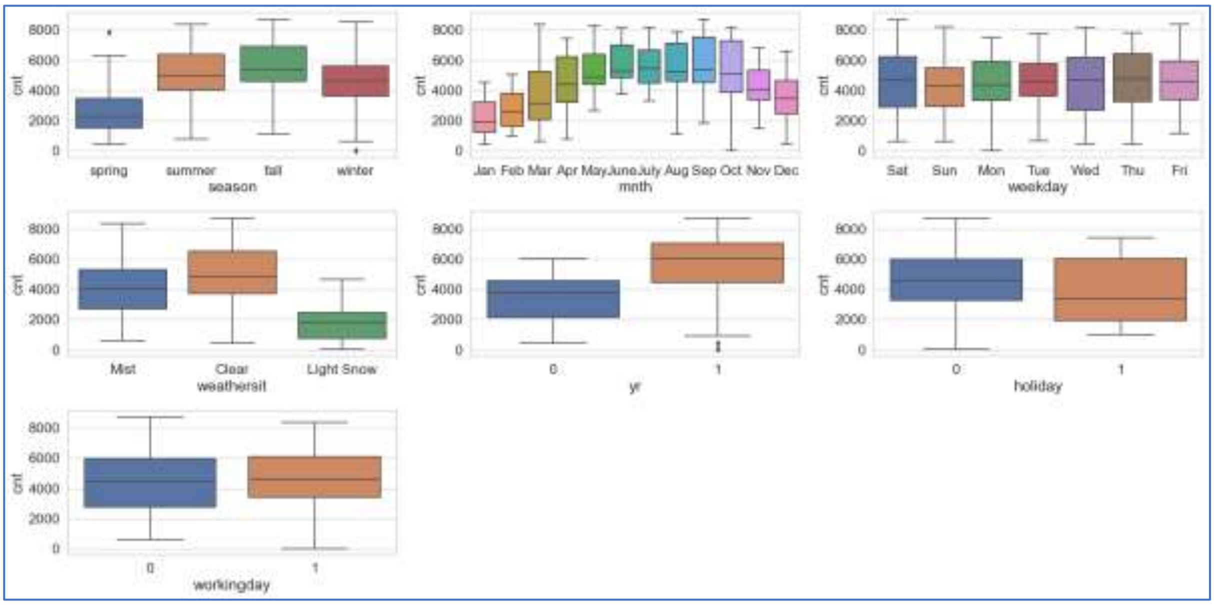
Subjective Questions

# I) Assignment-based Questions

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

   We can conclude the following based on the analysis of categorical variables from the dataset.



   Season: In fall & summer season the demand for shared bikes is high and in spring season shows least demand for shared bikes.

   Month: From Dec to Feb the demand for shared bike is very less in this season. With Sep month having highest while Jan has the lowest demand for shared bikes.

   Weekday: Wed, Thu, Fri and Sat show high trend in demand for shared bikes.

   Weather Situation: The demand for shared bikes is more when the weather is clear and there is least demand when there is snow fall.

   Year: In 2019, the demand for shared bikes increased compared to 2018. This suggests that as the situation returns to normal and quarantines and lockdowns are lifted, we can expect even higher demand for shared bikes in the coming year.

   Holiday: On holidays, bike demand tends to be less than on non-holidays.

   Working Day: Bike demand varies between working days and non-working days, with higher demand typically observed on working days.

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

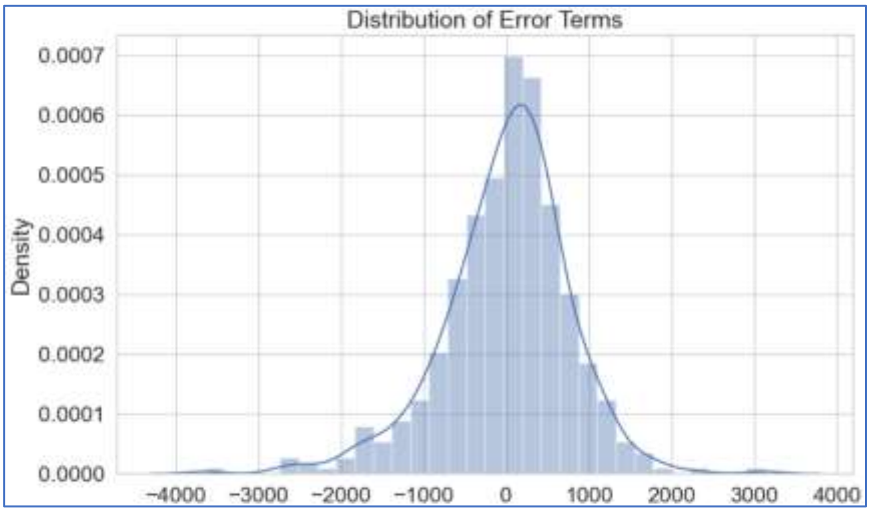
   When creating dummy variables for a categorical feature, each category is transformed into a separate binary column with 0s and 1s. If you have three categories, you end up with three columns. However, including all three columns can lead to issues because one column's values can be predicted from the others—a situation known as multicollinearity. By setting drop_first=True, you eliminate one of these columns. This simplifies your data, prevents confusion in your model, and helps avoid multicollinearity while reducing dimensionality.

