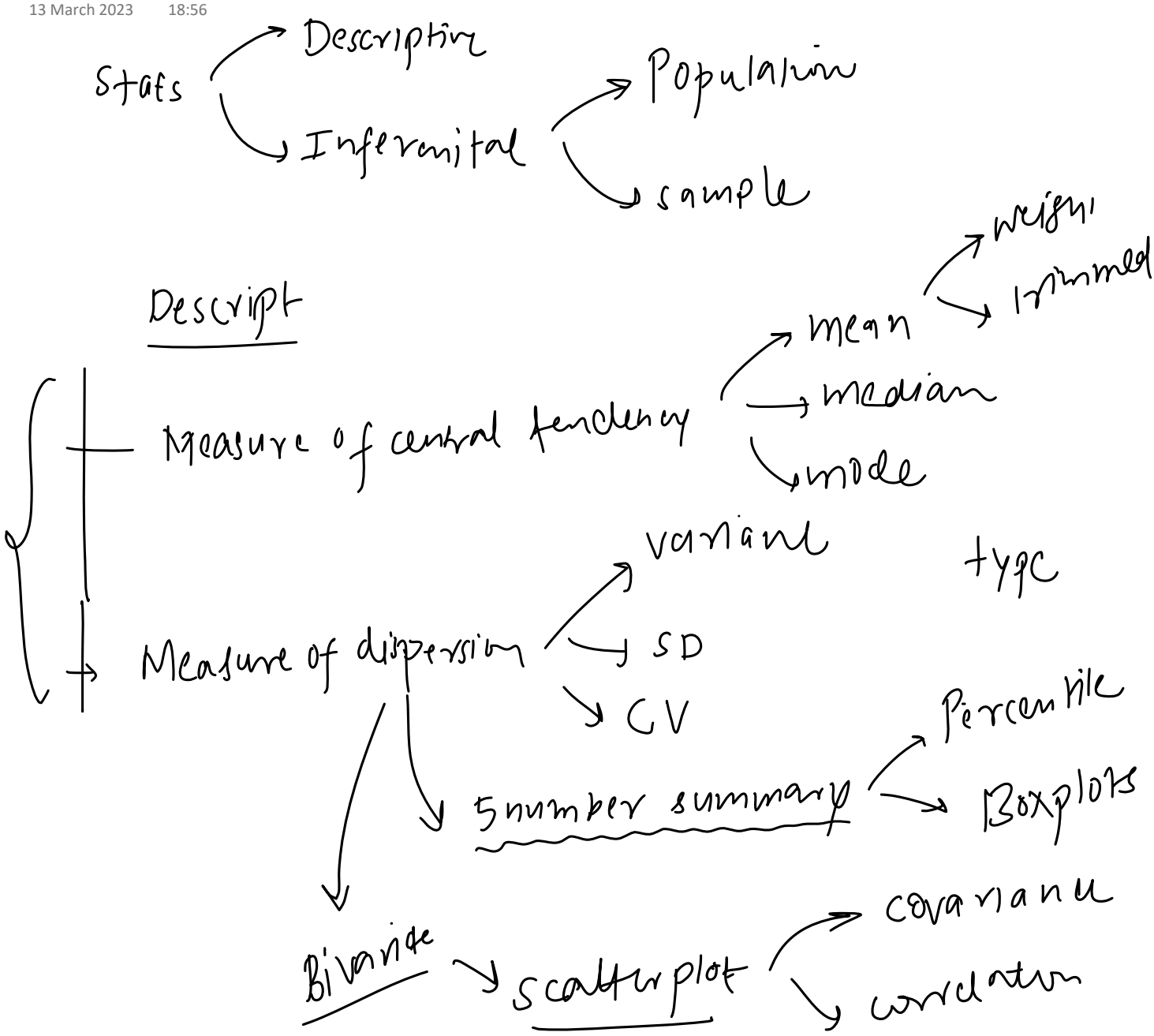
Recap

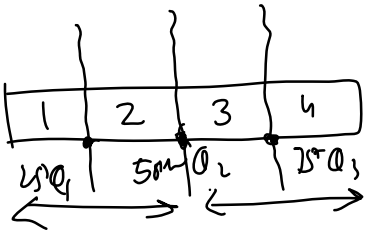
13 March 2023 18:56



Quantiles and Percentiles

13 March 2023 06:57

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



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:

a. Quartiles: Divide the data into four equal parts, Q1 (25th percentile), Q2 (50th percentile or median), and Q3 (75th percentile).

b. Deciles: Divide the data into ten equal parts, D1 (10th percentile), D2 (20th percentile), ..., D9 (90th percentile).

c. Percentiles: Divide the data into 100 equal parts, P1 (1st percentile), P2 (2nd percentile), ..., P99 (99th percentile).

d. 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

Percentile

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.

