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?

From the categorical variables from the dataset like Season, Weather, Month, and Year.

- In the year 2019 boom bikes had more business compared to 2018, the reason might post-pandemic.
- In the months of August and September, boom bikes had more business compared to other months, the reason might be the weather.
- In Summer and Winter seasons, boom bikes had more business compared to another season
- Snow and Mist weather had a drop in business, the reason might be people are not comfortable in bad weather situations.

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

To encode categorical data, one hot encoding is done, where a dummy variable is to be created for each discrete categorical variable for a feature. This can be done by using pandas.get_dummies() which will return dummy-coded data. Here we use the parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level. If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

Q3. Looking at the pair plot among the numerical variables, which one has the highest correlation with the target variable?

fTemp (Feel temperature) has the highest positive correlation for the target variable.

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

After building a linear regression model on the training set, several assumptions can be validated to ensure the model's reliability. Here are some common methods for validating assumptions in linear regression:

- Linearity: Check the scatter plot of the predicted values against the actual values to observe if they form a linear pattern. Alternatively, you can use techniques like residual plots or partial regression plots to assess linearity.

- Independence of residuals: Examine the residuals (the differences between predicted and actual values) to detect any patterns or correlations.

- Homoscedasticity: Evaluate the spread or dispersion of the residuals across the range of predicted values.

- Multicollinearity: Assess the presence of high correlation among predictor variables. Calculate the correlation matrix or use techniques like variance inflation factor (VIF) to identify variables that might be collinear.

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

The top three features which have the highest contribution in prediction are Temp, Year, and Season (Winter and Summer).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a popular algorithm used for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the input features and the target variable. Here's a brief explanation of the linear regression algorithm:

- Data Preparation: Start by collecting a dataset consisting of input features (independent variables) and the corresponding target variable (dependent variable). Ensure the data is cleaned and preprocessed, handling missing values, outliers, and normalization if required.
- Model Representation: In linear regression, the relationship between the input features (X) and the target variable (y) is represented by the equation: y = b0 + b1x1 + b2x2 + ... + bn*xn, where b0 is the intercept term and b1, b2, ..., bn are the coefficients associated with each input feature.
- Model Training: The goal is to find the optimal values for the coefficients (b0, b1, ..., bn) that minimize the difference between the predicted values and the actual target values in the training data. This is achieved by minimizing a cost function, typically the sum of squared differences (least squares).
- Estimation: Once the model is trained, the coefficients are determined. You can then use these coefficients to make predictions on new, unseen data. Simply plug in the values of the input features into the equation to obtain the predicted target variable.
- Evaluation: Assess the performance of the linear regression model by evaluating metrics such as mean squared error (MSE), root mean squared error (RMSE), or R-squared (coefficient of determination). These metrics indicate how well the model fits the training data and provide insights into its predictive accuracy.

- Model Interpretation: Since linear regression provides coefficient values for each input feature, you can interpret their magnitudes and signs. Positive coefficients indicate a positive relationship with the target variable, while negative coefficients indicate a negative relationship. The magnitude reflects the strength of the relationship.

2. Explain Anscombe’s quartet in detail.

Anscombe's quartet is a collection of four datasets that have similar statistical properties but exhibit different patterns when visualized. These datasets were created to emphasize the importance of data visualization in understanding and interpreting statistical analyses. Despite having similar summary statistics, the quartet highlights how different data distributions can lead to different conclusions. It serves as a reminder to not solely rely on numerical summaries and to visualize data to gain a more comprehensive understanding.

3. What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient or simply correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges from -1 to 1.

