Q-Q plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, like the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution. If the points in the Q-Q plot fall approximately along a straight line, it suggests that the data approximately follows the theoretical distribution being compared.

Use and Importance in Linear Regression:

1. Assumption Checking: In linear regression, it's crucial to check if the residuals (the differences between observed and predicted values) follow a normal distribution. Q-Q plots help in visually assessing whether the residuals are normally distributed. Departures from the straight line in the Q-Q plot may indicate deviations from normality.

2. Identifying Outliers: Q-Q plots can reveal outliers or extreme values in the data. Outliers may appear as points far away from the expected straight line in the plot.

3. Model Validity: Normality of residuals is an assumption in linear regression. If the residuals do not follow a normal distribution, it might affect the reliability of statistical inferences drawn from the model. Q-Q plots assist in validating this assumption.

Interpretation of a Q-Q Plot:

- Straight Line: If the points in the Q-Q plot form a straight line, it indicates that the data closely follows the assumed distribution.

- Deviation from Linearity: Departure from a straight line suggests deviations from the assumed distribution. Curves or patterns in the plot might indicate non-normality or other issues in the data.

- Outliers: Outliers appear as points that significantly deviate from the expected linearity. These points may represent extreme values or errors in the data.