Descriptive Statistics

What is Statistics

Statistics is a branch of mathematics that involves

- collecting,
- analysing,
- interpreting, and
- presenting data. It provides tools and methods to understand and make sense of large amounts of data and to draw conclusions and make decisions based on the data.

In practice, statistics is used in a wide range of fields, such as

- business,
- economics,
- social sciences,
- medicine, and
- engineering.

It is used to

- conduct research studies,
- analyse market trends,
- evaluate the effectiveness of treatments and interventions, and
- make forecasts and predictions.

Examples:

1. **Business** - Data Analysis(Identifying customer behavior) and Demand Forecasting
2. **Medical** - Identify efficacy of new medicines(Clinical trials), Identifying risk factor for diseases(Epidemiology)
3. **Government & Politics** - Conducting surveys, Polling
4. **Environmental Science** - Climate research

Types Of Statistics

2 Types of statistics

1. Descriptive
2. Inferential

1. Descriptive Statistics:

Descriptive statistics deals with the collection, organization, analysis, interpretation, and presentation of data. It focuses on summarizing and describing the main features of a set of data, without making inferences or predictions about the larger population.

> Descriptive Statistics is all about finding summary of your data, visualizing it and presenting in a more meaningful way.

2. Inferential statistics:

Inferential statistics deals with making conclusions and predictions about a population based on a sample. It involves the use of probability theory to estimate the likelihood of certain events occurring, hypothesis testing to determine if a certain claim about a population is supported by the data, and regression analysis to examine the relationships between variables

---

Population vs Sample

Population refers to the entire group of individuals or objects that we are

interested in studying. It is the complete set of observations that we want to make inferences about. For example, the population might be all the students in a particular school or all the cars in a particular city.

A **sample**, on the other hand, is a subset of the population. It is a smaller group of

individuals or objects that we select from the population to study. Samples are used to estimate characteristics of the population, such as the mean or the proportion with a certain attribute. For example, we might randomly select 100 students.

Examples

1. All cricket fans vs fans who were present in the stadium
2. All students vs who visit college for lectures

Things to be careful about which creating samples

1. Sample Size
2. Random
3. Representative

Parameter Vs Statistics

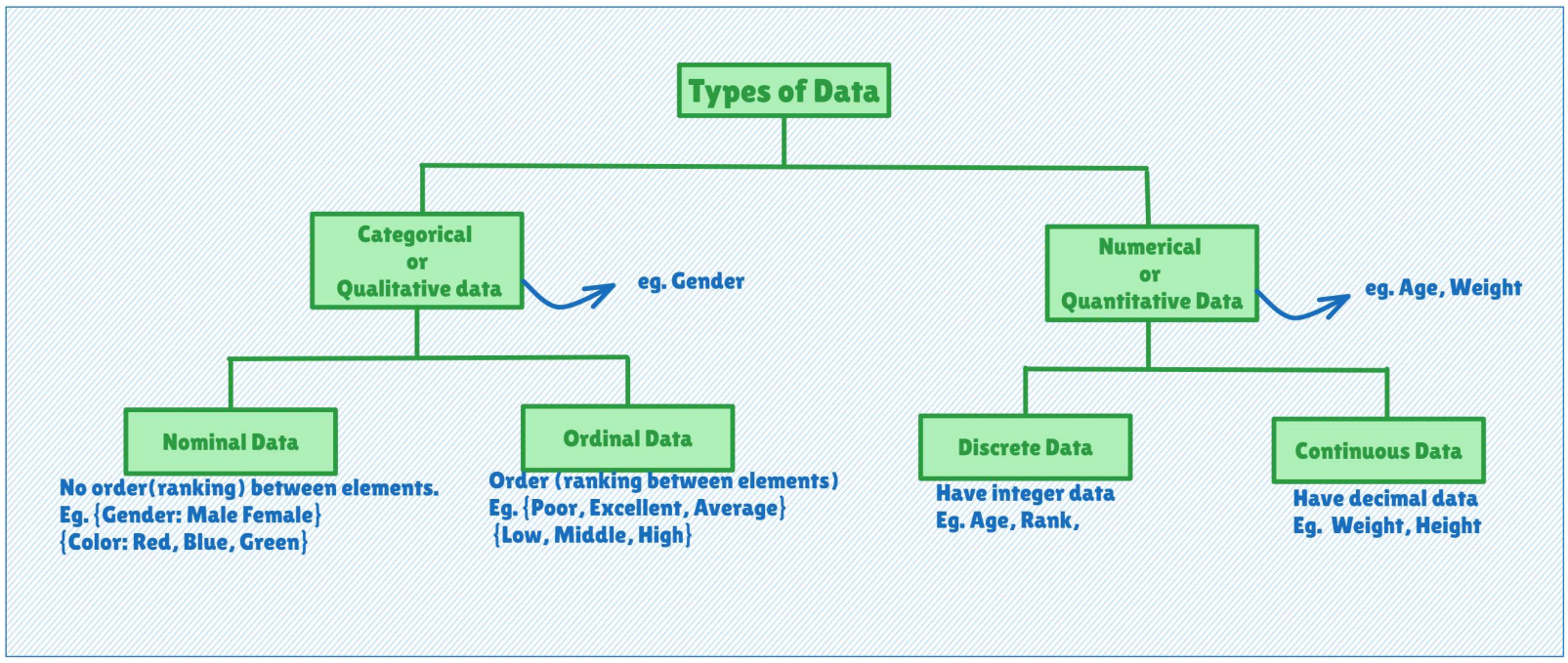
A parameter is a characteristic of a population, while a statistic is a characteristic of a sample. Parameters are generally unknown and are estimated using statistics. The goal of statistical inference is to use the information obtained from the sample to make inferences about the population parameters.

Inferential Statistics

Inferential statistics is a branch of statistics that deals with making inferences or predictions about a larger population based on a sample of data. It involves using statistical techniques to test hypotheses and draw conclusions from data. Some of the topics that come under inferential statistics are:

1. **Hypothesis testing:** This involves testing a hypothesis about a population parameter based on a sample of data. For example, testing whether the mean height of a population is different from a given value.
2. **Confidence intervals:** This involves estimating the range of values that a population parameter could take based on a sample of data. For example, estimating the population mean height within a given confidence level.
3. **Analysis of variance (ANOVA):** This involves comparing means across multiple groups to determine if there are any significant differences. For example, comparing the mean height of individuals from different regions.
4. **Regression analysis:** This involves modelling the relationship between a dependent variable and one or more independent variables. For example, predicting the sales of a product based on advertising expenditure.
5. **Chi-square tests:** This involves testing the independence or association between two categorical variables. For example, testing whether gender and occupation are independent variables.
6. **Sampling techniques:** This involves ensuring that the sample of data is representative of the population. For example, using random sampling to select individuals from a population.
7. **Bayesian statistics:** This is an alternative approach to statistical inference that involves updating beliefs about the probability of an event based on new evidence. For example, updating the probability of a disease given a positive test result.

Why ML is closely associated with statistics?



Short Note:

1. **Descriptive Statistics:** From given data, we will generate a summary or description without making any predictions.
2. **Inferential Statistics:** The statistics used to make **prediction** is called inferential statistics.
3. **Population** is entire data. **Sample** is part of data.
4. **Inferential Statistics** uses sample of data to predict conclusion. As, it is resource extensive to predict on whole population.

Measure of Central Tendency

1. Mean :

- Mean is sum of all values in the dataset divided by the number of values.

Population Mean (μ)

The **population mean** is the average of all values in the entire population:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

- X_i = each data value in the population
- N = total number of values in the population

Sample Mean ($\bar{X}$)

The **sample mean** is the average of values from a subset (sample) of the population:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- x_i = each data value in the sample
- n = number of values in the sample

Problem with Mean:

- Mean is **prone to outlier**: Just one extreme value (the outlier) can change the average significantly.
- **Example:** A student scored `(95,72,85,100, 0)` in 5 subjects his average of first 4 subject would be `(95+72+100+85/4) = 88` . But the average of all 5 dropped to `70` . This is all becuase of an outlier.

2. Median:

- Median is the middle value in the dataset when the data is arranged in order.
- It is a good alternative for mean, in case of outliers.

3. Mode:

- Mode is the value that appears most frequent in the dataset.
- Generally, we use it when we have categorical data. To tell which category is more frequent in our data set.
- **Example:** From a set of students, find out from which state most no of students came.

4. Weighted Mean:

- The weighted mean is the sum of the product of each value and its weight, divided by the sum of the weights. It is used to calculate a mean when the values in the dataset have different impotance or frequency.

Formula for Weighted Mean:

$$\bar{X}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Where:

- x_i = each data value
- w_i = weight (importance) assigned to x_i
- n = number of data points

Person	Correctness (Weight)	Observation (Value)
A	0.2	10
B	0.3	15
C	0.5	12

$$\bar{X}_w = \frac{(0.2 \times 10) + (0.3 \times 15) + (0.5 \times 12)}{0.2 + 0.3 + 0.5}$$

$$\bar{X}_w = \frac{2 + 4.5 + 6}{1.0}$$

$$\bar{X}_w = \frac{12.5}{1.0}$$

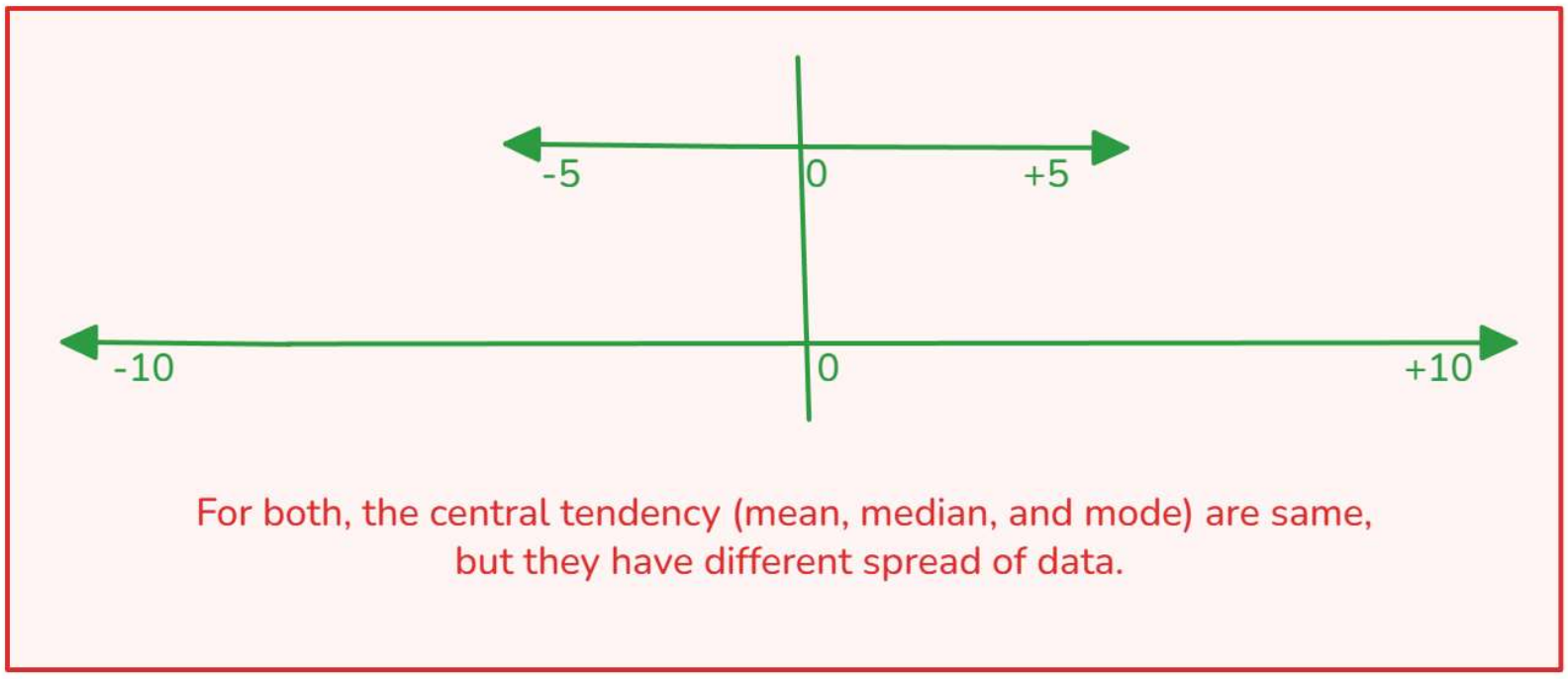
$$\bar{X}_w = 12.5$$

5. Trimmed Mean:

- A trimmed mean is calculated by removing a certain percentage of the smallest and largest values form the dataset and then taking the mean of the remaining values. The percentage of values removed is called the timmed percentage.
- **Example:**
  - `20,22,23,28,30,32,35,50,80 = 36.5`
  - `25,28,30,32,35 = 30`
- Trimmed mean, is used to reduce the impact of outliers.

Measure of Dispersion:

- A measure of dispersion is a statistical measure that describe the spread or variablility of a dataset. It provided information about how the data is distributed around the central tendency (mean, median, mode) of the dataset.
- Dispersion means how spread out the data is. Two sets of numbers can have the same mean but look very different in terms of spread.



1. Range

The range is the difference between the maximum and minimum values in a dataset. It is a simple measure of dispersion that is easy to calculate but can be affected by outliers.

- Formula:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

2. Mean Deviation (MD)

- Average of absolute differences from the mean (or median).
- Formula:

$$MD = \frac{\sum |x_i - \bar{X}|}{n}$$

> **Note:** Mean Deviation are less prone to outliers compared to Variance.
> But a downside is that, we cannot use it to calculate the **inference**

3. Variance (σ^2 or s^2)

The variance is the average of the squared differences between each data point and the mean. It measures the average distance of each data point from the mean and is useful in comparing the dispersion of datasets with different means.

- Or simply, Variance is (Summation of (sqaure of (Spread of all elements from the mean))).

-	X-mean	(X-mean)^2
3	3-3	0
2	2-3	1

-	X-mean	(X-mean)^2
1	1-3	4
5	5-3	4
4	4-3	1

- For population:

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$$

- For sample:

$$s^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$$

$$Variance = \frac{0+1+4+4+1}{5} = \frac{10}{5} = 2$$

> **Note:** Variance is not exactly the spread, but it is proportional to the spread. Because it calculate, the element is how far from the mean (spread from mean).

> **Why are negative deviations also added?** Becuase, in case of spread, If we simply subtracted negative values, they would cancel out the positive ones and give a misleading result.. That's why, in measures of dispersion, deviations are either taken as absolute values or squared — so that all spreads are counted positively.

> **Note:** Variance is highly sensitive to outliers—much more than the mean deviation. This happens because variance uses the square of the deviations, so large differences become even larger in the calculation.

Question: While calculating variance for Samples, why we divide it by `(n-1)` and not `(n)` .

---

4. Standard Deviation (SD)

- Standard Deviation, is the Square root of variance.
- It is widely used measure of dispersion in describing the shape of a distribution.
- Population:

$$\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$$

- Sample:

$$s = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}}$$

> **Why SD, if we have variance?** Because the standard deviation is expressed in the same unit as the data, which makes it easier to interpret. Variance, on the other hand, involves squared deviations $(X_i - \bar{X})^2$, so if data is in cm we will get variance in cm^2

5. Coefficient of Variation (CV)

The coefficient of Variation (CV) is the ratio of the standard deviation to the mean expressed as a percentage.

- It is used to compare the variability of datasets with differnt means, and is commonly used in fields such as biology, chemistry, and engineering.
- The coefficient of variation (CV) is a statistical measure that express the amount of variability in a dataset relative to the mean.
- It is a dimensionless quanity that is expressed as a percentage.
- **Standard deviation as a percentage of the mean.**
- Formula:

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100\%$$

> **Note:** It is useful for comparing datasets with different scales.

Example: How can we compare `Salary` and `Experience` when they are measured in different units? By using the CV, we convert both into percentages, which removes the unit issue and allows direct comparison.

> Types of graphs we can plot, based on the given data.

