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:

Categorical variables like yr, season, mnth and weathersit are significant with dependent variable.

More details:

yr category

As per the box plot between yr and cnt,

2019 has good amount cnt values, means total rental bikes demand is high in 2019.

As per the model having good R-squared value,

yr has positive coefficient significance.

season category

As per the box plot between season and cnt,

Summer and fall seasons are having good cnt values, means total rental bikes demand is high in summer and fall seasons.

As per the model having good R-squared value,

winter season has positive coefficient significance whereas spring has negative coefficient significance.

mnth category

As per the box plot between mnth and cnt,

Sept month has high cnt value, means total rental bikes demand is high in September month compare to other months.

As per the model having good R-squared value,

Sept month has positive coefficient significance whereas July as well as November has negative coefficient significance.

weathersit category

As per the box plot between wethersit and cnt,

Clear, Few clouds, Partly cloudy, Partly cloudy has high cnt values.

As per the model having good R-squared value,

both below types has negative coefficient significance

Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

Other categories

As per the box plot between each below category and cnt,

Holiday , weekday and workingday not showing any meaningful significance to the dependent variable cnt.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
Answer:

When creating dummy variables from categorical variables in linear regression, it's important to use the drop_first=True parameter to avoid multicollinearity issues and improve model interpretability.

If one of the dummy variable is not dropped, it will lead to multicollinearity.
Means, one dummy variable can be predicted from the others.
SO, itss important to use drop_first = true.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer:
both temp and atemp are shown same 0.65 as the correlation value.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer:
Linear relationship between X and y:
scatterplots of each X against the y. Each scatterplot has shown linear pattern, indicating a linear relationship between that predictor and the target variable.

Normal distribution of error terms:
Plotting a dist plot of the residuals and visually inspecting for symmetry around zero could indicate normal distribution.

Independence:
VIF of each independent variable explained all the variables are independent.

Constant variance of error terms:
With the help of Residual plot, Its evident that variance in the residuals are constant against change in predicted variable values..

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer:
temp (temperature): 0.47
yr (year): 0.23
weathersit_3: -0.3

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Answer:
Linear regression is a fundamental supervised learning algorithm used for predictive analysis. It's a statistical method used to model the relationship between one or more predictor variables (independent variables) and a target variable (dependent variable) by fitting a linear equation to

observed data. The goal of linear regression is to find the best-fitting straight line (or hyperplane in higher dimensions) that minimizes the difference between the observed and predicted values of the target variable.
Here's a detailed explanation of the linear regression algorithm:
Assumptions:

- Linear relationship: Assumes that there is a linear relationship between the predictor variables and the target variable.
- Independence: Assumes that the observations are independent of each other.
- Homoscedasticity: Assumes that the variance of the errors (residuals) is constant across all levels of the predictor variables.
- Normality: Assumes that the errors follow a normal distribution.

Model Representation:
The linear regression model is represented by the equation:
$y=\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn+\epsilon$

Model Training:

- To train the linear regression model, we use a dataset with *m* observations (rows) and *n* predictor variables (columns).The model learns the coefficients $\beta 0,\beta 1,...,\beta n$ that minimize the sum of squared differences between the observed and predicted values of the target variable. This is typically done using the method of least squares or optimization algorithms like gradient descent.

Model Evaluation:

- Once the model is trained, it's important to evaluate its performance on unseen data. Common evaluation metrics for linear regression include mean squared error (MSE), root mean squared error (RMSE), R- Squared score, and adjusted *R- Squared* score.
- Additionally, diagnostic plots such as residual plots, Q-Q plots, and scatterplots of observed vs. predicted values can be used to assess the model's assumptions and identify any issues.

Prediction:

- Once the model is trained and evaluated, it can be used to make predictions on new or unseen data by simply plugging in the values of the predictor variables into the linear regression equation.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer:

Anscombe's quartet is a famous dataset consisting of four sets of x and y values that have nearly identical descriptive statistics but vastly different graphical representations. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before analyzing it and to highlight the limitations of relying solely on summary statistics.

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

The Dataset:

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

Despite having different (x, y) values, each dataset has nearly identical descriptive statistics, including mean, variance, correlation coefficient, and linear regression parameters.

Graphical Representations:

When plotted, each dataset in Anscombe's quartet produces a different graphical pattern.

Dataset I: Exhibits a linear relationship with some scatter around the line.

Dataset II: Demonstrates a non-linear relationship with a clear quadratic pattern.

Dataset III: Appears to have a linear relationship with one outlier point significantly deviating from the rest.

Dataset IV: Appears to have a linear relationship until one point substantially influences the regression line.

Importance of Visualizing Data:

Anscombe's quartet emphasizes the importance of visualizing data before drawing conclusions or performing analyses.

While summary statistics can provide valuable insights into the central tendency and variability of data, they may not reveal the full complexity or underlying patterns present in the data.

By visually inspecting the data, analysts can identify outliers, non-linear relationships, and other patterns that may influence statistical analyses or model interpretations.

Statistical Lessons:

Despite having identical summary statistics, the datasets in Anscombe's quartet have distinct graphical representations, highlighting the limitations of summary statistics in characterizing datasets accurately.

