Linear Regression

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

- mnth : month (1 to 12)
- weathersit :
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- weekday : day of the week
- season : season (1:spring, 2:summer, 3:fall, 4:winter)

for above variables dummy variables are created because actual number does not meaningful in this context.

linear model parameters

const	0.116369
yr	0.232232
temp	0.551183
windspeed	-0.162267
season_4	0.123801
mnth_3	0.057427
mnth_4	0.099956
mnth_5	0.096907
mnth_8	0.040396
mnth_9	0.104841
mnth_10	0.030138
weekday_6	0.023881
weathersit_2	-0.079954
weathersit_3	-0.285931

Categorical variable effect:
yr : 0.232232, season_4 : 0.123801, mnth_3 : 0.057427, mnth_4 : 0.099956, mnth_5 : 0.096907, mnth_8 : 0.040396, mnth_9 : 0.104841, mnth_10 : 0.030138, weekday_6 : 0.023881, weathersit_2 : -0.079954, weathersit_3   -0.285931

Categorical variables effects final prediction values as per their coefficient values.

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

Using `drop_first=True` when creating dummy variables (also known as one-hot encoding) is important to avoid multicollinearity, specifically the dummy variable trap, in regression models.

Dummy Variable Trap

The dummy variable trap occurs when the dummy variables created from a categorical variable are perfectly collinear. This means that one dummy variable can be perfectly predicted from the others. In other words, if you include all categories as dummy variables, you introduce perfect multicollinearity, which can lead to issues in regression analysis because:

1. **Singular Matrix**: The design matrix XXX becomes singular, meaning it doesn't have full rank. This makes the matrix non-invertible, and as a result, the regression coefficients cannot be uniquely estimated.
2. **Inflated Variance**: Perfect multicollinearity leads to infinite or very high Variance Inflation Factor (VIF) values, making the estimates of the regression coefficients highly unstable and unreliable.

Using `drop_first=True` when creating dummy variables is important to:

1. **Avoid Perfect Multicollinearity**: Prevents the dummy variable trap, ensuring the design matrix is invertible.
2. **Stable and Reliable Estimates**: Helps in obtaining stable and reliable estimates of regression coefficients by avoiding inflated variances.
3. **Simpler Interpretation**: Provides a clear reference category for interpretation of the regression coefficients.

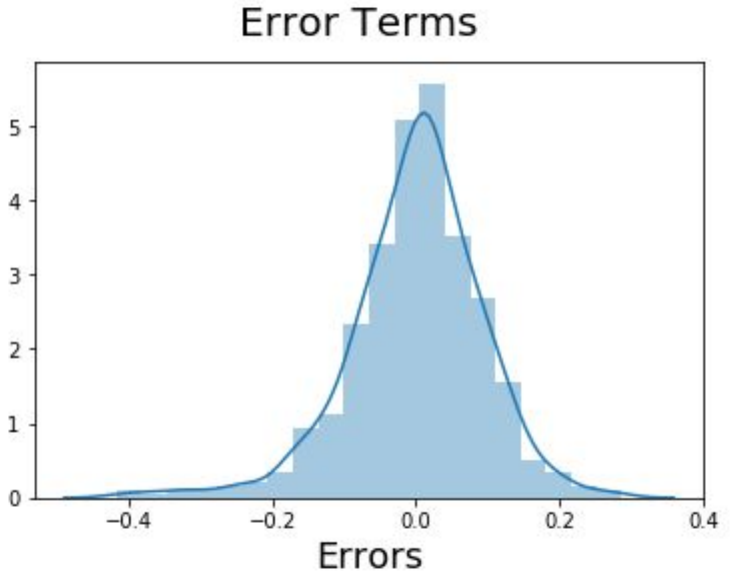
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temp: 0.63 atemp: 0.63 yr:0.57
These variables are having highest similarity in order.

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

Validating the assumptions of linear regression is crucial to ensure that the model is appropriate for the data and that the results can be trusted. The key assumptions of linear regression are:

1. **Linearity**: The relationship between the predictors and the response variable is linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: Constant variance of the residuals (errors).
4. **Normality**: The residuals are normally distributed.
5. **No Multicollinearity**: Predictors are not too highly correlated with each other.



Normality: residuals normally distributed

No Multicollinearity : through VIF factor

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

linear model parameters

const	0.116369
yr	0.232232
temp	0.551183
windspeed	-0.162267
season_4	0.123801
mnth_3	0.057427
mnth_4	0.099956
mnth_5	0.096907
mnth_8	0.040396
mnth_9	0.104841
mnth_10	0.030138
weekday_6	0.023881
weathersit_2	-0.079954
weathersit_3	-0.285931

Top 3 features:

yr	0.232232
temp	0.551183
windspeed	-0.162267

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable (continuous variable) and one or more independent variables. The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that describes the relationship between these variables.

