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 :

- The year box plots indicates that more bikes are rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during september month.
- The weekday box plots indicates that more bikes are rent during saturday.
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

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

Answer :

is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer :

By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

Answer :

Considering the VIFs and p-values both are within an acceptable range. So we go ahead and make our predictions using this model only.

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

Answer :

I think the below 3 feature are contributing significantly,

1. weathersit_Light_Snow(negative correlation).
2. yr_2019(Positive correlation).
3. temp(Positive correlation).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer :

Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable (often denoted as y) and one or more independent variables (often denoted as x). The basic idea behind linear regression is to find the best-fitting straight line that represents the relationship between the variables.

two types of linear regression are getting used mostly:

● Simple linear regression

● Multiple linear regression

2. Explain the Anscombe’s quartet in detail.

Answer :

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, including means, variances, correlations, and linear regression lines. However, when visualized, each dataset appears significantly different. This quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphical data exploration and the potential pitfalls of relying solely on summary statistics.

Here's a detailed explanation of Anscombe's quartet:

The Datasets:

Anscombe's quartet consists of four datasets, each containing 11 (x, y) pairs:

Dataset I: A simple linear relationship with some random noise.

Dataset II: A non-linear relationship between x and y.

Dataset III: A strong linear relationship with one outlier.

Dataset IV: A dataset where all x values are the same except for one.

Statistical Properties:

Despite the visual differences among the datasets, they share nearly identical statistical properties:

- For each dataset, the mean and variance of both x and y values are very close.
- The correlation coefficient between x and y is approximately 0.816 for all datasets.
- The linear regression line for each dataset has the same slope and intercept.

Visual Representation:

When plotted, the datasets reveal their visual differences:

- Dataset I: Shows a clear linear relationship between x and y.
- Dataset II: Exhibits a non-linear relationship, where a simple linear regression line would not be appropriate.
- Dataset III: Appears to have a strong linear relationship, but with one outlier that significantly affects the regression line.
- Dataset IV: Demonstrates the importance of checking for influential points, as one outlier drastically changes the regression line.

Implications:

Anscombe's quartet highlights several important points:

- Summary statistics alone may not provide a complete understanding of the data.
- Visualizing the data can reveal patterns, outliers, and relationships that summary statistics might obscure.
- Relying solely on linear regression or correlation coefficients without assessing the data graphically can lead to incorrect conclusions about the underlying relationships.

Educational and Analytical Tool:

Anscombe's quartet is often used as a teaching tool in statistics and data analysis courses to emphasize the importance of graphical exploration and to caution against over-reliance on summary statistics. It underscores the need for researchers to thoroughly examine their data using visualization techniques before drawing conclusions or making decisions based on statistical analyses alone.

In summary, Anscombe's quartet is a powerful illustration of how datasets with similar statistical properties can exhibit vastly different visual patterns, highlighting the importance of graphical exploration and the limitations of summary statistics in data analysis.

3. What is Pearson's R?

Answer :

Pearson's r (often referred to as the Pearson correlation coefficient or Pearson's r) is a measure of the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1

Pearson's r provides valuable insights into the relationship between variables, but it assumes that the relationship is linear and that both variables are normally distributed. Additionally, Pearson's r measures only the strength and direction of the linear relationship and may not capture other types of relationships, such as non-linear or monotonic relationships. Therefore, it's essential to interpret Pearson's r in conjunction with visualizations and consider the context of the data when assessing relationships between variables.

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

Answer :

The presence of infinite VIF values indicates severe multicollinearity in the regression model, making it difficult to estimate the regression coefficients accurately. In such cases, it is essential to address multicollinearity by either removing one of the highly correlated variables, combining them into a single variable, or using techniques like ridge regression or principal component analysis (PCA) to mitigate multicollinearity effects.

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

Answer :

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a set of data follows a specific probability distribution, such as the normal distribution. It compares the quantiles of the data distribution to the quantiles of a theoretical distribution (e.g., normal distribution). If the points in the Q-Q plot lie approximately along a straight line, it indicates that the data distribution is similar to the theoretical distribution being tested.