Graphs for Univariate Analysis (Analysis of one variable).

Graph for Univariate Analysis, are plotted based on the type of column.

1. **Categorical**
2. **Numerical**

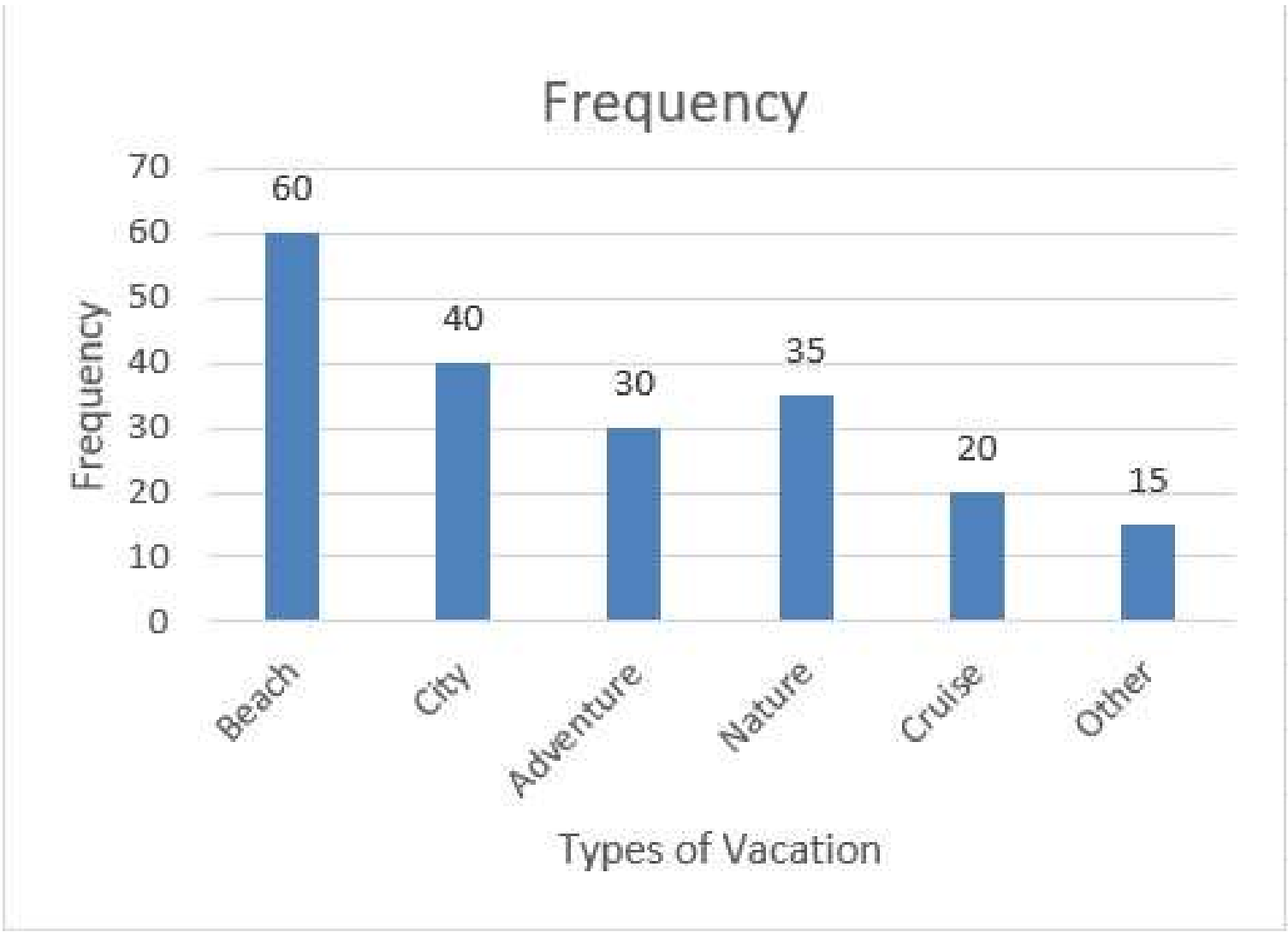
1 **Categorical** - Frequency Distribution Table & Cumutative Frequency.

1.1) A **frequency distribution table** is a table that summarizes the number of time (or frequency) that each value occurs in a dataset.

Let's say we have survey of 200 people and we ask them about their favourite type of vacation, which could be one of six categories.

Type of Vacation	Frequency
Beach	60
City	40
Adventure	30
Nature	35
Cruise	20
Other	15

> using this we can make bar graph.

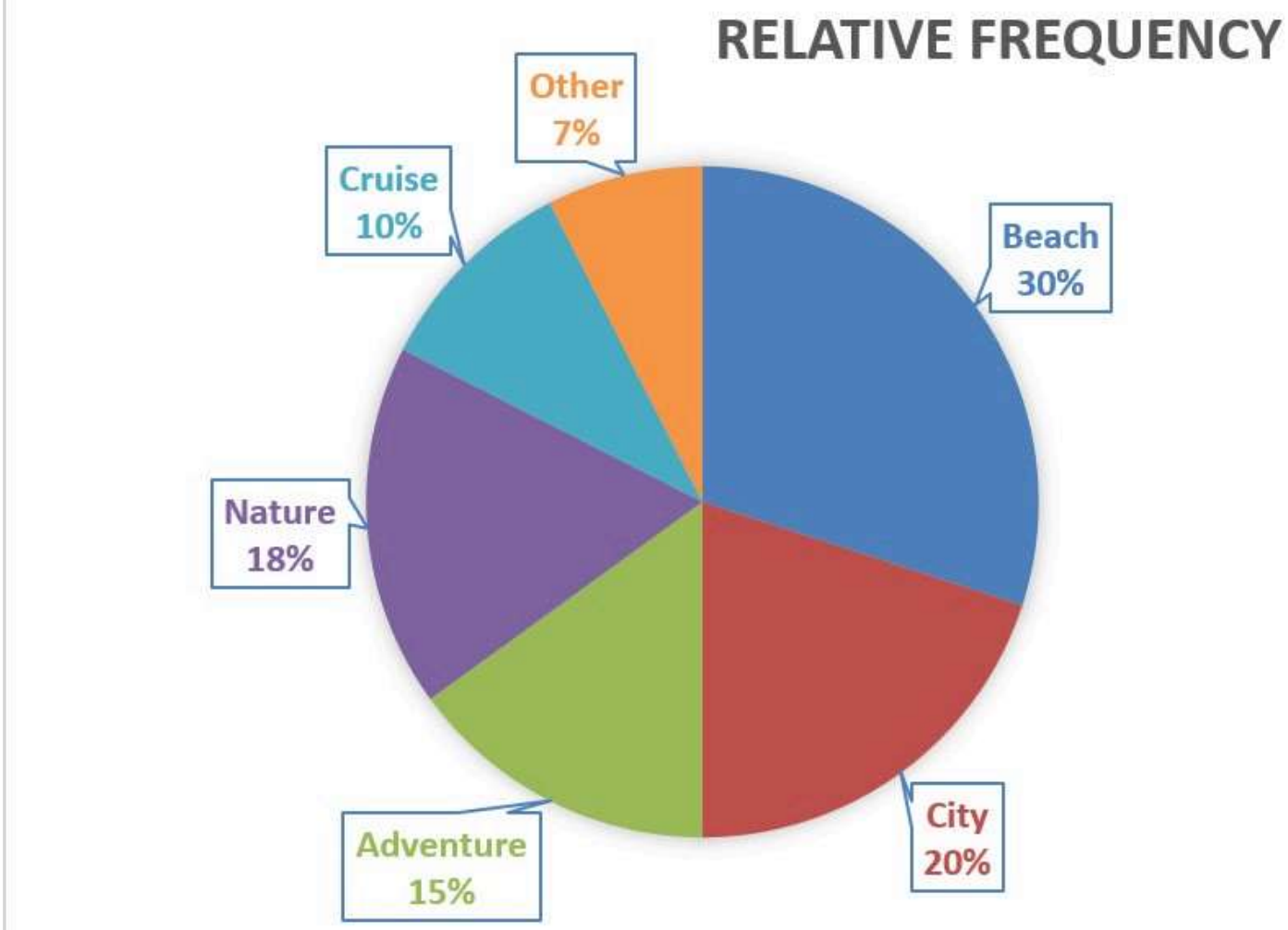


1.2) **Relative Frequency** is the proportion or percentage of a category in a dataset or sample. It is calculated by dividing the frequency of a category by the total number of observations in the dataset or sample.

Type of Vacation	Frequency	Relative Frequency
Beach	60	0.3
City	40	0.2

Type of Vacation	Frequency	Relative Frequency
Adventure	30	0.15
Nature	35	0.175
Cruise	20	0.1
Other	15	0.075

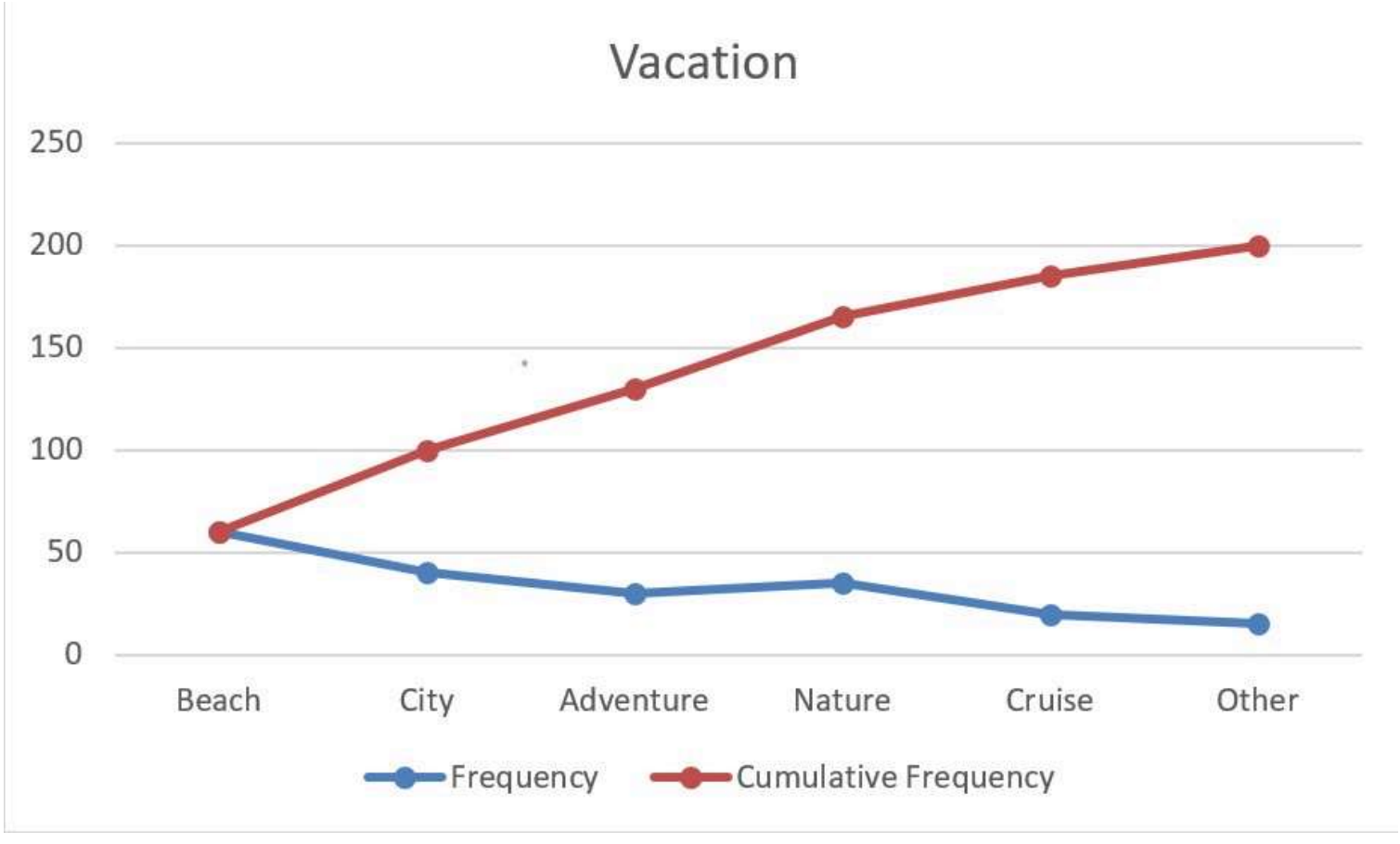
> using this we can make pie chart.



1.3) **Cumulative Frequency** is the running total of a frequency of a variabble or category in a dataset or sample. It is calculated by adding up the frequency of the current category and all previous categories in the dataset or sample.

Type of Vacation	Frequency	Relative Frequency	Cumulative Frequency
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
Other	15	0.075	200

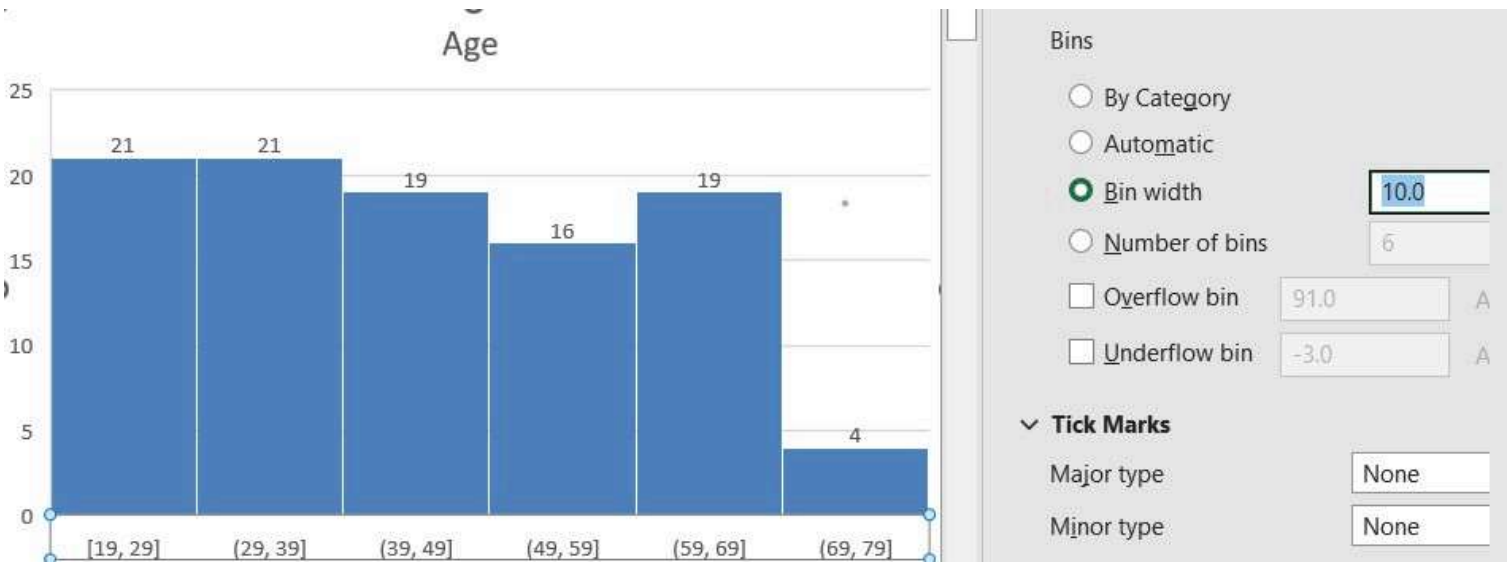
> we can make line graph



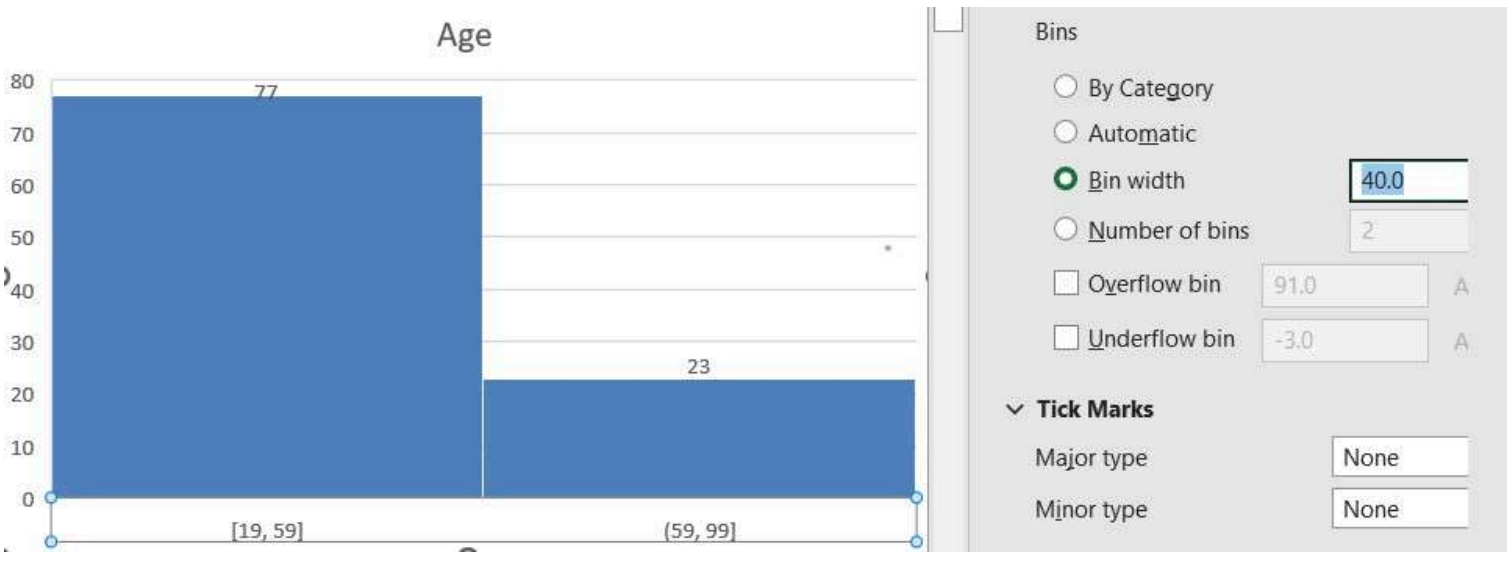
2. **Numerical** - Frequency Disribution Table & Histogram.

