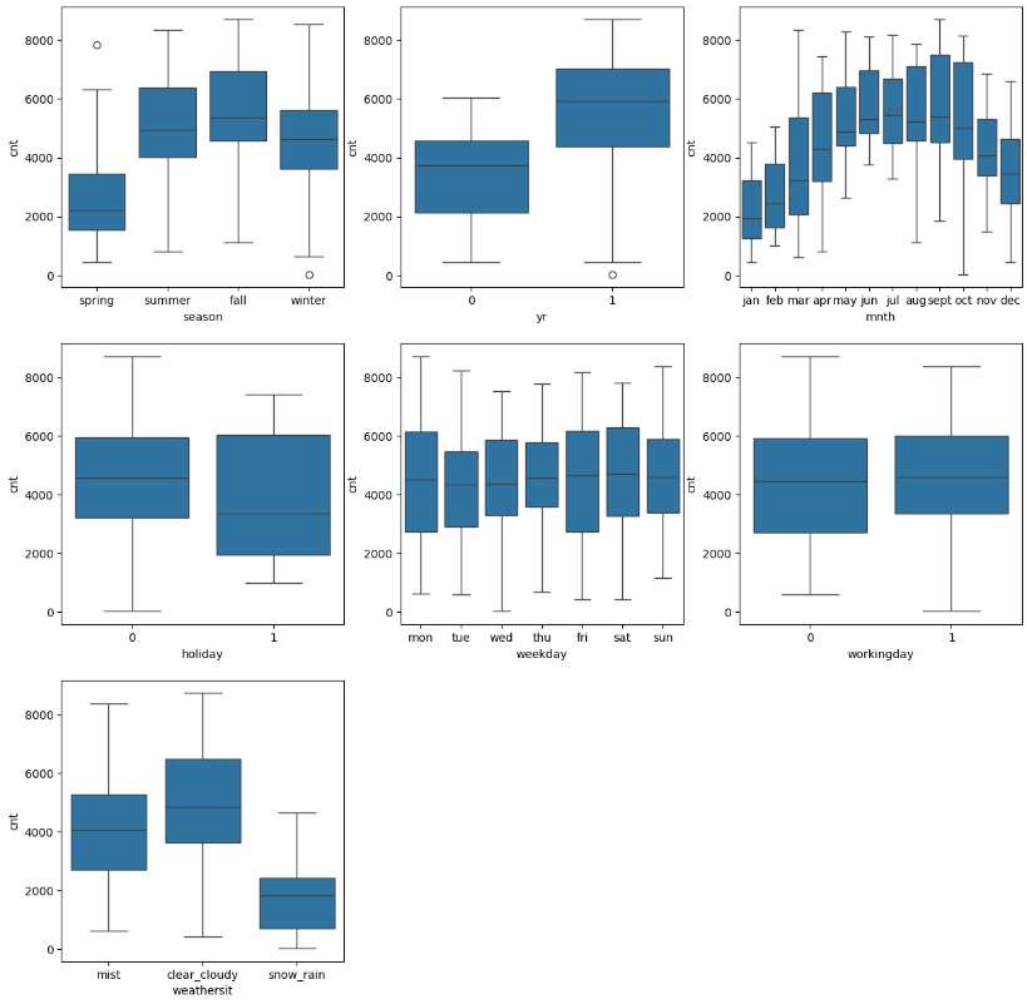
Linear Regression Subjective Questions

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:

With the analysis of categorical variables from dataset below conclusion can be drawn



1. Season: Fall (Season 3) shows the highest demand for rental bikes.
2. Yearly Trend: Demand has increased compared to the previous year.
3. Monthly Trend: Demand grows continuously each month until June, with September exhibiting the highest demand. After September, demand begins to decrease.
4. Holidays: Demand decreases during holiday periods.
5. Weekdays: Demand patterns do not show a clear trend based on weekdays.
6. Weather: Clear weather conditions correlate with the highest demand for bikes.
7. Seasonal Variation: Bike sharing is more prevalent in September, while it decreases towards the end and beginning of the year, possibly due to extreme weather conditions.