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

   Temp and atemp is highly correlated with the target variable demand for shared bike(cnt).

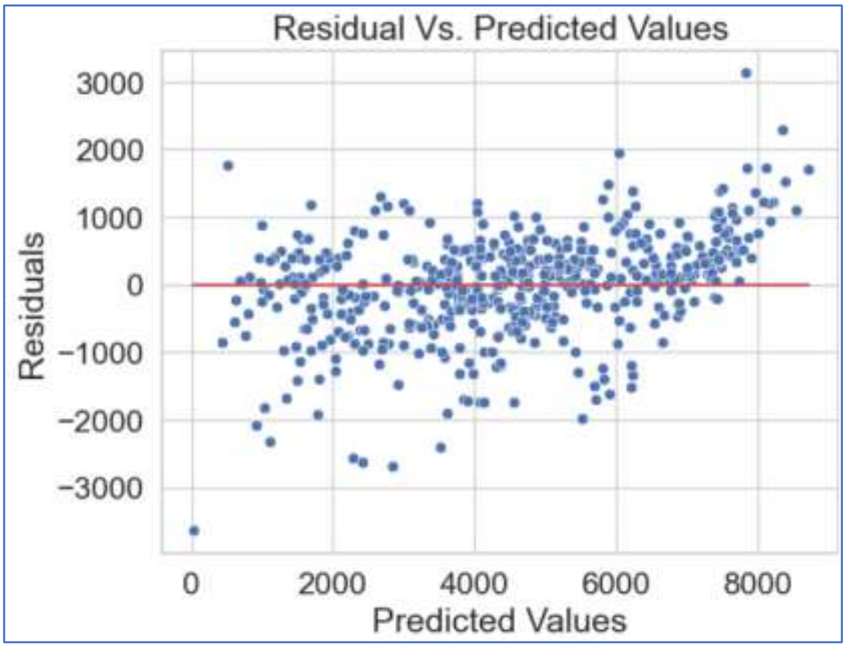
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Residual Analysis



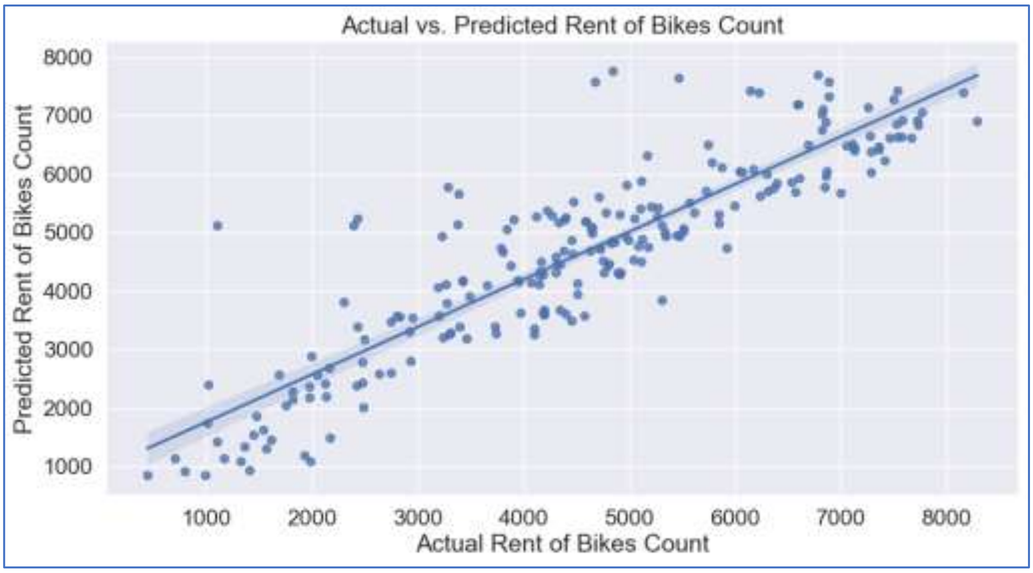
We can observe that the error terms follow a normal distribution.