> **Note:** Here we have to make categories, which we call **bins** or **bucket**

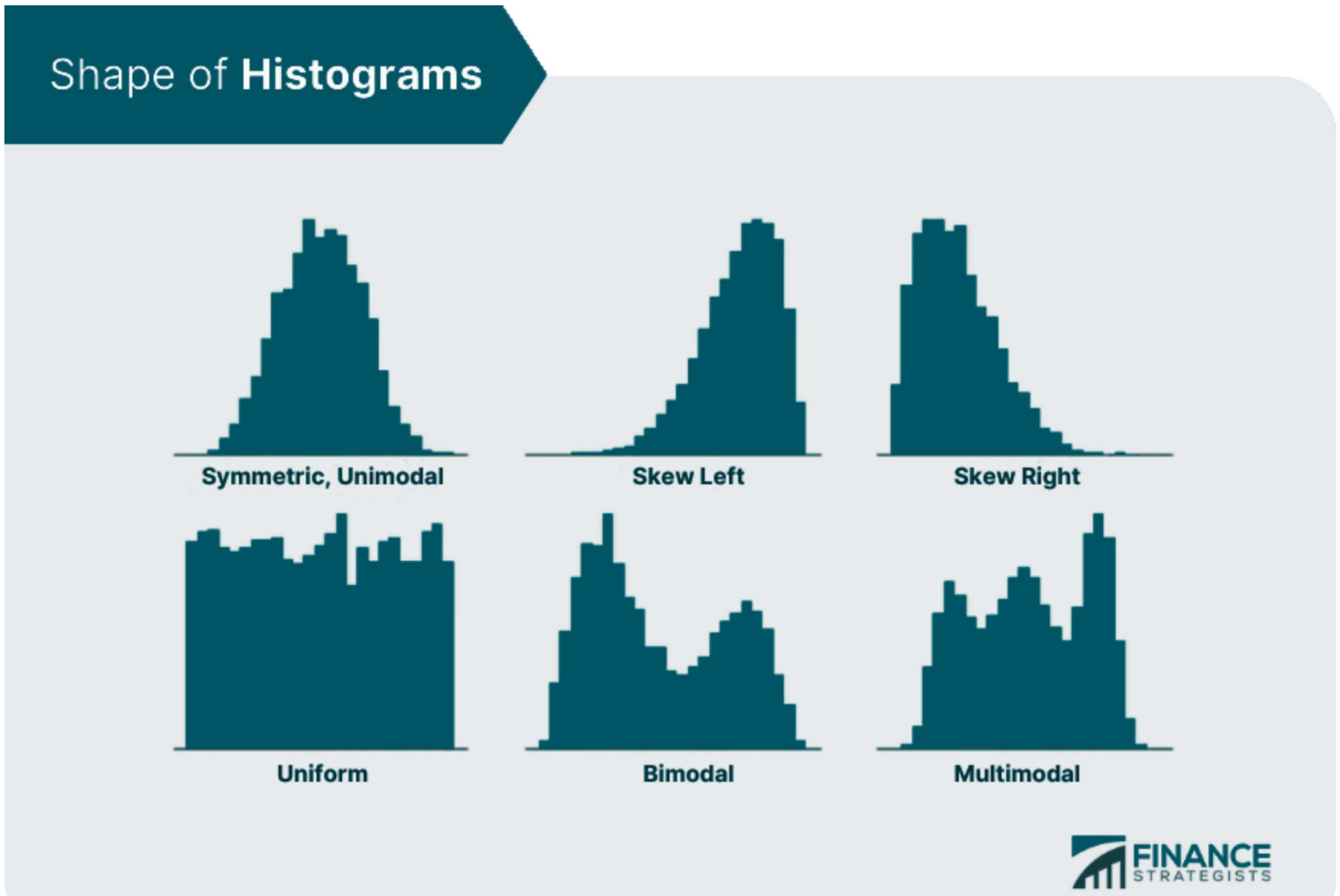
- Here, we have data of age between 18 and 70. We have 100 data elements.
- Lets make histogram with bin size = 10.



- Lets make histograph with bin size = 40.



Types of Histogram



> **Note:** When bin size increased, generally it became **Uniform**. When bin size decreased, generally it became **multimodal(No Pattern)**

Graph For bivariate Analysis (Analysis of relationship between two variables).

1. When both column is **Categorical**
2. When both column is **Numerical**
3. When one column is **Categorical** and one column is **Numerical**

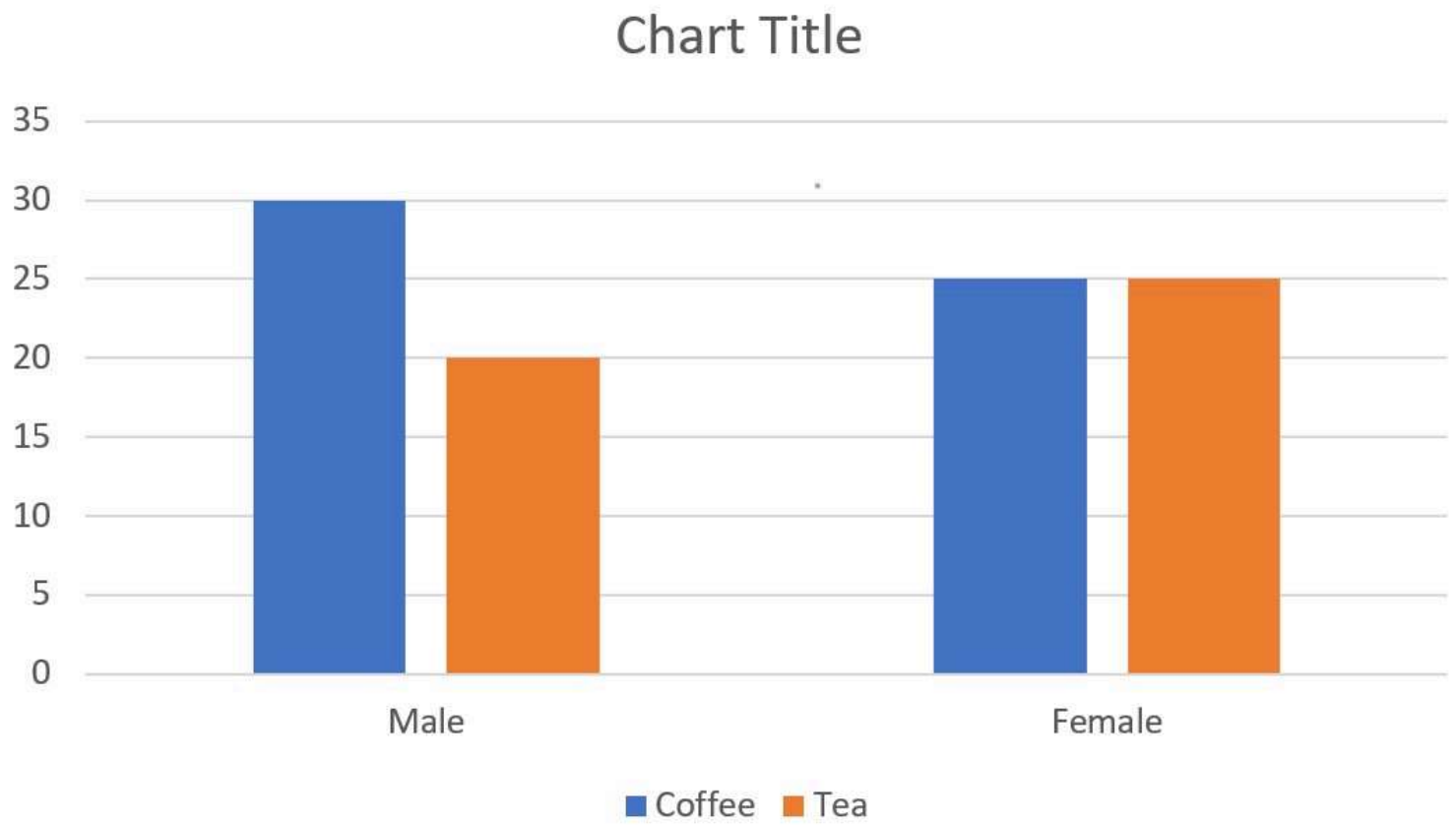
1. Categorical - Categorical

Contingency Table

- A contingency table, (also known as a cross-tabulation or crosstab), is a type of table used in statistics to summarize the relationship between two categorical variables.
- A contigency table displays the frequencies or relative frequencies of the observed value of the two variables, organized into rows and columns.
- It helps us understand the relationship between those variables.

	Coffee	Tea	Total
Male	30	20	50
Female	25	25	50

	Coffee	Tea	Total
Total	55	45	100



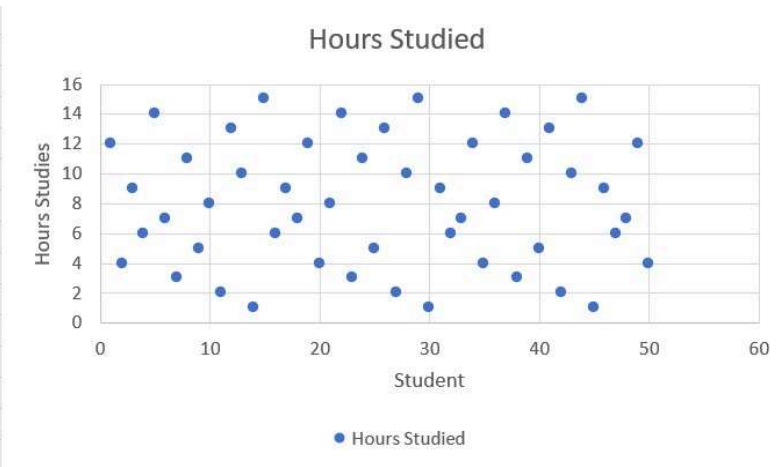
2. Numerical - Numerical

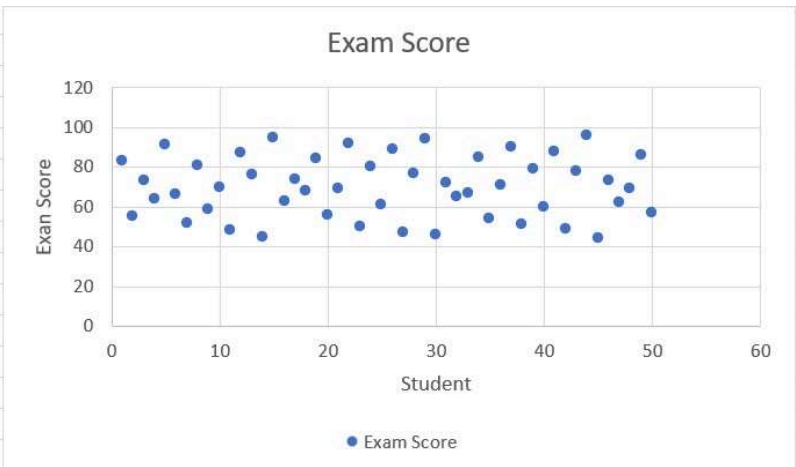
If we have 2 numerical column, the chart we plot is a scattered plot.

Student	Hours Studied	Exam Score
1	12	83
2	4	55
3	9	73
4	6	64
5	14	91
6	7	66
7	3	52
8	11	81
9	5	59
10	8	70
11	2	48
12	13	87
13	10	76
14	1	45
15	15	95

Student	Hours Studied	Exam Score
16	6	63
17	9	74
18	7	68
19	12	84
20	4	56
21	8	69
22	14	92
23	3	50
24	11	80
25	5	61
26	13	89
27	2	47
28	10	77
29	15	94
30	1	46
31	9	72
32	6	65
33	7	67
34	12	85
35	4	54
36	8	71
37	14	90
38	3	51
39	11	79
40	5	60
41	13	88
42	2	49
43	10	78
44	15	96
45	1	44
46	9	73
47	6	62

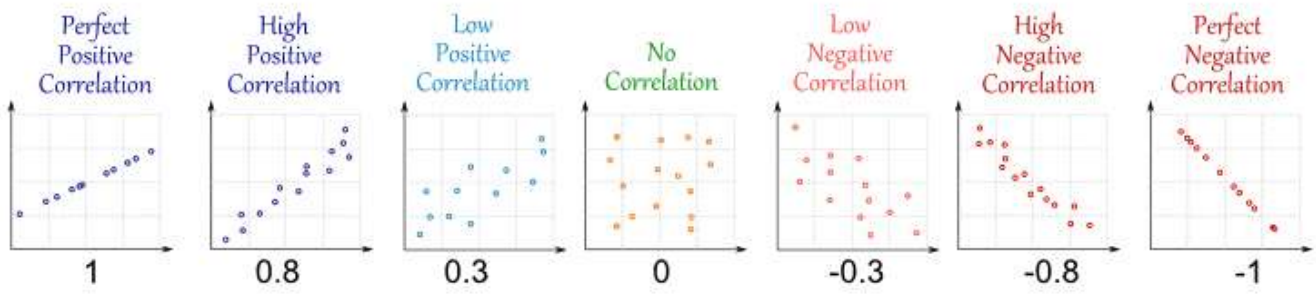
Student	Hours Studied	Exam Score
48	7	69
49	12	86
50	4	57





