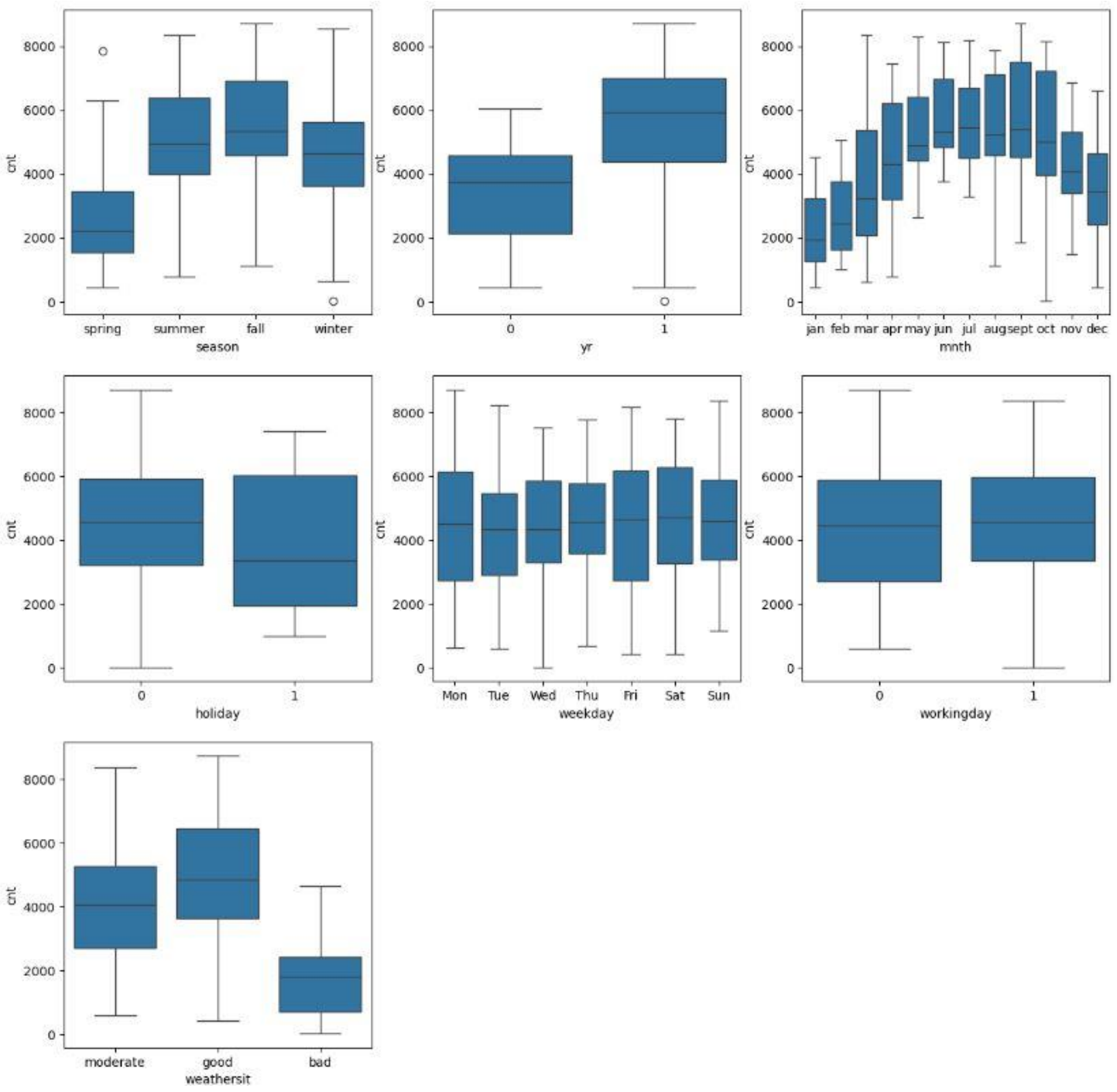
Assignment-based Subjective Questions

- Abhijit Wable

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:



I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization:

- Fall season attracted more Bike booking than other seasons.
- Year 2019 got more bike booking than year 2018. Which indicates that business is growing with number of days from start of business.