Concept

Linear regression assumes that the relationship between the dependent variable yyy and the independent variables XXX can be approximated by a linear equation:

y=β0+β1x1+β2x2+…+βnxn+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilony=β0+β1x1+β2x2+…+βnxn+ϵ

- yyy is the dependent variable (the outcome we are trying to predict).
- x1,x2,…,xnx_1, x_2, \ldots, x_nx1,x2,…,xn are the independent variables (the features or predictors).
- β0\beta_0β0 is the intercept of the regression line.
- β1,β2,…,βn\beta_1, \beta_2, \ldots, \beta_nβ1,β2,…,βn are the coefficients (slopes) of the independent variables.
- ϵ\epsilonϵ is the error term (the difference between the actual and predicted values).

Assumptions

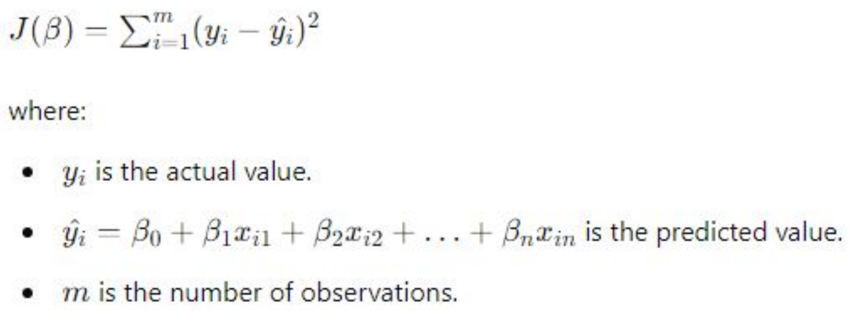
Linear regression relies on several key assumptions:

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: The observations are independent of each other.
3. **Homoscedasticity**: The residuals (errors) have constant variance at every level of the independent variables.
4. **Normality**: The residuals of the model are normally distributed.

3. The Objective

The objective of linear regression is to find the coefficients β\betaβ that minimize the sum of the squared differences between the observed values and the values predicted by the linear model. This method is known as **Ordinary Least Squares (OLS)**.

The cost function for OLS is:



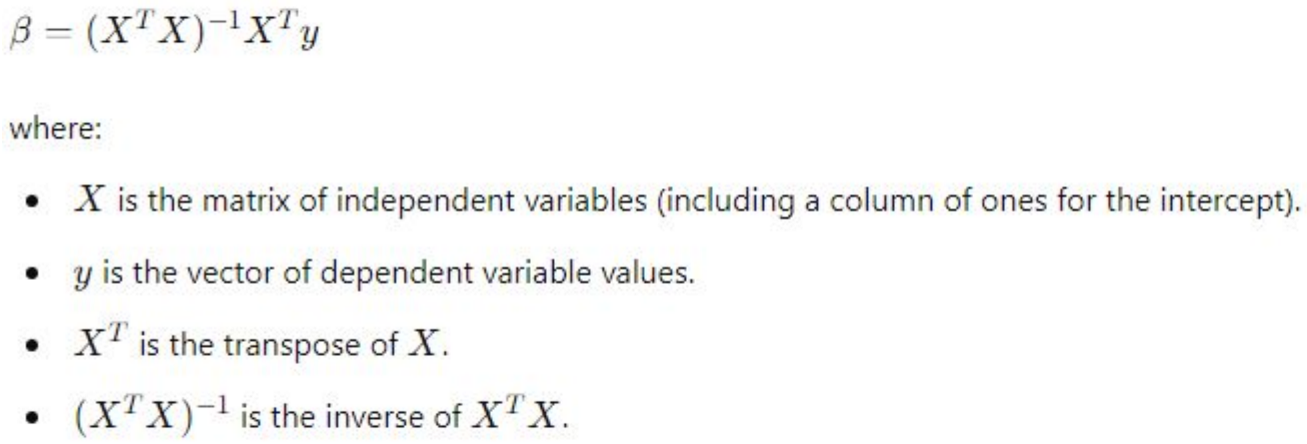
$$J(\beta) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

where:

- y_i is the actual value.
- $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in}$ is the predicted value.
- m is the number of observations.

Solving for Coefficients

The coefficients are estimated using the following formula:



$$\beta = (X^T X)^{-1} X^T y$$

where:

- X is the matrix of independent variables (including a column of ones for the intercept).
- y is the vector of dependent variable values.
- X^T is the transpose of X.
- $(X^T X)^{-1}$ is the inverse of $X^T X$.

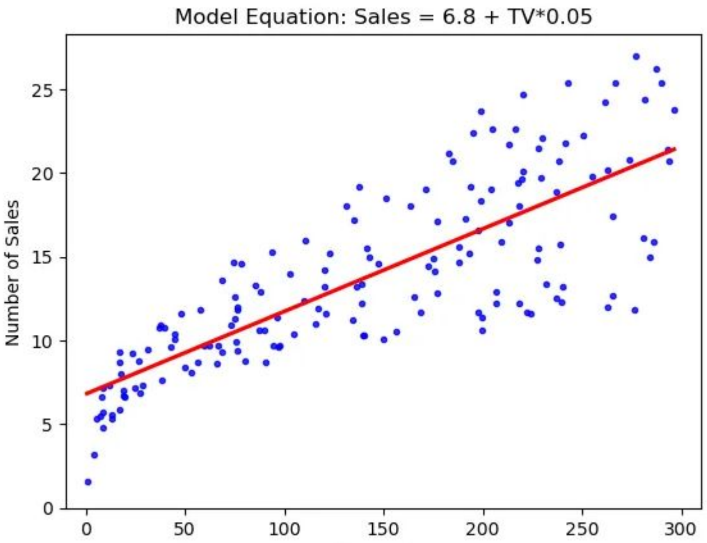
Interpretation of Coefficients

- **Intercept (β0\beta_0β0)**: Represents the expected value of yyy when all xix_ixi are 0.
- **Slope (βi\beta_iβi)**: Represents the change in yyy for a one-unit change in xix_ixi, holding all other variables constant.

