Linear Regression: Bike Sharing Assignment

Subjective questions:

Submitted by: Shamseena VM

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?

Answer:

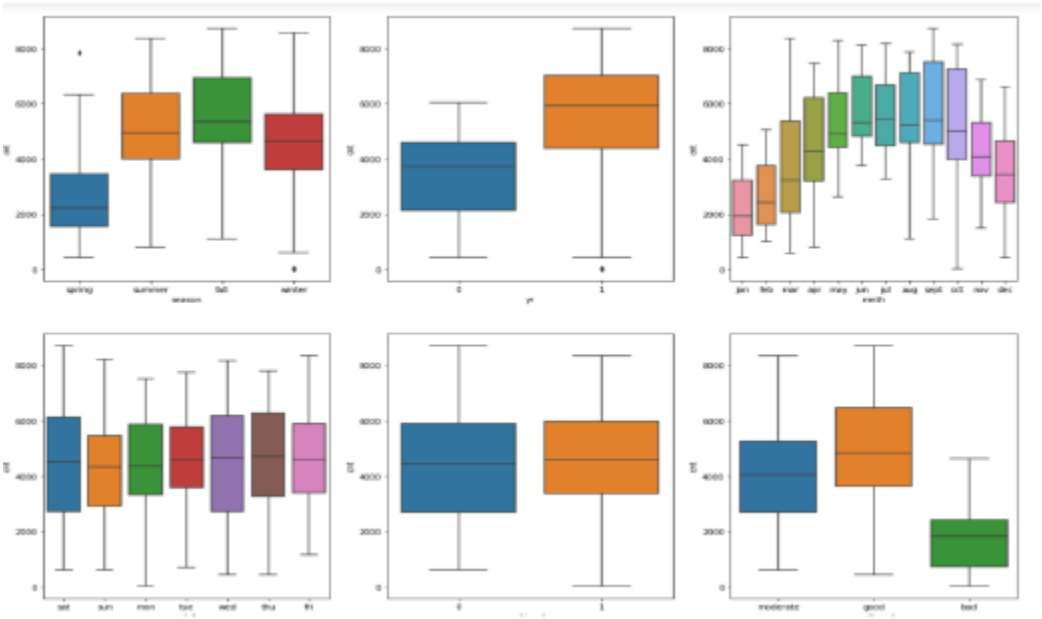
1.The effect of temperature has high significance with bike rentals.

2.The demand of bike rental has increased from 2018 to 2019.

3.Demand of bike rentals in Fall, Summer and Winter is higher.

4.Demand of bike rentals on holidays are less.

4.Bad weather conditions will affect the bike booking



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

Answer:

Dummy variables are used in model building

1.To simplify the model

2.To assign normalized values for the column values and to provide numerical stability.

When dummy variable encoding is done, it will provide n-1 values.

For eg: if 3 values are there , after dummy variable encoding it will provide n-l values.

‘drop_first’=True helps to ensure that the model is reliable and interpretable to machine learning.

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

The ‘temp’ and ‘atemp’ variable has got highest correlation to the target variable cnt.

They are highly correlated with each other and hence will produce multicollinearity. So during model building, ‘atemp’ column will be removed.

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

Answer:

The validation on the model is done using the below:

1.Linearity

2.Multicollinearity

2.Residual analysis

3. Homoscedasticity

4.Using pvalue and coefficients

5. Using R-squared, adjusted R squared values

6.Using VIF

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

Answer:

Top 3 significant features are:

1.temperature

2.year

3.season

General Subjective Questions

1. Explain the linear regression algorithm in detail ?

Answer:

Linear regression is a fundamental and widely used supervised learning algorithm in machine learning and statistics. It is used for predicting a continuous target variable (also known as the dependent variable) based on one or more predictor variables (independent variables). The basic idea behind linear regression is to model the relationship between the predictors and the target variable as a linear equation. There are two primary types of linear regression: simple linear regression and multiple linear regression.

Simple Linear Regression:

n simple linear regression, there is only one predictor variable (X) that is used to predict a single target variable (Y). The relationship between the predictor and the target is modeled as a straight line:

$Y=\beta 0+\beta 1X+\varepsilon$

Y is the target variable.