Types of Correlation Graph

1. Positive Correlated Data
2. negative Correlated Data
3. No Correlation



3. Categorical - Numerical

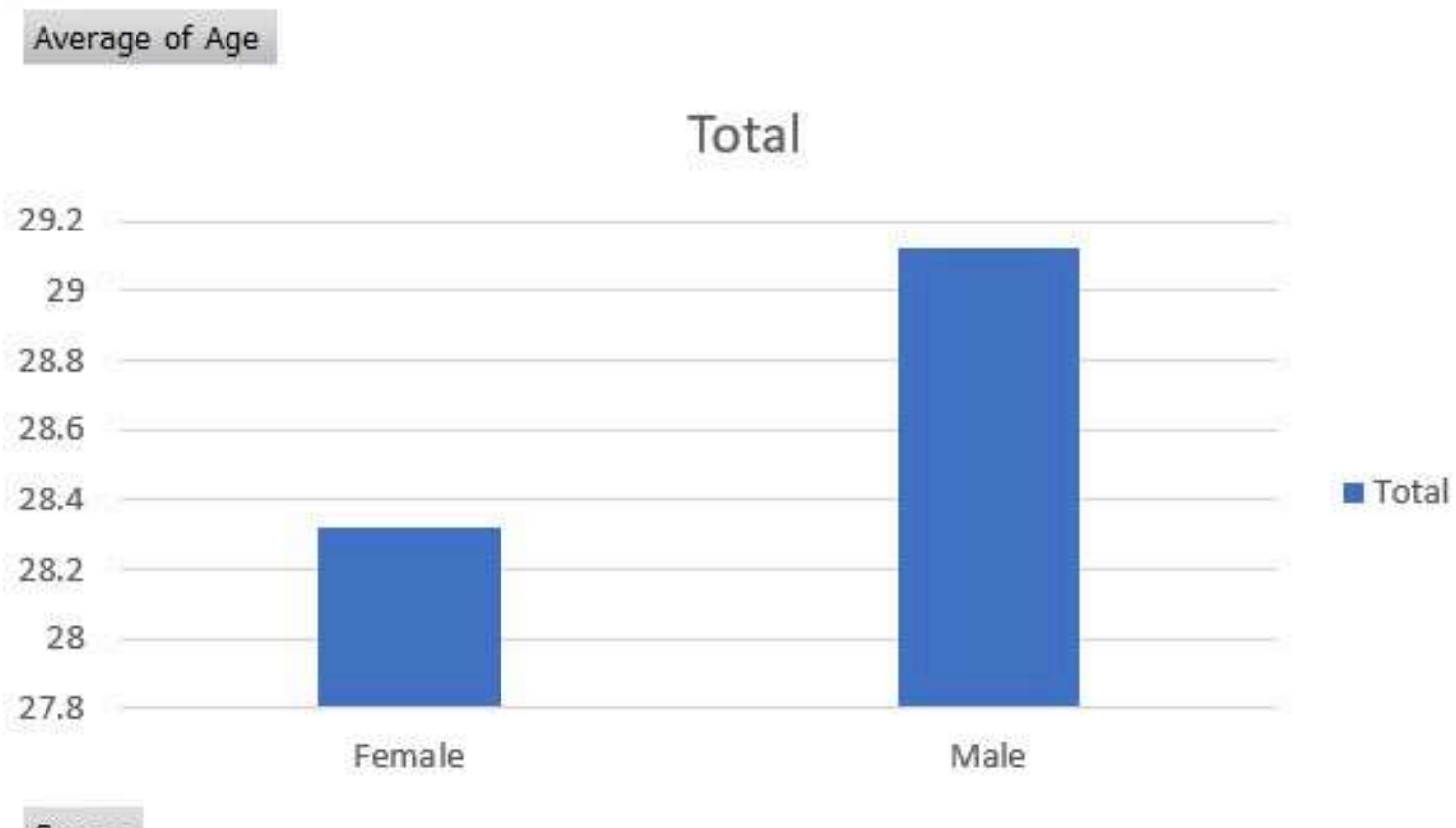
Person	Sex	Age
1	Male	25
2	Female	22
3	Male	30
4	Female	27
5	Male	35
6	Female	29
7	Male	24
8	Female	31
9	Male	28
10	Female	26

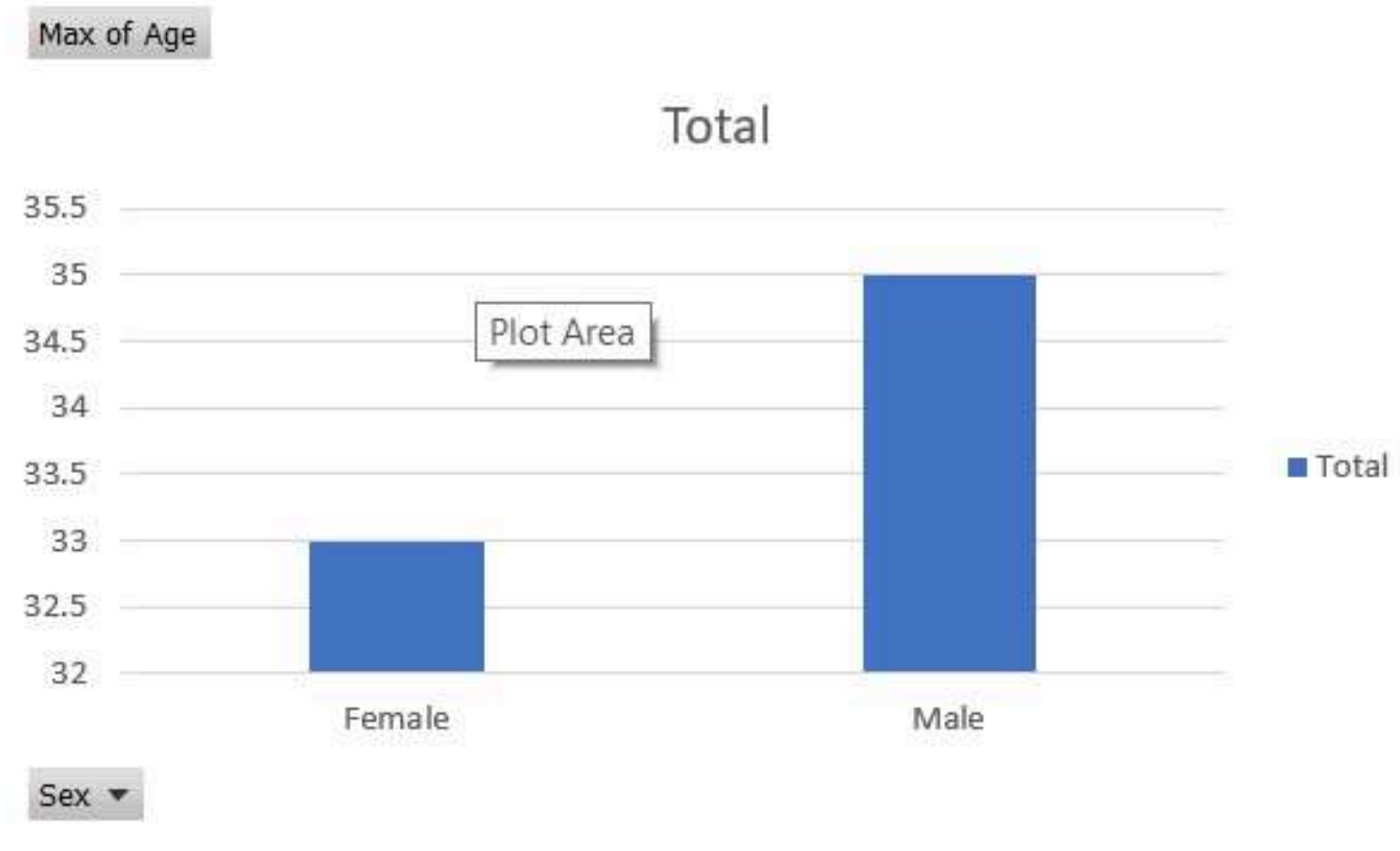
Person	Sex	Age
11	Male	32
12	Female	30
13	Male	27
14	Female	24
15	Male	29
16	Female	28
17	Male	26
18	Female	32
19	Male	31
20	Female	25
21	Male	34
22	Female	33
23	Male	23
24	Female	29
25	Male	28
26	Female	30
27	Male	27
28	Female	26
29	Male	35
30	Female	31
31	Male	24
32	Female	28
33	Male	29
34	Female	27
35	Male	33
36	Female	32
37	Male	30
38	Female	25
39	Male	26
40	Female	29
41	Male	31
42	Female	28

Person	Sex	Age
43	Male	27
44	Female	33
45	Male	34
46	Female	26
47	Male	28
48	Female	30
49	Male	32
50	Female	27

> In this category, we are flexible, we can make bar chart or contigency table(histogram)

Bar Chart





Quantiles and Percentiles.

Quantiles

Quantiles are statistical measure used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.

> General way of splitting data into equal chunks.

Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

- There are several types of quantiles used in statistical analysis, including:

1. Quartiles: Divide the data into four equal parts Q1(25 percentile), Q2(50 percentile or median), Q3 (75 percentile)
2. Deciles: Divide the data into ten equal parts, D1(10th percentile), D2(20 percentile) ... D9 (90th percentile).
3. Percentiles: Divide the data into 100 equal parts, P1 (1st percentile), P2(2nd percentile), ... P99 (99th percentile)
4. Quintiles: Divides the data into 5 equal parts.

Things to remember while calculating these measures:

1. Data should be sorted from low to high
2. You are basically finding the location of an observation

3. They are not actual values in the data
4. All other tiles can be easily derived from Percentiles (percentiles can be used to derive all other quantiles).

Percentiles.

A percentile is a statistical measure that represents the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile is the value below which 75% of the observations in the dataset fall.

> Specific case where data is split into 100 chunks.

Formula

$$P_L = \text{Value at rank } \left(\frac{p}{100} \times (N+1)\right)$$

Where:

- P_L = the desired percentile location.
- p = the percentile you want or get.
- N = total number of data points

Example

Find the 75th percentile score from the below data.

78,82,84,88,91,93,94,96,98,99

Step 1: Sort the data in ascending order.

78,82,84,88,91,93,94,96,98,99

Step 2: Apply the formula

$$P_L = \text{Value at rank } \left(\frac{p}{100} \times (N+1)\right)$$

$$P_L = \text{Value at rank } \left(\frac{75}{100} \times (10+1)\right)$$

$$P_L = \text{Value at rank } \left(\frac{3}{4} \times (11)\right)$$

$$P_L = \text{Value at rank } (\frac{33}{4})$$

$$P_L = \text{Value at rank } 8.25$$

That means, our `75` percentile value, is between `8th or 9th location` i.e. `between 96 and 98`.

`= 96 + 0.25(98-96) = 96 + 0.25*2 = 96.5`

`Answer = 96.5` This means, if you have scored 96.5 marks in this exam. You would have get 75 percentile.

Percentile of a value.

Previously, we started with a percentile (e.g., the 75th percentile) and calculated the corresponding score. Now, we'll do the opposite: given a specific score, we will determine which percentile it falls into.

Formula

$$PercentileRank = \frac{X + 0.5Y}{N}$$

```
txt
where,
X = number of values below the given value
Y = number of values equal to given value
N = total number of values in the dataset
```

Example: `78,82,84,88,91,93,94,96,98,99.` Find percentile of `88`

$$PercentileRank = \frac{X + 0.5Y}{N}$$

$$PercentileRank = \frac{3 + 0.5 * 1}{10}$$

$$PercentileRank = 0.35$$

$$PercentileRank = 35^{\text{th}}\ Percentile$$

Example: `78,82,84,88,91,93,94,96,98,99.` Find percentile of `99`

$$PercentileRank = \frac{X + 0.5Y}{N}$$

$$PercentileRank = \frac{9 + 0.5 * 1}{10}$$

$$PercentileRank = 0.95$$

$$PercentileRank = 95^{\text{th}}\ Percentile$$

> **Note:** Quantile is a broad term, and under it we have different types such as quartiles, deciles, percentiles, and quintiles.

Measure of Dispersion.

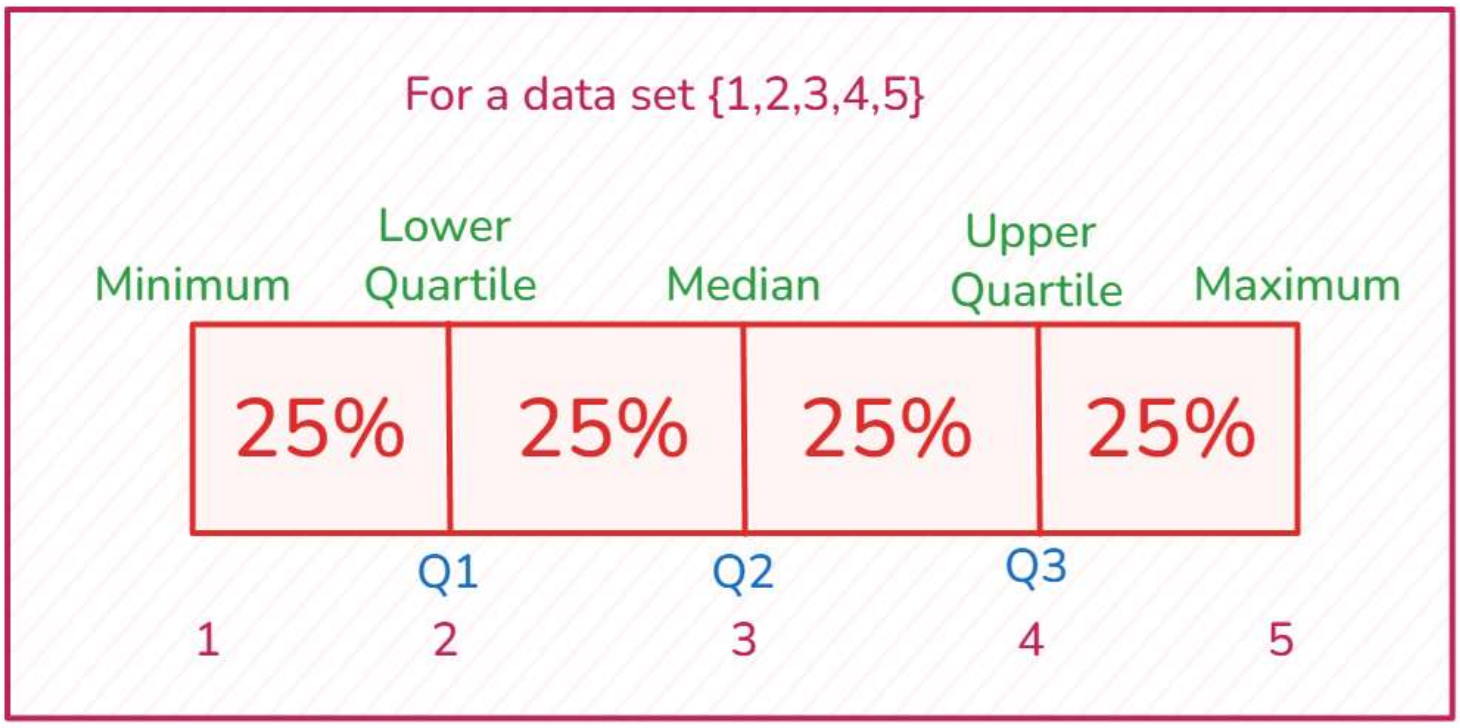
We studied about: Vaiance, Standard Deviation, Coefficient of Variation. But there is one more measure of Dispersion. (5 number summary)

5 number summary

The five-number summary is a descriptive statistic that provides a summary of a dataset. It consists of five values that divide the dataset into four equal parts, also known as quartiles. The five-number summary includes the following values:

1. **Minimum value:** The smallest value in the dataset.
2. **First quartile (Q1):** The value that separates the lowest 25% of the data from the rest of the dataset.
3. **Median (Q2):** The value that separates the lowest 50% from the highest 50% of the data.
4. **Third quartile (Q3):** The value that separates the lowest 75% of the data from the highest 25% of the data.
5. **Maximum value:** The largest value in the dataset.