- Range: The correlation coefficient can take values between -1 and 1. A value of -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.
- Strength: The absolute value of the correlation coefficient indicates the strength of the relationship. Values close to -1 or 1 suggest a strong linear relationship, while values close to 0 suggest a weak or no linear relationship.
- Direction: The sign of the correlation coefficient (+ or -) indicates the direction of the relationship. A positive value indicates a positive correlation, meaning that as one variable increases, the other tends to increase as well. A negative value indicates a negative correlation, meaning that as one variable increases, the other tends to decrease.
- Assumptions: Pearson's R assumes that the relationship between the variables is linear, the variables are normally distributed, and there are no outliers or influential data points.
- Interpretation: The correlation coefficient provides a measure of the linear association between two variables but does not imply causation. It is important to consider the context, domain knowledge, and potential confounding factors when interpreting the correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of transforming variables to a specific range or distribution. It is performed to bring all variables to a comparable scale, making them easier to interpret and analyze. Scaling is commonly used in data preprocessing before feeding data into machine learning algorithms.

Scaling is performed for the following reasons:

- Range Adjustment: Variables often have different scales or units. Scaling ensures that all variables have a similar scale, preventing any particular variable from dominating the analysis due to its larger magnitude.
- Model Performance: Many machine learning algorithms are sensitive to the scale of variables. Scaling helps in avoiding biased results and can improve the performance of the models.
- Convergence and Efficiency: Optimization algorithms, such as gradient descent, often converge faster and more efficiently when variables are on a similar scale. Scaling can speed up convergence and improve efficiency.

The two common scaling techniques are normalized scaling and standardized scaling:

- Normalized Scaling (Min-Max Scaling): Transforms variables to a specific range, typically between 0 and 1. It uses the minimum and maximum values of each variable to perform the scaling. Preserves the relative relationships between values but may be sensitive to outliers.
- Standardized Scaling (Z-Score Scaling): Transforms variables to have zero mean and unit standard deviation. It uses the mean and standard deviation of each variable to perform the scaling. Helps in comparing variables on a common scale and is less sensitive to outliers.

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 a linear regression model, an infinite value can occur in certain scenarios. The occurrence of infinite VIF values is typically due to perfect multicollinearity. Multicollinearity refers to a high correlation or linear dependence between predictor variables in a regression model. It can cause issues in the model, such as inflated standard errors, unreliable coefficient estimates, and difficulties in interpreting the individual effects of the correlated variables. VIF is a measure used to quantify multicollinearity. It assesses how much the variance of the estimated regression coefficient for a particular predictor variable is inflated due to multicollinearity. The formula for VIF involves dividing the variance of the estimated coefficient by its expected variance under no multicollinearity. When the

correlation between variables is perfect (r = 1 or -1), it leads to a situation called "perfect multicollinearity."

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

A Q-Q plot, short for the quantile-quantile plot, is a graphical tool used to assess the distributional similarity between a given dataset and a theoretical distribution, typically the normal distribution. It helps in understanding whether the data follows a particular distribution or if there are departures from it. In linear regression, a Q-Q plot is particularly useful for examining the normality assumption of residuals, which is one of the key assumptions in linear regression analysis. The normality assumption states that the residuals (the differences between predicted and actual values) should follow a normal distribution.

Here's the use and importance of a Q-Q plot in linear regression:

Assessing Normality: A Q-Q plot allows you to visually compare the observed residuals against the expected values if they were normally distributed. If the points on the plot align closely along a straight line, it suggests that the residuals approximate a normal distribution, supporting the normality assumption. Deviations from the straight line indicate departures from normality.

Detecting Skewness and Outliers: In a Q-Q plot, departures from the straight line indicate skewness or heavy-tailedness in the distribution of residuals. If the points deviate in the tails, it suggests the presence of outliers or extreme values. These deviations can provide insights into potential issues that might affect the linear regression model.

Model Diagnostics: A Q-Q plot is a useful diagnostic tool for evaluating the goodness-of-fit of a linear regression model. If the residuals violate the normality assumption, it might impact the validity of statistical tests, confidence intervals, and the interpretation of coefficient estimates. Detecting and addressing departures from normality can help ensure the reliability of the linear regression model.