Evaluation Metrics

Common metrics to evaluate the performance of a linear regression model include:

- **R-squared (R2R^2R2)**: Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Mean Squared Error (MSE)**: The average of the squared differences between the observed and predicted values.
- **Root Mean Squared Error (RMSE)**: The square root of MSE, providing a measure of the average magnitude of the errors.



2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

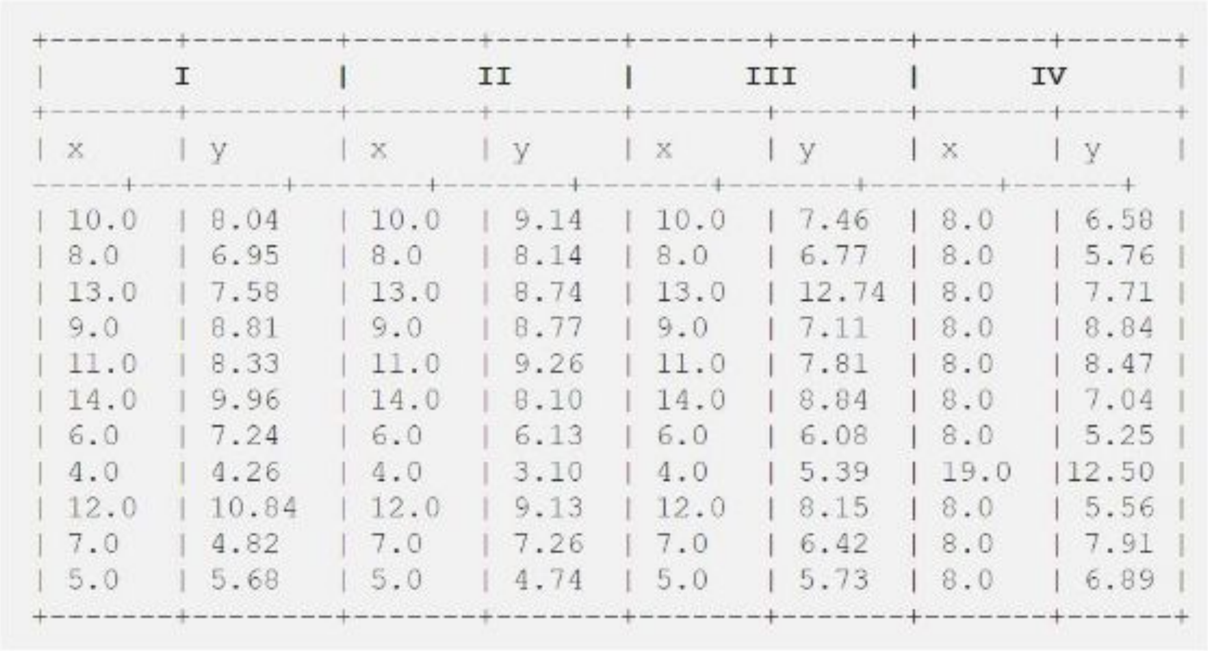
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

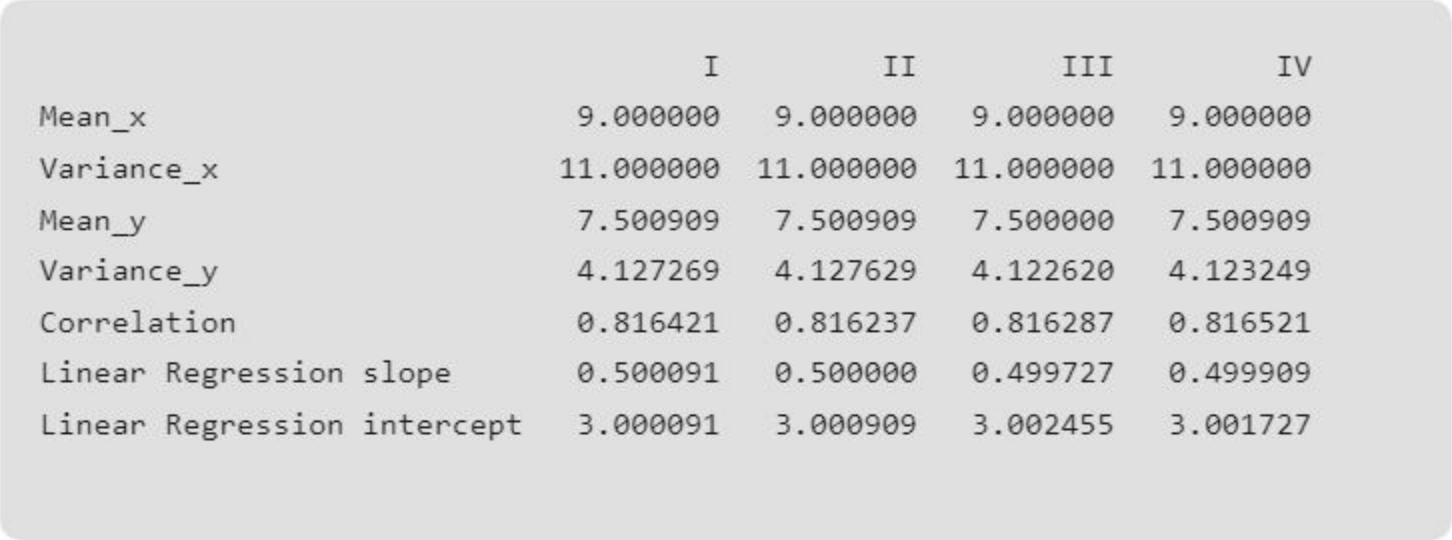
Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe’s Quartet Dataset