- From Months April to October, season and weather is good for biking, hence more booking has observed.
- Good weather attracted more bookings and Bad weather less bookings.
- Holidays have shown less bookings.
- Working day have less effect on number of bookings.

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

Syntax :

drop_first: bool, default False, which implies whether to get k-1 dummies out of k
categorical levels by removing the first level.

Importance:

Using drop_first=True during dummy variable creation is important to avoid the dummy variable trap, which occurs when the dummy variables are highly correlated (multicollinear). This can lead to issues in regression models where the independent variables are not truly independent. By dropping the first dummy variable, you ensure that the remaining variables provide the same information without redundancy, thus improving the model's performance and interpretability.

Example:

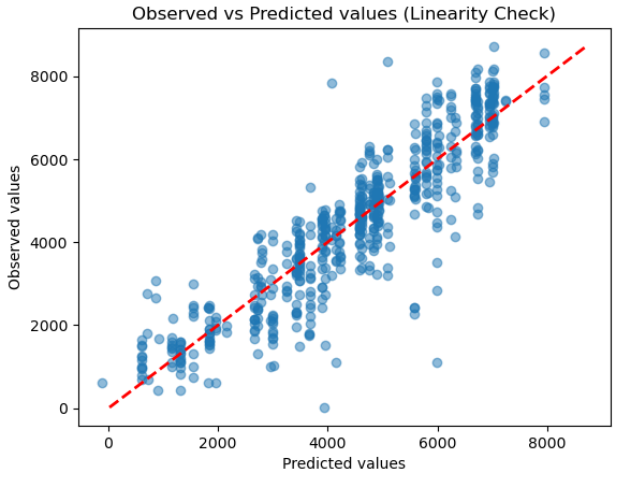
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
Answer:
'temp' variable has the highest correlation (0.63) with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
(3 marks)
Answer:
I have validated the assumption of Linear Regression Model based on below 5 assumptions -

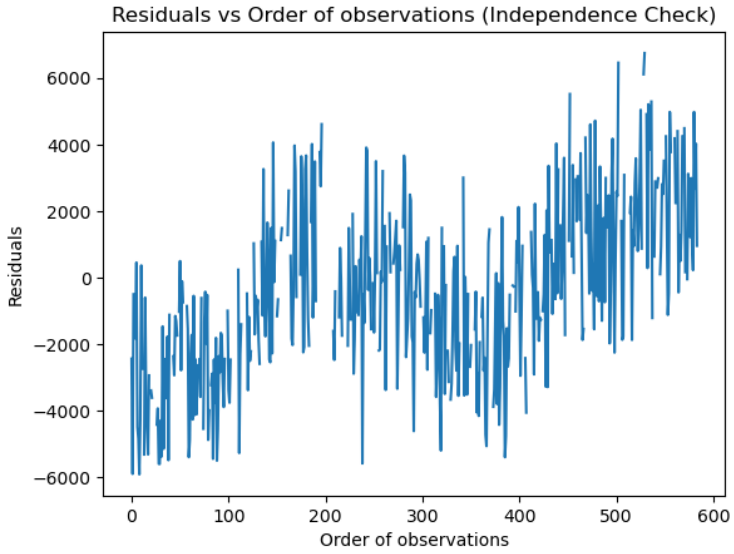
1. **Linearity:** Check if the relationship between the independent and dependent variables is linear.

Plot observed vs predicted value. Plot shows linear behavior between the dependent variable