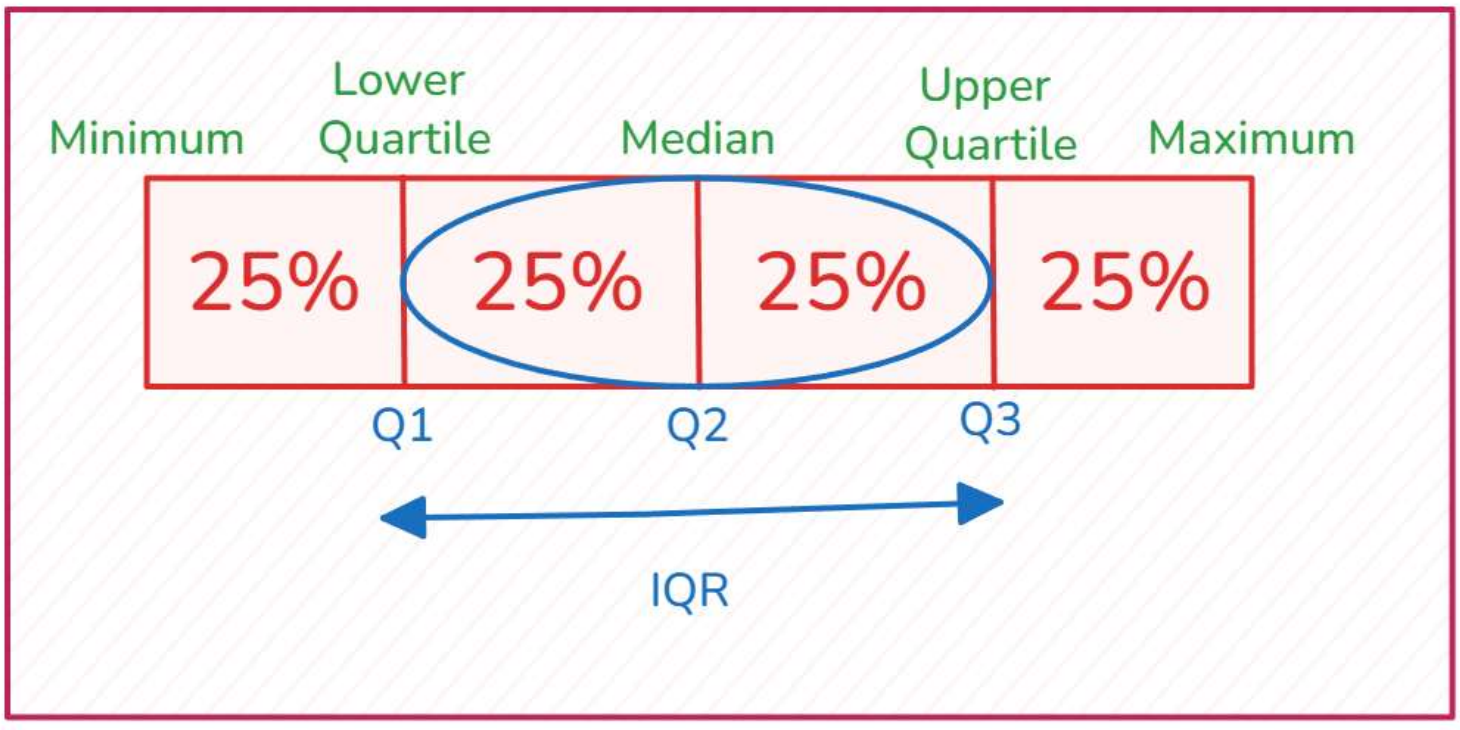
The five-number summary is often represented visually using a box plot, which displays the range of the dataset, the median, and the quartiles. The five-number summary is a useful way to quickly summarize the central tendency, variability, and distribution of a dataset.



> 5 number summary is very importqant statistical measure for ML algorithms.

IQR: Inter-Quartile Range

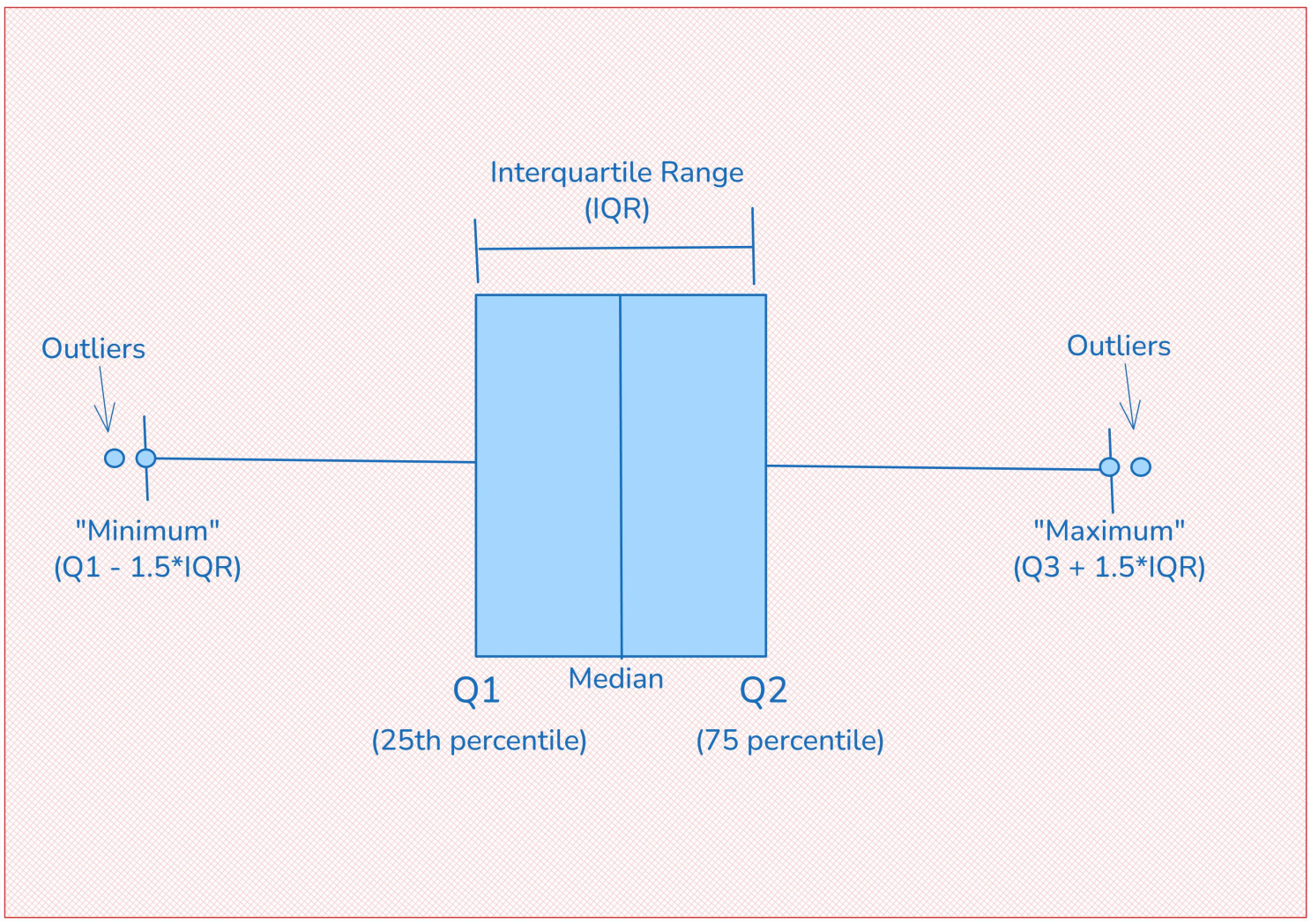
The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.



Boxplots

What is boxplot

A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset that shows the distribution of the data. The box plot displays a summary of the data, including the minimum and maximum values, the first quartile (Q1), the median (Q2), and the third quartile (Q3).



1. Benefits of a Boxplot

- Easy way to see the distribution of data
- Tells about skewness of data
- Can identify outliers
- Compare 2 categories of data

2. How to create a boxplot with example.

Calculate Q1, Q2, Q3

Let's say we have a data of 10 numbers (arranged in ascending number)

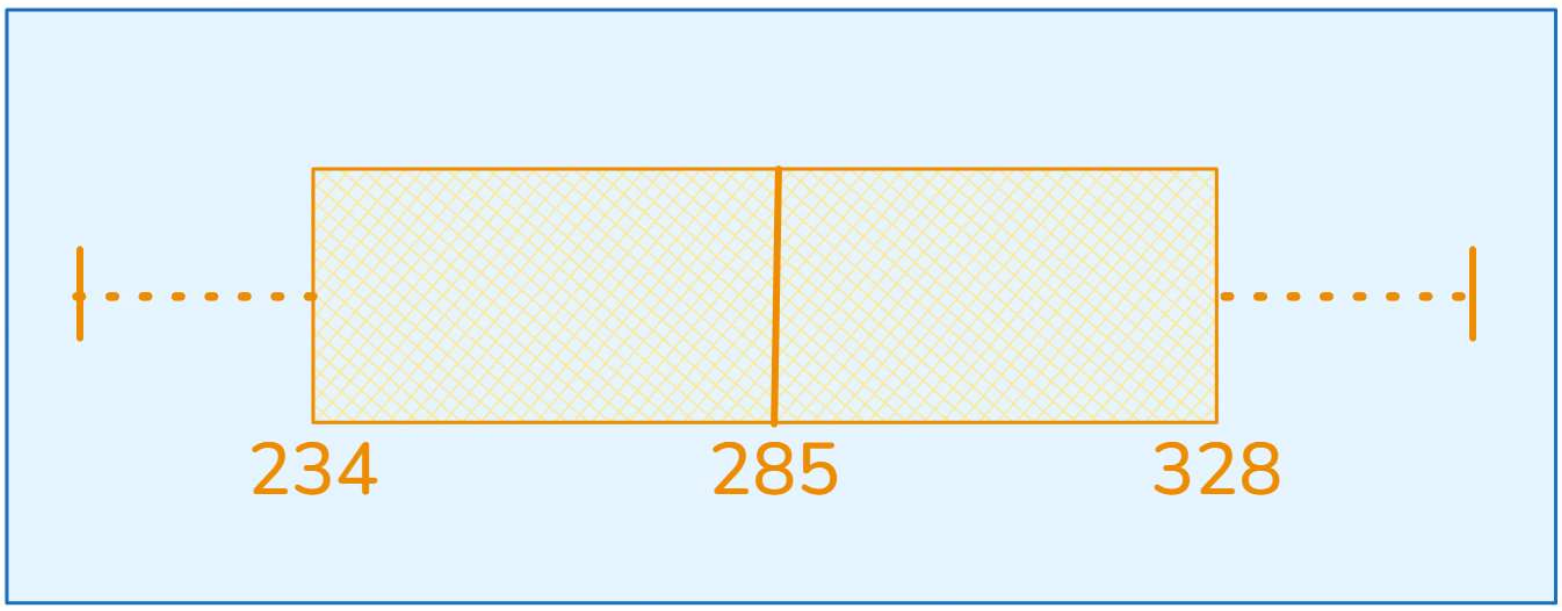
```
6, 213 , 241, 260 , 290 , 314, 321, 350, 1500
```

1	2	3	4	5	6	7	8	9	10
6	213	241	260	281	290	314	321	350	1500

Q2 = `50/100 (11) = 5.5 = 285.5` (281+0.5(290-281))

Q1 = `25/100(11)  = 2.75 = 234` (213 + 0.75(241-213))

Q3 = `75/100(11) = 8.25 = 328.25` (321 + 0.25(350-321))



3. Calculate Minimum and Maximum

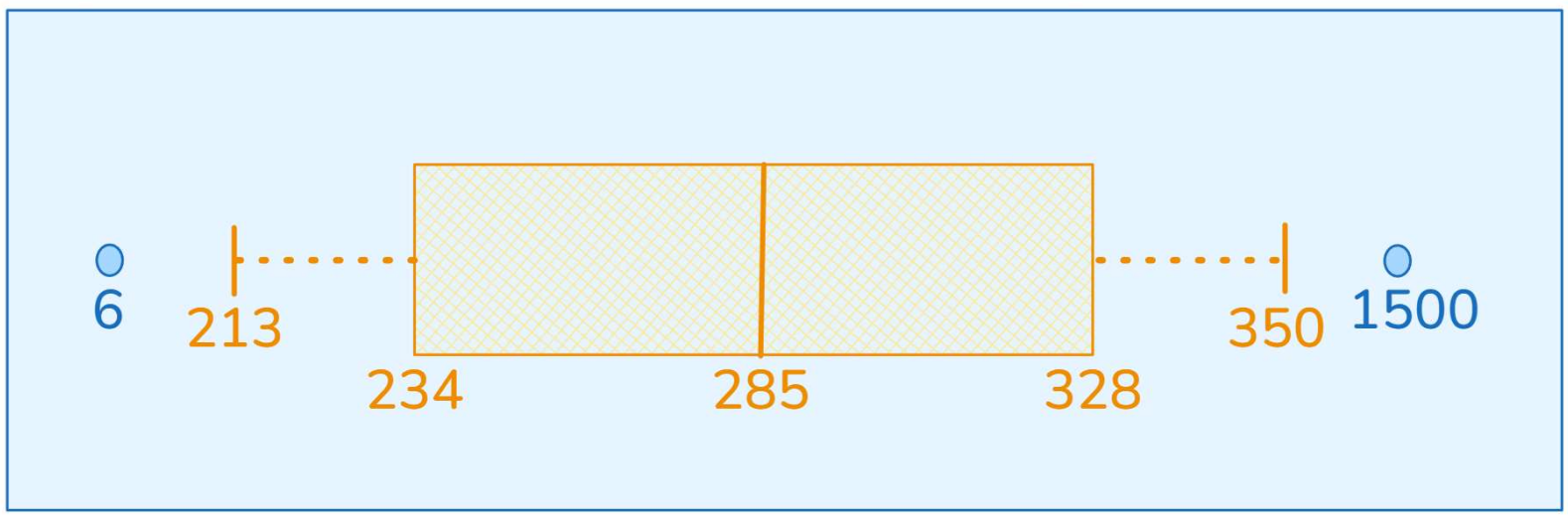
- Now, lets calculate minimum and maximum value.

MINIMUM = `Q1 - 1.5(IQR)` **MAXIMUM =** `Q3 + 1.5(IQR)`

IQR = `328-234 = 94`

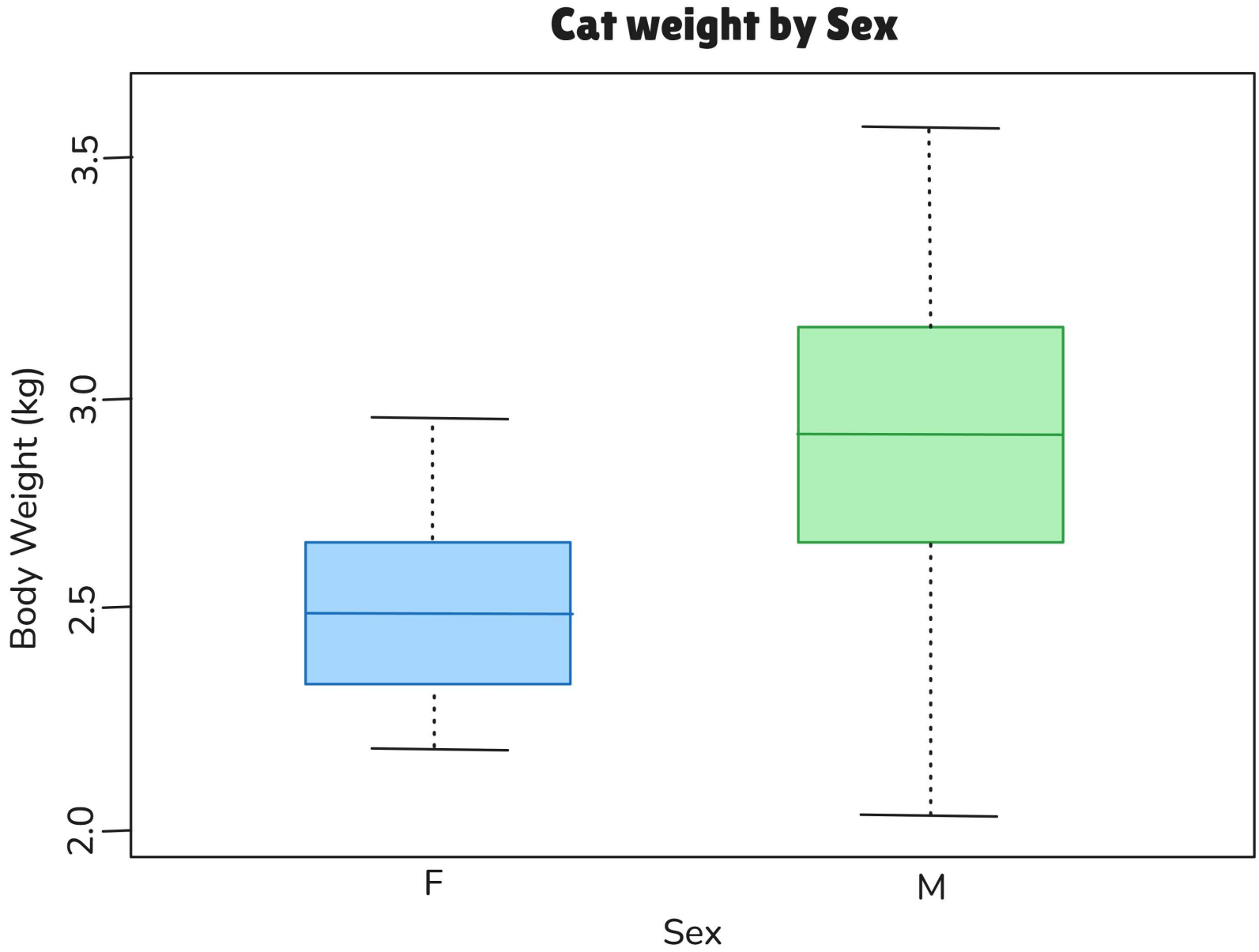
MINIMUM = `285 - 1.5(94) = 93` **MAXIMUM =** `328 + 1.5(94) = 469`

> **Note:** We got 93 as minimum and 469 as maximum. But we will stop at 213 and 350 as they are the given value. And consider 6 and 1500 as outliers.



