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?

Ans: Based on the analysis of categorical variables in the dataset, we can infer the following about their effect on the dependent variable: -

Season: The season variable shows a significant effect on bike demand. Certain seasons, such as fall and summer, tend to have higher bike usage compared to other seasons. This indicates that there is a seasonal pattern in bike demand, with higher usage during warmer months.

Weather: The weather conditions have an impact on bike demand. Clear and cloudy weather conditions are associated with higher demand, indicating that people are more likely to use bikes when the weather is favorable. On the other hand, adverse weather conditions like rain and snow can lead to lower bike usage.

Holiday: The presence of holidays affects bike demand. It is observed that bike usage decreases during holidays.

Working day: The working day variable also influences bike demand. On working days (non-weekend days), bike usage is higher, possibly due to commuting or work-related travel. Weather, on non-working days (weekends), bike demand tends to decrease.

Month: The month can impact bike demand. It is likely that bike usage is higher during April and November month.

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

Ans: It is very important to use **drop_first=True**. It helps in achieving model efficiency, interpretability, and avoiding multicollinearity issues. It ensures that the regression model does not include redundant information and allows for proper interpretation of the coefficients associated with the dummy variables.

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

Ans: Temperature

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

Ans: After building a Linear Regression model on the training set, we can validate the assumptions of the model by conducting several diagnostic tests and analyses as follows:

1.Checking Linearity of Relationships

2. Multicollinearity

3. Residual Analysis

4. Normality of Residuals

5. Homoscedasticity

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

Ans: Based on the final model, we can say that Temp, Year, Weather clear and Season Summer are top feature variable which contributing significantly.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a supervised learning algorithm used for predicting continuous numeric values based on input features. It establishes a linear relationship between the input variables (also called independent variables, features, or predictors) and the output variable (also known as the dependent variable or target variable). The algorithm aims to find the best-fitting line that minimizes the difference between the predicted values and the actual values.

Here's a detailed explanation of the linear regression algorithm:

1. Data Preparation: First, the input data is collected and organized into a dataset. The dataset consists of a set of observations or instances, where each instance has multiple input features and a corresponding output value.
2. Assumptions: Linear regression assumes that there is a linear relationship between the input features and the output variable. It also assumes that the errors or residuals (the differences between the predicted and actual values) follow a normal distribution, have constant variance (homoscedasticity), and are independent of each other.
3. Model Representation: In linear regression, the relationship between the input features (x) and the output variable (y) is represented by the equation:

$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_r x_r$

where y is the predicted value, β_o is the y-intercept or bias term, β_1, β_2, ..., β_r are the coefficients (also known as weights or slopes) associated with each feature, and x_1, x_2, ..., x_r are the corresponding input feature values.

4. Cost Function: The goal of linear regression is to find the coefficients (β_o, β_1, β_2, ..., β_r) that minimize the difference between the predicted values and the actual values. This is achieved by defining a cost function, typically the mean squared error (MSE), which calculates the average squared difference between the predicted and actual values over the entire dataset.
5. Parameter Estimation: The coefficients are estimated using a method called Ordinary Least Squares (OLS). OLS aims to minimize the cost function by adjusting the coefficients to find the line that best fits the data. The coefficients can be calculated analytically using matrix operations or numerically using optimization algorithms.
6. Model Evaluation: Once the coefficients are estimated, the model's performance is evaluated using various metrics such as the R-squared value, which indicates the proportion of variance in the target variable explained by the model. Other evaluation metrics may include Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to assess the prediction accuracy.
7. Prediction: Once the model is trained and evaluated, it can be used for making predictions on new, unseen data by plugging in the input feature values into the regression equation.

Linear regression is a widely used algorithm due to its simplicity, interpretability, and effectiveness in scenarios where a linear relationship exists between the input features and the output variable. However, it may not be suitable for complex relationships or when the assumptions of linearity and independence are violated. In such cases, other regression techniques or nonlinear models may be more appropriate.

2. Explain the Anscombe’s quartet in detail

Ans: Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but exhibit significantly different patterns when plotted. These datasets were created by the statistician Francis Anscombe in 1973 to highlight the importance of data visualization and the limitations of relying solely on summary statistics.

The four datasets in Anscombe's quartet have the same mean, variance, correlation, and regression line, but they demonstrate the importance of visually exploring data to understand its underlying patterns. Here's a detailed explanation of each

The purpose of Anscombe's quartet is to emphasize the importance of data visualization and the dangers of relying solely on summary statistics. Despite having identical statistical properties, the datasets exhibit different patterns and relationships, highlighting the need to explore and understand the data through visualizations.