Homoscedasticity



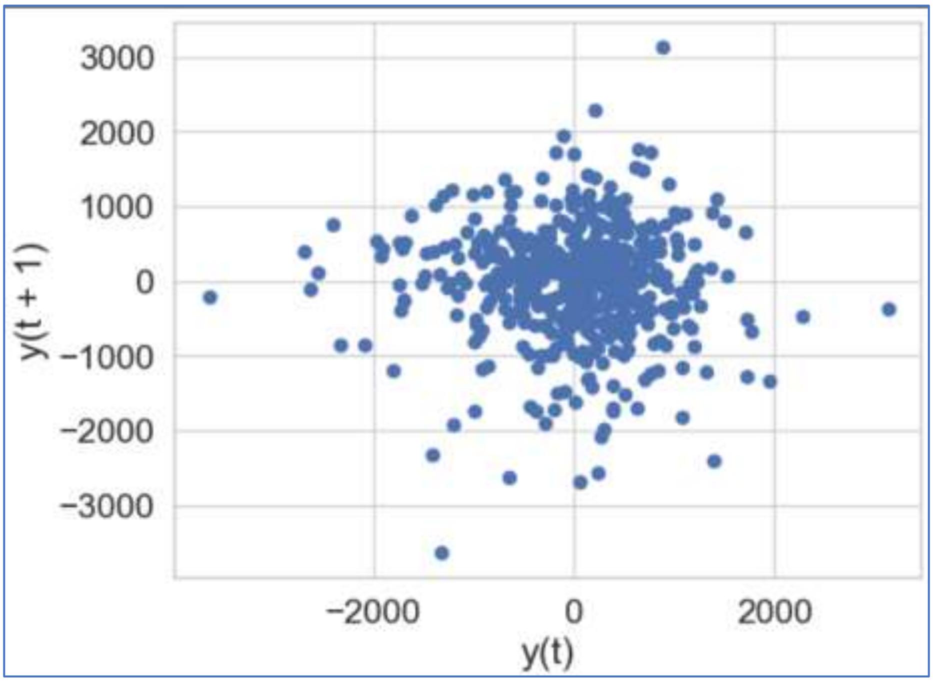
We can observe that there is a constant deviation from the zero line, and there are no visible patterns in the error terms.

Linearity



Most the data points are around the Actual Vs Predicted Line

Error term Trend verification



Lagplot of residuals shows no trend. Hence the error terms have constant variance.

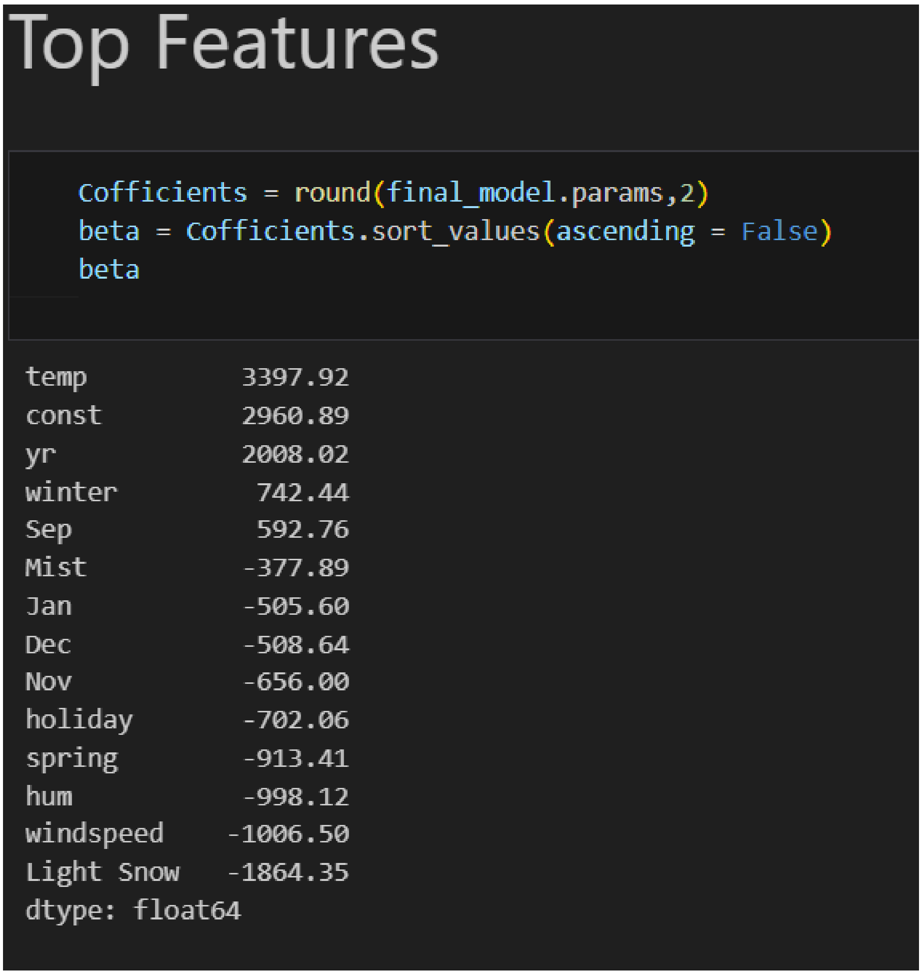
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top 3 features contributing significantly towards explaining the demand of the shared bikes are temperature(temp), year (yr), Spring season (season_spring)

Temperature (temp): Temperature has the most significant positive impact on bike rental demand. As the temperature rises, the demand for bike rentals increases significantly.

Year (yr): The year has a positive impact on demand. Over time, there has been an increasing trend in bike rentals.

Winter Season (winter): The winter season is associated with increased demand, due to recreational winter activities and holidays.