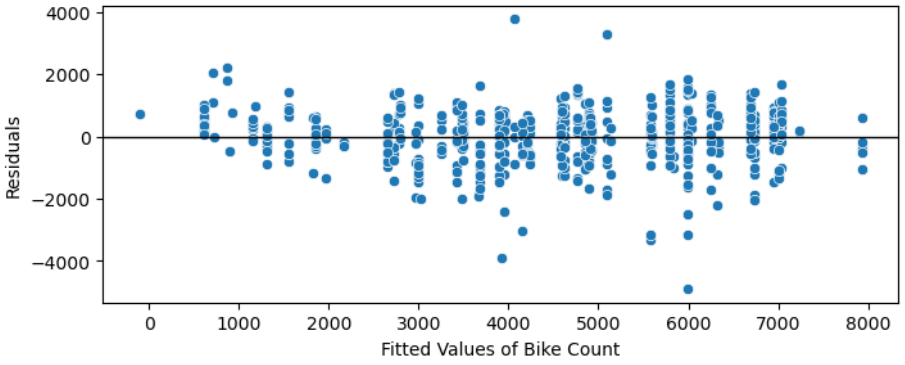


and independent variables.

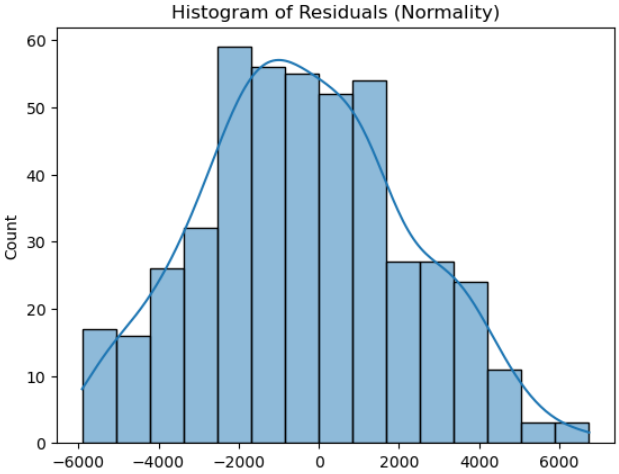
2. **Independence:** Ensure that the residuals (errors) are independent.
   Plot residuals vs order of observations.



3. **Homoscedasticity:** Check if the residuals have constant variance.
   Plot residuals vs predicted values



4. **Normality:** Verify that the residuals are normally distributed.
   Plot histogram and Q-Q plot of residuals



5. **No Multicollinearity:** Ensure that the independent variables are not highly correlated.
   Calculate Variance Inflation Factor (VIF). VIF should be below 5 in general.

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:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- year
- season winter
- weathersit_good

General Subjective Questions

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

Answer:

To calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

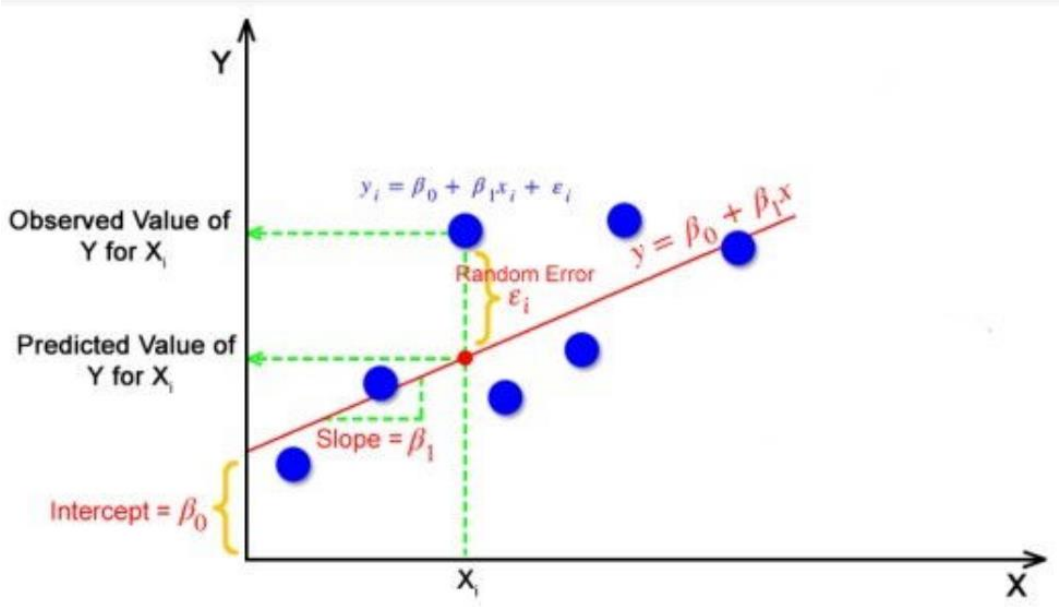
Y i = β 0 + β 1 X i

where Y i = Dependent variable, β 0 = constant/Intercept, β 1 = Slope/Intercept, X i = Independent variable.

