Recap

Stats → Descriptive

Stats → Inferenital → Population

Inferenital → sample

Descript

Measure of central tendency → mean → weight

mean → trimmed

→ median

→ mode

Measure of dispersion → variance

→ SD

→ CV

type

5 number summary → Percentile

5 number summary → Boxplots

Bivariate → scatterplot → covariance

scatterplot → correlation

multivariate analysis

Quantiles and Percentiles

We are just trying to find out the location of an observation
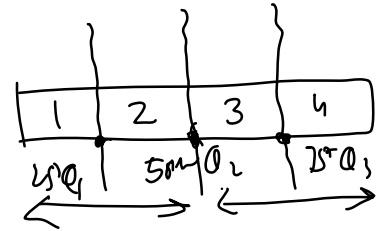
Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.

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

1. Value of percentile.
2. Percentile of value

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:

1. Value of percentile.

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

1. find the location or index. 2. then particular element or value (may be or may not be present in the column)

where:

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

Ques - Value of 25th percentile, Value of 15th percentile

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

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

$$\downarrow \text{ Index}$$

$$8^{th} \quad 9^{th}$$

$$\leftarrow 96 - 98$$

$$= 96 + 0.25(98-96)$$

75th percentile is $= 96.5$

Solution: $\dfrac{20}{100} \times (10+1) = \dfrac{11}{5} = 2\cdot2 \rightarrow 2-3$

Find Value 20th percentile.

$$82 - 84$$

$$20^{th}\text{ percentile} = 82 + 0\cdot2(2) \checkmark$$
$$= 82\cdot4$$

2. Percentile of Value

The percentage of values in a dataset that are equal to or less than that value.

Ques - Percentile of 88, Percentile of 94.

Percentile rank = $\dfrac{x}{N} \times 100$

X = number of values below and equal the given value

N = total number of values in the dataset

78, 82, 84, 88, 91, 93, 94, 96, 98, 99
  1   2   3

For 88 $\rightarrow$
$$= \dfrac{4}{10} \times 100 = 40^{th}\text{ percentile}$$

For 99 $\rightarrow$
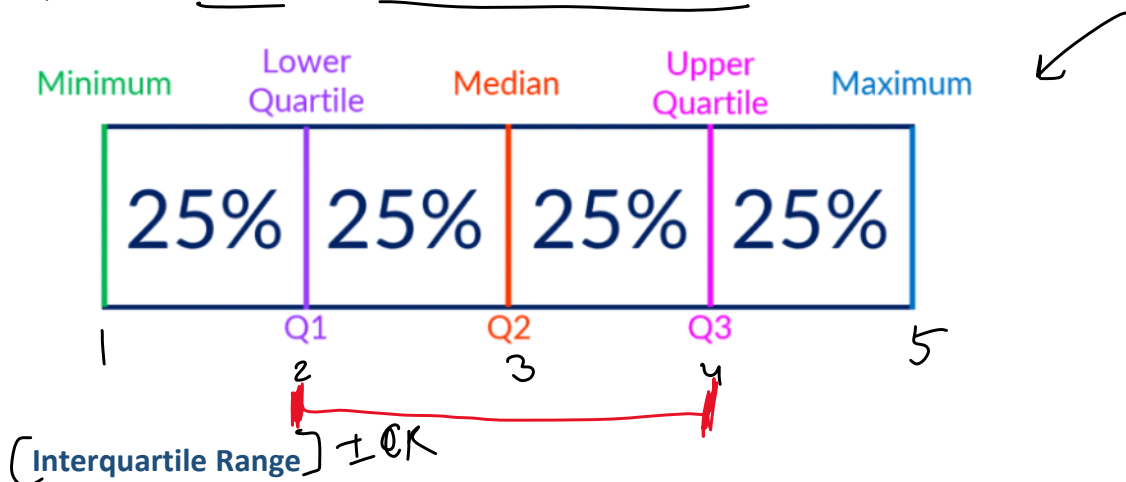$$= \dfrac{10}{10} \times 100 = 100^{th}\text{ percentile}$$

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:

(describe)

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.

Minimum	Lower Quartile	Median	Upper Quartile	Maximum
25%	25%	25%	25%	
	Q1	Q2	Q3	
1	2	3	4	5