Top Features

```
Cofficients = round(final_model.params,2)
beta = Cofficients.sort_values(ascending = False)
beta
```

```
temp          3397.92
const         2960.89
yr            2008.02
winter         742.44
Sep            592.76
Mist          -377.89
Jan           -505.60
Dec           -508.64
Nov           -656.00
holiday       -702.06
spring        -913.41
hum           -998.12
windspeed    -1006.50
Light Snow   -1864.35
dtype: float64
```

# II) General Subjective Questions

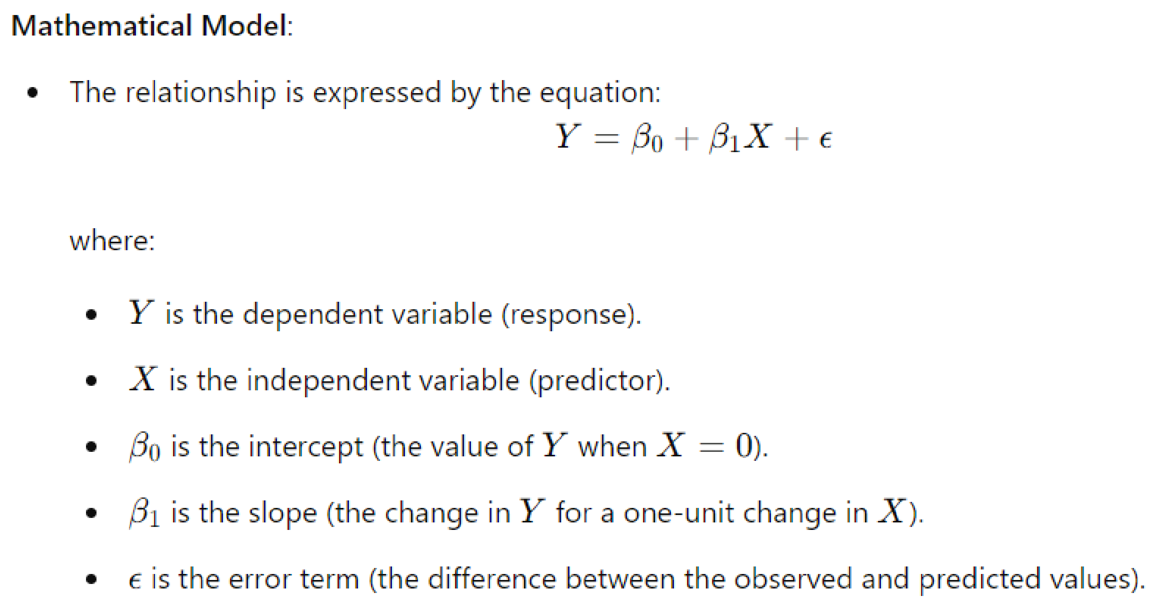
1. Explain the linear regression algorithm in detail

Linear regression is a statistical method used to model and analyse the relationships between a dependent variable and one or more independent variables. Its goal is to find a linear relationship that best explains or predicts the dependent variable based on the values of the independent variables.

It uses the following model representations:

1. Simple Linear Regression

Simple Linear Regression involves modelling the relationship between a single independent variable X and a dependent variable Yusing a straight line.



Mathematical Model:

- The relationship is expressed by the equation:

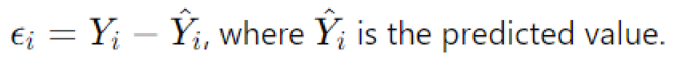
$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y is the dependent variable (response).
- X is the independent variable (predictor).
- β_0 is the intercept (the value of Y when $X = 0$).
- β_1 is the slope (the change in Y for a one-unit change in X).
- ϵ is the error term (the difference between the observed and predicted values).

Objective Function:

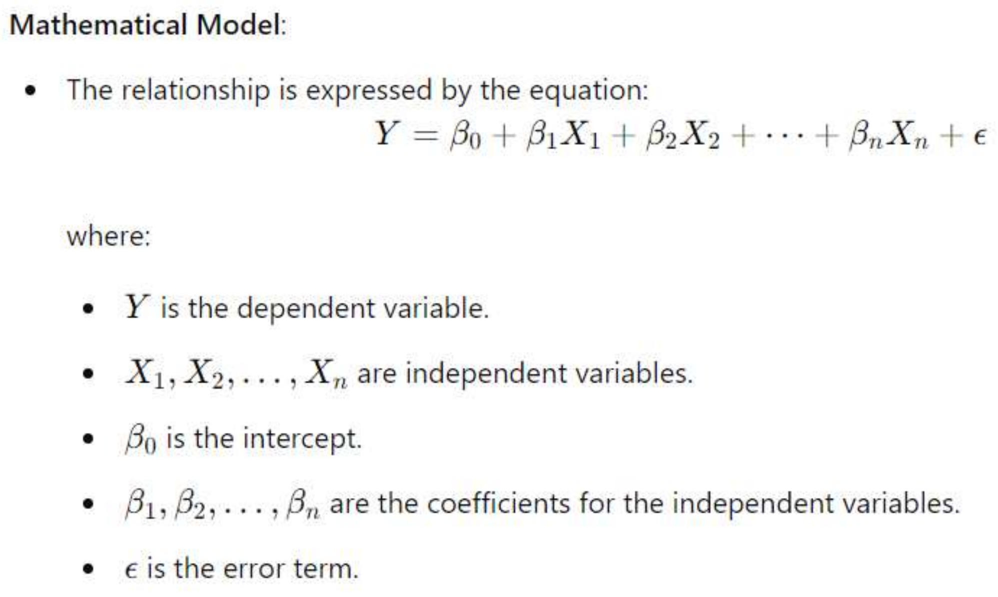
The objective is to find the values of β0 and β1 that minimize the sum of squared residuals (errors). The residual for each observation is



$\epsilon_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i$ is the predicted value.