Side by Side Box Plot

Example: We have a plot of "Cat weight by sex" with 2 columns, Bodyweight and Gender(M/F).

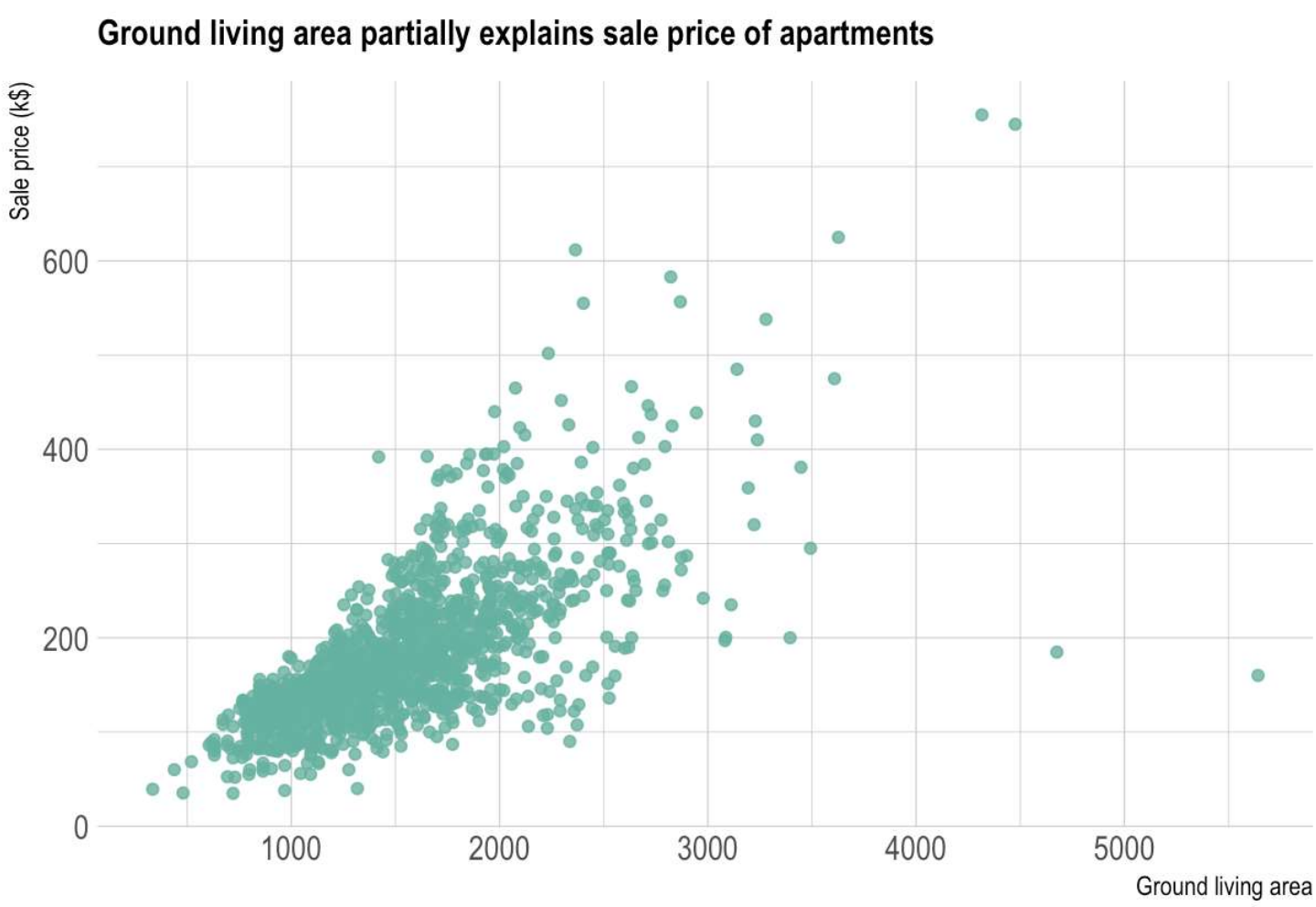


Using this plot, we can draw several conclusions.

1. We are comparing data of cats by their body weight and sex.
2. Median value (age) of male cats is greater than female cats
3. Central tendancy if males is higher than females.
4. Variability (variable values - फैलाव) in male cats are greater than female cats. (Female cats value vary in very small range, compare to male cats).
5. Female cat's weight are wkewed toward lower values. Whereas male cat's weight are uniformally distributed.

ScatterPlots

When the both data in x-axis and y-axis are numerical (numerical-numerical type) we draw scatter plot.



> In a scatter plot, to study the relationship between two numericals, we have to study 2 statistical measures

1. Covariance
2. Co-relation

1. Covariance

In mathematics and statistics, covariance is a measure of the relationship between two random variables.

What problem does Covariance solve?

To study spread of 2 datasets having same mean but different spread we used **"Variance"**

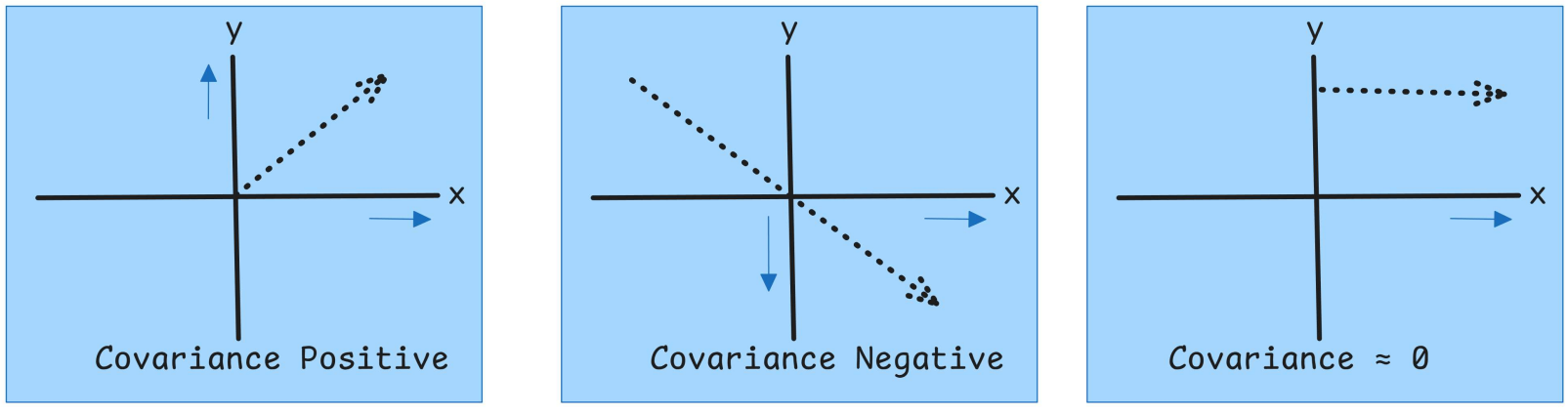
However, variance only works for a single variable. When we want to study the relationship between two variables at the same time (like study hours and exam scores), variance isn't enough. That's where covariance comes in — it measures how two variables change together.

What is covariance

- Covariance is a statistical measure that describes the degree to which two variables (numerical variables) are linearly related. It measures how much two variables change together, such that when one variable increases, does the other variable also increase, or does it decrease?

Types of Covariance

It shows whether the variables are positively related, negatively related, or not related at all.



1. If the covariance between two variables is positive, it means that the variables tend to move together in the same direction.
2. If the covariance is negative, it means that the variables tend to move in opposite directions.
3. A covariance of zero indicates that the variables are not linearly related.

How is it calculated?

Covariance Formula	
Population	**Sample**
$\sigma_{xy} = \frac{\sum(X-\mu_x)(Y-\mu_y)}{N}$	$s_{xy} = \frac{\sum(X-\bar{x})(Y-\bar{y})}{n-1}$
X, Y – The Value of X and Y in the Population *μ_x, μ_y – The population Mean of X and Y* *N – Total Number of Observations*	*X, Y – The Value of X and Y in the Sample Data* *$\bar{x}, \bar{y}$ – The Sample Mean of X and Y* *n - Total Number of Observations*

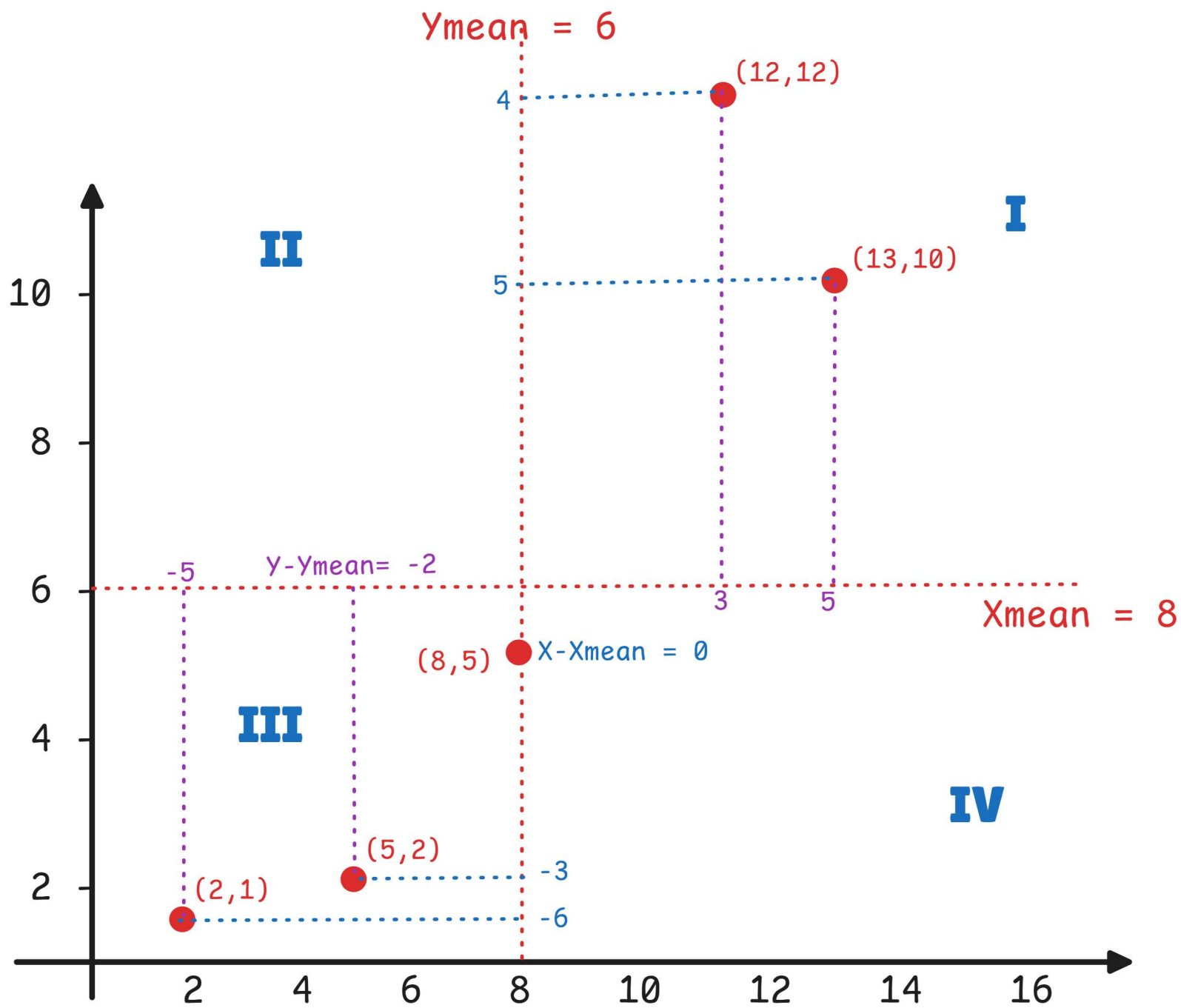
Ques.1:Let's we have 2 data columns Experience(x) and Salary(Y). Let's calculate the relation between them.

Exp(x)	Salary(y)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
2	1			
5	2			
8	5			
12	12			
13	10			

Solution:

Step-1: Let's calculate the mean first. $\bar{X}=8$ $\bar{Y}=6$

Exp(x)	Salary(y)	X – X̄	Y – Ȳ	(X – X̄)(Y – Ȳ)
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	0
12	12	4	6	24
13	10	5	4	20
Σ				**86**



Here the relationship is positive

> Note: If the point lies in `Ist or IIIrd Quadrant`. The value of `(X – X̄)(Y – Ȳ)` is always positive. Whereas in `IInd and IVth quadrant` the value of `(X – X̄)(Y – Ȳ)` is always negative.