Formula to calculate the percentile value:

$$PL = \frac{p}{100}(N+1)$$

where:

- PL = the desired percentile value location
- N = the total number of observations in the dataset
- p = the percentile rank (expressed as a percentage)

Example:

Find the 75th percentile score from the below data

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

Step1 - Sort the data (Asc)

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

1 2 3 4 5 6 7 8 9 10

$$PL = \frac{75}{100}(10+1) = \frac{3}{4} \times 11 = \frac{33}{4} = 8.25$$

96 —— 98 Q75 91 —— 93

100

96 —— 98 (0.75) 91 —— 93

8 —— 9 5 6

↑

$96 + 0.25(98 - 96) = 96 + 0.25 \times 2 = 96.5$

↘

75th percentile = 96.5

$91 + 0.5(93 - 91)$

$91 + 0.5 \times 2 = 92$

$P_L = \frac{50}{100}(10+1) = \frac{1}{2} \times 11 = 5.5$

Percentile of a value

Percentile rank = $\frac{x + 0.5y}{n}$ $\left(\frac{n}{N}\right)$

X = number of values below the given value

Y = number of values equal to the given value

N = total number of values in the dataset

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

1 2 3

$= \frac{3 + 0.5 \times 1}{10} = \frac{3.5}{10}$

0.35 → 35th

$\frac{9 + 0.5 \times 1}{10} = \frac{9.5}{10} = 0.95$

$\frac{9}{10}$ 0.9

5 number summary

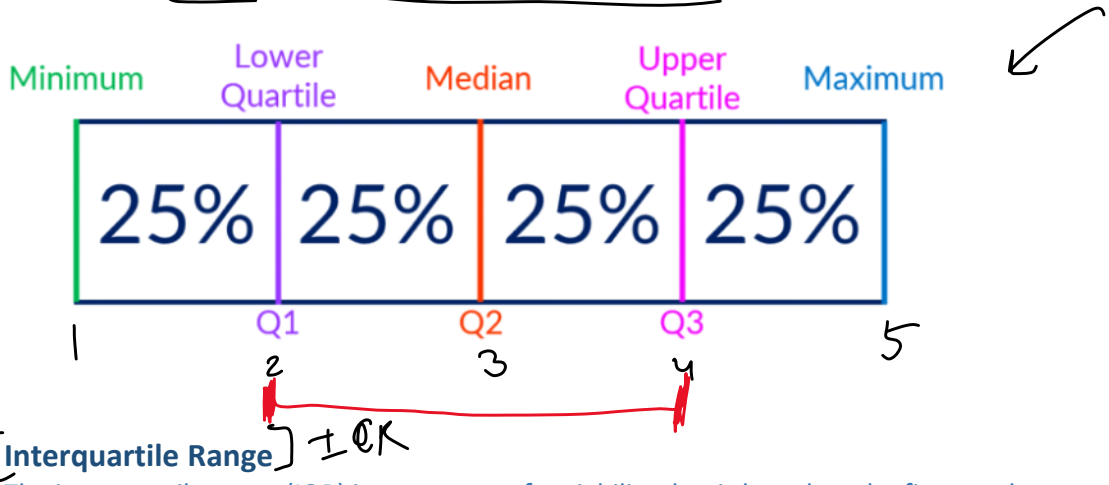
13 March 2023 06:57

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

(descriptive)

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.



Interquartile Range

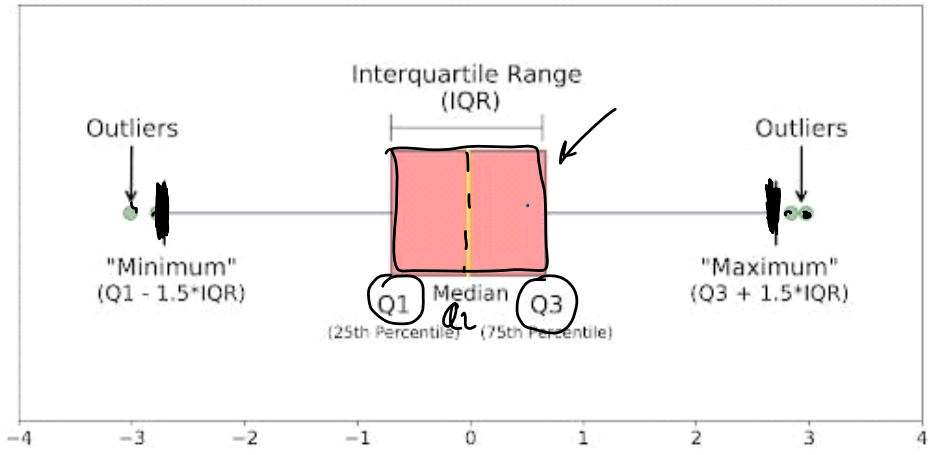
The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.

Boxplots

13 March 2023 06:57

1. What is a 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).



2. How to create a boxplot with example

6 213 241 260 281 290 314 321 350 1500

1 2 3 4 5 6 7 8 9 10

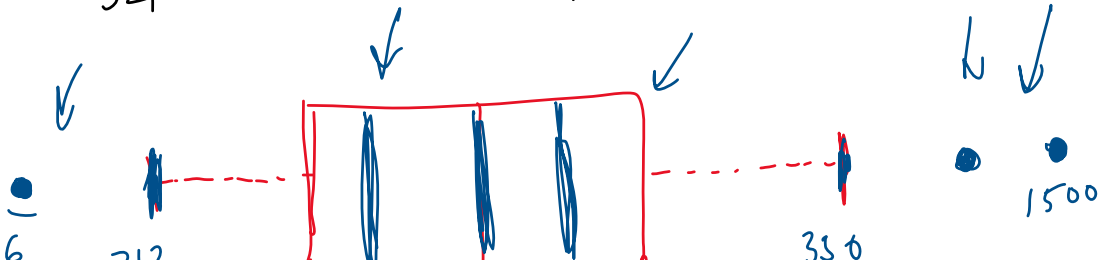
$Q_2 = \frac{50}{100}(11) = 5.5 = 285.5$