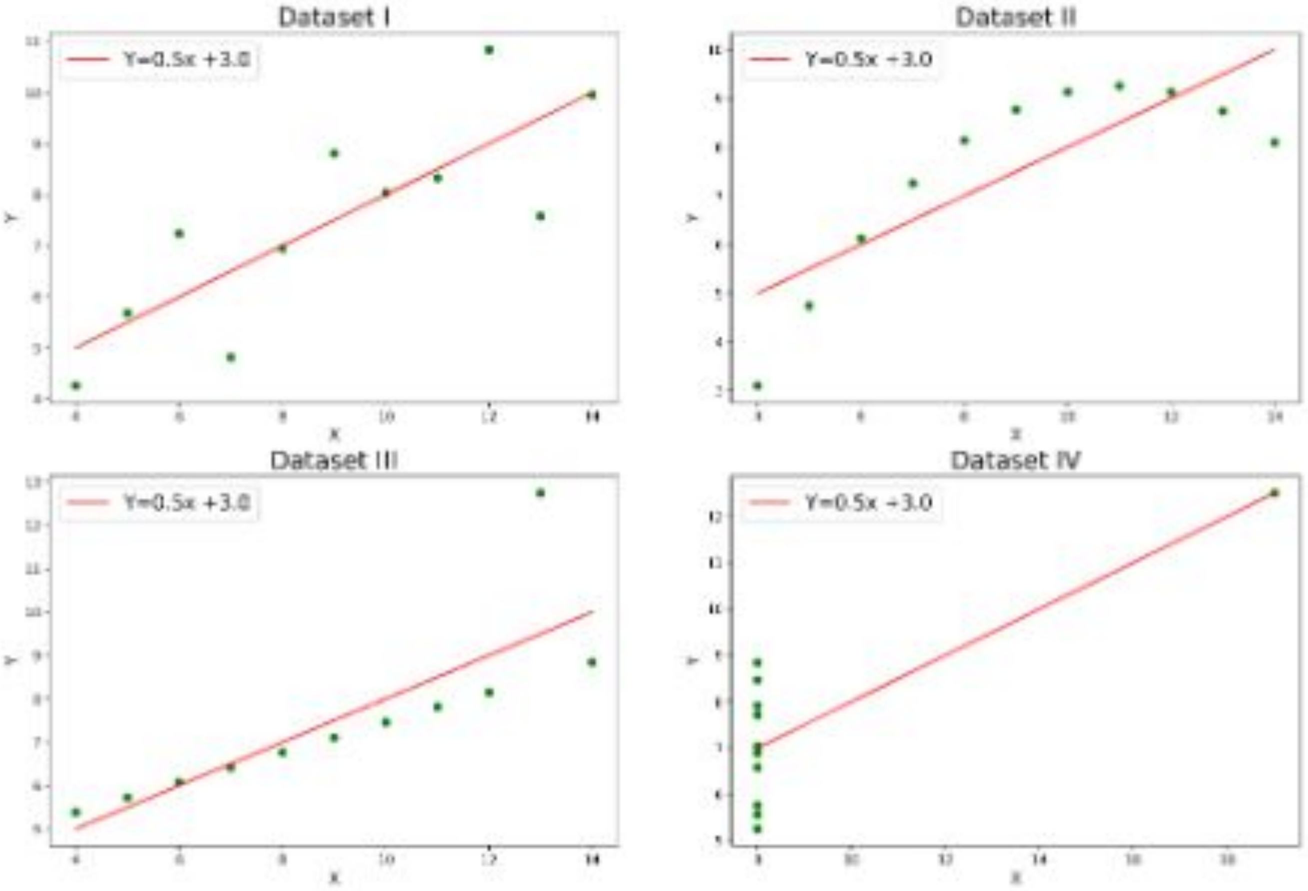
The four datasets of **Anscombe’s quartet.**