Ques.2:

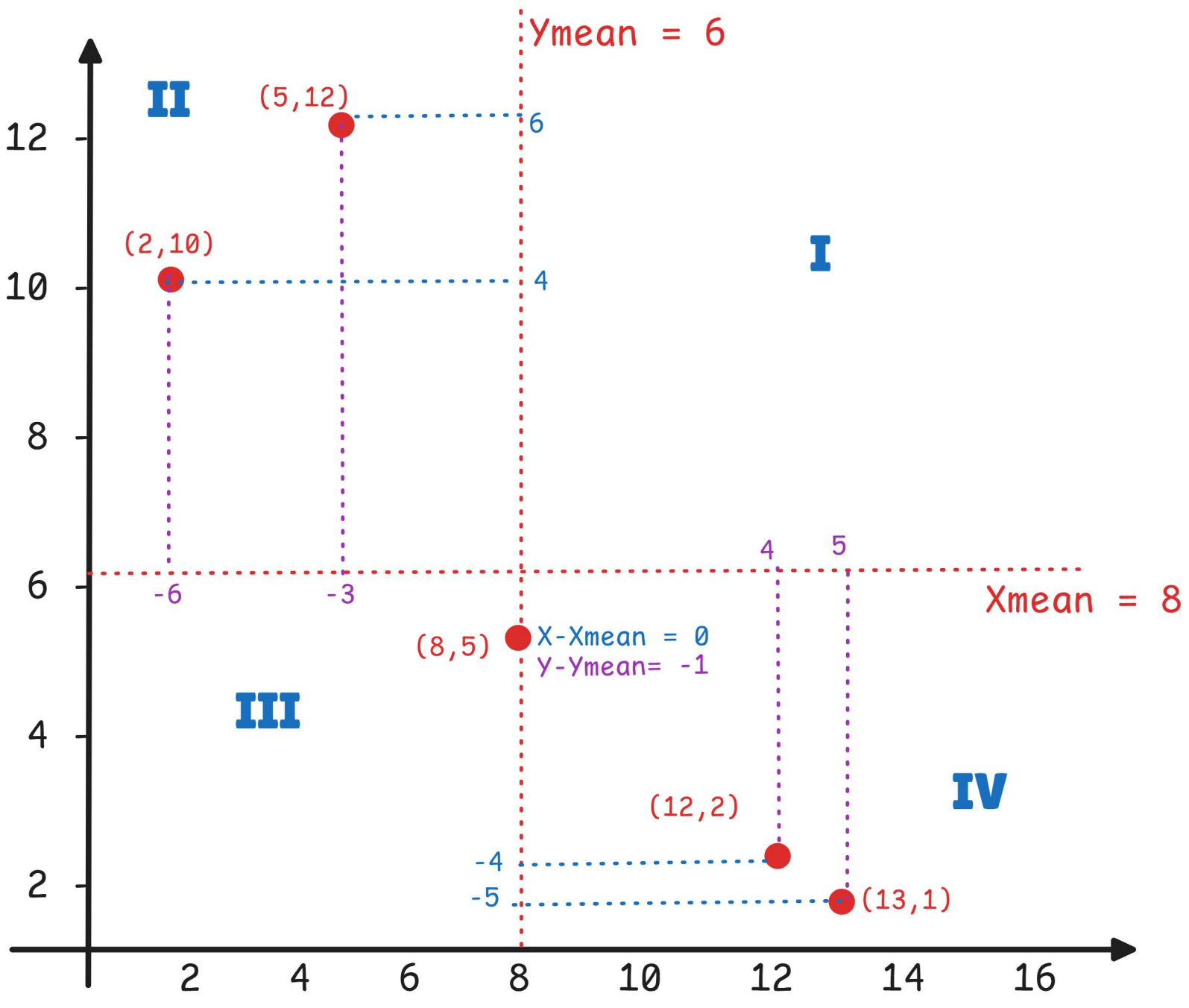
Backlogs (x)	Package (y)	X – X̄	Y – Ȳ	(X – X̄)(Y – Ȳ)
2	10			
5	12			

Backlogs (x)	Package (y)	X − X̄	Y − Ȳ	(X − X̄)(Y − Ȳ)
8	5			
12	2			
13	1			

Solution:

Step-1: Let's calculate the mean first. $\bar{X}=8$ $\bar{Y}=6$

Backlogs (x)	Package (y)	X − X̄	Y − Ȳ	(X − X̄)(Y − Ȳ)
2	10	-6	4	-24
5	12	-3	6	-18
8	5	0	-1	0
12	2	4	-4	-16
13	1	5	-5	-25
			Σ	**-83**



Here the relationship is positive

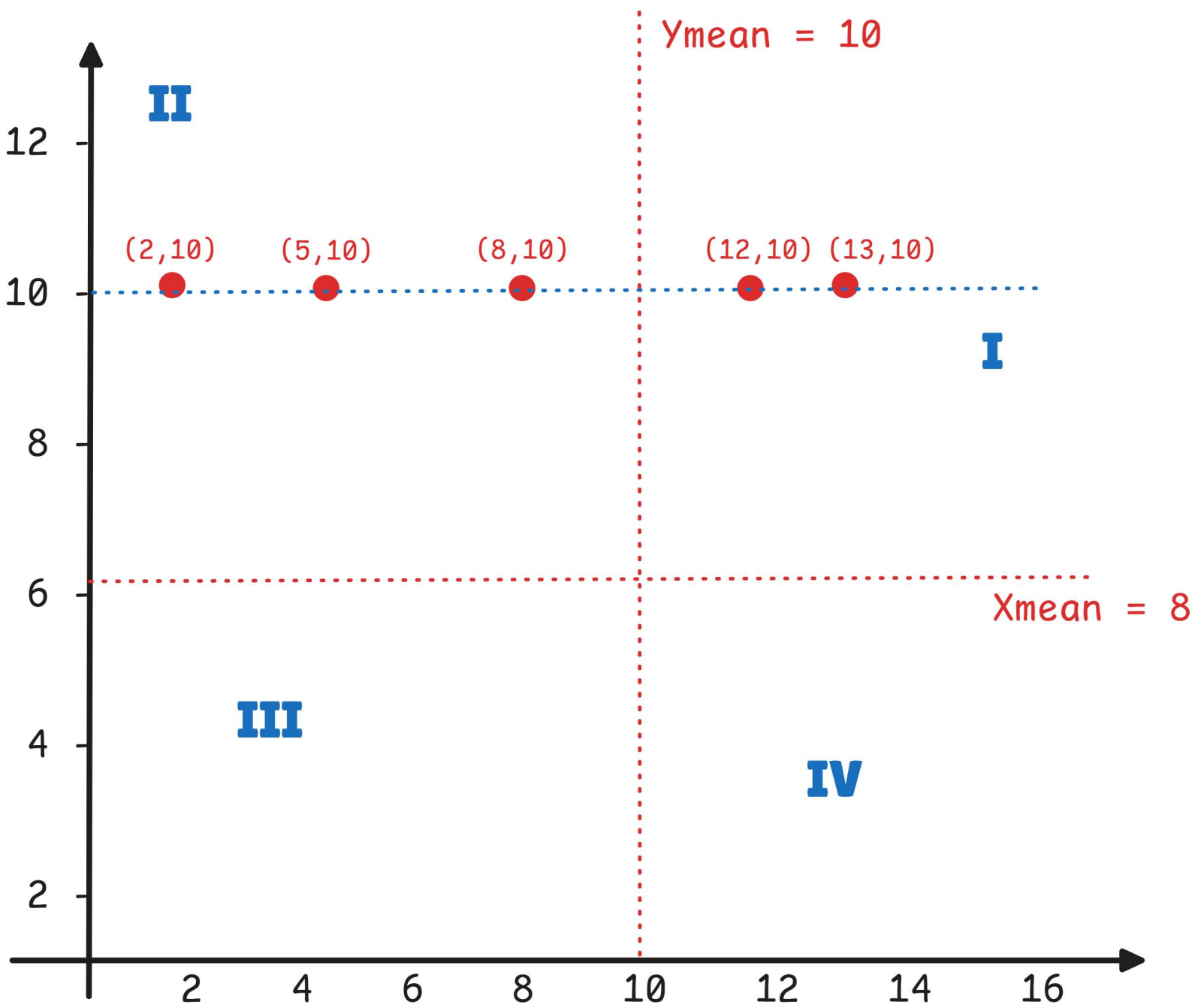
Ques.3:

Backlogs (x)	Package (y)	X − X̄	Y − Ȳ	(X − X̄)(Y − Ȳ)
2	10			
5	10			
8	10			
12	10			
13	10			

Solution:

Step-1: Let's calculate the mean first. `X̄=8` `Ȳ=10`

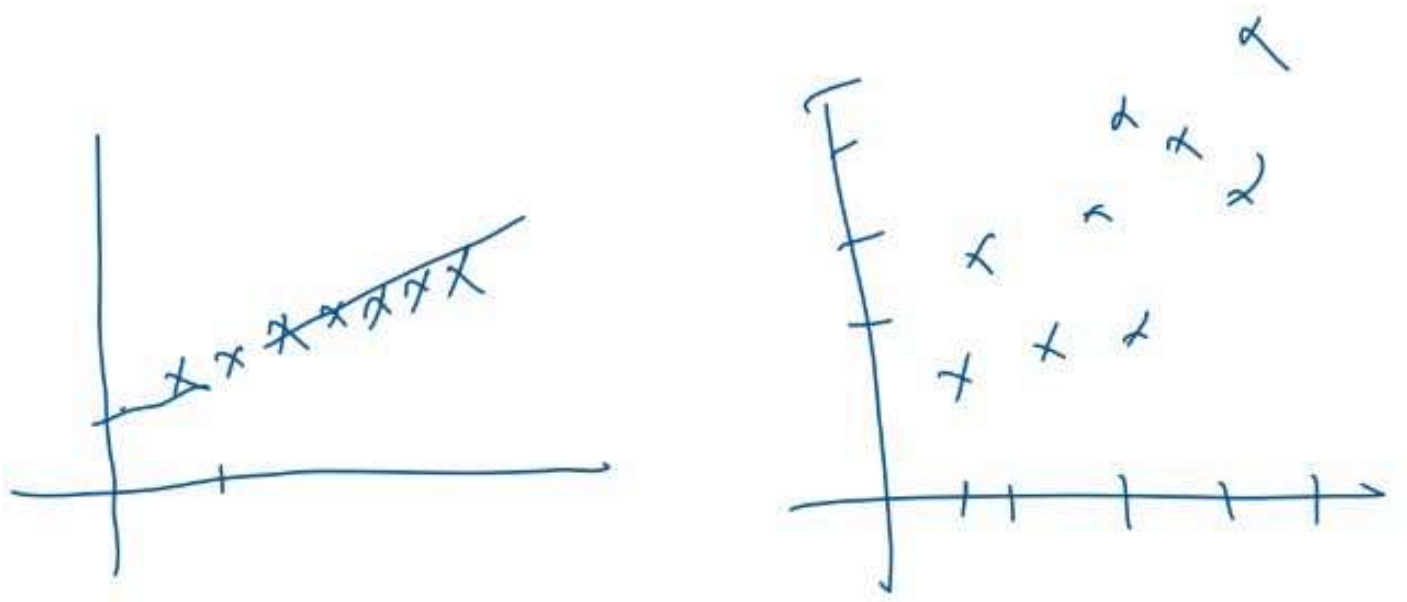
Backlogs (x)	Package (y)	X − X̄	Y − Ȳ	(X − X̄)(Y − Ȳ)
2	10	-6	0	0
5	10	-3	0	0
8	10	0	0	0
12	10	4	0	0
13	10	5	0	0
				0



Here no relationship is there

Disadvantages of Covariance

One limitation of covariance is that it only tells about if relationship is positive or negative, but it does not tell about the strength of the relationship between two variables.



In this figure, both the relationship is positive, but both are differnt in terms of strength.

> **Note:** The covariance we calculated after adding all quantity, that does not tell us about the strength, it changes as we change the scale and the relationship remains same.

> By changing scale we mean, if we multiply all x value by 2 and y value by 2. The relationship will look same, but the covariance changes.

Covariance of a variable by itself became the **Variance**

$$Covariance = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Let's We have only 1 variable, that is X

$$Covariance = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

It will become the variance:

$$Covariance = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = Variance$$

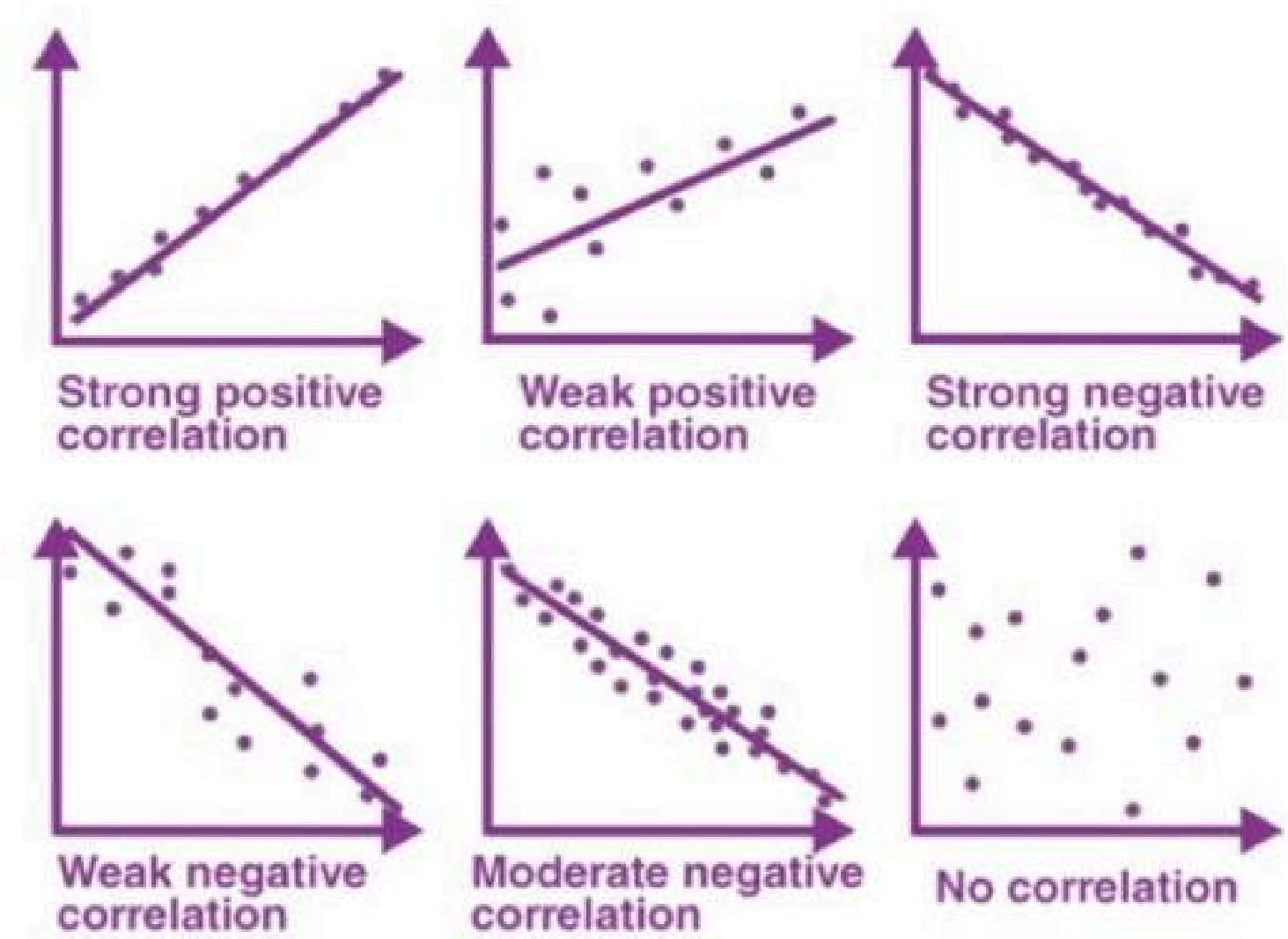
Variance = Covariance of 1 variable:

$$Variance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Co-relation

1. What problem does Co-relation Solve

- It quantified the relationship of linear relationship.
- In short, the shortcoming of Co-variance is that, it tells us about relationship but not about the strength of relationship. But Co-relation tells us about the strength(magnitude) of relationship between 2 variables in a linear graph.



2. What is Correlation?

- Correlation refers to a statistical relationship between two or more variables. Specifically, it measures the degree to which two variables are related and how they tend to change together.
- Correlation is often measured using a statistical tool called the correlation coefficient, which ranges from -1 to 1.
    - A correlation coefficient of -1 indicates a perfect negative correlation,
    - a correlation coefficient of 0 indicates no correlation, and
    - a correlation coefficient of 1 indicates a perfect positive correlation.

> **Note:** The more toward 0 is less strong. The more towards -1 & +1 is more strong negative or positive correlation

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$

For finding out Corelations, We divide the Co-variance by standard deviation of X and Y

> **Note:** Co-relation is a very reliable statistical measure, because it is No affected by scaling (even if we multiply values of X and Y by same number i.e. increase the scale by 2) the co-realtion does not change.

> > That is why, when we will study Linear Regression or try to understand the linear relatioship between 2 Numerical Quantity, we would always perform a co-relation analysis and not co-variance analysis (because co-variance is not reliable.)