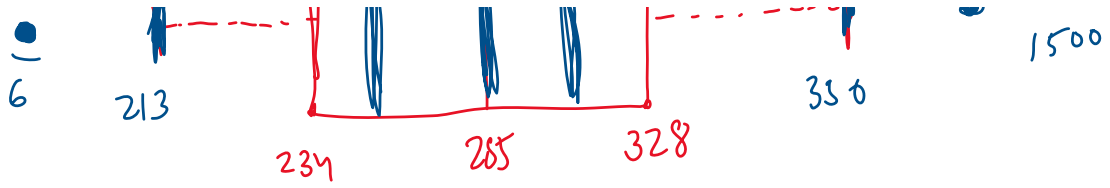
$Q_1 = \frac{25}{100} \times 11 = \frac{11}{4} = 2.75$

$213 + 0.75(241 - 213) = 234$

$Q_3 = \frac{75}{100} \times 11 = \frac{33}{4} = 8.25$

$321 + 0.25(350 - 321) = 328.25$





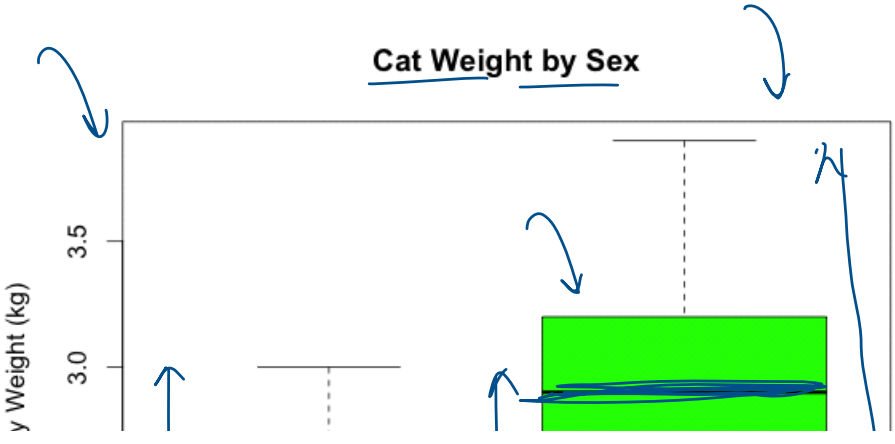
min and max

IQR = 94

$\min = Q_1 - 1.5(IQR) = 93$

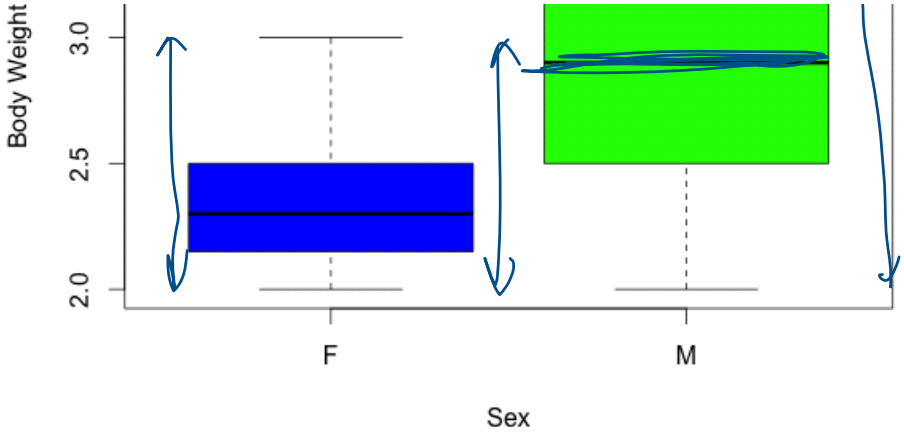
$\max = Q_3 + 1.5(IQR) = 469$

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. Side by side boxplot



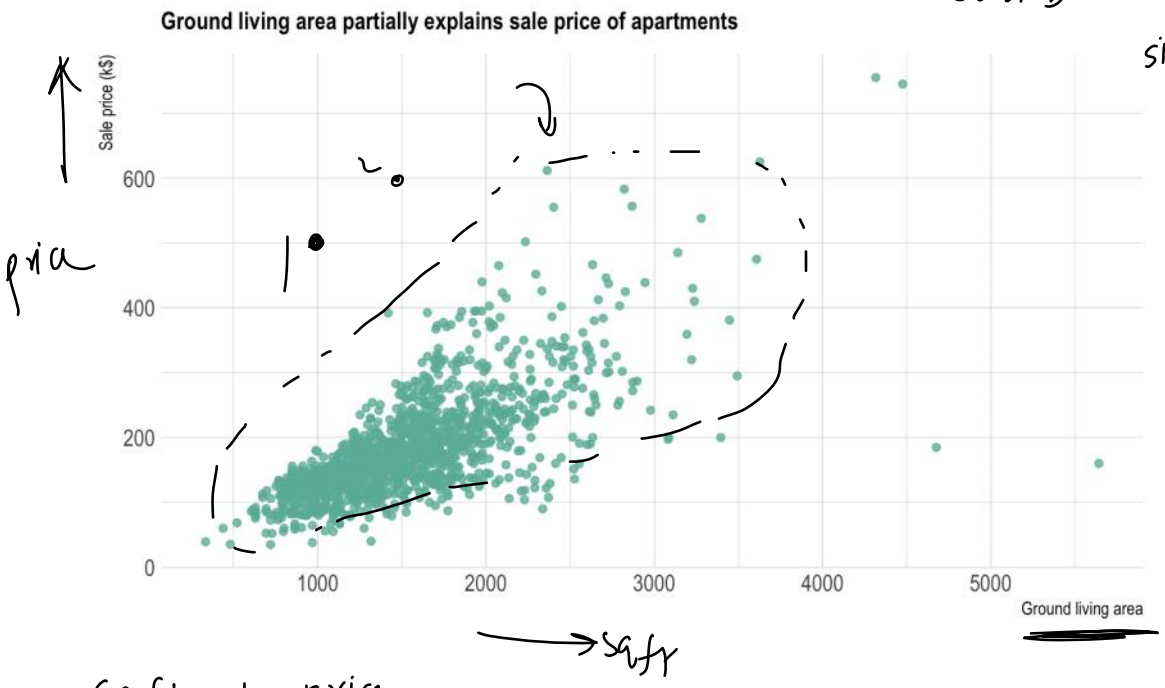
Weight

Age	gender
21	M
34	F
43	M


Body Weight
3.0
2.5
2.0
F
M
Sex

Scatterplots

13 March 2023 06:58

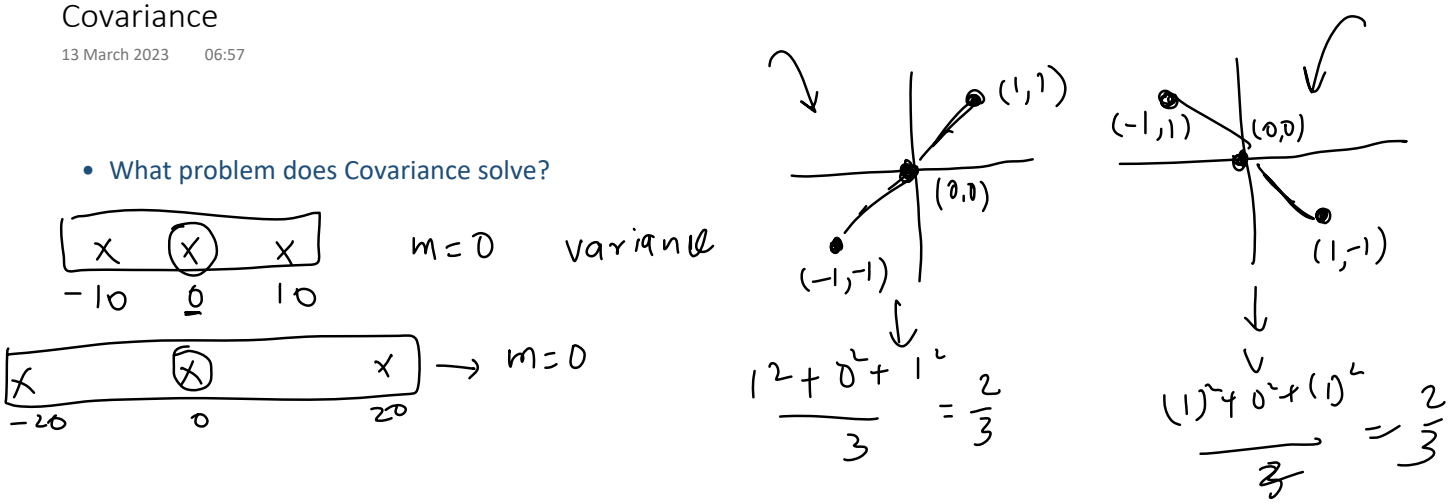


Sqft	price
1500	60L
1700	80L
1200	35L

Covariance

13 March 2023 06:57

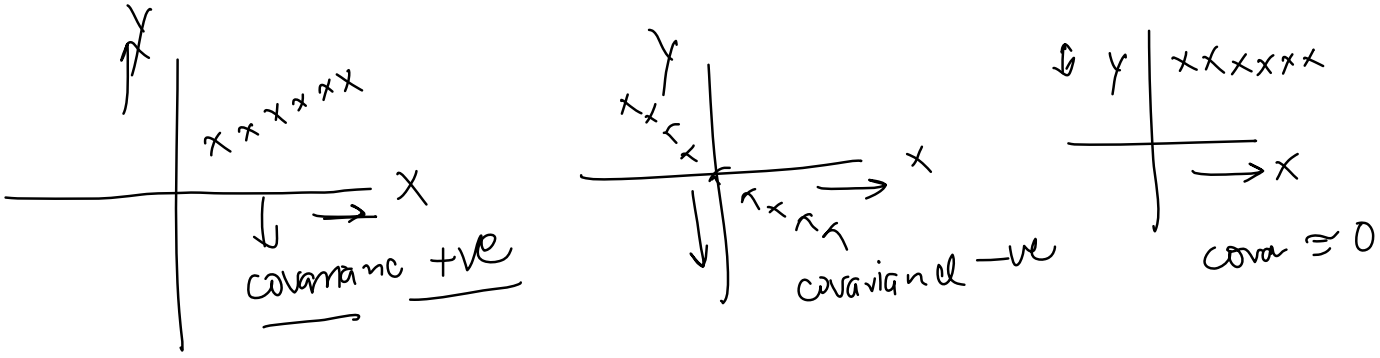
- What problem does Covariance solve?



- What is covariance and how is it interpreted?