This algorithm explains the linear relationship between the dependent(output) variable y and the independent(predictor) variable X using a straight line Y= B 0 + B 1 X.

The goal of the linear regression algorithm is to get the best values for B 0 and B 1 to find the best-fit line.

The best-fit line is a line that has the least error which means the **error between predicted values and actual values should be minimum.**



Objective : Minimize error between predicted values and actual

In regression, the difference between the observed value of the dependent variable(y i) and the predicted value(predicted) is called the residuals.

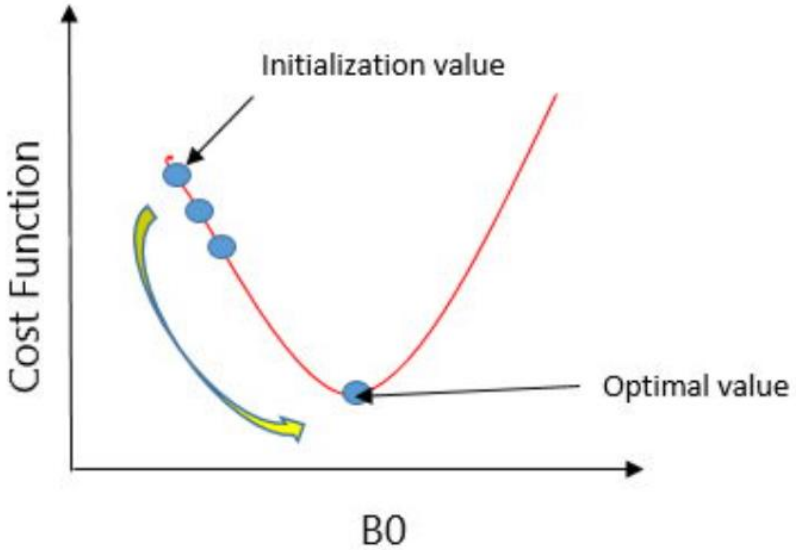
ε i = y predicted – y i

where y predicted = B 0 + B 1 X i

Method: This can be solved by gradient descent

Gradient Descent is one of the optimization algorithms that optimize the cost function (objective function) to reach the optimal minimal solution. To find the optimum solution, we need to reduce the cost function (MSE) for all data points. This is done by updating the values of the slope coefficient (B1) and the constant coefficient (B0) iteratively until we get an optimal solution for the linear function.

A regression model optimizes the gradient descent algorithm to update the coefficients of the line by reducing the cost function by randomly selecting coefficient values and then iteratively updating the coefficient values to reach the minimum cost function.



To update B 0 and B 1, we take gradients from the cost function. To find these gradients, we take partial derivatives for B 0 and B 1.

After solving partial derivatives , we get values of B0 and B1.

Multiple Linear Regression:

Multiple linear regression is a technique to understand the relationship between a single dependent variable and multiple independent variables.

The formulation for multiple linear regression is also similar to simple linear regression with

the small change that instead of having one beta variable, you will now have betas for all the variables used. The formula is given as:

$$Y = B0 + B1X1 + B2X2 + \ldots + BpXp + \varepsilon$$

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

Answer:

