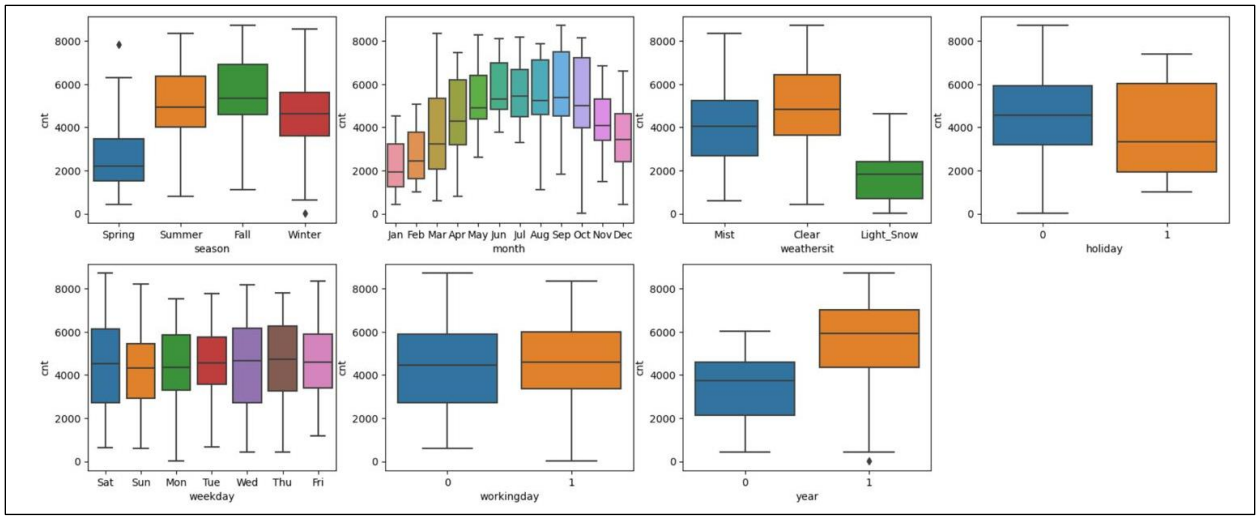
Assignment-based Subjective Questions

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

Answer:

As per my analysis, variables (season, month, weathersit, holiday) have major effect on 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, it is important to use ‘drop_first=True’ because it helps in reducing the extra column created during dummy variable creation. This reduces the correlations created among dummy variables .

For instance, if we have a categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables . By dropping the first column, we can avoid the dummy variable trap and reduce the dimensionality of the dataset .

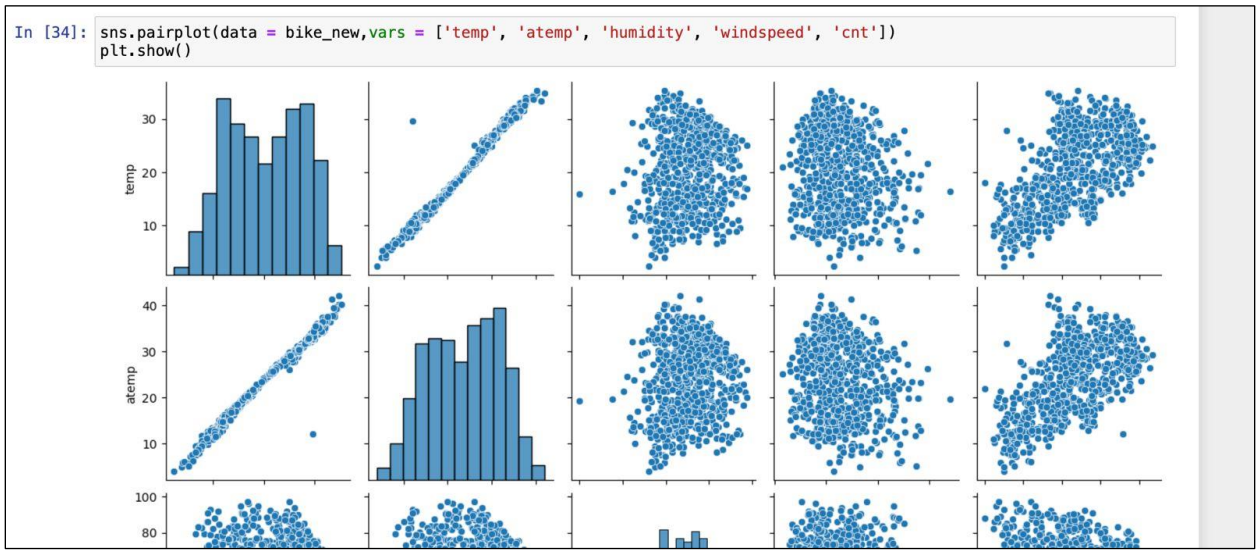
In other words, `drop_first=True` drops the first column during dummy variable creation. Suppose you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown". We do NOT need another column for "Unknown" .