Covariance is a statistical measure that describes the degree to which two 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?

If the covariance between two variables is positive, it means that the variables tend to move together in the same direction. If the covariance is negative, it means that the variables tend to move in opposite directions. 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*

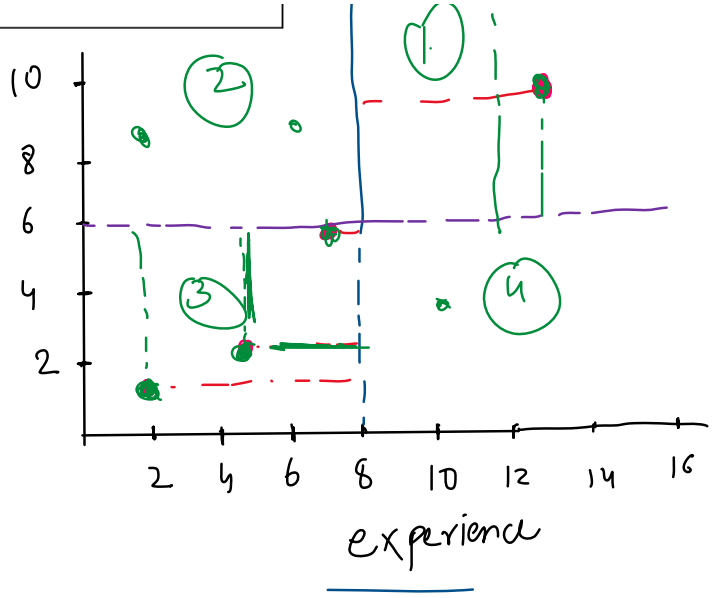
Exp(x)	Salary(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	-1
12	12	4	6	24
13	10	5	4	20

$\bar{X} = 8 \quad \bar{Y} = 6$

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

$n-1$

$\frac{85}{4} =$ COV = 21.5

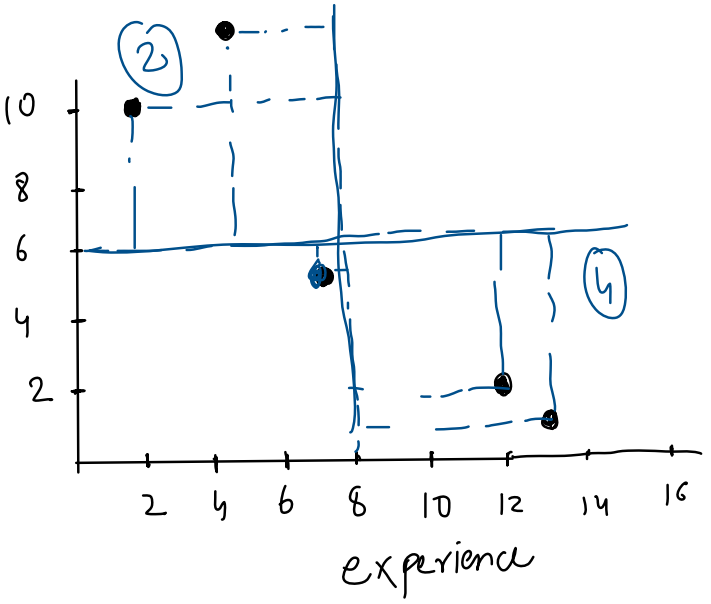


Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10	-6	4	-24
5	12	-3	6	-18
8	5	0	1	0
12	2	4	-4	-16
13	1	5	-5	-25