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

Answer:

Using drop_first=True when creating dummy variables is important to avoid multicollinearity and improve model interpretability. It drops one category (the reference category) to prevent redundancy and ensures that the remaining variables provide meaningful comparisons to the baseline.
Syntax:
drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.
Example:
Suppose we have a categorical variable with three categories: A, B, and C. Creating dummy variables without drop_first=True will result in three dummy variables:

- A, B, and C.

If we include all three in the model, the value of C can be determined when both A and B are 0, causing multicollinearity. Using drop_first=True, only two dummy variables (e.g., A and B) are included, and the third category C becomes the reference category.
This ensures that our model is well-specified and easier to interpret.

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

Answer:
"temp" has the highest correlation with the target variable.

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

Answer:
To validate the assumptions of Linear Regression after building the model on the training set, we follow these steps:
1. **Linearity**: The residuals should show no discernible patterns, indicating a linear relationship between the independent and dependent variables.
2. **Independence of residuals**: Residuals should be independent, with no significant autocorrelation.
3. **Homoscedasticity**: The residuals should display constant variance across all levels of predicted values. If the plot shows a funnel shape or other patterns, this indicates heteroscedasticity, which violates the assumption of homoscedasticity.
4. **Normality of Residuals**: The residuals should be approximately normally distributed
5. **Multicollinearity**: VIF values should generally be below 5, indicating low multicollinearity. High VIF values point to potential multicollinearity, which can distort the estimation of coefficients.

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

Answer:
The top three features significantly impacting the demand for shared bikes are temperature, year, and season.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:
Linear regression is a supervised machine learning technique that models the relationship between a dependent variable and one or more independent features by fitting a linear equation to observed data.

- **Simple Linear Regression** is used when there is a single independent feature. It models the relationship with a straight line.

The equation for simple linear regression is:
$y=\beta0 + \beta1X$
Where:

- *Y is the dependent variable*
- *X is the independent variable*
- *β0 is the intercept*
- *β1 is the slope*

- **Multiple Linear Regression** involves multiple independent features and models the relationship with a hyperplane in multidimensional space.

  The equation for multiple linear regression is:
  $y=\beta0+\beta1X1+\beta2X2+\ldots\ldots\beta nXn$
  Where
  - *Y is the dependent variable*
  - *X1, X2, …, Xn are the independent variables*
  - *β0 is the intercept*
  - *β1, β2, …, βn are the slopes*

Additionally:

- **Univariate Linear Regression** refers to the case where there is one dependent variable and one or more independent variables.
- **Multivariate Regression** involves multiple dependent variables being predicted from one or more independent variables.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet is a famous set of four datasets that were constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphical analysis in statistics. The quartet illustrates how datasets with identical statistical properties can have very different distributions and relationships when visualized. This underscores the idea that descriptive statistics alone may not fully capture the characteristics of the data.

Anscombe's Quartet consists of four datasets, each with the following properties:

- Mean of x values
- Mean of y values
- Variance of x values
- Variance of y values
- Correlation coefficient between x and y
- Regression line (least squares fit)

For each dataset in the quartet:

- Mean of x values: 9.0
- Mean of y values: 7.5
- Variance of x values: 11.0
- Variance of y values: 4.12
- Correlation coefficient between x and y: 0.816
- Regression line equation: y=3+0.5x

Despite having the same statistical summary, the datasets exhibit different patterns when plotted.