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



```
                                     I         II        III         IV
Mean_x                        9.000000   9.000000   9.000000   9.000000
Variance_x                   11.000000  11.000000  11.000000  11.000000
Mean_y                        7.500909   7.500909   7.500000   7.500909
Variance_y                    4.127269   4.127629   4.122620   4.123249
Correlation                   0.816421   0.816237   0.816287   0.816521
Linear Regression slope       0.500091   0.500000   0.499727   0.499909
Linear Regression intercept   3.000091   3.000909   3.002455   3.001727
```


Dataset I
Y=0.5x +3.0
Y
X
4
5
6
7
8
9
10
11
4
6
8
10
12
14
Dataset II
Y=0.5x +3.0
Y
X
3
4
5
6
7
8
9
10
4
6
8
10
12
14
Dataset III
Y=0.5x +3.0
Y
X
5
6
7
8
9
10
11
12
13
4
6
8
10
12
14
Dataset IV
Y=0.5x +3.0
Y
X
5
6
7
8
9
10
11
12
8
10
12
14
16
18

3. What is Pearson’s R?

The Pearson correlation coefficient is a statistical measure used to evaluate the linear relationship between two variables. Ranging from -1 to 1, it indicates the strength and direction of the relationship. When X and Y are identical (X = Y), the correlation coefficient, identified by the symbol “r,” assumes a value of 1.

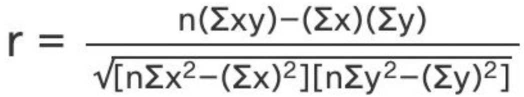
The Pearson correlation coefficient is a way to indicate the extent to which a change in one variable corresponds to a change in another variable. Its values fall within the range of -1 to 1:

- A correlation of **1** indicates a perfect **positive linear relationship.**
- A correlation of **-1** signifies a perfect **negative linear relationship.**
- A correlation of **0** denotes **no linear relationship** between the variables.

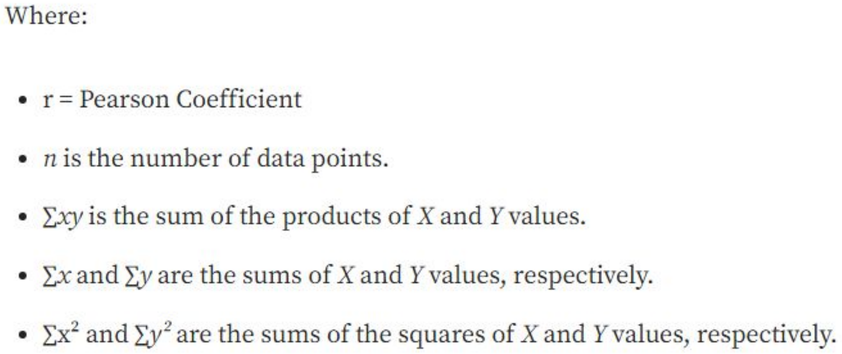
X = Y: Exploring Perfect Correlation

When two variables, X and Y, are identical (X = Y), they have a perfect positive linear relationship. This means that for every unit increase in X, there is a corresponding unit increase in Y. Graphically, these variables align perfectly along a straight line with a slope of 1, passing through the origin.

The formula for the Pearson correlation coefficient is given by:



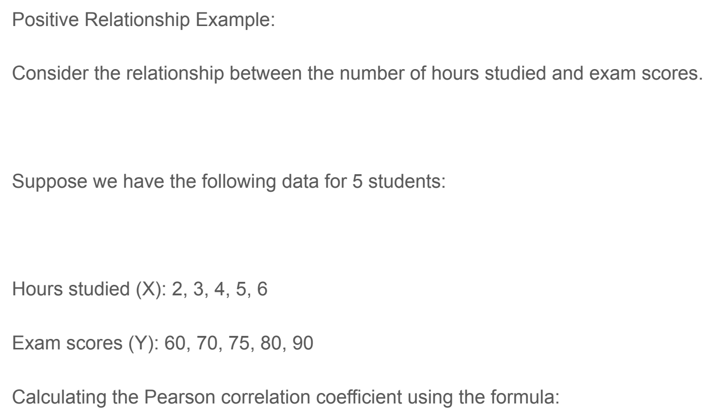
$$r = \frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2-(\Sigma x)^2][n\Sigma y^2-(\Sigma y)^2]}}$$



Where:

- r = Pearson Coefficient
- *n* is the number of data points.
- $\sum xy$ is the sum of the products of *X* and *Y* values.
- $\sum x$ and $\sum y$ are the sums of *X* and *Y* values, respectively.
- $\sum x^2$ and $\sum y^2$ are the sums of the squares of *X* and *Y* values, respectively.

When X and Y are equal, the numerator in the Pearson correlation coefficient formula becomes the sum of squares of the deviation from the mean, resulting in a perfect positive correlation. Therefore, the correlation coefficient, in this case, is 1.



Positive Relationship Example:

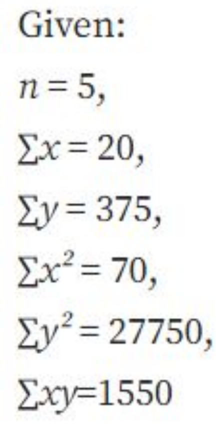
Consider the relationship between the number of hours studied and exam scores.