2. Multiple Linear Regression

Multiple Linear Regression extends simple linear regression to include two or more independent variables. It models the relationship between the dependent variable and several predictors.



Mathematical Model:

- The relationship is expressed by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

where:

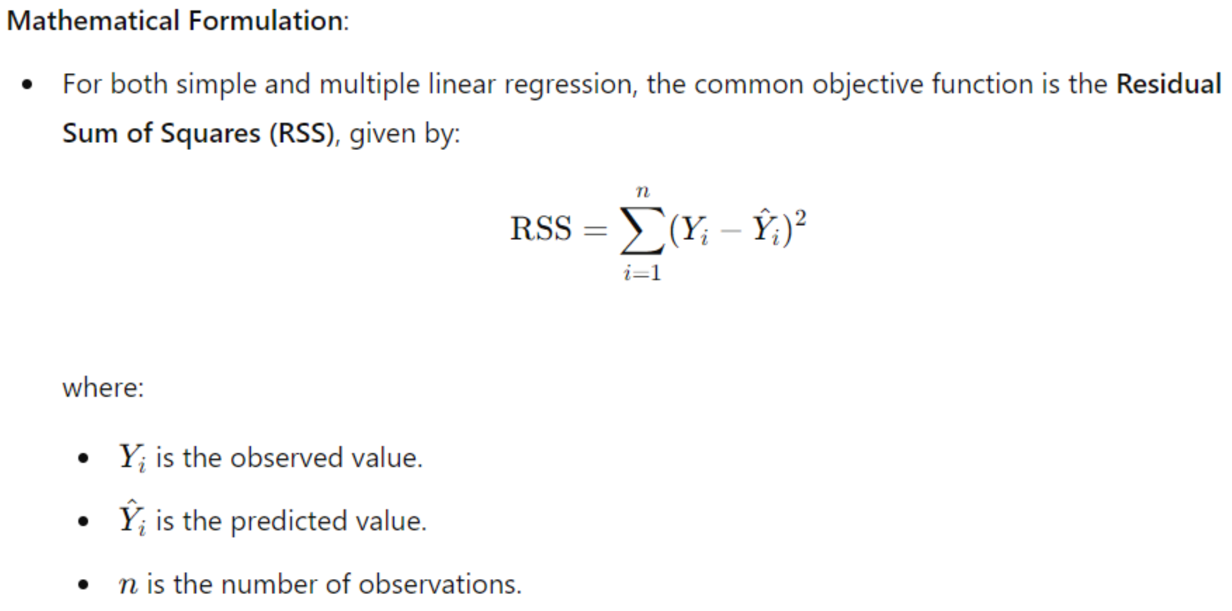
- Y is the dependent variable.
- $X_1, X_2, \ldots, X_n$ are independent variables.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for the independent variables.
- ϵ is the error term.

Objective Function:

The goal is to find the coefficients β0,β1,…,βn that minimize the sum of squared residuals. The residuals are the differences between the observed values and the predicted values based on the model.

3. Objective Function

The Objective Function in linear regression is a mathematical function that measures the discrepancy between the observed data and the values predicted by the model. The goal is to minimize this function to find the best-fitting model.



Mathematical Formulation:

- For both simple and multiple linear regression, the common objective function is the **Residual Sum of Squares (RSS)**, given by:

$$\text{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where:

- Y_i is the observed value.
- $\hat{Y}_i$ is the predicted value.
- n is the number of observations.

Optimization:

The coefficients β0,β1,…,βn are determined by minimizing the RSS. This is typically done using techniques such as Ordinary Least Squares (OLS), which provides the best linear unbiased estimates of the coefficients.

Assumptions

The relationship between the independent variables and the dependent variable is linear.

The error terms (ϵ) are independent of each other.

The variance of the error terms is constant across all levels of the independent variables.

The error terms follow a normal distribution.

Independent variables are not highly correlated with each other.

Interpretation

Linear regression allows for the interpretation of the coefficients β0,β1,β2,…,βp. For example, β1 represents the change in the dependent variable Y for a one-unit change in X1, holding all other variables constant.

Evaluation

To assess the performance of a linear regression model, various metrics can be used, such as Mean Squared Error (MSE), R-squared (R2), and others, to measure how well the model fits the data and makes predictions.

Predictions

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

Extensions

Linear regression has several extensions and variations, including ridge regression, lasso regression, and polynomial regression, which address issues like multicollinearity and allow for more flexible modeling.