- X is the predictor variable.
- $\beta 0$ is the y-intercept, representing the value of Y when X is 0.
- $\beta 1$ is the slope, representing the change in Y for a one-unit change in X.
- ε represents the error term.

The goal of simple linear regression is to estimate the values of $\beta 0$ and $\beta 1$ that minimize the sum of squared errors (residuals) between the observed Y and the predicted Y values.

Multiple Linear Regression:

In multiple linear regression, there are multiple predictor variables (X1, X2, X3, ..., Xn) used to predict a single target variable (Y). The relationship is modeled as a linear equation with multiple predictor variables:

$Y=\beta 0+\beta 1X1+\beta 2X2+\beta 3X3+\ldots+\beta nXn+\varepsilon$

- Y is the target variable.
- $X1, X2, \ldots, Xn$ are the predictor variables.
- $\beta 0$ is the y-intercept.

- $\beta 1, \beta 2, \ldots, \beta n$ are the slopes associated with each predictor variable.
- ε represents the error term.

2.Explain the Anscombe's quartet in detail?

Answer:

Anscombe's Quartet is a set of four small datasets that have nearly identical simple descriptive statistics (such as mean, variance, correlation, and regression coefficients) but display very different characteristics when graphically visualized. The quartet was created by the British statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and the limitations of relying solely on summary statistics.

The key idea behind Anscombe's Quartet is to illustrate that datasets with the same statistical properties can exhibit dramatically different patterns when plotted.

3. What is Pearson's R?

Answer:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

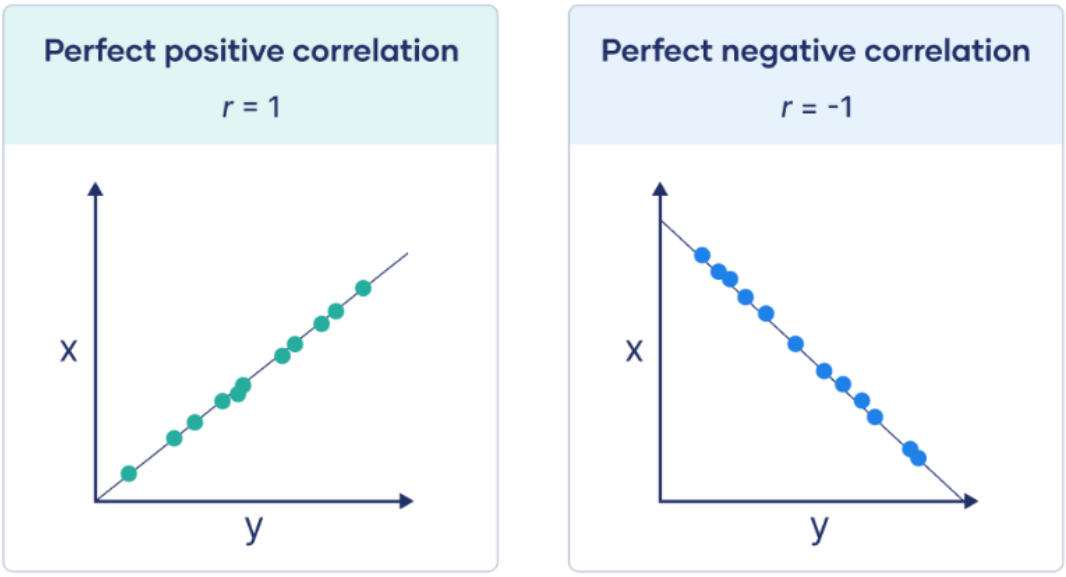
The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Visualizing the Pearson correlation coefficient

Another way to think of the Pearson correlation coefficient (*r*) is as a measure of how close the observations are to a line of best fit.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.

When *r* is 1 or –1, all the points fall exactly on the line of best fit:



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

Answer:

Scaling in the context of data preprocessing refers to the process of transforming data in a way that all features (or variables) have similar scales or ranges. The primary goal of scaling is to ensure that no single feature dominates the learning algorithm due to its larger magnitude, potentially causing the algorithm to be biased towards that feature.