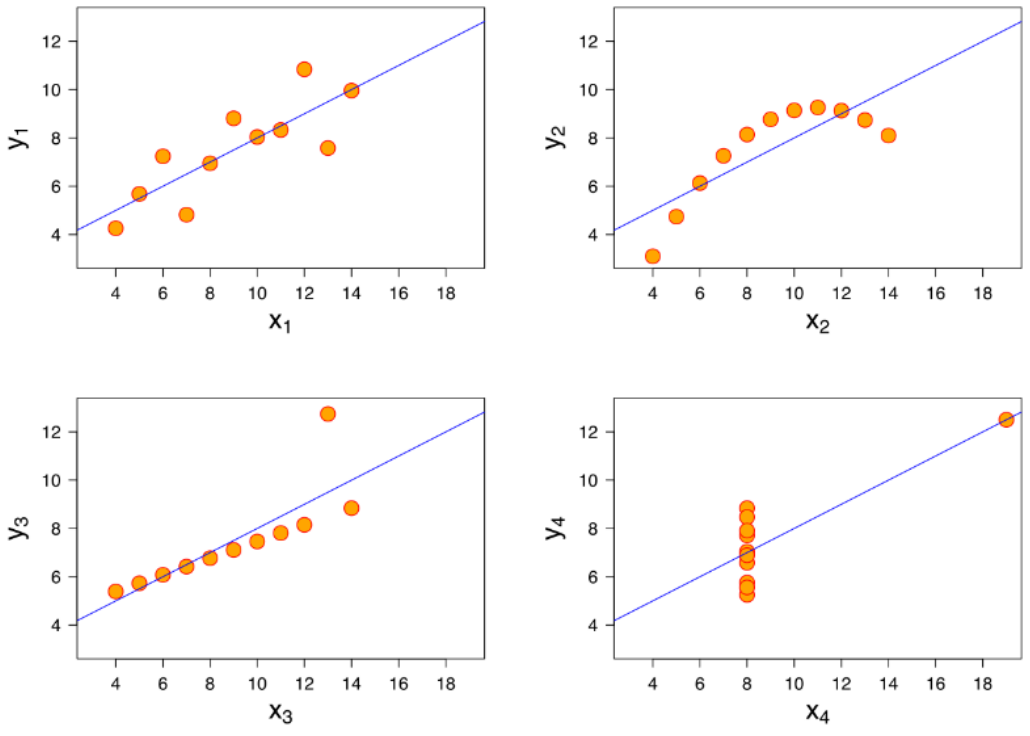
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets
is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



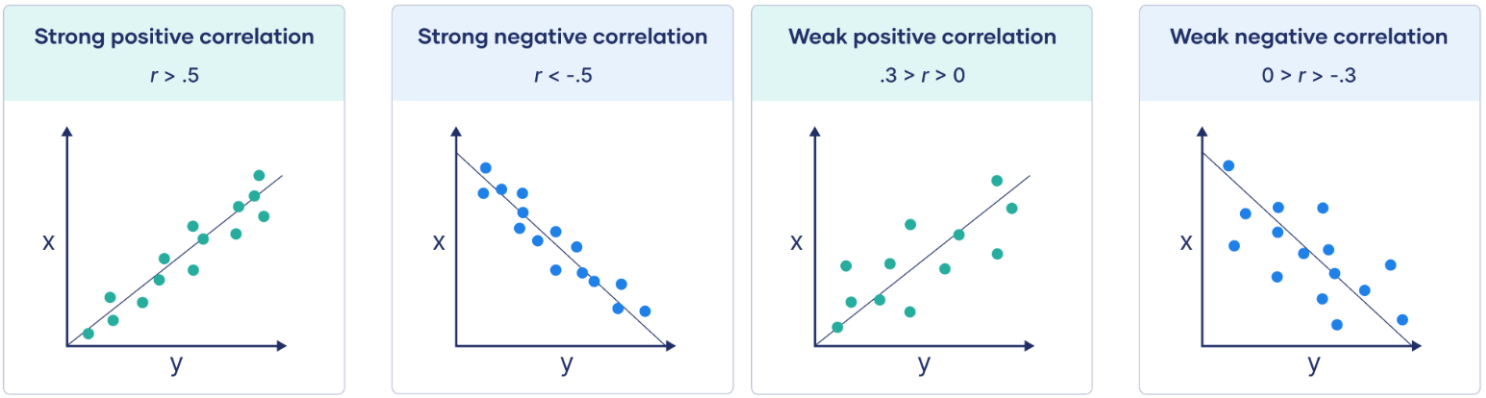
- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



When to use the Pearson correlation coefficient

The Pearson correlation coefficient (*r*) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when **all** of the following are true:

Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.