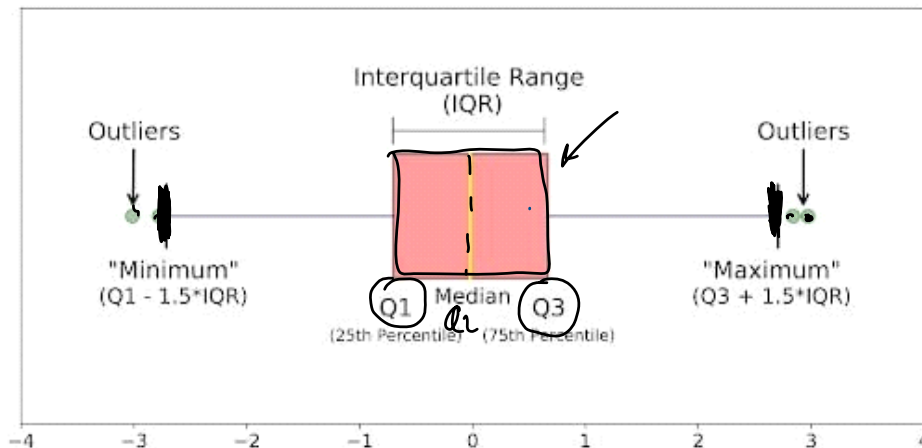**Interquartile Range**] $\pm$ IQR

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.
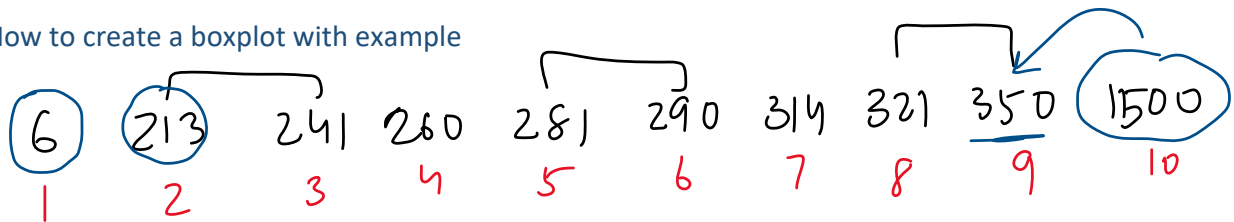
Boxplots

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).
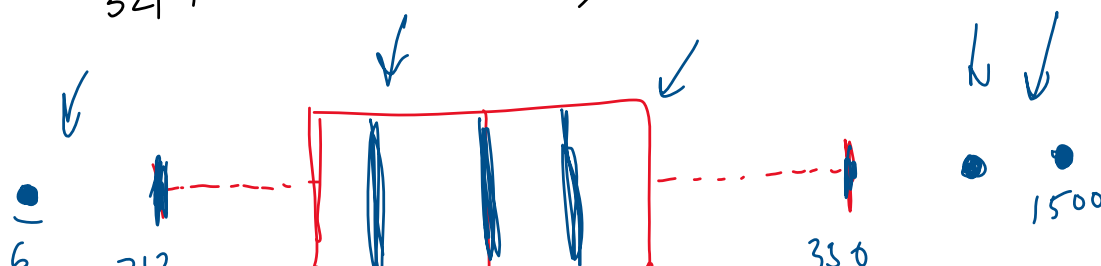

2. How to create a boxplot with example

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

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

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

$$213 + 0.75(241-213) = \boxed{234}$$

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

$$321 + 0.25(350-321) = \boxed{328.25}$$

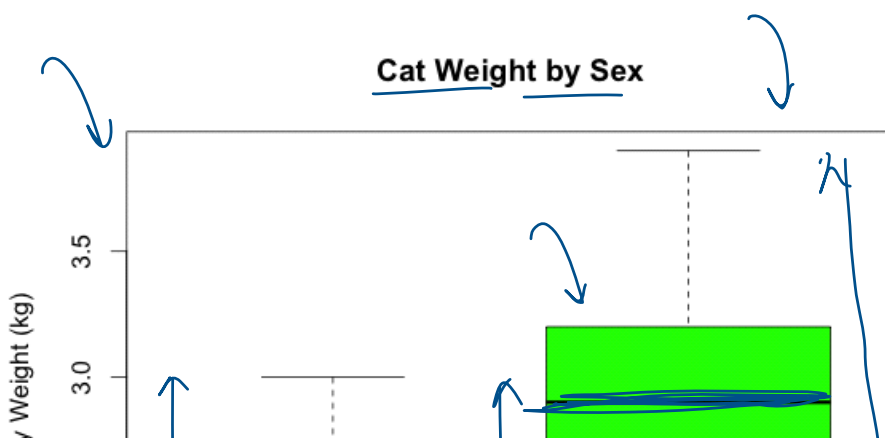6   213   330   1500

6 213
234 285 328 350 1500
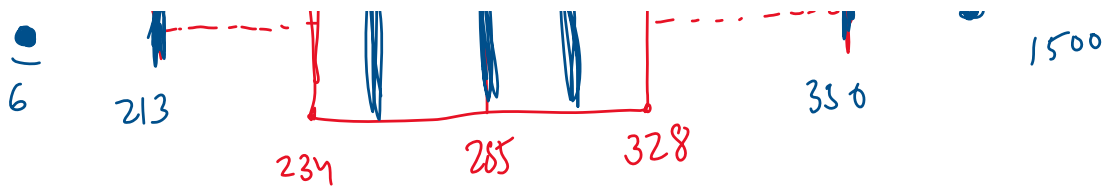
IQR = 94

min and max

$$min = Q_1 - 1.5(IQR) = \boxed{93}$$
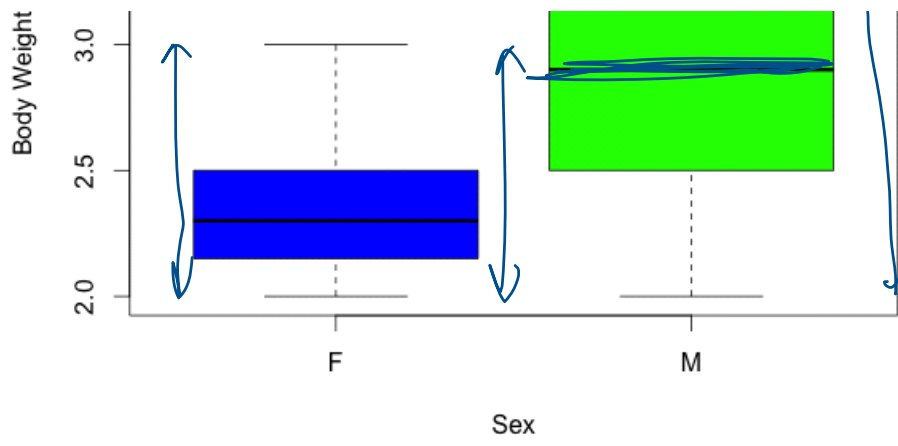$$max = Q_3 + 1.5(IQR) = \boxed{469}$$