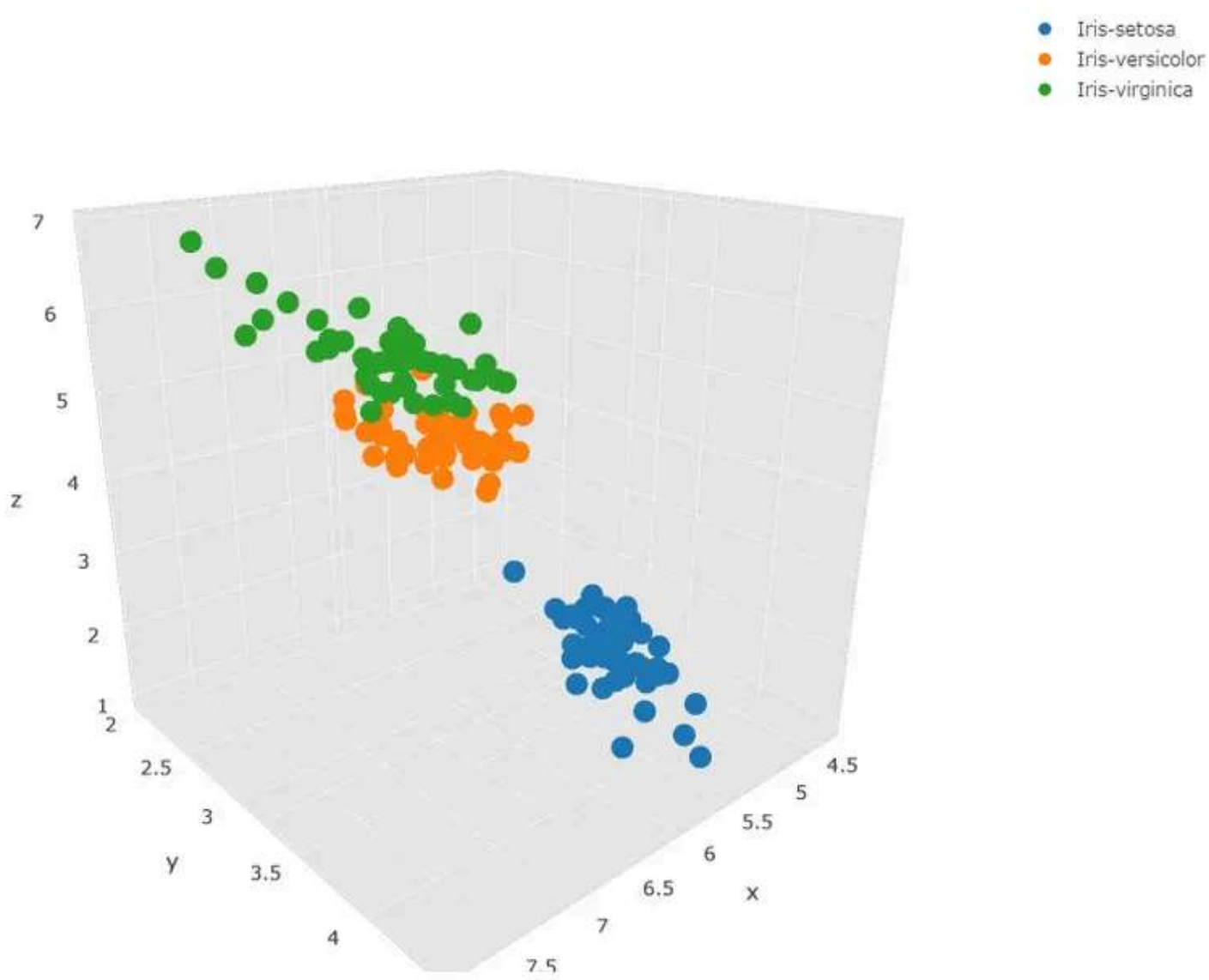
Co-relation and Causation

- The phrase **"correlation does not imply causation"** means that even if two variables are related to each other, it does not necessarily mean that one causes the other. They ar erelated but it doesnot mean 1 variable affect behaviour of other variable.
- For example, imagine there is a positive correlation between the **number of firefighters** at a fire and the amount of **damage caused**. It would be wrong to conclude that firefighters create more damage. The real explanation is a third variable: the severity of the fire. Larger fires both need more firefighters and cause more damage.
- So while correlations are useful for spotting relationships between variables, they cannot prove causality. To establish causation, we need stronger evidence- like controlled experiments, randomized trials, or carefully designed observational studies.
- Few more examplees, where co-relation doesnot mean causation is.

- **Salary VS Experience:** Other factors also decide salary (maybe company type, employee skill etc.)
- **Education Level vs. Income:** Higher education is correlated with higher income, but education alone doesn't cause high income. Family background, networking, economic conditions, or even geography also play a role.

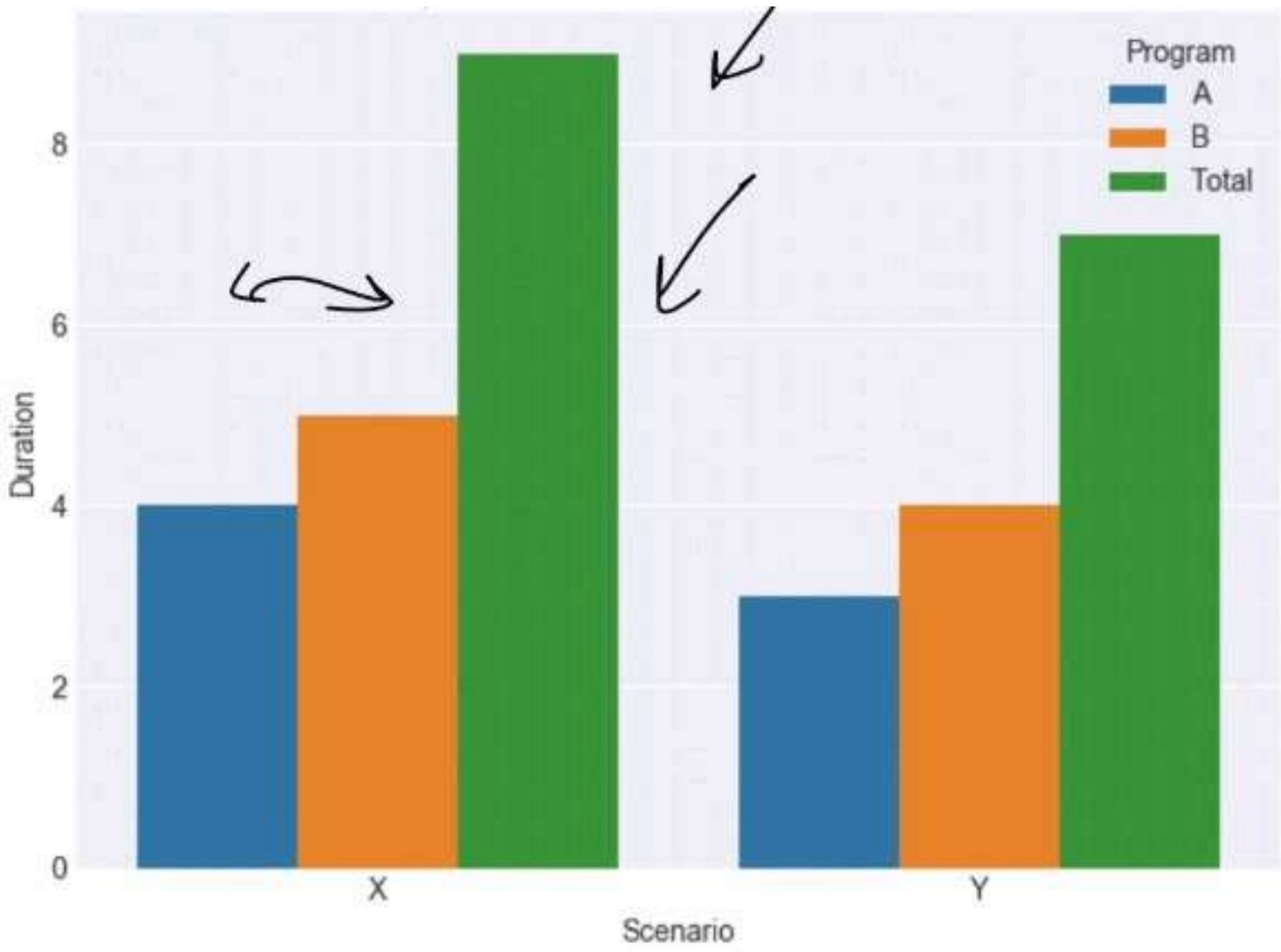
Visualizing Multiple Variables:

1. 3D Scatter Plots



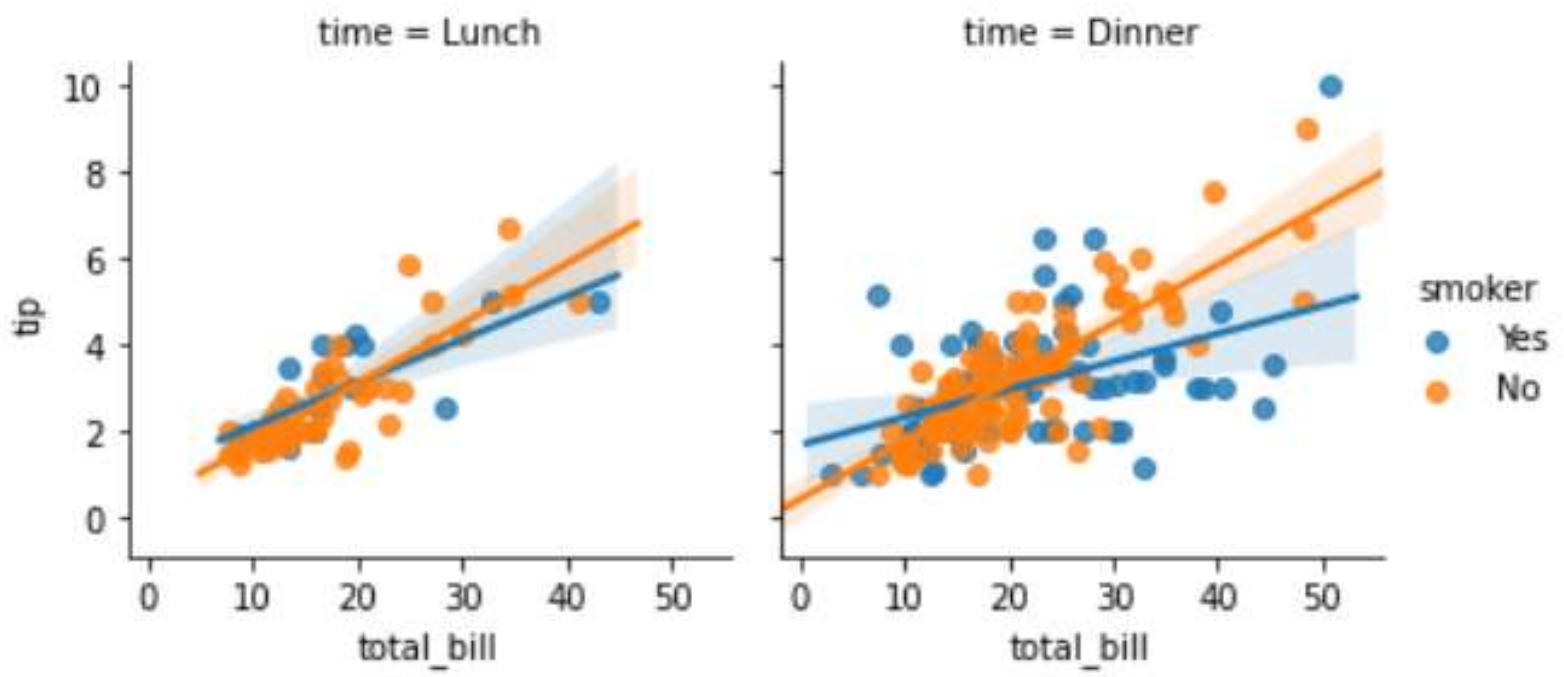
2. Hue Parameters.

Analysis between Category-Category-Numerical

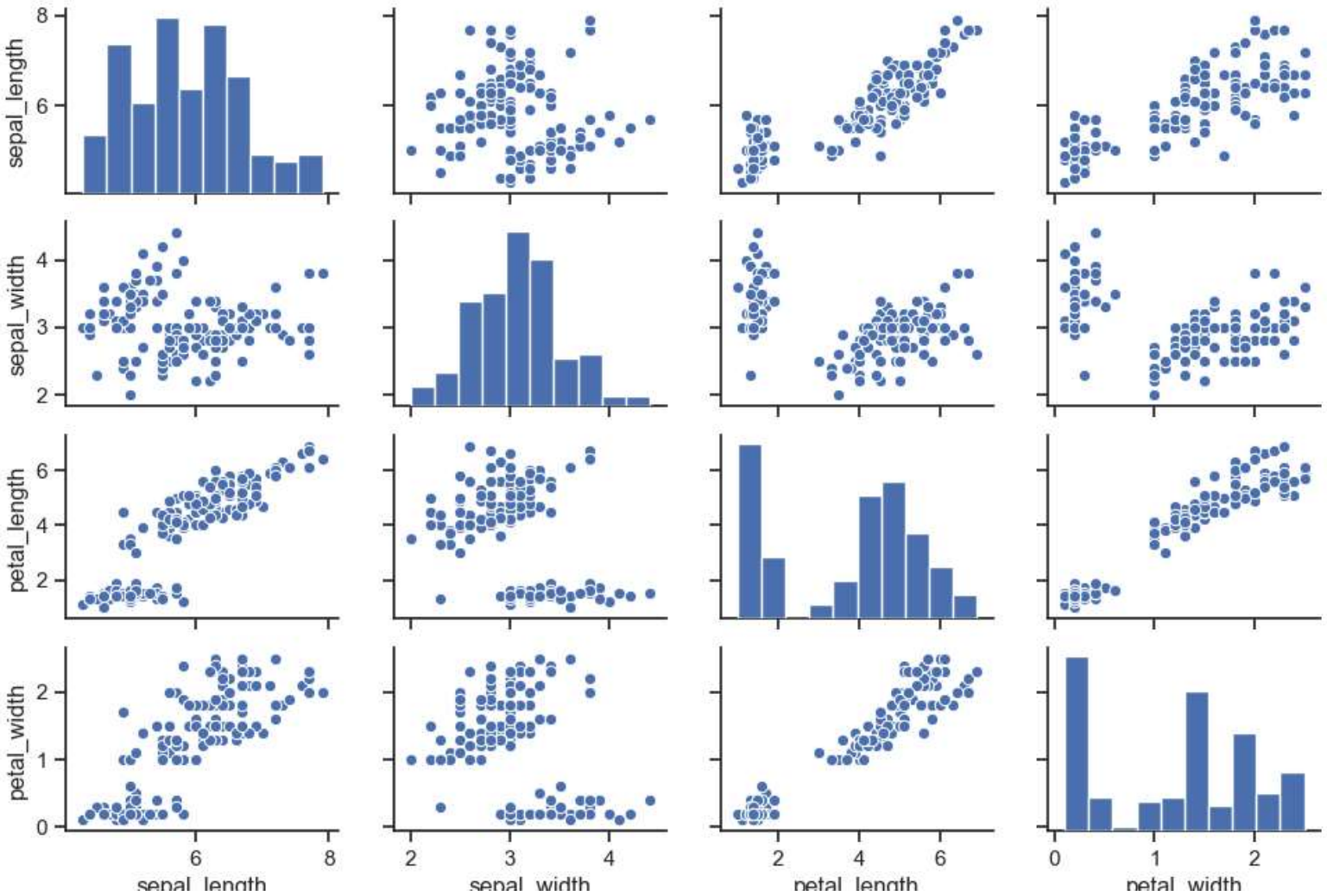


3. Facet Grid

- Studying 2 scattered plot in 1 Facet Grid (4 columns at a time)



4. Pair Plots



5. Bubble Chart