The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.

The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.

The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

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

Scaling is the process of transforming the features of a dataset to a similar scale. This is done to ensure that the features have a comparable influence on the model, particularly when using algorithms that are sensitive to the scale of the input features. Scaling is performed to improve the performance, convergence, and stability of various machine learning algorithms.

The main reasons for performing scaling are:

Algorithm Sensitivity: Some machine learning algorithms, such as support vector machines, k-nearest neighbors, and neural networks, are sensitive to the scale of the input features. Without scaling, features with larger scales may dominate the learning process and lead to suboptimal performance.

Convergence Speed: Scaling can help algorithms converge more quickly by ensuring that the optimization process is not overly influenced by features with larger scales.

Regularization: Regularization techniques, such as L1 and L2 regularization, assume that all features are on a similar scale. Scaling helps in achieving this and ensures that regularization penalties are applied uniformly across all features.

There are two common methods for scaling features: normalized scaling and standardized scaling.

Normalized Scaling (Min-Max Scaling):

In normalized scaling, the values of the features are scaled to fit within a specific range, typically 0 to 1.

Standardized Scaling (Z-Score Normalization):

Standardized scaling, also known as z-score normalization, transforms the features to have a mean of 0 and a standard deviation of 1.

.Example**:** If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give
wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.No.	Normalized scaling	Standardized scaling
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Answer:

Variance Inflation Factor (VIF):
VIF is a measure used to detect multicollinearity among predictor variables in a regression model.
It quantifies how much the variance of a coefficient estimate is inflated due to multicollinearity.

The Formula: VIF = 1 / (1 — R2)

A high VIF suggests that the variable in question is highly correlated with others, making its contribution to the model less distinct. In the context of VIF, infinity represents perfect correlation, and as VIF values increase, the reliability of the regression results decreases. On the flip side, a low VIF indicates that the variable is relatively independent and doesn't suffer from multicollinearity concerns.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

An infinite VIF occurs when a predictor variable is nearly perfectly predictable from the other predictor variables in the model, indicating severe multicollinearity. Identifying and addressing multicollinearity, such as through variable selection, regularization techniques, or domain knowledge, is crucial to ensure the stability and reliability of the regression model.

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 plot, short for quantile-quantile plot, is a graphical tool used to assess if a dataset follows a particular theoretical distribution, such as the normal distribution. It is a scatterplot created by plotting two sets of quantiles against each other. The main purpose of a Q-Q plot is to visually compare the distribution of a dataset to a theoretical distribution, allowing for the assessment of deviations from the expected distribution.

Here's a detailed explanation of the use and importance of a Q-Q plot in the context of linear regression:

Use of Q-Q Plot:

A Q-Q plot is commonly used to assess whether the residuals (i.e., the differences between observed and predicted values) from a linear regression model are normally distributed.
By plotting the quantiles of the residuals against the quantiles of a normal distribution, the Q-Q plot provides a visual assessment of how closely the residuals follow a normal distribution.
Interpretation:

If the residuals closely follow a diagonal line in the Q-Q plot, it indicates that they are approximately normally distributed.
Deviations from the diagonal line suggest departures from normality, such as skewness or heavy tails in the distribution of residuals.

Importance in Linear Regression:

Assessing the normality of residuals is important in linear regression because many statistical tests and inference procedures assume that the residuals are normally distributed.
Deviations from normality in the residuals can affect the validity of statistical inferences, such as hypothesis testing, confidence intervals, and model predictions.
Identification of Outliers and Anomalies:

Q-Q plots can also reveal outliers and anomalies in the dataset by highlighting deviations from the expected distribution.
Outliers in the dataset may appear as points that deviate significantly from the diagonal line in the Q-Q plot.
Validation of Model Assumptions:

Checking the normality of residuals through Q-Q plots is an essential step in validating the assumptions of the linear regression model.
If the Q-Q plot shows substantial departures from normality, it may indicate issues with the model's assumptions and the need for further investigation or model refinement.