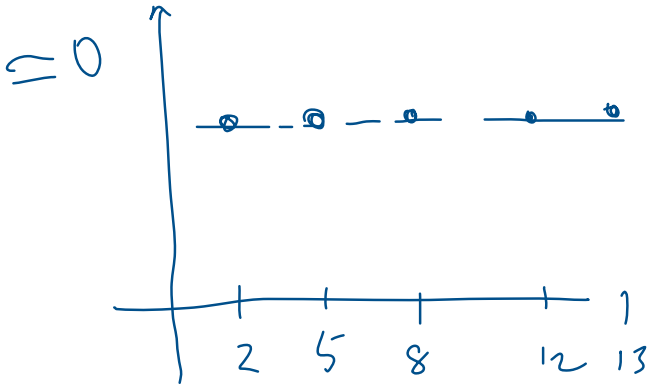
$\bar{X} = 8 \quad \bar{Y} = 6$

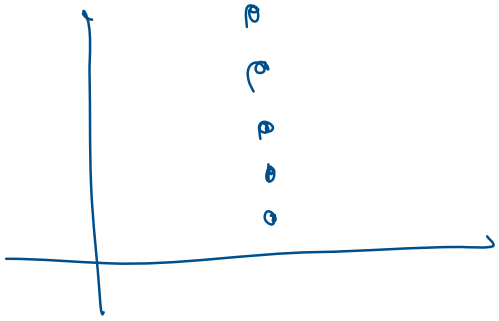
$\frac{-83}{4}$

-21.3



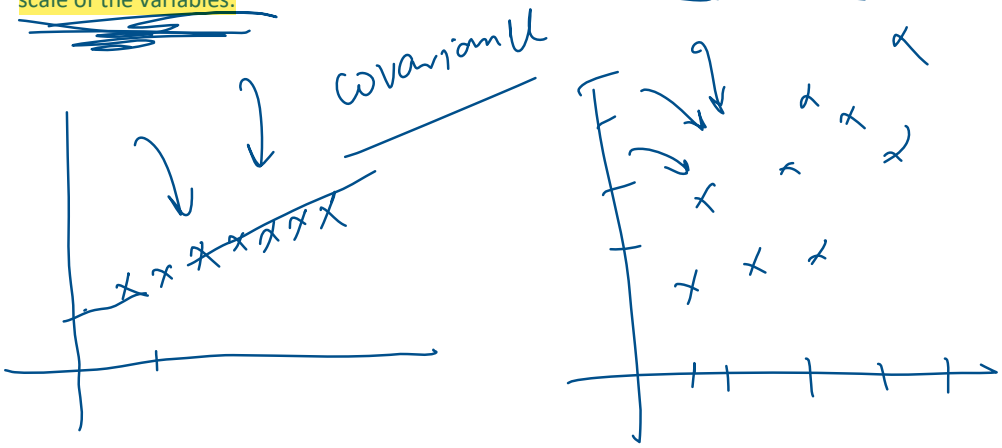
Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10			
5	10			
8	10			
12	10			
13	10			





- Disadvantages of using Covariance

  One limitation of covariance is that it does not tell us about the strength of the relationship between two variables, since the magnitude of covariance is affected by the scale of the variables.



- Covariance of a variable with itself

$$\sum (x-\bar{x})(y-\bar{y})$$

$$\sum_{i=1} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

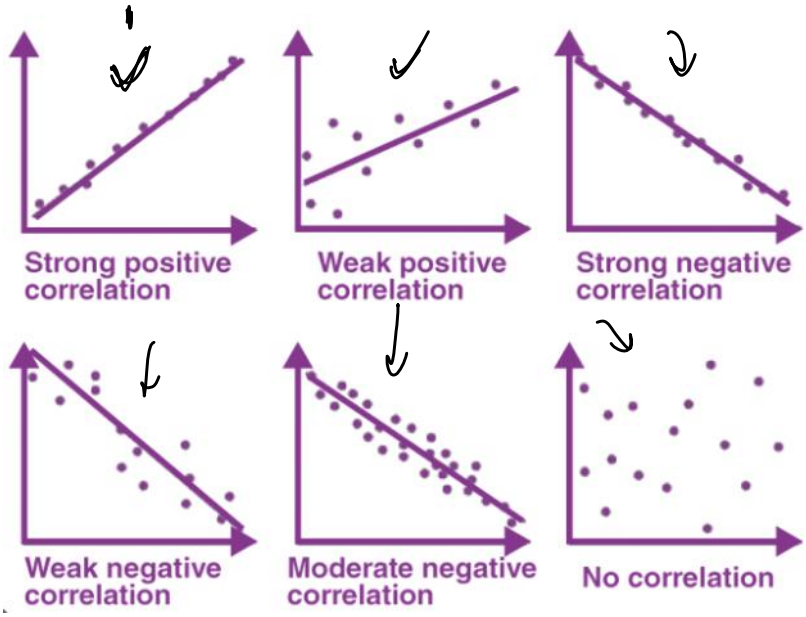
$$= \sum_{i=1}^{n} \frac{(x-\bar{x})(x-\bar{x})}{n-1}$$

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

Correlation

13 March 2023 06:58

1. What problem does Correlation solve?



Can we quantify this weak and strong relationship?

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.

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

Correlation and Causation

13 March 2023 18:31

The phrase "**correlation does not imply causation**" means that just because two variables are associated with each other, it does not necessarily mean that one causes the other. In other words, a correlation between two variables does not necessarily imply that one variable is the reason for the other variable's behaviour.