Suppose we have the following data for 5 students:

Hours studied (X): 2, 3, 4, 5, 6

Exam scores (Y): 60, 70, 75, 80, 90

Calculating the Pearson correlation coefficient using the formula:



Given:

$n = 5,$

$\sum x = 20,$

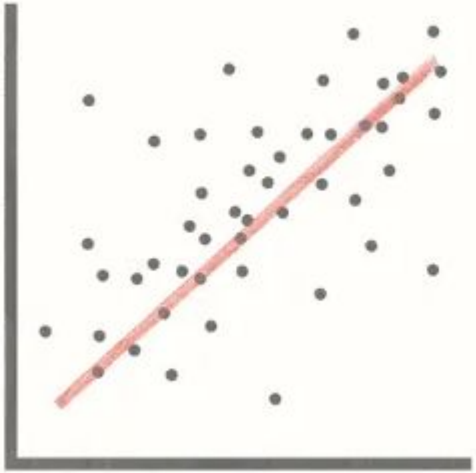
$\sum y = 375,$

$\sum x^2 = 70,$

$\sum y^2 = 27750,$

$\sum xy = 1550$

After calculation, let’s say the Pearson correlation coefficient (R) is approximately 0.96.A value close to 1 indicates a ***strong positive linear relationship*** between hours studied and exam scores, indicating that as the number of hours studied increases, the exam scores also tend to increase, showing a ***strong positive correlation***.

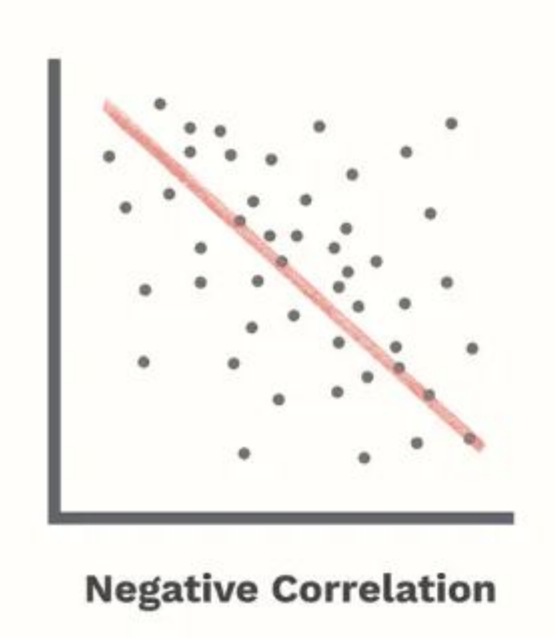


Positive Correlation

Negative Relationship Example:

Now, let’s consider a negative relationship between product price and quantity demanded. Suppose we have the following data for 5 products: Product price (X): 10, 15, 20, 25, 30, Quantity demanded (Y): 100, 80, 60, 40, 20 For this data set, after the calculations, let’s say the Pearson correlation coefficient (R) is approximately -0.95.

A value close to -1 indicates a ***strong negative linear relationship*** between product price and quantity demanded. This indicates that as the product price increases, the quantity demanded decreases, showing ***a strong negative correlation***.



In summary, the Pearson correlation coefficient is a valuable tool for understanding how two variables are related. When the coefficient is close to 1, it means they usually change in the same way. Conversely, when it’s near -1, they tend to change in opposite directions. And if it’s around 0, there might not be any clear connection between their changes. Understanding the properties and associations of the Pearson correlation coefficient helps in interpreting and analyzing the relationships between different variables in different fields such as statistics, economics, and science.

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

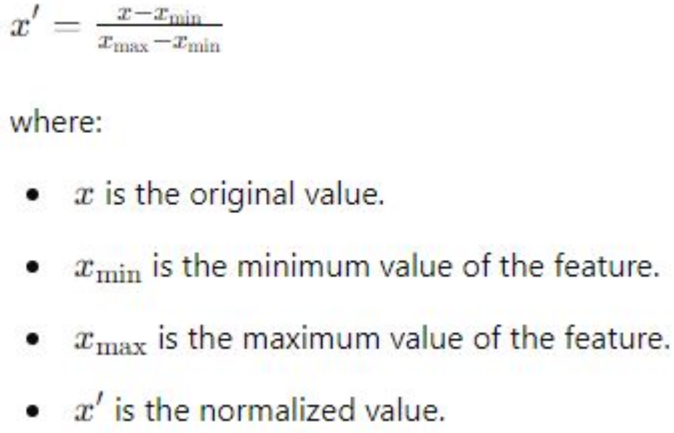
Scaling is a data preprocessing technique used in machine learning and statistical analysis to adjust the range and distribution of data features. The primary goal of scaling is to ensure that features contribute equally to the analysis, particularly when using algorithms that rely on distance measures (e.g., k-nearest neighbors, support vector machines) or gradient-based optimization (e.g., neural networks, linear regression).