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

Cat Weight by Sex

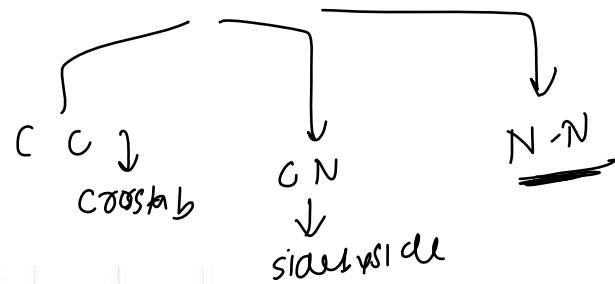y Weight (kg)
3.5
3.0

weight

Age ‖ gender

21 ← M
39   I
43   M

43   M

54   F

Scatterplots

Scatter

C C → crosstab

C N → side by side

N -N

Ground living area partially explains sale price of apartments
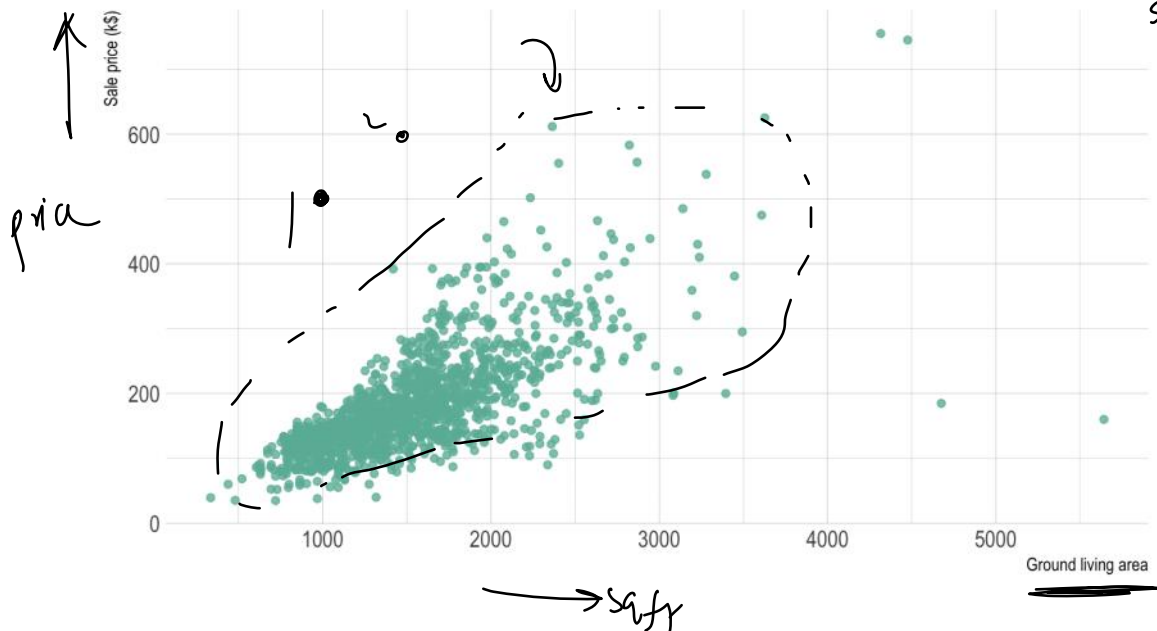
price

sqft

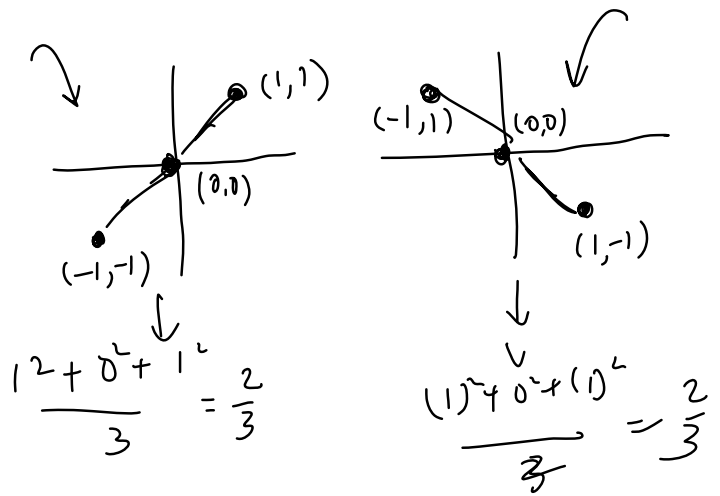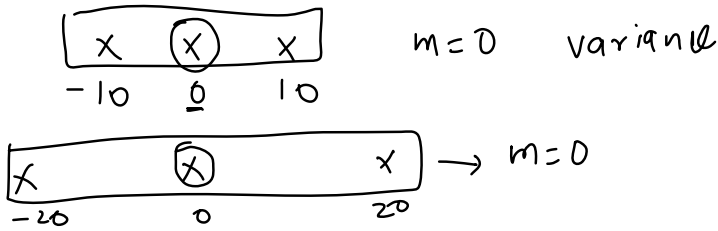| sqft	price
1500	60L
1700	80L
1200	35L

Covariance

Direction

- the variance will become fails to talk about the type of linearity.
- Give same result on positive and negative linearity b/w variables.