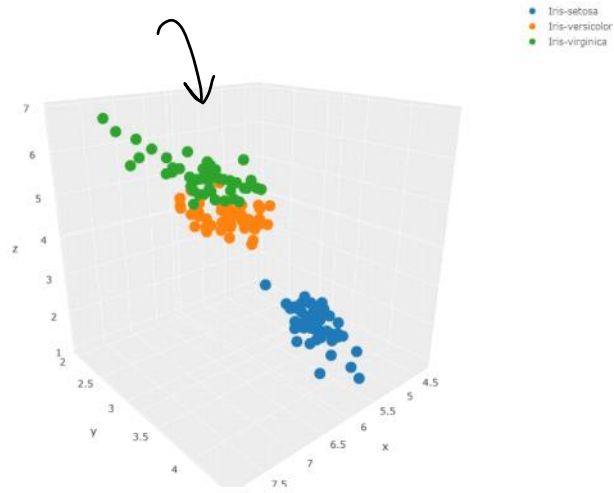
Suppose there is a positive correlation between the number of firefighters present at a fire and the amount of damage caused by the fire. One might be tempted to conclude that the presence of firefighters causes more damage. However, this correlation could be explained by a third variable - the severity of the fire. More severe fires might require more firefighters to be present, and also cause more damage.

Thus, while correlations can provide valuable insights into how different variables are related, they cannot be used to establish causality. Establishing causality often requires additional evidence such as experiments, randomized controlled trials, or well-designed observational studies.