The quartet illustrates the dangers of blindly applying statistical methods without considering the data's context or examining its graphical representations.

It underscores the importance of exploratory data analysis (EDA) and data visualization techniques in understanding and interpreting datasets effectively.

In summary, Anscombe's quartet serves as a powerful reminder of the necessity of visualizing data and conducting exploratory analyses before drawing conclusions or making decisions based solely on summary statistics. It emphasizes the complementary roles of statistical analysis and graphical visualization in understanding and interpreting data accurately.

3. What is Pearson’s R? (3 marks)

Answer:

- Pearson's r measures the degree to which two variables are linearly related to each other. It ranges from -1 to 1.
- A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is a preprocessing step in data analysis and machine learning that involves transforming the features (variables) of a dataset to a similar scale. The goal of scaling is to ensure that all features have comparable magnitudes, which can improve the performance and effectiveness of certain algorithms, particularly those based on distance metrics or gradient descent optimization.

Here's a breakdown of scaling, its purpose, and the difference between normalized scaling and standardized scaling:

Purpose of Scaling:

Improved Algorithm Performance: Many machine learning algorithms, such as support vector machines (SVM), k-nearest neighbors (KNN), and gradient descent-based algorithms, are sensitive to the scale of the input features. Scaling helps these algorithms converge faster and prevents them from being dominated by features with larger magnitudes.

Interpretability: Scaling ensures that the coefficients or weights assigned to different features in linear models are comparable, making it easier to interpret their relative importance.

Types of Scaling:

a. Normalized Scaling:

Normalization, also known as min-max scaling, transforms the features to a common scale within a specified range (usually between 0 and 1).

This type of scaling is sensitive to outliers because it uses the minimum and maximum values of the feature.

b. Standardized Scaling:

Standardization, also known as z-score scaling, transforms the features to have a mean of 0 and a standard deviation of 1.

This type of scaling is less sensitive to outliers because it uses the mean and standard deviation of the feature, which are less affected by extreme values.

Difference between Normalized Scaling and Standardized Scaling:

Range of Values: Normalized scaling transforms the values of features to a fixed range (typically 0 to 1), while standardized scaling does not impose any specific range on the transformed values.

Sensitivity to Outliers: Normalized scaling can be sensitive to outliers because it uses the minimum and maximum values, while standardized scaling is less affected by outliers because it uses the mean and standard deviation.

Interpretability: Normalized scaling preserves the original distribution and range of values, while standardized scaling centers the data around 0 and rescales it based on the standard deviation, making it easier to compare the relative importance of features.

In summary, scaling is performed to ensure that all features have comparable magnitudes, thereby improving algorithm performance and interpretability. Normalized scaling and standardized scaling are two common methods used for this purpose, each with its own advantages and considerations.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer:
Sometimes, the Variance Inflation Factor (VIF) can become infinite. This occurs when there is perfect multicollinearity among the predictor variables in the regression model.

VIF measures how much the variance of an estimated regression coefficient is inflated due to multicollinearity among the predictor variables. Mathematically, the VIF for the
ith predictor variable is calculated as:
1/(1- ith R-Squared value)

When perfect multicollinearity exists, the coefficient of determination R - Suqare for the affected predictor variable(s) becomes 1. This is because the variation in the predictor variable(s) is completely explained by the other predictor variables in the model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer:
A Q-Q plot, short for Quantile-Quantile plot, is a graphical technique used to assess whether a dataset follows a particular probability distribution. It compares the quantiles of the empirical data to the quantiles of a theoretical distribution, typically a normal distribution, by plotting them against each other.

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

How a Q-Q Plot Works:

In a Q-Q plot, the x-axis represents the quantiles of a theoretical distribution (e.g., a normal distribution), while the y-axis represents the quantiles of the observed data.
If the data follow the theoretical distribution closely, the points on the Q-Q plot will fall approximately along a straight line.
Any deviations from the straight line indicate departures from the assumed distribution.
Use and Importance in Linear Regression:

Normality Assumption: One of the key assumptions of linear regression is that the residuals (errors) follow a normal distribution. Q-Q plots are commonly used to assess the normality of residuals in linear regression models.
Diagnostic Tool: Q-Q plots provide a visual way to evaluate whether the residuals are normally distributed. If the points on the Q-Q plot fall approximately along a straight line, it suggests that the residuals are normally distributed. Deviations from the straight line indicate departures from normality.
Model Assessment: Assessing the normality of residuals is important for evaluating the validity and reliability of a linear regression model. Departures from normality may indicate that the model assumptions are violated, which can lead to biased estimates and incorrect inferences.
Model Improvement: If the Q-Q plot reveals non-normality in the residuals, corrective actions such as data transformation or using robust regression techniques may be necessary to improve the model's performance.
In summary, Q-Q plots are valuable tools for assessing the normality of residuals in linear regression models. They provide a visual way to evaluate whether the residuals follow a normal distribution, which is essential for ensuring the validity and reliability of the regression analysis. By examining Q-Q plots, researchers can identify departures from normality and take appropriate steps to improve the model's performance.