By using `drop_first=True`, we can reduce the number of dummy features, which makes it easier for the algorithm to fit and prevents overfitting .

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

Answer:

Before the data cleaning, the variable registered has higher correlation value. However, after the data cleaning, atemp & temp variables have highest correlation with target variable 'cnt'



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

Answer:

After building the model on the training set, I carried out the following analysis: -

Assumptions of Linear Regression:

1. There is a linear relationship between X and Y
2. Error terms are normally distributed with mean zero (not X, Y)
3. Residual Analysis of Training Data proves that the Residuals are normally distributed.
4. Hence our assumption for Linear Regression is valid.
5. Eliminations and inclusion of independent variables into each model based on VIF and p-values to avoid multi collinearity.

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:

Top 3 features that has significant impact towards explaining the demand of the shared bikes are **temperature, year, and season**.

General Subjective Questions

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

   Answer:

   Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also vary accordingly.

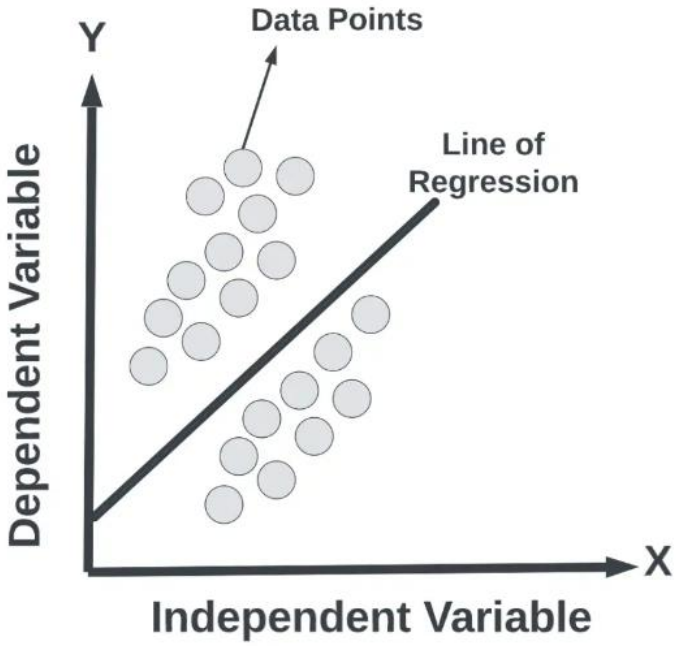
   Mathematically the relationship can be represented with the help of equation: Y = mX + c

   Where, Y is the dependent variable to predict.

   X is the independent variable being used to make predictions.

   m is the slope of the regression line which represents the effect X has on Y

   c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

The linear regression model can be represented by the following equation:

$Y = \theta_o + \theta_1 x_1 + \theta_1 x_1 + \ldots\ldots + \theta n\ xn$

where,

Y is the predicted value

θ_o is the constant term.

$\theta_1,\ldots,\theta n$ are the model parameters

$x_1, x_2,\ldots,xn$ are the feature values.

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 but differ significantly when graphed or analyzed further. This quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data and visually inspecting it before drawing conclusions based solely on summary statistics. The quartet consists of four sets of (x, y) data points, and each set has 11 data points.

The characteristics of each dataset in Anscombe's quartet are as follows:

1. Dataset I: Linear Relationship
   - The relationship between x and y is linear.
   - The correlation coefficient is close to 1.
   - A simple linear regression model would be appropriate.
2. Dataset II: Non-Linear Relationship
   - The relationship is not linear but follows a clear curve.
   - Again, a correlation coefficient may not reveal the nature of the relationship.
   - A different model might be more suitable, such as a quadratic regression.
3. Dataset III: Outlier Influence
   - The dataset has a similar linear relationship as Dataset I, but with an influential outlier.
   - The outlier significantly affects the correlation coefficient and the regression line.

- This highlights the impact of outliers on statistical analysis.

4. Dataset IV: No Clear Relationship

- There is no apparent linear relationship between x and y.