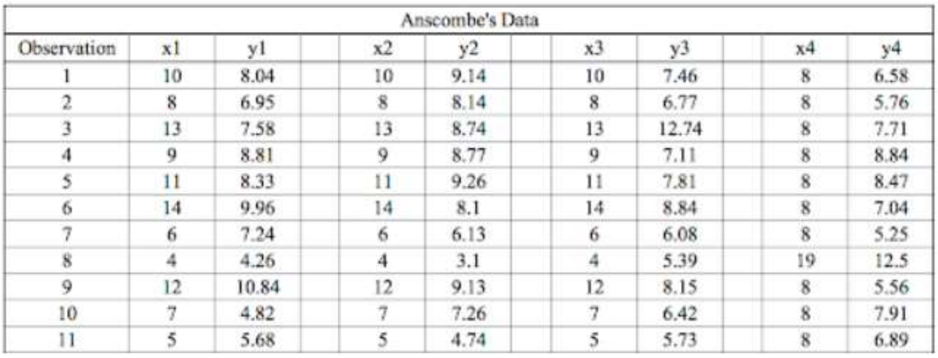
Linear regression is a powerful and interpretable tool commonly used for tasks such as predicting house prices, analyzing the impact of variables on an outcome, and understanding relationships in data. However, it has its limitations, and its effectiveness depends on the assumptions being met and the nature of the data being modelled.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that illustrate the importance of graphing data before analyzing it, highlighting how different datasets can have the same statistical properties but very different distributions and relationships when visualized.

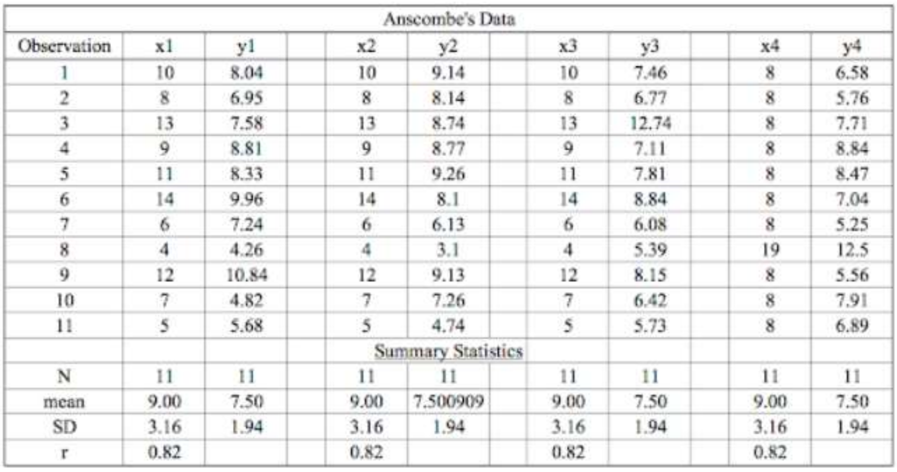
Anscombe's quartet was created by the statistician Francis Anscombe in 1973. The quartet consists of four datasets, each with 11 pairs of (x,y)(x, y)(x,y) values. Despite having nearly identical statistical properties, such as mean, variance, and correlation, the datasets exhibit very different distributions and relationships when plotted.

Anscombe's quartet emphasizes the need to visualize data before modeling. Plotting data features helps detect anomalies and ensures we understand its characteristics. Additionally, linear regression is suitable only for linear relationships; other data types require different approaches. We can define these four plots as follows:



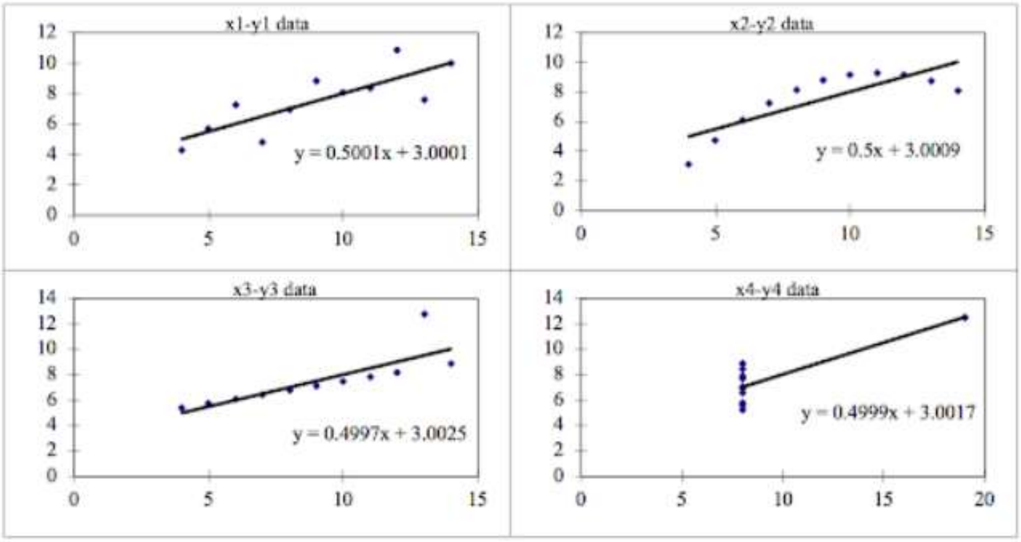
					Anscombe's Data						
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for these four data sets are approximately similar. We can compute them as follows:



					Anscombe's Data						
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Anscombe's quartet highlights the importance of visualizing data. Even when statistical summaries appear identical, visual exploration reveals hidden differences in data distributions and relationships.

