Assignment-based Subjective Questions

Q1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

There were 6 categorical variables in the dataset.

We used Box plot to study their effect on the dependent variable ('cnt') .

The inference derived is as follows:

- **season**: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
- **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
- **weathersit**: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
- **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
- **workingday**: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer:

- Using drop_first=True during dummy variable creation is important because it prevents multicollinearity, ensuring the regression model remains statistically valid and the coefficients can be accurately estimated.

- This practice also simplifies the interpretation of the model by comparing the effects of each category relative to a reference category.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

Among numerical variables in pair-plot “temp”, “atemp” has the highest correlation with the target variable “cnt” .

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

After building the model on the training set, the following steps are taken to validate the assumptions:

1. Linearity

Checked relationship between the independent and dependent variables is linear using scatter plot of each independent variable against the dependent variable to visually inspect linear relationships.

2. Normality

Histogram of Residuals**: Residuals formed a bell-shaped curve.**

3. No Multicollinearity

Made sure that the independent variables are not highly correlated with each other.

- **Variance Inflation Factor (VIF)**: VIF values are less than 5 indicate high no multicollinearity.

4. No Endogeneity

Verify that there are no endogeneity issues (independent variables are uncorrelated with the error term).

5. Homoscedasticity

Made sure that residuals have constant variance by plotting residual and y_train.

6. Independence

Verified that the residuals are independent.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

1.spring

2.workingday

3.sep

General Subjective Questions

Q1. Explain the linear regression algorithm in detail

Answer:

Linear regression is a fundamental algorithm in machine learning for **predicting a continuous value** (dependent variable) based on one or more independent variables. It works by finding the best-fitting straight line through a set of data points.

1. **Model Representation:**
    - We have a dataset with **independent variables (X)** and a **dependent variable (Y)**.
    - The linear regression model assumes a linear relationship between X and Y, represented by the equation: $\mathbf{Y = \theta_0 + \theta_1 X + \varepsilon}$
        - θ_0 (theta-nought) is the intercept (y-axis value where the line crosses).
        - θ_1 (theta-one) is the slope of the line.
        - ε (epsilon) represents the error term (difference between actual Y and predicted Y).
2. **Learning the Model:**
    - The algorithm doesn't directly learn the equation. It learns the optimal values for θ_0 and θ_1 that minimize the error between the predicted Y values and the actual Y values in the training data.
3. **Minimizing Error:**
    - A common approach is using **least squares** to minimize the error. This involves calculating the sum of squared errors (SSE) between predicted and actual Y values.
    - The algorithm uses an optimization technique like **gradient descent** to iteratively adjust θ_0 and θ_1. It adjusts them in the direction that reduces

the SSE the most. This process continues until the change in SSE becomes negligible.

4. **Evaluation and Prediction:**
    - Once trained, the model can be used to predict the dependent variable for new unseen data points based on their independent variable values.
    - The model's performance is evaluated using metrics like R-squared, which indicates how well the model fits the data.

- There are two main types of linear regression:
    - **Simple linear regression:** Uses one independent variable.
    - **Multiple linear regression:** Uses multiple independent variables.

In conclusion, linear regression is a powerful tool for uncovering linear relationships in data and making predictions. It's a foundational concept for understanding more complex machine learning algorithms.

Q2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression line) but differ significantly when graphed. It highlights the importance of graphing data before analyzing it and the potential pitfalls of relying solely on statistical measures.

1. **Identical Descriptive Statistics**:
    - All four datasets in Anscombe's quartet have the same mean, variance, and correlation.
    - The linear regression lines for all four datasets are also nearly identical.
2. **Different Graphical Representations**:
    - **Dataset 1**: A simple, linear relationship between xxx and yyy.
    - **Dataset 2**: A clear nonlinear relationship that a linear model poorly represents.
    - **Dataset 3**: A linear relationship with an outlier that significantly affects the regression line.
    - **Dataset 4**: Vertical alignment of most points with one high leverage point that drives the regression line.

Anscombe's quartet demonstrates that statistical properties alone do not provide a complete understanding of data. Visualizing data can reveal underlying patterns, anomalies, and relationships that statistical summaries may miss, underscoring the need for both graphical and numerical analysis in data interpretation.

Note:

Anscombe's quartet consists of four datasets with nearly identical statistical properties but distinctly different graphs, emphasizing the importance of visual data inspection in addition to numerical analysis.

Q3. What is Pearson's R?

Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables. It ranges from -1 to 1, where:

- **+1** indicates a perfect positive linear correlation.
- **-1** indicates a perfect negative linear correlation.
- **0** indicates no linear correlation.

Pearson's R is calculated using the formula:

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

where:

- x_i and y_i are individual data points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables.

Note:

Pearson's R quantifies the strength and direction of the linear relationship between two variables, providing insight into how changes in one variable are associated with changes in another.

Q4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Scaling is the process of transforming features to ensure they contribute equally in analysis. It's performed to prevent features with larger magnitudes from dominating and to improve algorithm performance

Why Scaling is Performed?

1. **Equal Contribution**: Ensures that features with larger scales do not dominate those with smaller scales.
2. **Improved Convergence**: Helps gradient descent converge more quickly by making the cost function more symmetrical.
3. **Enhanced Performance**: Improves the performance of distance-based algorithms by ensuring that all features are treated equally.

Normalized scaling (Min-Max) transforms data to a [0, 1] range. Preferred when the distribution is not Gaussian or when the data has a fixed minimum and maximum.

Standardized scaling (Z-score) adjusts data to have a mean of 0 and standard deviation of 1. Preferred when the data follows a Gaussian distribution or when outliers are present.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

The value of the Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity in the regression model, meaning that one predictor variable is an exact linear combination of one or more of the other predictor variables. This occurs because the VIF formula is given by:

$$\text{VIF} = \frac{1}{1-R^2}$$

where R2 is the coefficient of determination of the regression of the predictor on the other predictors. When R2=1, it indicates that the predictor is perfectly predicted by the other predictors, leading to the denominator of the VIF formula becoming zero. Since division by zero is undefined, the VIF value becomes infinite. This situation reveals that there is redundant information in the model and significant multicollinearity, causing instability in the estimation of regression coefficients and making them unreliable.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a particular distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution (like the normal distribution).

Use and Importance in Linear Regression

1. **Distribution Assessment**: Q-Q plots help visualize whether the residuals (the differences between observed and predicted values) of a linear regression model follow a normal distribution. In linear regression, normally distributed residuals are a key assumption for valid statistical inference.

2. **Identifying Departures from Normality**: By examining the points on the Q-Q plot, you can identify deviations from the theoretical line (which represents normality). If the points deviate significantly from this line, it indicates that the residuals may not be normally distributed.
3. **Assumption Checking**: Linear regression assumes that residuals are normally distributed with a mean of zero and constant variance (homoscedasticity). Deviations from normality observed in the Q-Q plot suggest potential issues with these assumptions, prompting further investigation or modification of the model.

Summary

A Q-Q plot is a powerful tool in linear regression analysis as it visually assesses the normality assumption of residuals. It helps to validate the model's statistical assumptions and ensures the reliability of the regression results.