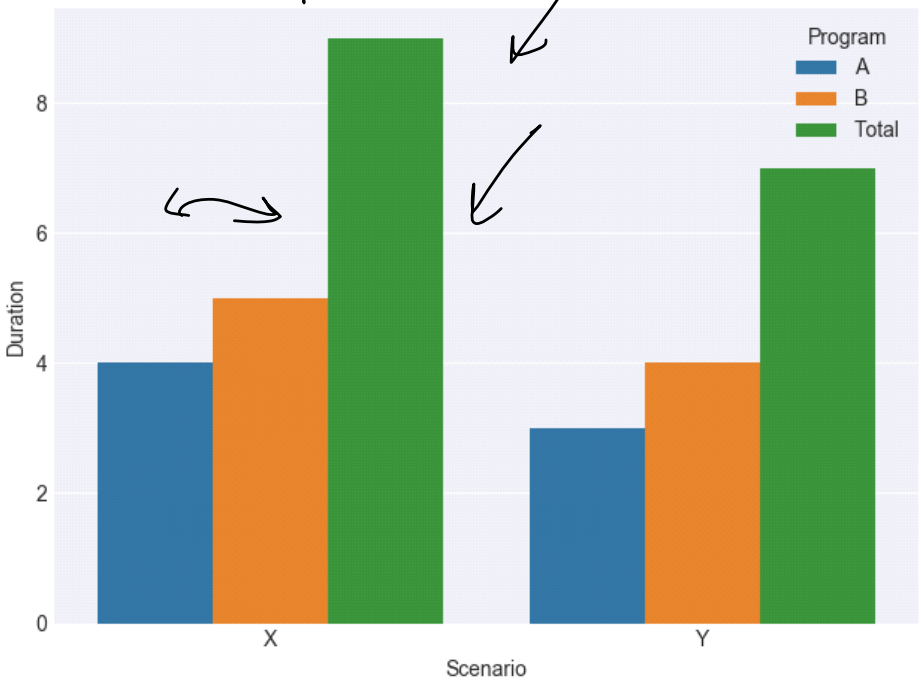
Visualizing Multiple Variables

13 March 2023 06:58

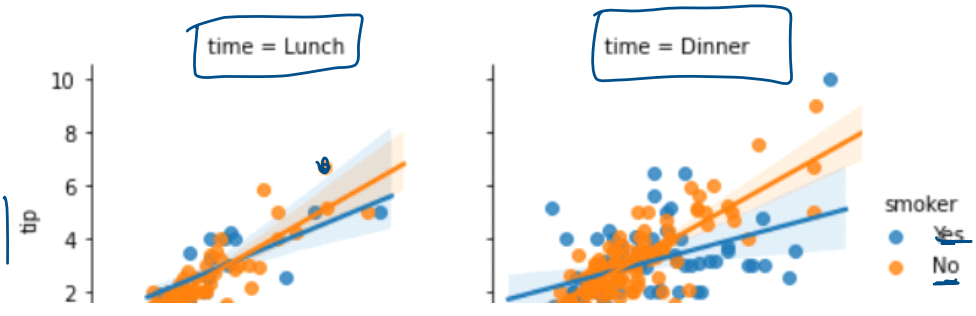
1. 3D Scatter Plots

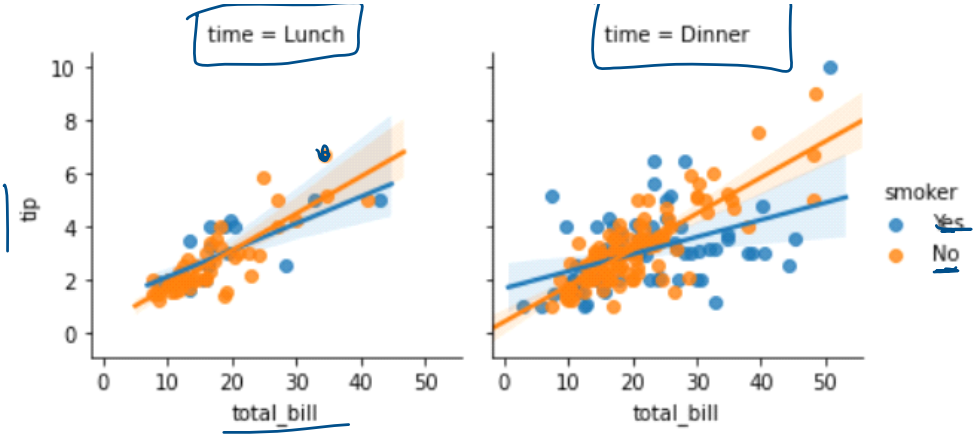