3. What is Pearson's R?

The Pearson correlation coefficient, often denoted as R for samples is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is one of the most widely used correlation coefficients in statistics.

Range

The value of R ranges from −1 to 1.

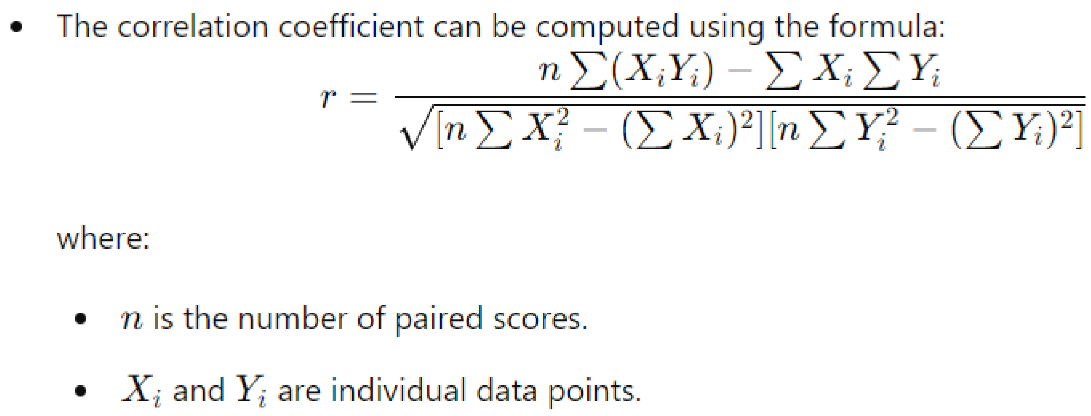
- r=1 indicates a perfect positive linear relationship (as one variable increases, the other variable increases proportionally).
- r=−1 indicates a perfect negative linear relationship (as one variable increases, the other variable decreases proportionally).
- r=0 indicates no linear relationship between the variables.

Interpretation

- **Positive Correlation**: R values greater than 0 indicate a positive linear relationship, where higher values of one variable tend to be associated with higher values of the other variable.
- **Negative Correlation**: R values less than 0 indicate a negative linear relationship, where higher values of one variable tend to be associated with lower values of the other variable.

- **Strength of Relationship**:
  - **0.1 to 0.3** (or −0.1to−0.3-0.1 to -0.3−0.1to−0.3): Weak correlation
  - **0.3 to 0.5** (or −0.3to−0.5-0.3 to -0.5−0.3to−0.5): Moderate correlation
  - **0.5 to 1.0** (or −0.5to−1.0-0.5 to -1.0−0.5to−1.0): Strong correlation

Calculation



- The correlation coefficient can be computed using the formula:

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

where:

- n is the number of paired scores.
- X_i and Y_i are individual data points.

Example:

Imagine you have data on students' study hours and their test scores. Calculating Pearson's correlation coefficient (r) between these variables helps reveal whether more study hours correlate with higher test scores.

In summary, r quantifies the linear relationship between continuous variables, but it's best used alongside other methods and visualizations for a comprehensive data understanding.

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

Scaling is a method used to make different features in a dataset comparable. It adjusts the values of each feature so they're in a similar range. This is important because:

1. It helps prevent features with larger numbers from dominating those with smaller numbers.
2. It makes it easier for machine learning algorithms to understand and use the data.
3. It can improve the performance and accuracy of many models.

Why is Scaling Performed?

1. Improving Model Performance: Many machine learning algorithms, especially those based on distance metrics (e.g., K-Nearest Neighbors, Support Vector Machines, and clustering algorithms), perform better when features are on a similar scale. This is because large differences in feature scales can disproportionately affect the distance calculations and, subsequently, the model performance.