- **What problem does Covariance solve?**

$m = 0$ variance

$m = 0$

$(1,1)$ $(0,0)$

$(-1,-1)$

$\dfrac{1^2 + 0^2 + 1^2}{3} = \dfrac{2}{3}$

$(-1,1)$ $(0,0)$

$(1,-1)$

$\dfrac{(1)^2 + 0^2 + (1)^2}{3} = \dfrac{2}{3}$

- **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.
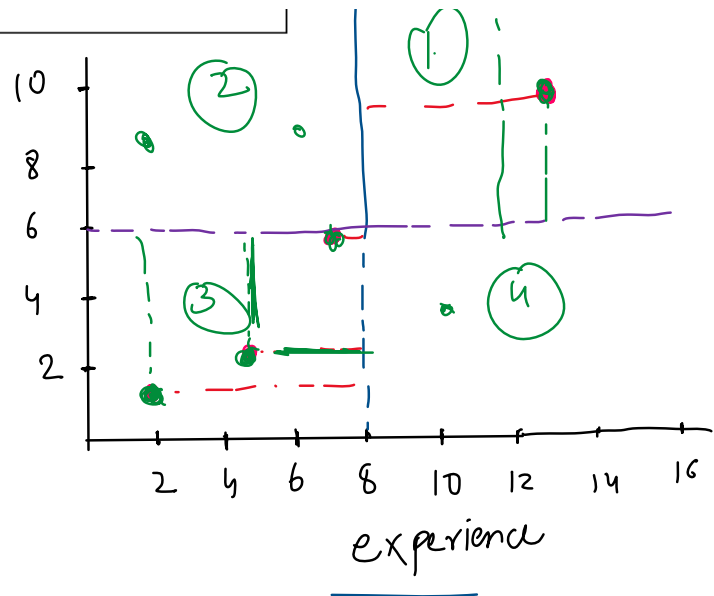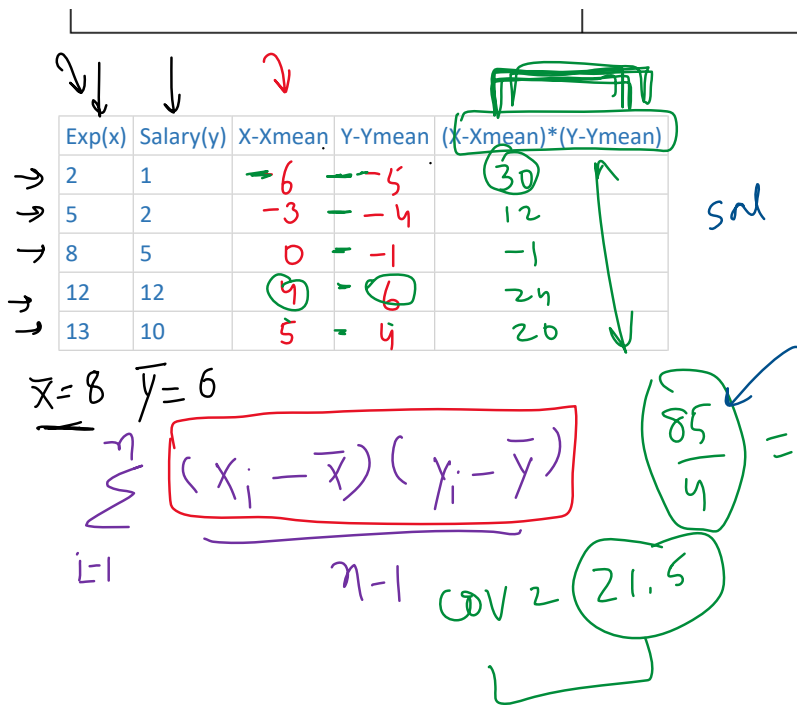