- The correlation coefficient is close to zero.

- This dataset illustrates the importance of not assuming a relationship based solely on the lack of correlation.

The key lesson from Anscombe's quartet is that relying solely on summary statistics like mean, variance, and correlation can be misleading. Even when these statistics are similar, the underlying data patterns may be vastly different. Visualization, through techniques like scatter plots, is essential to understand the distribution and relationships within the data. Anscombe's quartet serves as a powerful reminder to always explore and visualize data before drawing conclusions based on summary statistics alone.

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

Answer:

Pearson's R, also known as the Pearson Correlation Coefficient, is a measure of the linear correlation between two variables. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

1. **Values**: Pearson's R can take on values from -1 to +1. A value of +1 indicates a perfect positive linear relationship, a value of -1 indicates a perfect negative linear relationship, and a value of 0 indicates no linear relationship.

2. **Interpretation**: The strength and direction of the correlation are indicated by the value of R. For example, a value greater than 0.5 indicates a strong positive correlation, a value between 0.3 and 0.5 indicates a moderate positive correlation, a value between 0 and 0.3 indicates a weak positive correlation, and so on.

3. **Usage**: Pearson's R is used to quantify the strength of the linear association between two variables. It is commonly used in statistics and machine learning for feature selection, data analysis, and model evaluation.

4. **Limitations**: Pearson's R only measures linear relationships. Therefore, it may not accurately capture relationships that are not linear.

The formula for Pearson's correlation coefficient (r) between variables X and Y with n data points is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Xi & Yi are the individual data points, $\bar{X}$ and $\bar{Y}$ & are the means of X and Y, and the summations are taken over all data points.

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

Answer:

Scaling is a process in data preprocessing that involves transforming the values of variables (features) in a dataset to a specific range. The purpose of scaling is to ensure that the variables contribute equally to the analysis, particularly in algorithms that are sensitive to the scale of the input features. Two common scaling techniques are normalization and standardization.

Normalized Scaling (Min-Max Scaling):

- **Formula:** X normalized = X-Xmin/(Xmax-Xmin)
- Scales the data to a specific range, typically [0, 1].
- Preserves the original distribution of the data.
- Sensitive to outliers when the range is determined by the minimum and maximum values in the dataset.

Standardized Scaling (Z-score normalization):

- **Formula:** X standardized= $(X - \mu)/\sigma$
- Transforms the data to have a mean (μ) of 0 and a standard deviation (σ) of 1.
- Less sensitive to outliers compared to normalization.
- Does not have a predefined range, making it suitable for algorithms that do not assume a specific data range.

Why Scaling is Performed:

Equalizing Influence: Scaling ensures that all variables contribute equally to the analysis. This is particularly important in machine learning algorithms that use distance-based metrics, such as k-nearest neighbors or clustering algorithms.

Gradient Descent Convergence: In optimization algorithms like gradient descent, scaling helps the algorithm converge faster by avoiding oscillations and reaching the minimum more efficiently.

Regularization: In regularization techniques, such as ridge regression or lasso regression, scaling ensures that all features are penalized equally.

Some Algorithms Requirement: Certain algorithms, like support vector machines and neural networks, often perform better when input features are on a similar scale.

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

(3 marks)

Answer:

The Variance Inflation Factor (VIF) measures the extent to which the variance of an estimated regression coefficient increases due to collinearity in the model. It quantifies how much the variance of the estimated regression coefficients is inflated as compared to when the predictor variables are not correlated.

The formula for VIF for a variable *Xi* in a regression model is given by:

VIF *i* = 1/1−*Ri*^2

where *Ri*^2 is the coefficient of determination obtained by regressing *Xi* against all other predictor variables in the model.

A VIF close to 1 indicates little or no multicollinearity, whereas higher values indicate a problematic degree of collinearity.

Now, if the VIF for a variable is calculated to be infinite, it typically implies perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in a regression model are a perfect linear combination of other variables. In other words, there is an exact linear relationship among some of the variables in the model.

Perfect multicollinearity leads to a singularity in the correlation matrix of the predictor variables, making it impossible to invert and calculate the VIF. The singularity arises because the determinant of the matrix becomes zero, and the matrix becomes non-invertible.

Common reasons for perfect multicollinearity include:

1. Duplicate Variables: If two or more variables are identical or nearly identical, their correlation coefficient is 1, leading to perfect multicollinearity.
2. Linear Relationships: A variable that can be expressed as a linear combination of other variables in the model.