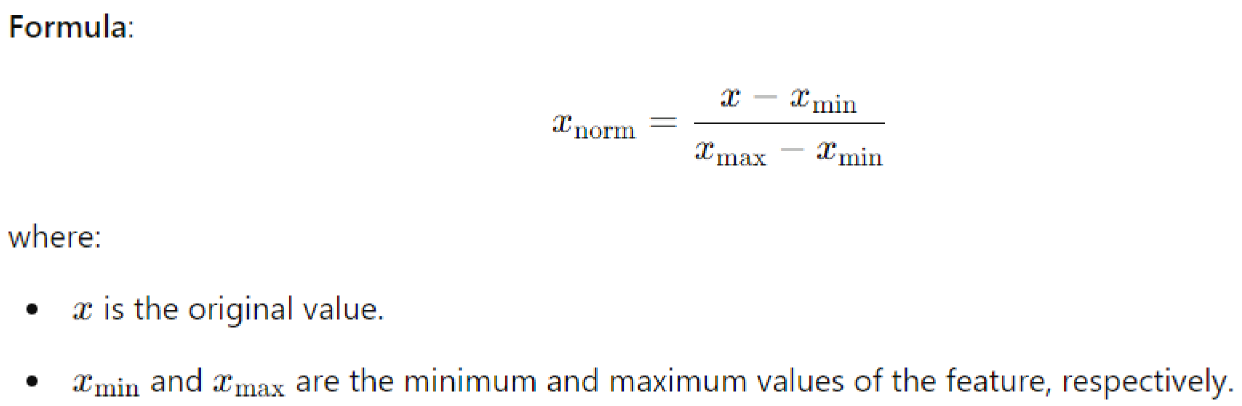
2. Convergence in Gradient Descent: For algorithms that use gradient descent (e.g., linear regression, neural networks), scaling can help in faster convergence. Features on different scales can cause the cost function to have an uneven surface, making it harder for the algorithm to find the optimal parameters efficiently.

3. Interpretability: Scaling can make the coefficients of a model more interpretable. When features are on similar scales, the magnitudes of the coefficients can be compared to understand the relative importance of each feature.

Types of Scaling

Normalized Scaling

Normalization (or Min-Max Scaling) transforms features to a specific range, usually between 0 and 1. This is done by adjusting the data based on its minimum and maximum values.



Formula:

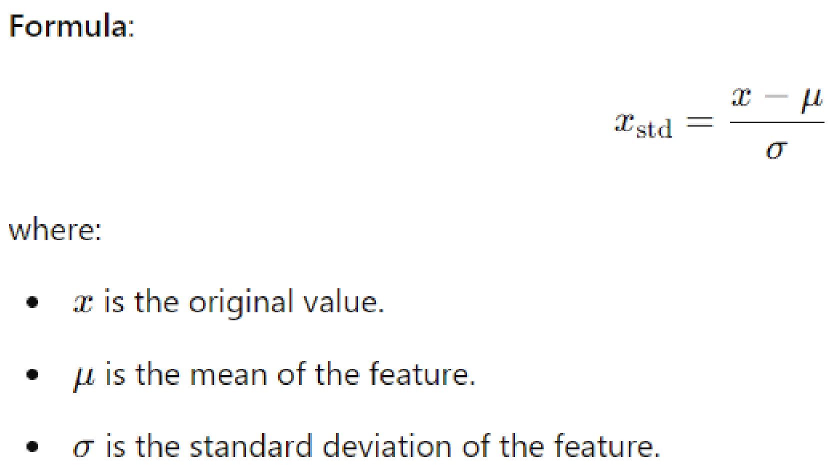
$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

where:

- x is the original value.
- x_{min} and x_{max} are the minimum and maximum values of the feature, respectively.

Standardized Scaling

Standardization (or Z-score normalization) transforms features to have a mean of 0 and a standard deviation of 1. This is done by adjusting the data based on its mean and standard deviation.



Formula:

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

where:

- x is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

Feature	**Normalization (Min-Max)**	**Standardization (Z-score)**
Scaling Method	Uses min and max values	Uses mean and standard deviation
Range	Bounded to [0, 1]	Unbounded
Sensitivity to Outliers	Highly affected	Less affected
Use Case	When data has no outliers	When data has outliers

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

The value of VIF = infinity, shows that a perfect correlation between two independent variables.

In case of perfect correlation, we get R-square =1, which leads to 1/(1-R-square) infinity.

An infinity value of VIF indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Before starts building a multiple linear regression we will do many assumptions, in that "issue of multicollinearity" is very important.

We will assume that there is no multicollinearity, that means the selected independent variables are nor correlated with any of the other selected independent variables.

But if there is perfect correlation between independent variables, the value of that particular variables VIF becomes infinity.

To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity or if it redundant or use regularization technique like ridge regression.

In short, an infinite VIF signals perfect multicollinearity among predictors. Fixing this issue is essential for reliable and interpretable regression models.

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

A Q-Q plot is a visual tool that compares a dataset's distribution to a theoretical one, usually the normal distribution. It does this by plotting the dataset's quantiles against the theoretical distribution's quantiles.

How to interpret a Q-Q plot:

1. Points following a straight line suggest the data matches the theoretical distribution.
2. Deviations from this line indicate differences:
   - Upward curve: Right-skewed data
   - Downward curve: Left-skewed data
   - S-shape: Heavy-tailed distribution

In linear regression, Q-Q plots are important for:

1. Checking if residuals are normally distributed, a key assumption in linear regression.
2. Spotting outliers that could affect the model.
3. Validating the model's appropriateness.

For example, if you plot the residuals from a linear regression model and they form a straight line on the Q-Q plot, it suggests the residuals are normally distributed, supporting the model's validity.

Q-Q plots offer a simple yet powerful way to visually assess if data follows a specific distribution, which is crucial for ensuring the reliability of linear regression models and their interpretations.