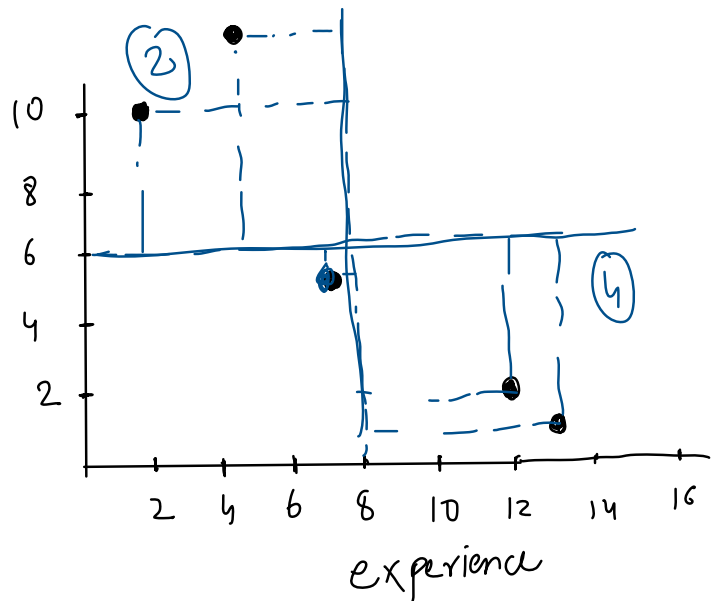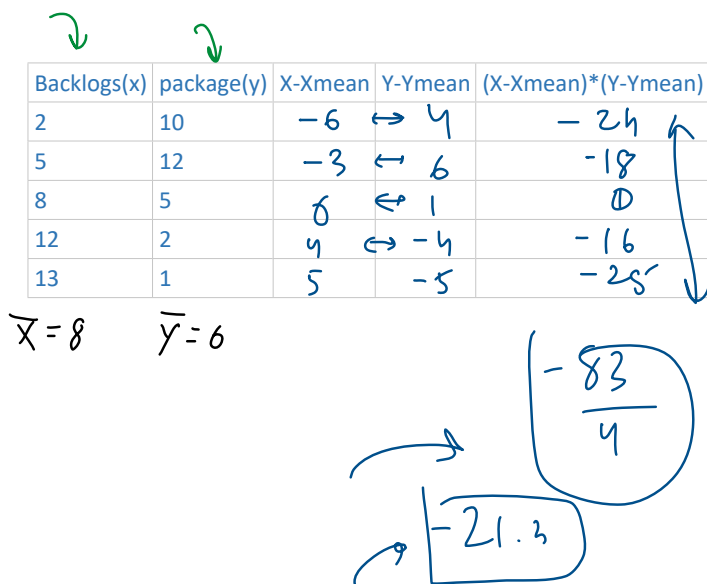
covariance +ve

covariance −ve

cova $\simeq 0$
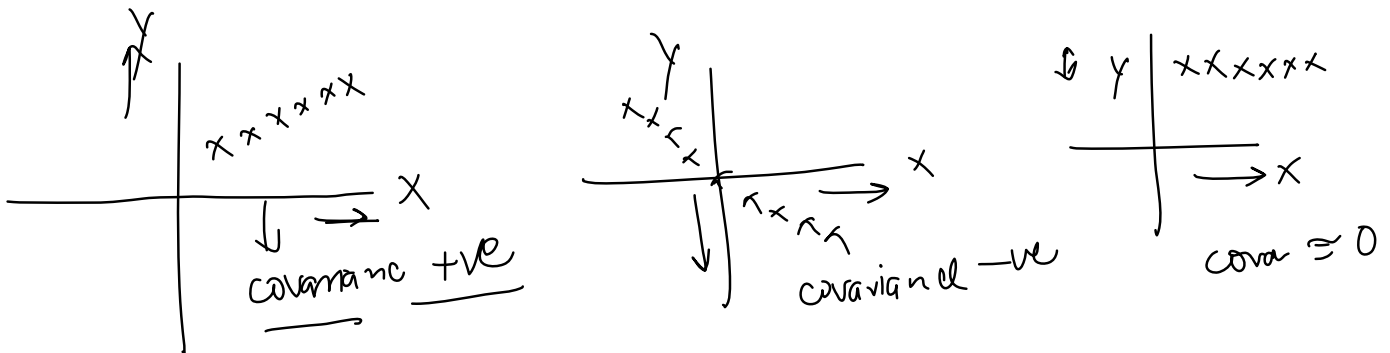
np.cov(df [x] ,df [y])

while this magnitude is not a true repr. of strength

$= -2418$

-- means, both variable are inversily depend on each other

- **How is it calculated?**

Covariance Formula	
Population	**Sample**
$\sigma_{xy} = \dfrac{\sum(X - \mu_x)(Y - \mu_y)}{N}$	$s_{xy} = \dfrac{\sum(X - \bar{x})(Y - \bar{y})}{n - 1}$
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

$\bar{x} = 8 \quad \bar{y} = 6$

sol

$$\sum_{i=1}^{n} \frac{(X_i - \bar{x})(Y_i - \bar{y})}{n-1}$$
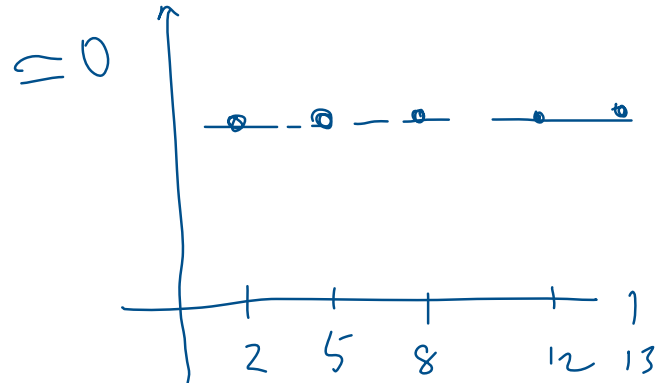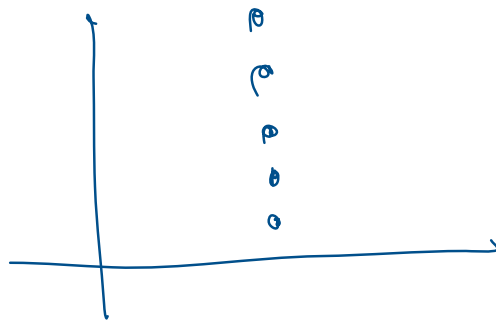
$\dfrac{85}{4} =$