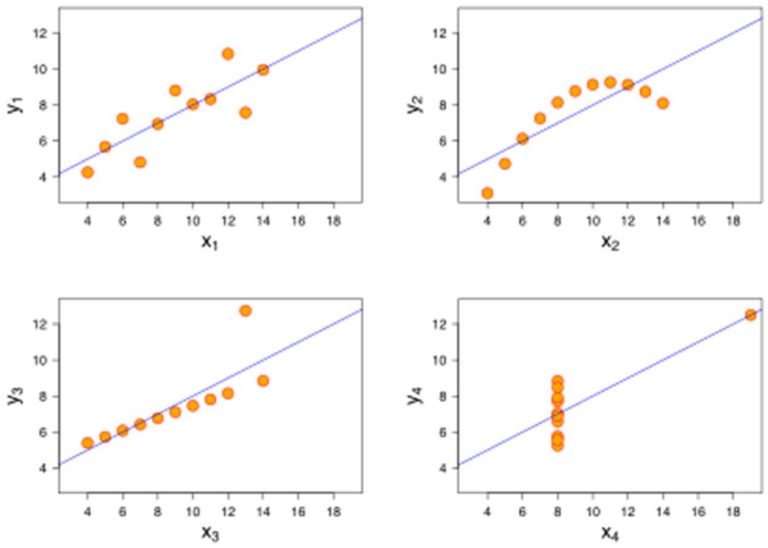
Datasets in Anscombe's Quartet

1. **Dataset I (Linear Relationship)**:

- **Description:** This dataset shows a linear relationship between x and y. The data points form a straight line when plotted.
- **Graph:** A scatter plot with a clear linear trend.

2. **Dataset II (Nonlinear Relationship)**:
   - **Description:** This dataset also has the same statistical properties as the first but shows a curved relationship. The points follow a parabolic pattern.
   - **Graph:** A scatter plot where data points follow a curve rather than a straight line.
3. **Dataset III (Outliers)**:
   - **Description:** This dataset contains a single outlier that significantly affects the distribution. All other data points form a linear relationship, but the outlier skews the results.
   - **Graph:** A scatter plot with a prominent outlier that disrupts the linear trend.
4. **Dataset IV (Horizontal Line with Outlier)**:
   - **Description:** This dataset has a strong vertical line with a single outlier. The outlier affects the slope of the regression line significantly, while most points are clustered horizontally.
   - **Graph:** A scatter plot where most data points lie on a horizontal line with one significant outlier affecting the regression.

The four datasets compose Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different



The four datasets compose Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

3. 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 quantifies the degree to which the variables are related and provides insight into the strength and direction of their relationship.

Mathematically, it is given by:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where:

- cov(X,Y) is the covariance between the variables X and Y.
- σX is the standard deviation of X.
- σY is the standard deviation of Y.

The formula for Pearson's R is:

$$r = \frac{n\sum(X_iY_i) - \sum X_i \sum Y_i}{\sqrt{[n\sum(X_i^2) - (\sum X_i)^2][n\sum(Y_i^2) - (\sum Y_i)^2]}}$$

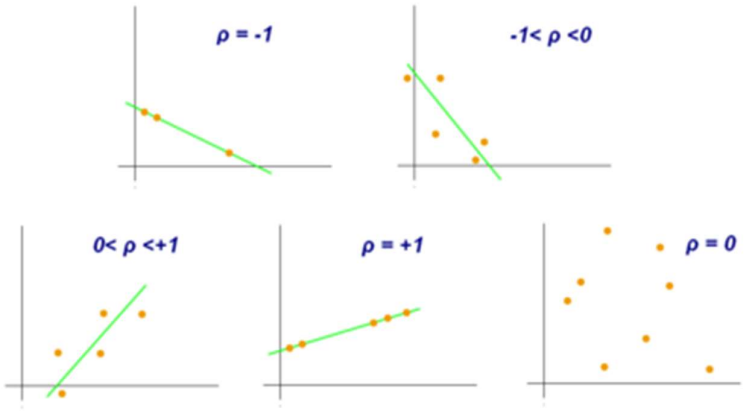
where:
n is the number of data points.
Xi and Yi are individual data points of variables X and Y respectively.