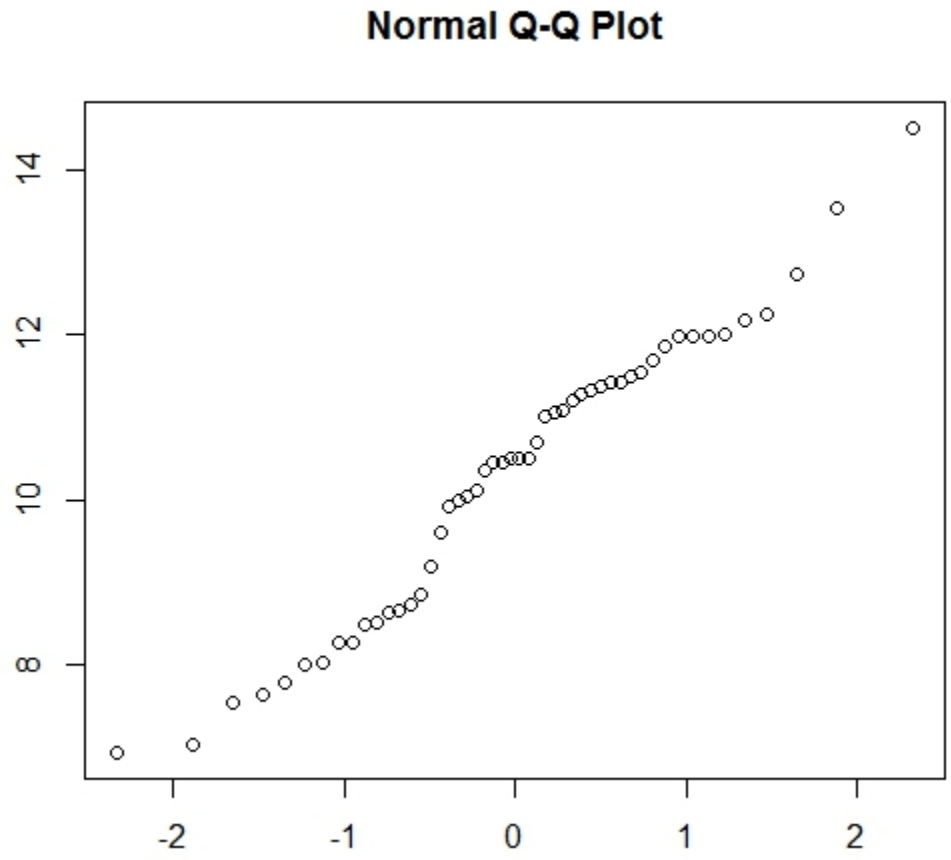
To address the issue of infinite VIF, it's essential to identify and resolve the root cause of multicollinearity, which may involve removing redundant variables, transforming variables, or reconsidering the model specification.

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

(3 marks)

Answer:

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution. It is often employed to check the normality assumption in statistical analysis, including linear regression. The Q-Q plot compares the quantiles of the observed data with the quantiles of a theoretical distribution, typically the normal distribution.



Here's how a Q-Q plot works:

Ordered Data:

Arrange the data in ascending order.

Associate each data point with its corresponding quantile in the theoretical distribution.

Plotting:

Plot the observed quantiles against the expected quantiles from the theoretical distribution.

If the points in the Q-Q plot fall approximately along a straight line, it suggests that the data follows the assumed distribution.

Use and Importance in Linear Regression:

Normality Assumption:

In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed.

Checking the normality assumption is important because it affects the validity of statistical inference and hypothesis testing based on regression results.

Detecting Departures from Normality:

The Q-Q plot visually displays whether the residuals deviate from a normal distribution.

Departures from a straight line in the Q-Q plot may indicate non-normality or suggest specific patterns, such as heavy tails or skewness.

Model Validity:

A linear regression model's results and conclusions can be misleading if the normality assumption is violated.

Q-Q plots help researchers and analysts identify potential issues with the distribution of residuals, allowing for adjustments or transformations if necessary.

Alternative Distributions:

Q-Q plots are not limited to normality checks; they can be used for other theoretical distributions as well.

If the data does not fit a normal distribution, analysts might explore alternative distributions that better describe the data.

In summary, Q-Q plots are valuable tools in linear regression analysis for assessing the normality assumption of residuals. By providing a visual representation of the distribution of residuals compared to a theoretical normal distribution, Q-Q plots aid in making informed decisions about the validity and reliability of regression model results.