COV = 21.5

experience

On the basis of their mean value , we create the quadrant
for -ve, points are in 2 and 4 th quadrant
for +ve, points are in 1 and 3 th quadrant

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

$\bar{x} = 8 \qquad \bar{y} = 6$

$\dfrac{-83}{4}$

-21.3

experience

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

$\simeq 0$

- 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.
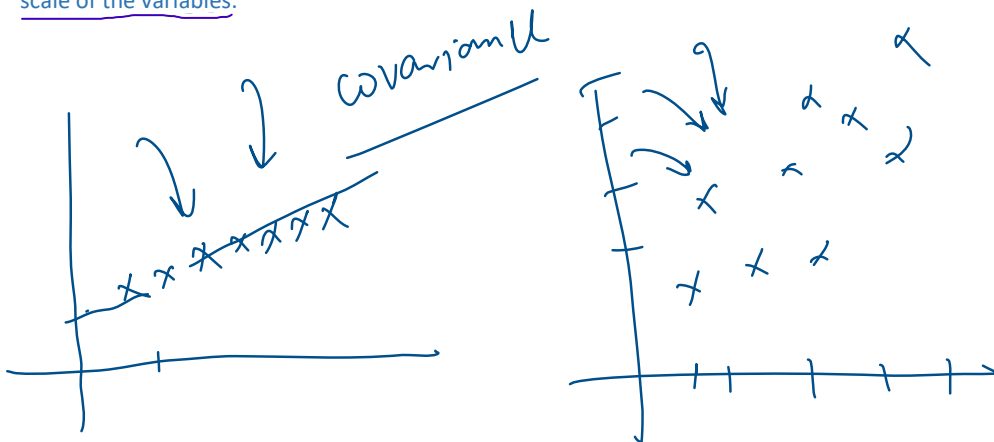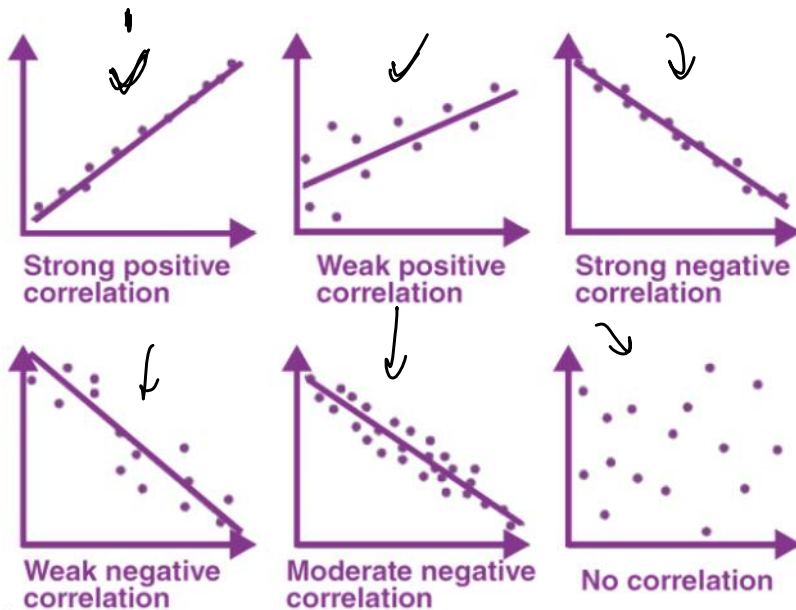
  covariance

- Covariance of a variable with itself

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

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

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

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

Relationship with variance $\longrightarrow$ Covariance of a variable with itself is variance

Correlation

direction(nature)+ strength

1. What problem does Correlation solve?

Strong positive correlation

Weak positive correlation

Strong negative correlation

Weak negative correlation

Moderate negative correlation

No correlation

Can we quantify this weak and strong relationship?

$12 \quad \backslash \quad 0$

$-1 \ to\ 1 \qquad 0-1$

$-1-0$

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.

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

$\sigma \rightarrow$ standard deviation
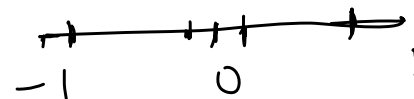
Correlation and Causation

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.
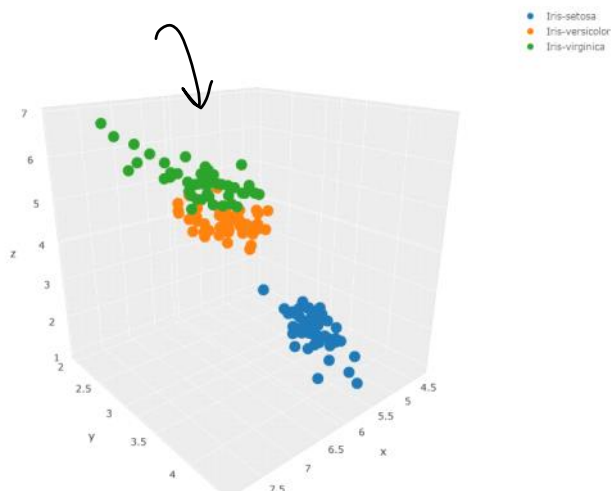
Visualizing Multiple Variables

