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)
Ans -
Below are the findings from study of categorical variables:
1. Season: Summer, Fall seasons have the highest rides compared to Spring and winter with Spring being significantly lower (less than half of other seasons)
2. Year: There is an increase in rides in 2019 compared to 2018, would be mostly due to increase in market capture
3. month: There is a visible trend across months which is inline with season. There is a chance of high collinearity between these 2 variables. We shoudl also check if teh same monthly trend is followed in both the years
4. Holiday: there is a slight trend across higher rides on non-holidays
5. Weekdays: There is a slight trend across days of the week but not highly significant
6. Weather: Highest rides are on clear sky days (1) followed by slight cloudy (2) and further drops with rains/snow (3)

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans -
Setting drop_first=True during dummy variable creation, especially in scenarios involving categorical variables, is important to avoid multicollinearity issues in regression analysis.
Multicollinearity Avoidance:

- *Including all dummy variables (without dropping one) introduces multicollinearity, where one dummy variable becomes a perfect linear combination of others.*
- *Dropping the first level alleviates this issue by creating linearly independent variables, preserving the necessary degrees of freedom and avoiding perfect multicollinearity.*

Interpretation of Coefficients:

- *Dropping one level does not affect the model's information; instead, it sets a reference category for interpretation.*
- *The dropped category becomes the reference against which the other categories are compared. This reference is inherent when interpreting the coefficients of remaining dummy variables.*

Model Efficiency:

- *By dropping one redundant dummy variable, you reduce redundancy in the model without losing information, thus improving computational efficiency.*

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

Temperature and Registered users (however this is an uncontrollable factor)

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

Assumption 1: Linearity

- *Residuals vs. Fitted Values Plot:*
    - *Check for a random scatter of residuals around zero across different levels of predicted values.*
    - *Patterns or trends in this plot might indicate non-linearity in the model.*

Assumption 2: Independence of Residuals

- *Durbin-Watson Statistic:*
    - *Measures autocorrelation in residuals. A value around 2 suggests no autocorrelation.*
    - *Significant deviation from 2 indicates autocorrelation, violating independence assumptions.*

Assumption 3: Homoscedasticity (Constant Variance)

- *Residuals vs. Fitted Values Plot (Homogeneity of Variance):*
    - *Look for an even spread of residuals across different levels of predicted values.*
    - *Cone-shaped or uneven patterns suggest heteroscedasticity (unequal variance).*

Assumption 4: Normality of Residuals

- *Q-Q (Quantile-Quantile) Plot:*
    - *Check if residuals follow a straight line against the theoretical quantiles of a normal distribution.*
    - *Departure from the straight line indicates deviations from normality.*

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

1. *Temperature,*
2. *Year,*
3. *Weathersit 3 i.e. light snow, rain or thunderstorm (negatively impacts)*

<u>General Subjective Questions</u>

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

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. Here's a step-by-step breakdown of the linear regression algorithm:

Simple Linear Regression (One independent variable):

1. *Data Collection: Gather a dataset consisting of paired observations for the dependent variable (Y) and the independent variable (X).*
2. *Data Preprocessing: Clean the data by handling missing values, outliers, or inconsistencies that might affect the analysis.*
3. *Model Representation: Assume a linear relationship between the independent variable (X) and the dependent variable (Y) as:*

$Y=\beta 0 + \beta 1 * X + \varepsilon$
B0 is the intercept,
B1 is the slope coefficient,
X is the independent variable, and
ε represents the error term.

4. *Fitting the Model: Calculate the coefficients that minimize the sum of squared differences between the actual Y values and the predicted values by using a method like Ordinary Least Squares (OLS).*
5. *Making Predictions: Once the coefficients are determined, use them to predict new values of the dependent variable (Y) based on new or existing values of the independent variable (X).*

Multiple Linear Regression (Multiple independent variables):

1. *Data Collection and Preprocessing: Similar to simple linear regression, collect data with multiple independent variables and preprocess it.*
2. *Model Representation: The model equation extends to accommodate multiple independent variables:*

$Y=\beta 0 + \beta 1 \cdot X1 + \beta 2 \cdot X2 + \ldots + \beta n \cdot Xn + \varepsilon$

3. *Fitting the Model: Use methods like OLS to estimate the coefficients (that minimize the difference between actual and predicted values.*
4. *Making Predictions: Use the obtained coefficients to predict the dependent variable based on new or existing values of the independent variables.*
5. Evaluation:
    a. Coefficient Interpretation: Interpret the coefficients to understand the impact of each independent variable on the dependent variable.
    b. Model Evaluation: Assess the model's goodness of fit using metrics like R-squared, adjusted R-squared, Mean Squared Error (MSE), or Root Mean Squared Error (RMSE) to understand how well the model fits the data.
    c. Assumption Checking: Validate assumptions such as linearity, independence of errors, homoscedasticity, and normality of residuals to ensure the model's reliability.

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

Ans -

Anscombe's quartet is a famous example in statistics that demonstrates the importance of visualization and the limitations of relying solely on summary statistics to understand datasets.