Anscombe's quartet serves as a reminder that descriptive statistics and numerical measures may not capture the full complexity of the underlying data and that visualizing the data can provide valuable insights that summary statistics alone cannot reveal.

3. What is Pearson’s R?

Ans: Pearson's R, also known as Pearson correlation coefficient or Pearson's correlation, 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 has a value between -1 and 1.

Pearson's R is used to assess the degree of association or correlation between two variables. The coefficient indicates the extent to which the variables move together in a linear fashion. Here's what different values of Pearson's R signify:

- A value of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable increases proportionally.
- A value of -1 indicates a perfect negative correlation, implying that as one variable increases, the other variable decreases proportionally.
- A value close to 0 suggests no linear correlation or a weak relationship between the variables.

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

Ans: Scaling refers to the process of transforming numerical features in a dataset to a specific range or distribution. It is commonly performed as a preprocessing step in machine learning and data analysis to bring the features to a similar scale and mitigate the impact of differing magnitudes or units.

There are two common types of scaling techniques: normalized scaling and standardized scaling.

1. Normalized Scaling: Also known as min-max scaling, this technique transforms the features to a specific range, typically between 0 and 1. It is achieved by subtracting the minimum value of the feature and dividing it by the range (maximum value minus minimum value). Normalized scaling preserves the relative relationships and distribution of the data but compresses it within the specified range.
2. Standardized Scaling: Standardization, also called z-score normalization, transforms the features to have a mean of 0 and a standard deviation of 1. It is achieved by subtracting the mean of the feature and dividing it by the standard deviation. Standardized scaling preserves the shape of the original distribution, but the resulting values are expressed in terms of standard deviations from the mean.

The main difference between normalized scaling and standardized scaling is the range of the transformed values. Normalized scaling constrains the values within a specified range (e.g., 0 to 1), while standardized scaling centers the values around 0 with a standard deviation of 1.

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

Ans: The occurrence of infinite values in the Variance Inflation Factor (VIF) is typically due to perfect multicollinearity among the independent variables in a regression model. Perfect multicollinearity refers to a situation where one or more independent variables can be perfectly predicted from a linear combination of other independent variables.

When perfect multicollinearity exists, the VIF calculation breaks down because it involves dividing by zero. Mathematically, the VIF for a particular independent variable is calculated as the ratio of the variance of that variable's coefficient estimate to the variance of the regression coefficient estimate when that variable is regressed on the other independent variables.

The formula for VIF is: $VIF = 1 / (1 - R^2)$

In the presence of perfect multicollinearity, the R^2 value for the affected variable becomes 1, resulting in the denominator becoming zero. As a result, the VIF value becomes infinite.

Perfect multicollinearity can occur due to various reasons, such as including redundant variables in the regression model, linear dependencies among the variables, or data manipulation issues. It is important to identify and handle multicollinearity in regression models because it can lead to unreliable coefficient estimates, unstable predictions, and difficulties in interpreting the relationships between variables.

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

Ans: A Q-Q (quantile-quantile) plot, also known as a quantile plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution. It compares the quantiles of the observed data against the quantiles of a specified reference distribution, typically the normal distribution. The Q-Q plot allows us to visually inspect the goodness-of-fit between the observed data and the theoretical distribution.

In linear regression, a Q-Q plot is often employed to check the assumption of normality for the residuals (the differences between the observed values and the predicted values). The Q-Q plot of residuals helps us assess if the residuals are normally distributed, which is a fundamental assumption of linear regression models. Here's how the Q-Q plot is used in linear regression:

1. Plotting the Q-Q plot: To create a Q-Q plot, we arrange the observed residuals in ascending order and calculate their corresponding quantiles. Then, we compare these quantiles against the quantiles of a normal distribution. The observed residuals are plotted on the y-axis, and the expected quantiles from the normal distribution are plotted on the x-axis.
2. Assessing normality: In a Q-Q plot, if the observed residuals closely follow the diagonal line (the line of equality), it suggests that the residuals are approximately normally distributed. On the other hand, deviations from the diagonal line indicate departures from normality.

3. Interpretation: In linear regression, a well-behaved Q-Q plot with residuals closely following the diagonal line suggests that the residuals are normally distributed. This confirms the assumption of normality, which is important for valid inference, hypothesis testing, and confidence intervals associated with the regression model. Deviations from the diagonal line may indicate issues such as skewness, heavy tails, or outliers in the residuals.

The Q-Q plot is an important diagnostic tool in linear regression to assess the assumption of normality for residuals. By visually comparing the observed residuals to the quantiles of a normal distribution, we can evaluate the validity of this assumption and make informed decisions about the regression model.