13 March 2023   06:58

wt |ht) age)

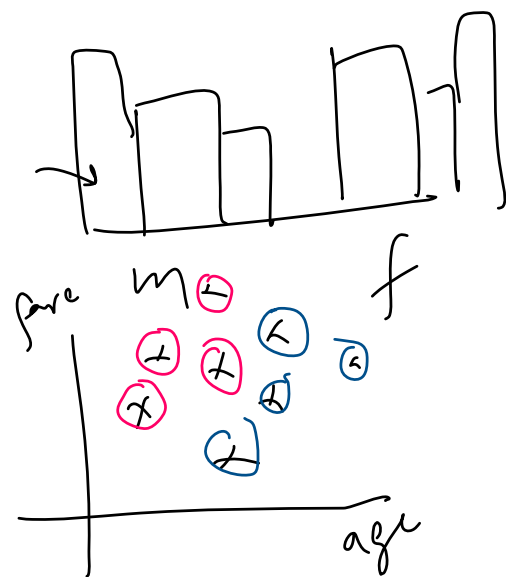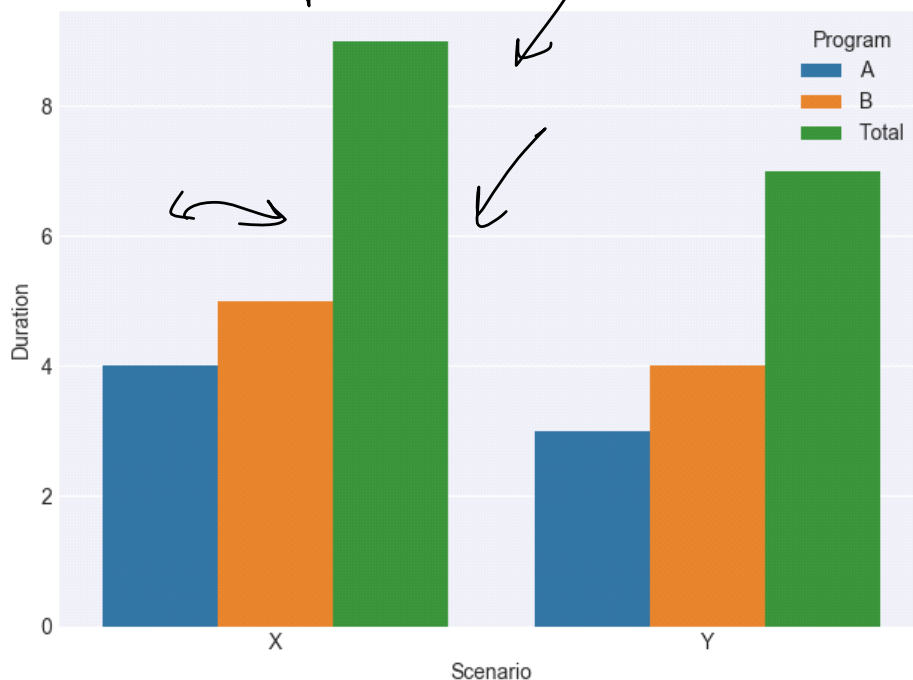### 1. 3D Scatter Plots