Characteristics:

- *Consists of four distinct datasets, each containing 11 (x, y) pairs.*
- *Each dataset, when analyzed using basic statistical measures like mean, variance, correlation, and regression coefficients, appears very similar or almost identical.*

The Four Datasets:

Dataset 1:

- *Shows a linear relationship between x and y.*
- *Fits perfectly to a linear regression model.*

Dataset 2:

- *Also displays a linear relationship but with one outlier.*
- *The presence of an outlier affects the regression line and correlation significantly.*

Dataset 3:

- *Exhibits a non-linear relationship between x and y.*
- *Fits better to a quadratic model rather than a linear one.*

Dataset 4:

- *Appears to have a strong linear relationship except for one point.*
- *The correlation is heavily influenced by this single outlier.*

Implications:

- *Similar Summary Statistics: Despite the starkly different relationships, these datasets share nearly identical summary statistics, which can mislead if used in isolation.*
- *Graphical Insights: Visualization reveals the nuances in relationships that summary statistics might obscure.*
- *Statistical Rigor: Emphasizes the importance of validating assumptions and exploring data beyond summary measures.*

3. What is Pearson’s R? (3 marks)

Ans -

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 from -1 to +1, where:

- *$r=1$ indicates a perfect positive linear relationship.*
- *$r=-1$ indicates a perfect negative linear relationship.*
- *$r=0$ indicates no linear relationship between the variables.*
- *The Pearson's Correlation Coefficient formula calculates the covariance of X and Y divided by the product of their standard deviations.*

Key Points:

- *Strength of Relationship: The magnitude of r indicates how strong the linear relationship is. Closer to ±1 suggests a stronger linear relationship.*
- *Direction of Relationship: The sign of r (positive or negative) indicates the direction of the relationship. Positive values signify a positive linear relationship, while negative values denote a negative linear relationship.*
- *Assumptions: Pearson's correlation assumes linearity and is sensitive to outliers. It measures only linear relationships and might not capture non-linear associations.*

Interpretation:

- *r=0: No linear relationship between variables.*
- *r close to ±1: Indicates a strong linear relationship. The closer to ±1, the stronger the association.*
- *Positive r : Indicates a positive linear relationship (as one variable increases, the other tends to increase).*
- *Negative r : Indicates a negative linear relationship (as one variable increases, the other tends to decrease).*

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?   (3 marks)
Ans -
Scaling is a preprocessing step in data analysis that involves transforming the values of variables to a standardized range. It's performed to bring different variables onto a similar scale, ensuring fair comparisons and improving the performance of certain algorithms that are sensitive to the scale of variables.

Why Scaling is Performed:

Algorithm Sensitivity: Some machine learning algorithms are sensitive to the scale of variables. For instance, algorithms like k-nearest neighbors (KNN) or support vector machines (SVM) calculate distances between data points, and having variables on different scales can disproportionately influence these distances.
Convergence Speed: Algorithms that use gradient descent (like neural networks or linear regression) converge faster when variables are on a similar scale, preventing one variable from dominating the optimization process.

Normalized Scaling vs. Standardized Scaling:

Normalized Scaling:

- *Also known as Min-Max scaling.*
- *Transforms values to a range between 0 and 1.*
- *Formula: (X−Xmin)/(Xmax−Xmin)*

- *Preserves the original distribution but compresses it into a specific range.*
- *Useful when the distribution and spread of data need to be preserved within a specific range.*

Standardized Scaling:

- *Also called Z-score normalization.*
- *Transforms values to have a mean of 0 and a standard deviation of 1.*
- *Formula: $(X-\mu)/\sigma$*
- *Centers the data around 0 and measures the number of standard deviations away from the mean.*
- *Maintains the shape of the original distribution but in a standardized form.*
- *Useful when algorithms require variables to have similar means and standard deviations.*

Differences:

- *Range: Normalized scaling compresses data into a specific range (0 to 1), while standardized scaling centers data around 0 with a standard deviation of 1.*
- *Preservation of Distribution: Normalized scaling preserves the original distribution, while standardized scaling maintains the distribution but in a standardized form.*
- *Use Cases: Normalized scaling is useful when data needs to be within a specific range, while standardized scaling is beneficial when algorithms require standardized variables.*

Both types of scaling have their utility depending on the context, the nature of the data, and the requirements of the algorithm being used.

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

Ans -

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. It quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity in the independent variables.

Causes of Infinite VIF:

1. *Perfect Collinearity: Infinite VIF occurs when there's perfect collinearity among the variables. This means one variable is an exact linear function of another, leading to a situation where the regression cannot be computed due to perfect multicollinearity.*
2. *Linearly Dependent Variables: When one variable can be expressed as a linear combination of other variables in the model, it causes a situation where the model matrix becomes singular, and VIF becomes infinite.*

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

Ans -

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution, typically a normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, usually the normal distribution. This comparison helps to visually assess whether the dataset deviates from the expected distribution.

Use of Q-Q Plot in Linear Regression:

Normality Checking:

- *In linear regression, the assumption of normality of residuals is crucial. Residuals are the differences between observed and predicted values.*
- *A Q-Q plot of residuals helps verify if they are normally distributed. If residuals follow a straight line in the Q-Q plot, it suggests that they approximately follow a normal distribution.*

Identifying Outliers or Skewness:

- *Q-Q plots can reveal outliers or skewness in data by showing deviations from the theoretical line.*
- *If the points on the Q-Q plot deviate significantly from the diagonal line, it indicates potential outliers or a lack of normality in the data.*

Assumption Validation:

- *Assessing the residuals' normality is crucial because violating the assumption of normality can impact the validity of statistical inferences derived from the regression model.*

How to Interpret a Q-Q Plot in Linear Regression:

- *Perfect Normality: If the points in the Q-Q plot fall approximately along the diagonal line, it indicates that residuals are normally distributed, validating the assumption for linear regression analysis.*
- *Departure from Normality: If the points deviate from the diagonal line:*
    - *Outliers or skewness might be present in the data.*
    - *Curvature or non-linearity in the plot suggests departures from normality.*

Importance:

- *Assumption Checking: Q-Q plots are vital for validating the assumption of normality of residuals, a key assumption in linear regression.*
- *Model Validity: Ensuring that residuals are normally distributed is crucial for accurate parameter estimates and valid statistical inferences from the regression model.*