Subjective Questions

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, we can infer that the categorical variables holiday, working day, month, and year may have an impact on the dependent variable. The variations in mean and quartile values for these variables say that potential seasonal, daily, and annual trends that are influencing the dependent variable.

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

Ans) Using drop_first=True during dummy variable creation is important to avoid multicollinearity in the dataset. When creating dummy variables from categorical variables, drop_first=True ensures that one of the dummy variables is dropped to prevent perfect multicollinearity, where one dummy variable can be perfectly predicted from the others. This helps in improving the accuracy in regression analysis, by preventing redundancy and improving the interpretability of the model coefficients.

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

Ans) The variable temperature has the highest correlation with the target variable. As indicated by the linear pattern in the scatter plot and histogram. The scatter plot of temperature and cnt shows linear relationship, showing high correlation.

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

Ans) After building the model, the assumptions of Linear Regression were validated by assessing the model's performance on the test set. The R-squared score on the test set was calculated to be 0.8387714141458165, indicating the proportion of the variance in the dependent variable that is predictable from the independent variable. This validation process helps to ensure that the model's assumptions, such as linearity, independence of errors, and normality of residuals, are met and that the model's predictive capability is reliable. Additionally, other diagnostic tests, such as residual analysis and checking for multicollinearity among the independent variables, have been performed to further validate the assumptions of the Linear Regression model.

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

Ans) The top three features contributing significantly towards explaining the demand for shared bikes, based on the final model, are "temp," "atemp," and "hum." These features have been identified as key contributors to the variation in the count of total rental bikes. The "temp" and "atemp" variables represent temperature and feeling temperature, while "hum" represents humidity. The mean values for these variables are approximately 20.32, 23.73, and 62.77, respectively, and they have standard deviations of 7.51, 8.15, and 14.24, indicating their significant impact on bike rental demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans) Linear regression is a supervised learning algorithm used in machine learning and statistics to predict a target variable, also known as a dependent variable, based on one or more predictor variables, also known as independent variables. It establishes a relationship between the dependent and independent variables by fitting the best line. This line of best fit is known as the regression line and represented by a linear equation:

Y=a*X+b

- Y is the dependent variable.
- a is the slope of the line.
- X is the independent variable.
- b is the y-intercept.

Here are the key steps involved in the linear regression algorithm:

1. Data Collection: The process begins with collecting data which includes a dependent variable and one or more independent variables.
2. Model Building: The algorithm then constructs a linear relationship model between the dependent and independent variables. The relationship is represented by the equation of a straight line.
3. Best Fit Line: The algorithm calculates the best fit line or the regression line. This line will be such that the vertical distances from the data points to the regression line would be minimal.
4. Prediction: Once the regression line is formed, it can be used to predict the values of the dependent variable for any given value of the independent variable.

There are two types of linear regression:

- Simple Linear Regression: When there is a single independent variable, the method is referred to as simple linear regression.
- Multiple Linear Regression: When there are more than one independent variables, the method is known as multiple linear regression.

Linear regression is widely used in machine learning and has numerous applications.

2) Explain the Anscombe's quartet in detail.

Ans) Anscombe's Quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. Each dataset consists of 11 x-y pairs. Despite having nearly identical descriptive statistical properties, including means, variance, correlations, and linear regression lines, they appear very different when visualized on a scatter plot.

Purpose of Anscombe's Quartet

The primary purpose of Anscombe's Quartet is to emphasize the importance of exploratory data analysis and the potential pitfalls of relying solely on summary statistics. It demonstrates that datasets with identical or similar statistical properties can indeed have very different distributions and relationships. This underscores the importance of visualizing data to spot trends, outliers, and other crucial details that might not be apparent from summary statistics alone.