Scaling is performed for several reasons:

There are two common methods for scaling data: **normalized scaling** and **standardized scaling**. Here's how they differ:

Normalized Scaling (Min-Max Scaling):

In **normalized scaling**, each feature is scaled to a specified range, typically [0, 1].

Standardized Scaling (Z-score Scaling):

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

Answer:

A Variance Inflation Factor (VIF) can become infinite for a predictor variable in the context of multiple linear regression when there is perfect multicollinearity between that predictor and other predictor variables. Perfect multicollinearity means that one predictor variable is an exact linear combination of other predictor variables.

Perfect multicollinearity can occur for several reasons:

Answer:

Duplication of Variables: When two or more predictor variables are identical or represent the same information. For example,'temp' and 'atemp' in bike sharing dataset.

Linear Dependency occurs When one predictor variable is a perfect linear combination of others. For example, if you have two variables, dependencies:

In our dataset, 'casual' and 'registered' have linear dependency with 'cnt' column since 'cnt' is the sum of casual and registered, it will have perfect multicollinearity

The presence of perfect multicollinearity can cause several issues in multiple linear regression:

Infinite VIF:

Unstable Coefficient Estimates:

To address the issue of perfect multicollinearity, you should take steps such as:

Dropping Redundant Variables: If you identify predictor variables that exhibit perfect multicollinearity, you can choose to drop one of them. The choice of which variable to drop should be based on your understanding of the problem and domain knowledge.

1. Variable Transformation: If possible, transform or combine variables to eliminate multicollinearity. For example, you might transform variables to orthogonal basis vectors to eliminate linear dependencies.
2. Regularization: Techniques like ridge regression or lasso regression can help mitigate the effects of multicollinearity by adding a penalty term to the coefficients, encouraging the model to reduce their magnitude.
3. Feature Selection: Consider using feature selection methods to automatically identify and remove redundant features.

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

A Q-Q plot, which stands for "Quantile-Quantile" plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It's a way to visually compare the quantiles of your data to the quantiles of the specified theoretical distribution. The Q-Q plot is particularly useful in linear regression and other statistical modeling to check the assumption of normality for the residuals.

Here's how a Q-Q plot works and its importance in linear regression:

How Q-Q Plot Works:

1. Data Sorting: Start by sorting the data in ascending order.
2. Theoretical Quantiles: Calculate the theoretical quantiles for the chosen distribution. For example, if you want to check for normality, you calculate the quantiles expected in a standard normal distribution.
3. Plotting: Plot the theoretical quantiles on the x-axis and the actual quantiles of your data on the y-axis. If the data follows the chosen distribution, the points on the Q-Q plot should form a roughly straight line.

Importance in Linear Regression:

1. Assumption of Normality: One of the key assumptions in linear regression is that the residuals (the differences between observed and predicted values) are normally distributed. Deviations from this assumption can affect the validity of statistical inference, confidence intervals, and hypothesis tests related to the model coefficients.
2. Residuals Normality Check: After fitting a linear regression model, you can create a Q-Q plot for the residuals to assess whether they follow a normal distribution. If the Q-Q plot deviates from a straight line, it suggests that the normality assumption may not hold, and you should investigate further.
3. Detecting Skewness and Outliers: Q-Q plots are useful for detecting deviations from normality, such as skewness (non-symmetry) and outliers. Skewness can cause the points in the Q-Q plot to bend, indicating a non-normal distribution. Outliers may also appear as points far from the expected line.
4. Model Diagnostics: Q-Q plots are an essential tool for diagnosing the adequacy of your linear regression model. If the residuals significantly deviate from the expected straight line in the Q-Q plot, it may suggest that **your model**

assumptions need reevaluation, and you might consider transformations or other adjustments to improve model performance.

5. Robustness of Inference: Ensuring that the residuals are normally distributed is essential for the validity of confidence intervals and hypothesis tests related to the regression coefficients. Violations of this assumption can lead to incorrect inferences.