2. Hue Parameter

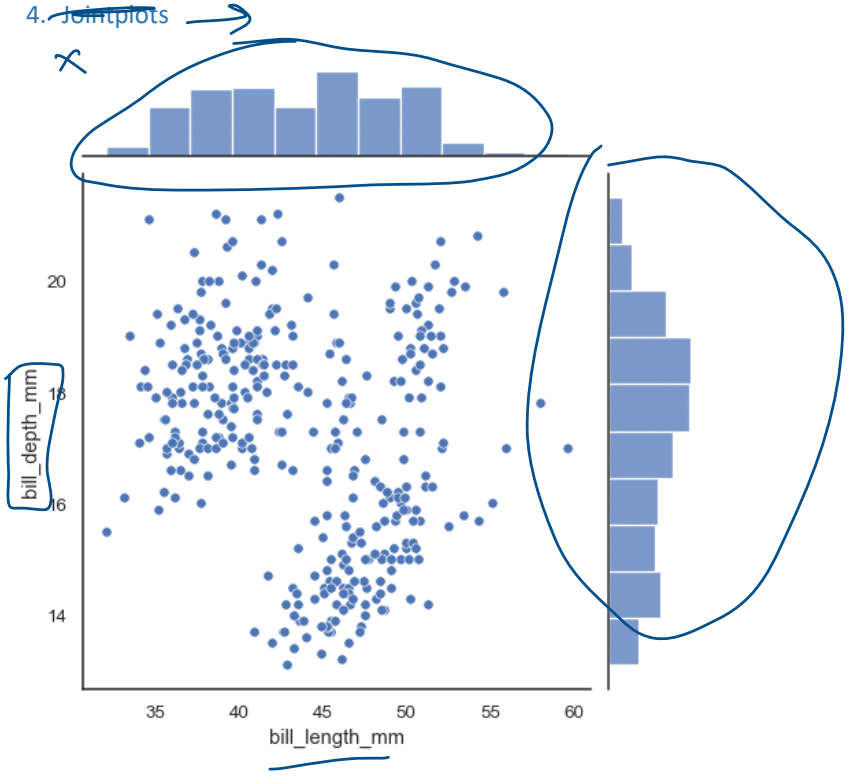


3. Facetgrids

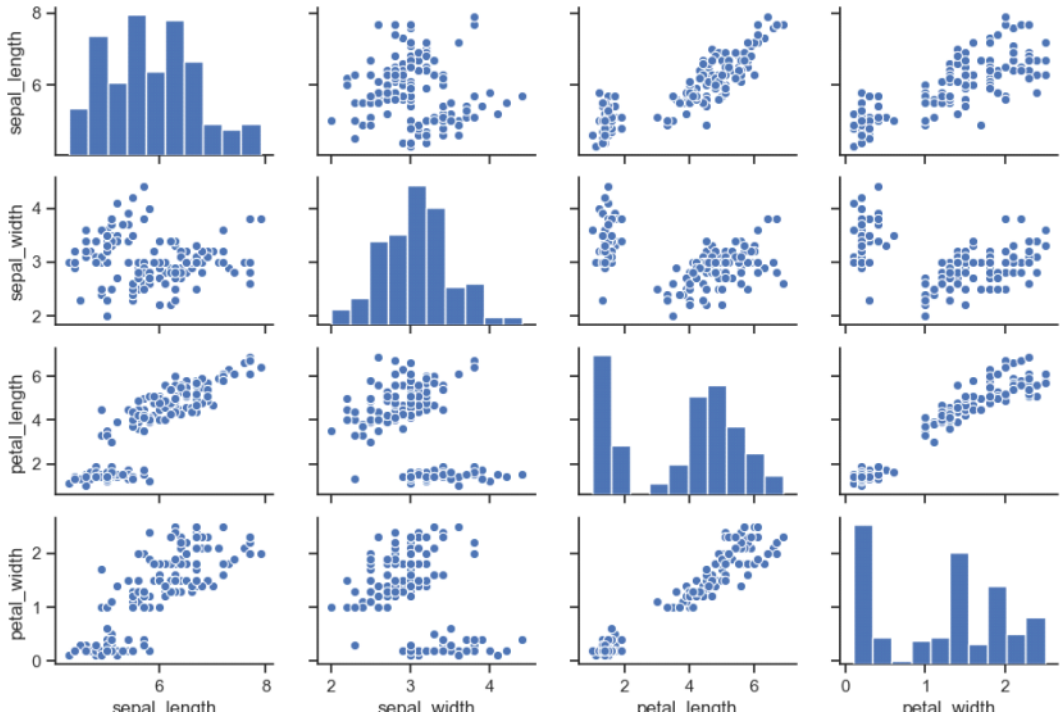




4. Jointplots



5. Pairplots



6. Bubble Plots