Pearson's R ranges from -1 to 1:

- **r = 1:** Perfect positive linear relationship.
- **r = -1:** Perfect negative linear relationship.
- **r = 0:** No linear relationship.

Interpretation:

- **Positive Correlation:** A positive Pearson's R value indicates that as one variable increases, the other variable also tends to increase.
- **Negative Correlation:** A negative Pearson's R value indicates that as one variable increases, the other variable tends to decrease.
- **Magnitude:** The closer the absolute value of Pearson's R is to 1, the stronger the linear relationship between the variables.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

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

Answer:
Scaling refers to the process of adjusting the range and distribution of features (variables) in a dataset. This is typically done to ensure that different features contribute equally to the analysis or model, especially when they are on different scales or units.

Scaling is performed for several reasons:

1. **Improve Model Performance:**

- Many machine learning algorithms, especially those that rely on distance metrics (e.g., k-nearest neighbors, support vector machines) or gradient-based optimization (e.g., linear regression, logistic regression), perform better when features are on similar scales.
- Algorithms may converge faster and perform more accurately when features are scaled properly.

2. **Ensure Equal Weight:**
   - Features with larger ranges or different units can disproportionately influence the model's performance. Scaling ensures that each feature contributes equally to the model.
3. **Normalize Data Distribution:**
   - Scaling can make the data distribution more uniform, which helps in meeting the assumptions of some statistical methods and models.
4. **Handle Different Units:**
   - When features are measured in different units (e.g., height in cm and weight in kg), scaling ensures that the model interprets these features on a comparable scale.

Difference between normalized scaling and standardized scaling

	Normalized scaling	Standardized scaling
1.	Scales data to a specific range, typically [0, 1]. The result is that the transformed features have minimum and maximum values defined by this range.	Transforms data to have a mean of 0 and a standard deviation of 1. There are no fixed bounds; the transformed values can be any real number.
2.	Does not alter the distribution shape; it simply rescales the data within a specific range.	Centers the data around the mean and scales it according to the standard deviation, making the data follow a standard normal distribution if the original data was normally distributed.
3.	Useful when you need bounded data or when features have different units and need to be scaled to a common range.	Useful for algorithms that assume normality or when you want to compare features on a common scale regardless of their original distribution.
4.	Sensitive to outliers because outliers can skew the minimum and maximum values.	More robust to outliers compared to normalization because it is based on mean and standard deviation.

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

Answer:

The VIF for a given predictor variable quantifies how much its variance is inflated due to multicollinearity with other predictors.

The formula for VIF is:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where Ri^2 is the R-squared value obtained by regressing the iii-th variable against all the other predictors.

This happen when Perfect multicollinearity occurs when one predictor variable is a perfect linear combination of other predictor variables. This means that there is an exact linear relationship among the predictors.

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

Answer:

A Q-Q plot is a scatter plot where the quantiles of a dataset are plotted against the quantiles of a theoretical distribution. For a normal Q-Q plot, the quantiles of the data are plotted against the quantiles of a standard normal distribution.

In the context of linear regression, a Q-Q plot is crucial for validating assumptions, particularly the normality of residuals. Here's why it is important:

1. **Assess Normality of Residuals:**
    - **Assumption Check:** Linear regression assumes that the residuals (errors) of the model are normally distributed. This assumption is important for making valid inferences and constructing confidence intervals for predictions.
2. **Detect Deviations from Normality:**
    - **Straight Line:** If the residuals follow a normal distribution, the points on the Q-Q plot will lie approximately along the 45-degree reference line.
3. **Identify Outliers and Influential Points:**
    - **Outliers:** Points that deviate significantly from the line can highlight outliers or influential data points that might affect the regression model.
4. **Enhance Model Interpretation:**
    - **Residuals Distribution:** Understanding the distribution of residuals helps in assessing the quality of the model fit and in interpreting the results more accurately.