Details of the Datasets

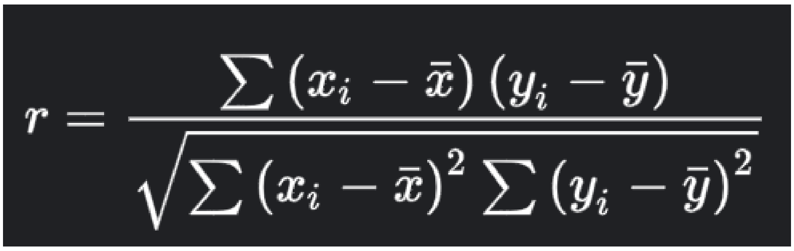
1. Dataset 1: Fits the linear regression model well.
2. Dataset 2: Cannot fit the linear regression model because the data is non-linear.
3. Dataset 3: Shows the outliers involved in the dataset, which cannot be handled by the linear regression model.
4. Dataset 4: Also shows the outliers involved in the dataset, which cannot be handled by the linear regression model.

   In conclusion, Anscombe's Quartet serves as a powerful reminder of the importance of data visualization and the potential misleading nature of summary statistics. It highlights the need for a comprehensive understanding of data, which includes both statistical analysis and visual examination.

3) What is Pearson's R?

Ans) Pearson's R, also known as the Pearson Correlation Coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is a number between -1 and 11:

- A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases.
- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases.
- A value of 0 indicates no linear relationship between the variables.



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

- x and y are the variables,
- n is the sample size, and
- ΣΣ represents a summation of all values.

It’s important to note that Pearson’s R measures the linear correlation between two variables, and it does not imply causation.

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

Ans) Scaling is a data preprocessing step applied to independent variables to normalize the data within a particular range. It is performed to handle highly varying magnitudes, units, and ranges in the data. If scaling is not done, an algorithm might only take magnitude into account, leading to incorrect modeling. Scaling helps in speeding up the calculations in an algorithm and makes machine learning algorithms train and converge faster.
There are two common types of scaling: Normalized Scaling and Standardized Scaling:

1. Normalized Scaling (Min-Max Scaling): This method transforms features to be on a similar scale, usually between 0 and 11. Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data4. However, it is sensitive to outliers.
2. Standardized Scaling (Z-Score Normalization): This method transforms features by subtracting the mean and dividing by the standard deviation. where μ is the mean and σ is the standard deviation. Standardization can be helpful in cases where the data follows a Gaussian distribution. It is much less affected by outliers.

In summary, the choice between normalization and standardization depends on the specific requirements of your data and the machine learning algorithm you are using.

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans) The Variance Inflation Factor (VIF) is a measure of multicollinearity in a set of multiple regression variables. A VIF value of infinity typically indicates perfect multicollinearity, meaning one independent variable can be expressed as a linear combination of other variables.

In mathematical terms, when the R-squared (R^2) value in a regression model is 1, indicating a perfect correlation between two independent variables, the VIF becomes infinite. This is because the formula for VIF is

VIF= 1/1- R^2

So, when R^2 equals 1, the denominator becomes 0, leading to an infinite VIF. In such cases, to resolve this issue, one of the variables causing this perfect multicollinearity is usually dropped from the dataset.

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans) A **Quantile-Quantile (Q-Q) plot** is a graphical tool used to assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential, or Uniform distribution. It is a plot of the quantiles of two distributions against each other. If the points in a Q-Q plot lie on a straight diagonal line, it suggests that the two datasets come from the same distribution.
In the context of **linear regression**, a Q-Q plot is particularly useful for checking the normality of residuals. The normality of residuals is an important assumption in linear regression. If the residuals are normally distributed, it validates the model. If they are not, it indicates that the model may need to be re-specified.
The Q-Q plot can also help determine if two data sets come from populations with the same distributions. This is useful when we have training and test data sets received separately, and we want to confirm that both data sets are from populations with the same distributions.
The Q-Q plot can detect many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers. Therefore, it is a powerful tool for visually checking the assumption of normality and identifying any potential issues with the data or the model.