Why Scaling is Performed

1. **Improves Convergence of Gradient Descent**: Algorithms like gradient descent converge faster when features are on a similar scale.
2. **Reduces the Effect of Outliers**: Scaling can reduce the influence of features with larger ranges, which can disproportionately affect the model.
3. **Improves Model Performance**: Many machine learning algorithms assume features are on a similar scale. Without scaling, models may perform poorly or give biased results.
4. **Ensures Equal Feature Contribution**: Features with larger values can dominate the learning process, leading to biased model training.

Normalization (Min-Max Scaling)

Normalization, also known as Min-Max scaling, rescales the features to a fixed range, usually [0, 1]. The formula for normalization is:



$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

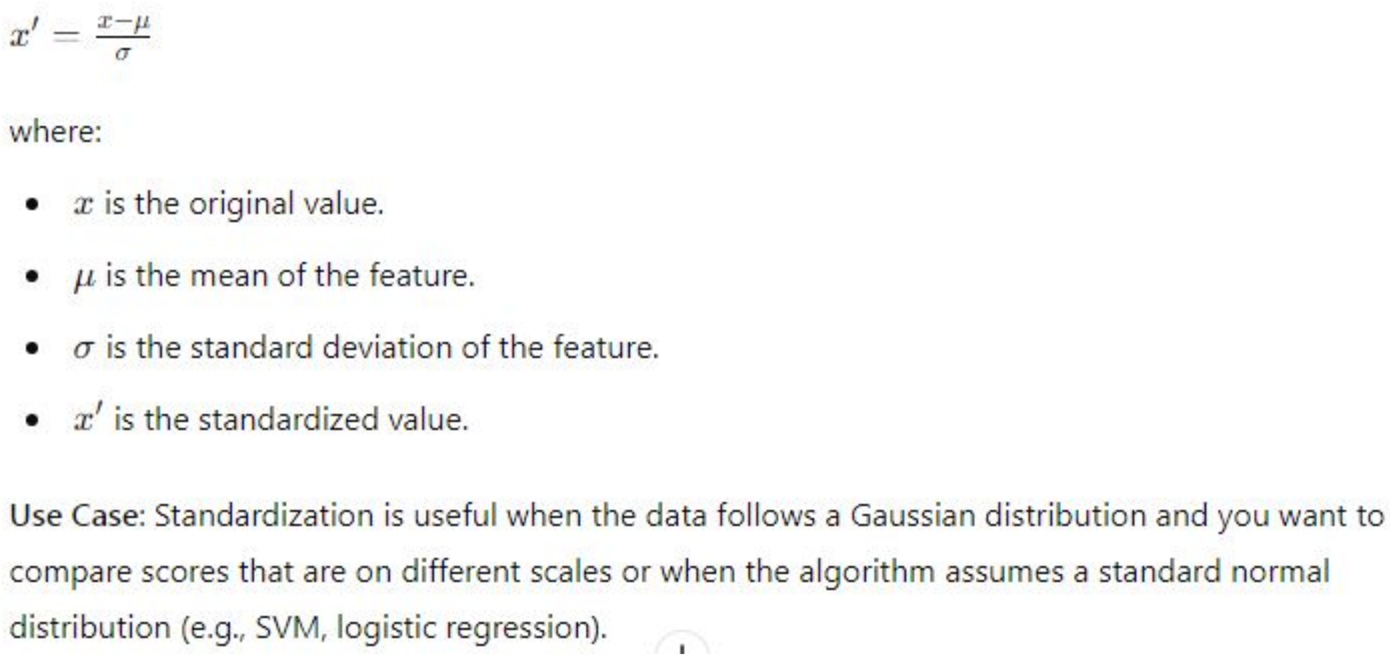
where:

- x is the original value.
- $x_{\min}$ is the minimum value of the feature.
- $x_{\max}$ is the maximum value of the feature.
- x' is the normalized value.

Use Case: Normalization is useful when the data does not follow a Gaussian distribution or when the features have different ranges and you want to bring them to a common scale.

Standardization (Z-score Scaling)

Standardization scales the features to have a mean of 0 and a standard deviation of 1. The formula for standardization is:



$$x' = \frac{x-\mu}{\sigma}$$

where:

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

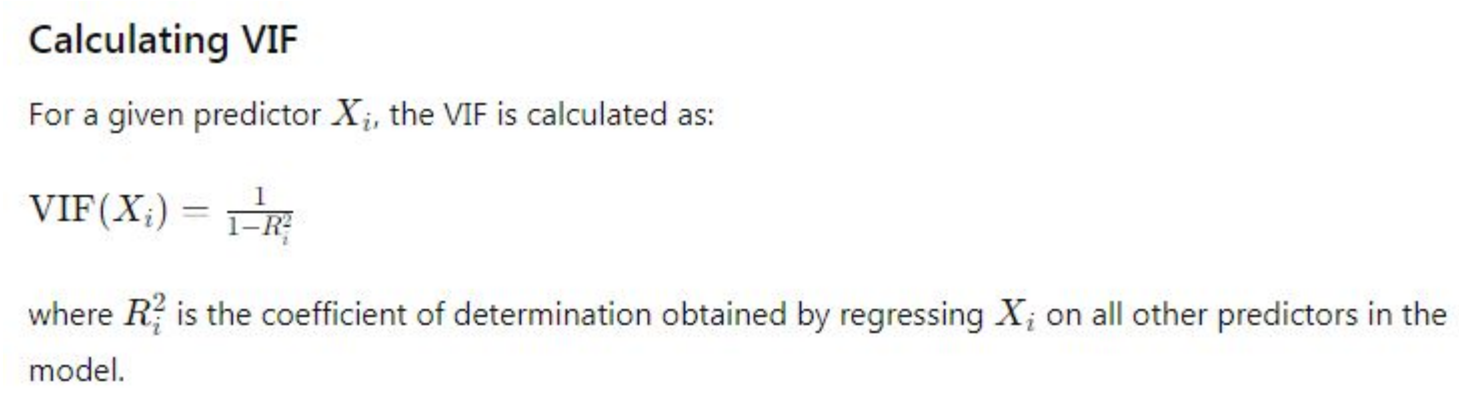
Use Case: Standardization is useful when the data follows a Gaussian distribution and you want to compare scores that are on different scales or when the algorithm assumes a standard normal distribution (e.g., SVM, logistic regression).

