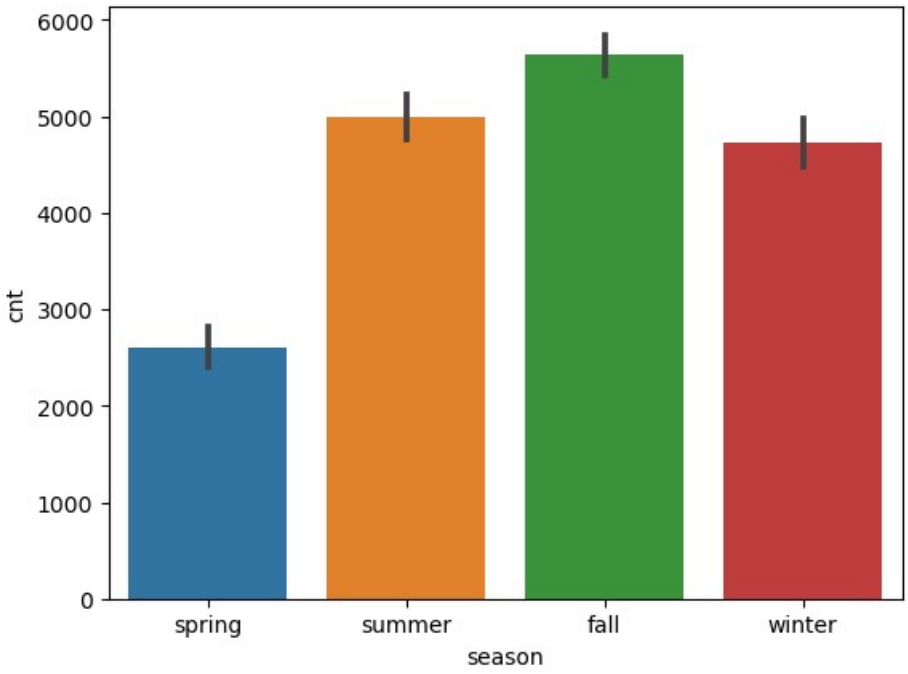
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

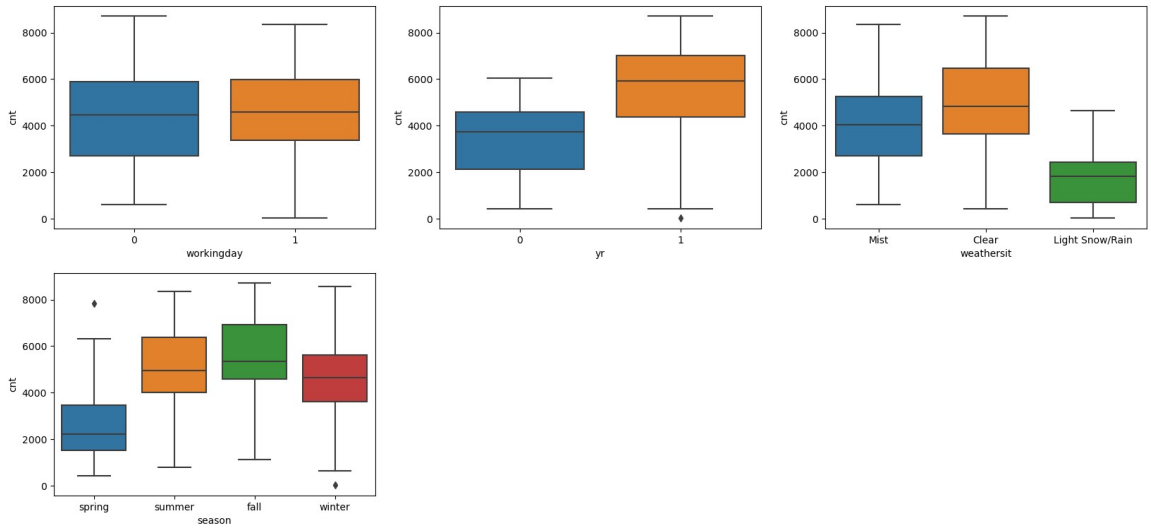
Answer:

In the dataset, variables like season, weathersit, holiday, weekday, and workingday could impact the count of the bike rentals (the dependent variable).

For example, in the below graph , one can see that the rental count was high during summers.



Also, the box plot showed the median and the high counts for the categorical variables



From the above graphs, one could infer each variable's impact which can help make operational decisions, like increasing bike availability in high-demand seasons or during clear weather.

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

Answer

One should always drop the first category, which will prevent multicollinearity, which occurs when two or more predictor variables are highly correlated. If all categories are included, the information in the first category can be perfectly predicted by the others, leading to redundancy in the model.

When one category is dropped, the remaining dummy variables will represent the effect of each category relative to the dropped one. This makes the model easier to interpret, as the coefficients of the remaining variables can be viewed as differences from the baseline category.

For example, in the dataset for clear, misty, rainy

Clear	Mist	Rainy
1	0	0
0	1	0
0	0	1

In the first row, values 100 can be accepted as 00 dropping the column clear. which will indicate clear.

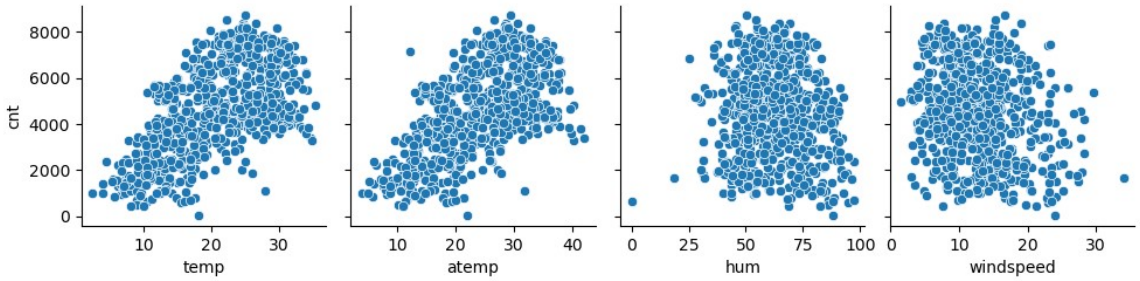
In the second row, 10, will indicate Mist

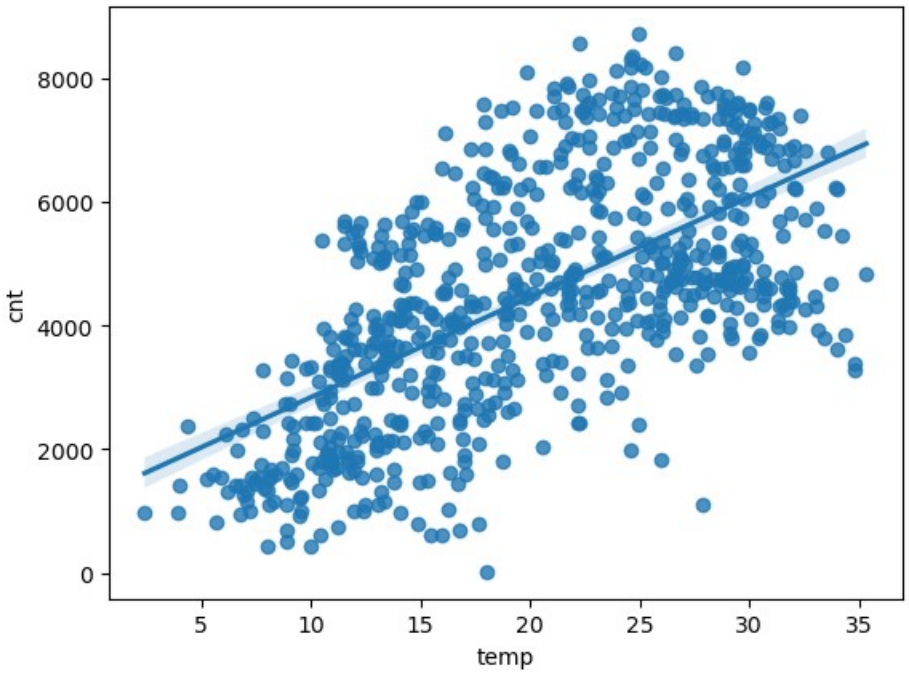
In the third row, 01 will indicate Rainy

This ensures the dataframe is not cluttered and makes it or easy reading.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer





Temp has the highest/strongest positive correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer

I made use of scatter plots of independent variables against the dependent variable to visually inspect for linear relationships. I also checked for multicollinearity and calculated their VIF(variation inflation factor) and their significance by making use of statsmodel summary method.

Noted the p-value for all the variables in the training set and their VIFs.

Dropped any variable that had a high p-value and a high VIF.(above 5)

Next, dropped any variable that has a high p-value(greater than 0) and checked for the effect that made on the VIFs of the remaining variables.

Lastly, dropped the variable that had a high VIF. Doing so, significantly reduced the VIF for other variables.

By looking at the r squared value that was 0.76, i can say that the linear regression model is robust and reliable.

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

The top 3 features that contributed significantly towards explaining the demand of the shared bikes are **temp, workingday and windspeed**.

The temp could also be seen making a strong positive correlationwith the cnt variable.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer

Linear regression is a fundamental statistical technique used for predicting a continuous dependent variable based on one or more independent variables.

Linear regression aims to establish a relationship between the dependent variable Y (the outcome) and one or more independent variables X (the predictors) by fitting a linear equation to observed data.

The simplest form is the simple linear regression, represented as:

$Y = B_0 + B_1X + \text{Error term}$

B_0 is calculated as Y - intercept and
B_1 is the slope of the line

In multiple linear regression, the equation extends to:

$Y = B_0 + B_1X_1 + ... + BnXn + \text{Error term}$

The objective of linear regression is to minimize the difference between the observed values and the values predicted by the model. This is typically done by minimizing the sum of squared residuals (errors):

sigma (Y - Ybar)^2

Now, once the data is obtained, one must
1. process the missing values.
2. Encode categorical variables (using techniques like one-hot encoding).
3. Scale numerical features if necessary.
4. Use a method such as Ordinary Least Squares (OLS) to estimate the coefficients
5. Use metrics like R-squared, Mean Squared Error (MSE), and residual analysis to assess model performance.

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

Answer

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but differ significantly in their distributions and visual representations.

Dataset A
Description: A linear relationship with a positive slope.
Visual Representation: Points cluster around a straight line, suggesting a linear relationship.

Dataset B
Description: A nonlinear relationship (quadratic) that resembles a U-shape.

Visual Representation: Points form a curve, indicating a clear nonlinear relationship.

Dataset C

Description: A linear relationship with a high degree of outlier influence.

Visual Representation: Most points cluster around a line, but there is a single outlier that has a significant effect on the slope.

Dataset D

Description: A vertical line, indicating no relationship between X and Y. The points are clustered at a fixed X value.

Visual Representation: Points are perfectly aligned vertically.

When you plot the four datasets, you will see stark differences in the scatter plots despite the identical summary statistics. This highlights the critical point that statistical measures alone can be misleading.

Importance of Anscombe's Quartet

Emphasizes Graphical Analysis:

The quartet demonstrates that visualizing data is essential for understanding the underlying patterns and relationships that summary statistics may conceal.

Reinforces Caution in Data Interpretation:

It warns against the dangers of making assumptions based solely on statistical metrics, especially in linear regression analysis.

Encourages Exploratory Data Analysis (EDA):

Anscombe's Quartet underscores the necessity of performing EDA to uncover hidden structures or anomalies in data.

Illustrates Different Relationships:

It exemplifies how different types of relationships (linear, nonlinear, and those influenced by outliers) can yield the same statistical results.

3. What is Pearson's R? (3 marks)

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 continuous variables. Here's a detailed overview of Pearson's R:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

n = number of data points
x and y = individual data points of the two variables
sigma x and y = sum of each variables score
sigme x^2 and y^2 = sum of the squares of each variable's scores

Range of Values:
1: Perfect positive linear correlation (as one variable increases, the other also increases perfectly).
-1: Perfect negative linear correlation (as one variable increases, the other decreases perfectly).
0: No linear correlation (no relationship between the variables).
Values between -1 and 1: Indicate varying degrees of correlation strength:
0 to 0.3 or 0 to -0.3: Weak correlation
0.3 to 0.7 or -0.3 to -0.7: Moderate correlation
0.7 to 1 or -0.7 to -1: Strong correlation

Pearson's R is a fundamental tool for understanding the linear relationships between variables. While it provides valuable insights, it should be used in conjunction with graphical methods and other statistical analyses to draw comprehensive conclusions about the data.

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 technique used in data analysis and machine learning to adjust the range and distribution of data features. It is important for ensuring that different features contribute equally to the analysis, especially when they are on different scales.

Scaling is performed for:
1. Equal Contribution: Many machine learning algorithms, particularly those that use distance calculations (like k-nearest neighbors and support vector machines), are sensitive to the scale of the data. Features with larger ranges can disproportionately influence the results.

2. Improved Convergence: In gradient descent optimization methods, scaling helps in faster convergence by ensuring that the cost function is minimized more effectively.

3. Enhances Model Performance: Scaling can improve the performance of models by reducing the risk of overfitting or underfitting due to varying feature magnitudes.

4. Facilitates Interpretation: It can make the coefficients of a model easier to interpret, especially in regression models where feature magnitudes may differ significantly.

Differences between:

Feature	Normalized Scaling (Min-Max)	Standardized Scaling (Z-score)
Range	[0, 1] (or any specified range)	Mean = 0, Standard Deviation = 1
Distribution	Preserves the shape of the original data distribution	Transforms to standard normal distribution
Sensitivity to Outliers	Sensitive; outliers can skew the scaling	Less sensitive; outliers affect mean and standard deviation
Use Cases	When data is not normally distributed and when maintaining relationships is important	When data is normally distributed and when algorithms assume normality

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

Answer

This occurs when one independent variable is a perfect linear combination of other independent variables. For instance, if you have two variables X1 and X2 such that X2 = k* X1 (where k is a constant), then you have perfect multicollinearity.In such cases, the denominator of the VIF formula becomes zero, leading to an infinite VIF value.

VIF = 1/ (1 - R^2)

If R^2 becomes 1, then VIF becomes infinite.

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 (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, such as the normal distribution. It helps assess whether the data follows a specified distribution by plotting the quantiles of the data against the quantiles of the theoretical distribution.

Structure of a Q-Q Plot

1. Axes:

- The x-axis represents the quantiles of the theoretical distribution (e.g., standard normal distribution).
- The y-axis represents the quantiles of the sample data.

2. Reference Line:

A reference line (usually a 45-degree line) is included to help visualize how closely the points follow the theoretical distribution.

1. Linear Pattern: If the points in the Q-Q plot lie approximately along the reference line, the data is likely to follow the specified theoretical distribution.

2. Deviation from Line:

S-shaped Curve: Indicates a heavy-tailed distribution (more outliers).

Concave Downward: Suggests a distribution with lighter tails than the normal distribution.

Concave Upward: Indicates a distribution with heavier tails than the normal distribution.

Use of Q-Q Plots in Linear Regression

1. Assessing Normality of Residuals: One of the key assumptions of linear regression is that the residuals (the differences between observed and predicted values) should be normally distributed. A Q-Q plot of the residuals can help evaluate this assumption.

2. Detecting Non-Normality: If the Q-Q plot shows that residuals deviate significantly from the reference line, it suggests that the normality assumption is violated. This can impact the validity

of hypothesis tests and confidence intervals derived from the model.

3. Identifying Outliers: Q-Q plots can help identify outliers. Points that are far from the reference line indicate potential outliers in the dataset.

4. Model Diagnostics: By examining the Q-Q plot of the residuals, analysts can determine whether the linear regression model is appropriate or if data transformations are needed.

Importance of Q-Q Plots

1. Visual Assessment: Q-Q plots provide a quick and intuitive way to visually assess the distribution of residuals.

2. Model Validity: Ensuring that the normality assumption holds is crucial for the validity of linear regression results. Deviations can lead to incorrect conclusions.

3. Data Insights: They can provide insights into the nature of the data distribution, which can inform further analysis or model adjustments.