- Iris-setosa
- Iris-versicolor
- Iris-virginica

2. Hue Parameter

barplot   1cat — 1Num
           1cat

fare   m   f

age

4 cols

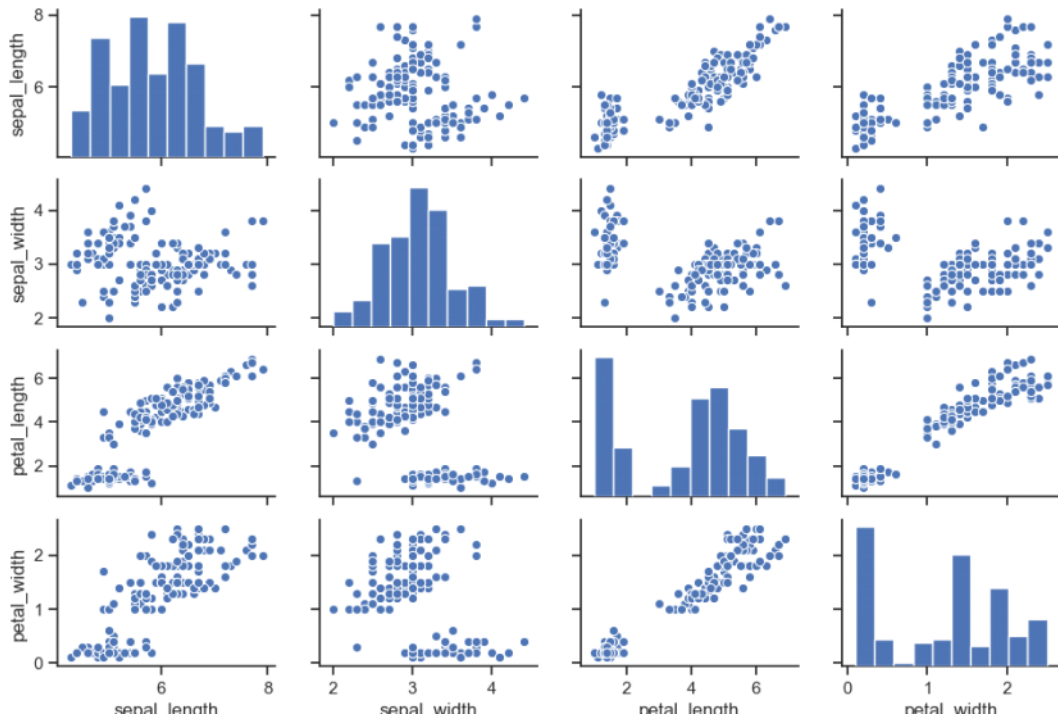### 3. Facetgrids

time = Lunch   time = Dinner

smoker
- Yes
- No

4. Jointplots
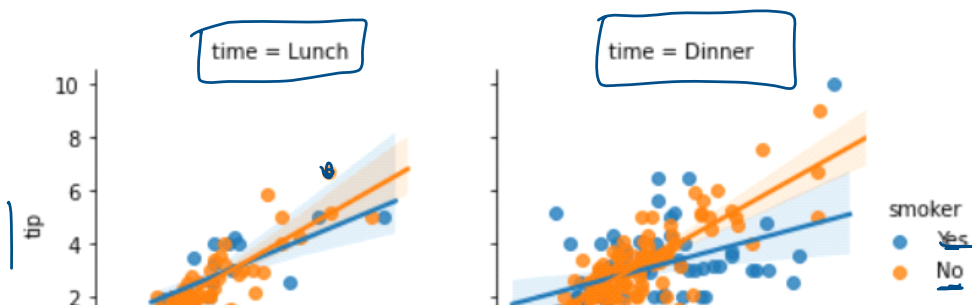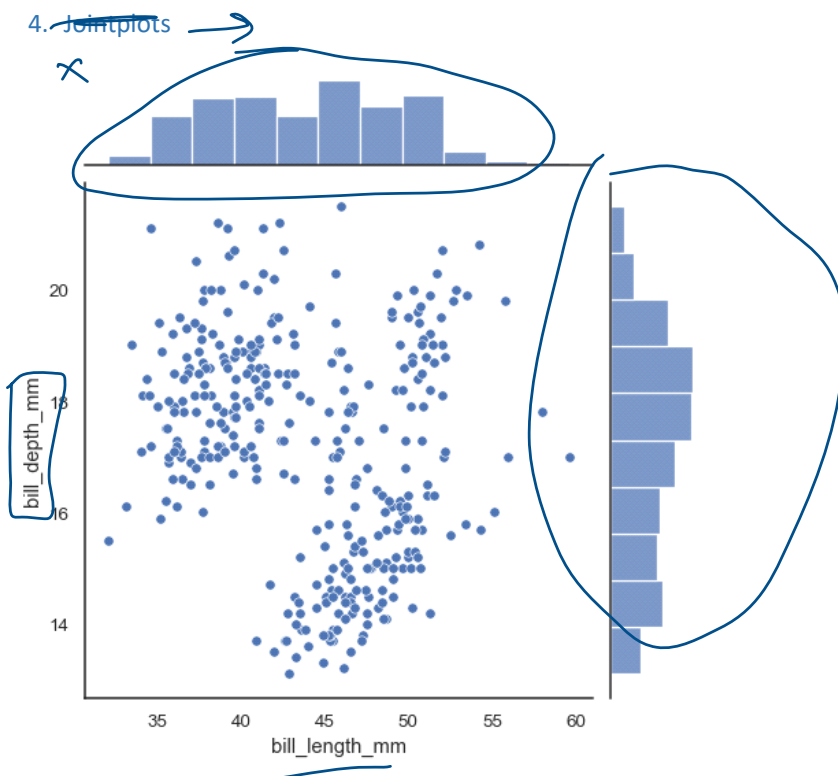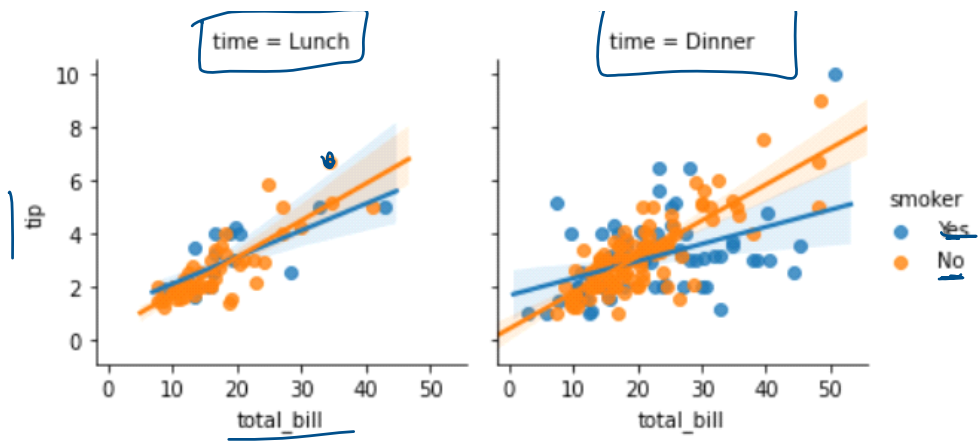
5. Pairplots

6. Bubble Plots

Bubble Chart (plotly like

pop

y

x   gdp

Legend:
- Series 1
- Series 2
- Series 3
- Series 4