Difference Between Normalized Scaling and Standardized Scaling

1. **Range**:
   - **Normalization**: Rescales features to a specific range, usually [0, 1].
   - **Standardization**: Transforms features to have a mean of 0 and a standard deviation of 1.
2. **Effect on Distribution**:
   - **Normalization**: Does not change the shape of the original distribution of data.
   - **Standardization**: Adjusts the distribution to be centered around 0 with a spread defined by the standard deviation.
3. **Use Cases**:
   - **Normalization**: Preferred when data does not follow a Gaussian distribution or when features are on different scales.
   - **Standardization**: Preferred when data follows a Gaussian distribution or when the algorithm assumes a standard normal distribution.
4. **Sensitivity to Outliers**:
   - **Normalization**: Can be heavily influenced by outliers since it uses min and max values.
   - **Standardization**: Less sensitive to outliers as it uses mean and standard deviation.

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

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity.



Calculating VIF

For a given predictor X_i, the VIF is calculated as:

$$\mathrm{VIF}(X_i) = \frac{1}{1-R_i^2}$$

where R_i^2 is the coefficient of determination obtained by regressing X_i on all other predictors in the model.

Infinite VIF Values

A VIF value becomes infinite when Ri2R_i^2Ri2 is equal to 1. This situation occurs when there is perfect multicollinearity, meaning one predictor is an exact linear combination of one or more of the other predictors. In this case, the denominator in the VIF formula becomes zero, leading to an infinite VIF value.

Reasons for Infinite VIF

1. **Perfect Linear Dependence**:
   - When one predictor variable is perfectly correlated with another (or a combination of other variables), Ri2R_i^2Ri2 becomes 1. This perfect linear relationship causes the VIF to be infinite.
2. **Dummy Variable Trap**:
   - In the case of dummy variables (categorical variables converted into binary indicators), if you include all categories without dropping one, you introduce perfect multicollinearity. For instance, if you have a categorical variable with 3 categories (A, B, and C), you should only include two dummy variables (e.g., A and B). Including all three will lead to perfect multicollinearity because the third can be perfectly predicted from the other two.

Addressing Infinite VIF

1. **Remove Perfectly Collinear Variables**:
   - Identify and remove one of the perfectly collinear variables. This can be done using correlation matrices or by inspecting VIF values.
2. **Drop One Dummy Variable**:
   - In the case of dummy variables, always drop one category to avoid the dummy variable trap.
3. **Principal Component Analysis (PCA)**:
   - Use PCA to transform the predictors into a set of uncorrelated components. This method is useful when you have multiple collinear predictors.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

Interpretation:

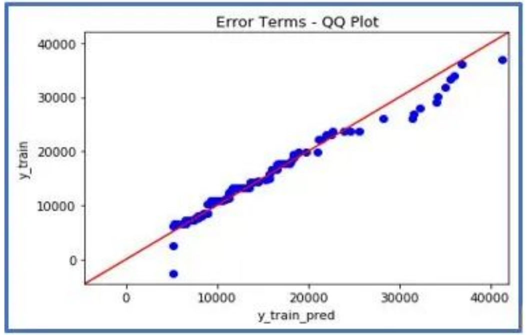
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.Below are the possible interpretations for two data sets

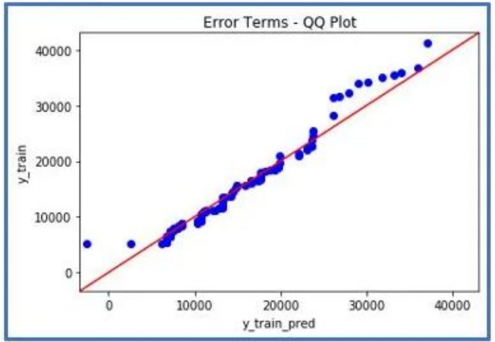
a) ***Similar distribution****: If all point of quantiles lies on or close to straight line at an angle of 45*

degree from x -axis

b) ***Y-values < X-values:*** *If y-quantiles are lower than the x-quantiles.*

c) ***X-values < Y-values:*** *If x-quantiles are lower than the y-quantiles.*





d) ***Different distribution:*** *If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis*