Standard Deviation

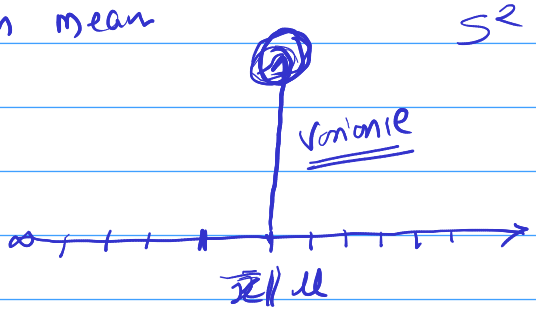
$\bar{x}$ - Sample mean

& μ → population mean

Variance $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$

$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$

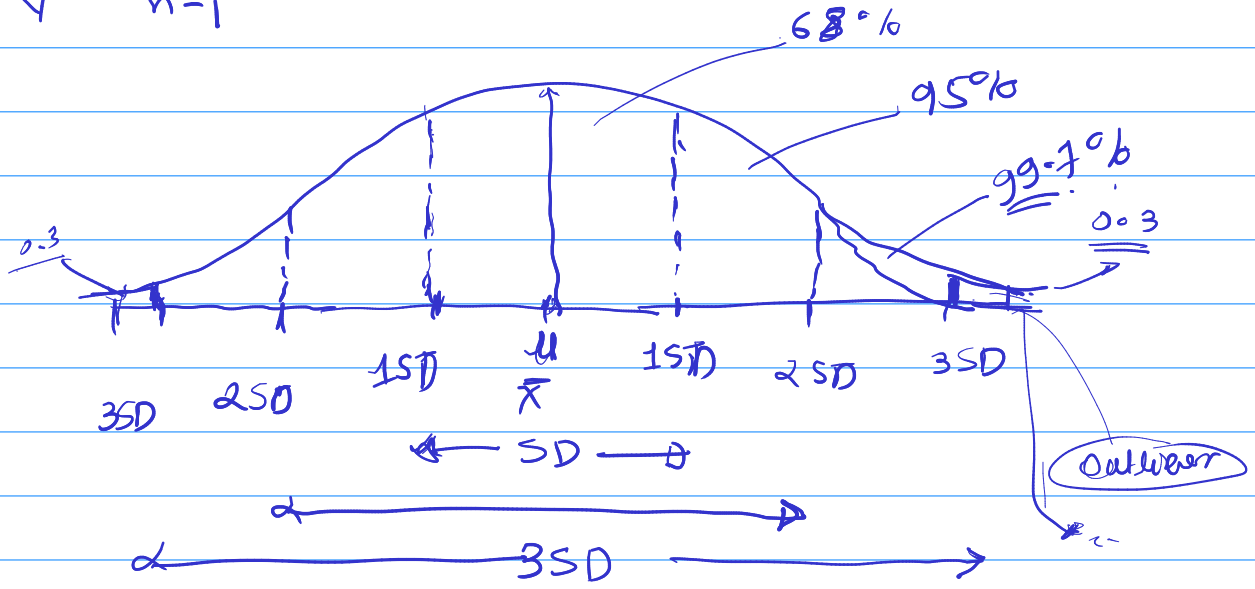
Variance –



Standard Deviation → is the square root of variance. It is widely used measure of dispersion that is useful in describing the shape of distribution

$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ = $\sigma = \sqrt{\sigma^2}$

$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ $s = \sqrt{s^2}$



Empirical Rule of the data ⇒

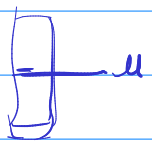
① Variance

② SD

③ Coeff. of Variation

③ **Coefficient of Variation**

$\text{Coef of var} = \frac{SD}{\text{mean}} \times 100$ (CV)



ratio of standard deviation to the mean expressed as a percentage

Univariate Analysis about the data

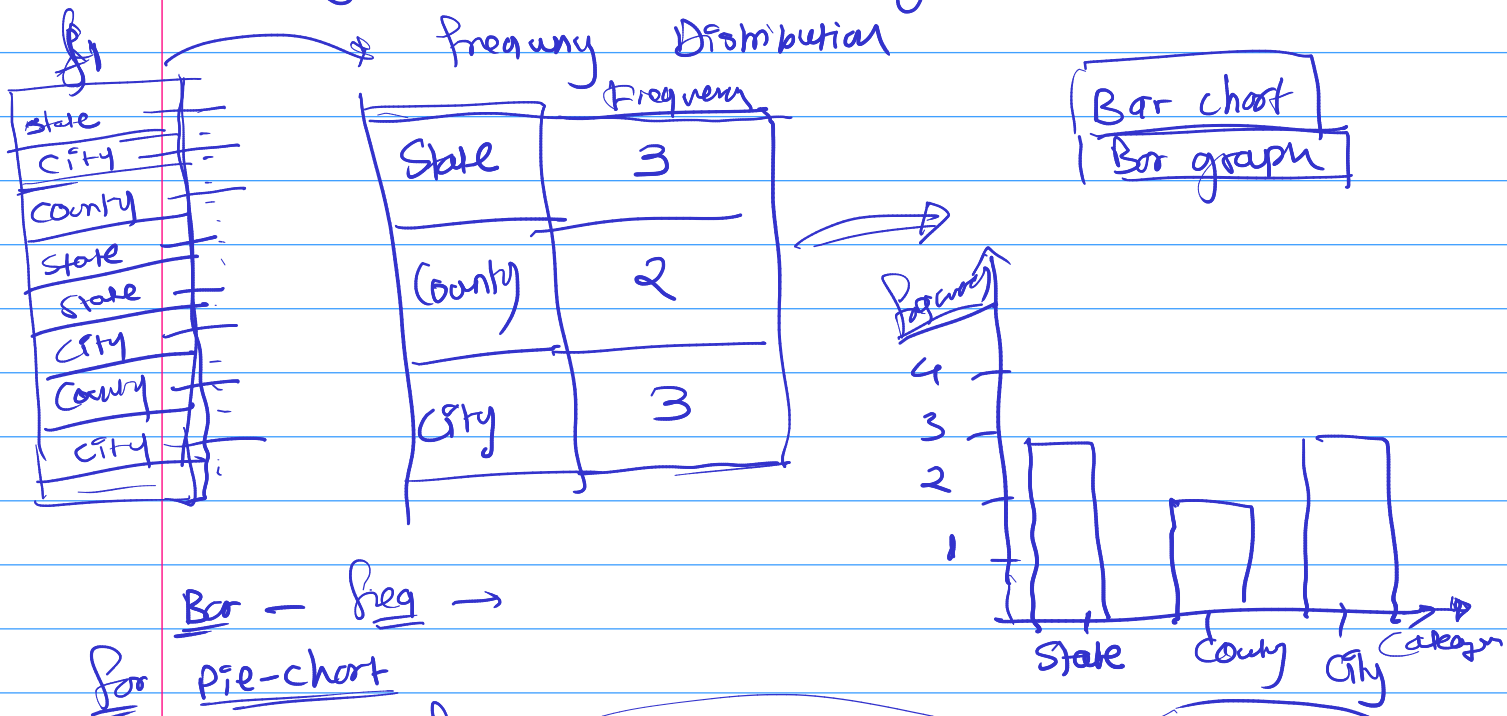
- univariate (single feature)
- Bivariate (2 features)
- multivariate (multiple feature)

f_1 → Categorical

f_2 → Numerical

Analyze the data (EDA) = Exploratory Data Analysis →

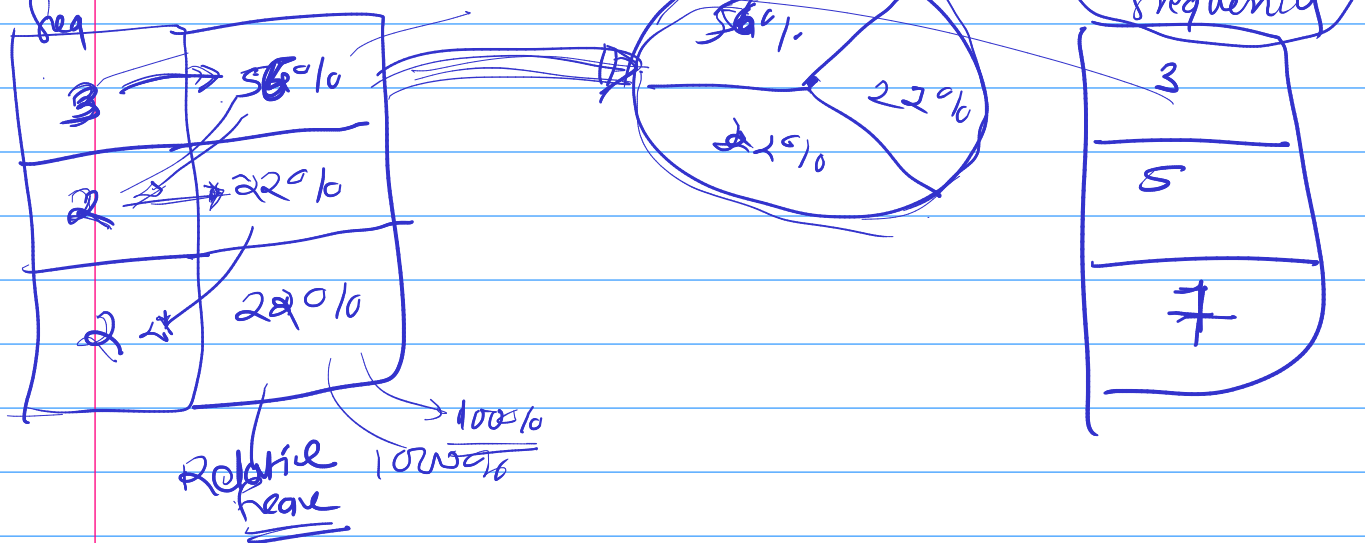
1> Categorical Data → Frequency Distribution Table



Bar – freq →

For pie-chart

Calculate Relative freq



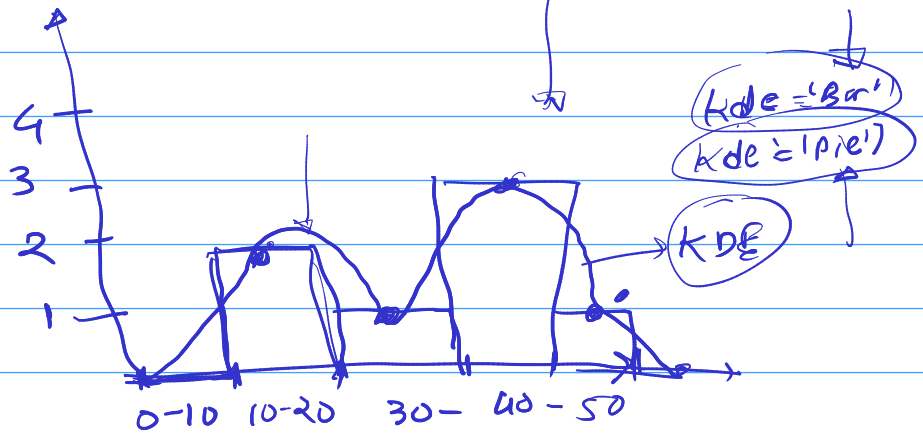
2) Numerical - Feature

Continuous

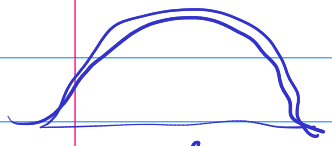
Bar chart - Histogram

Age
32
28
35
39
41
12
15

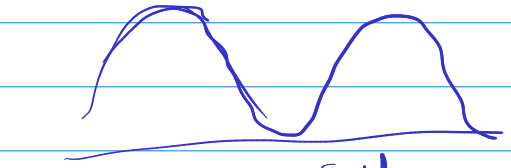
Freq	Freq
0-10	0
10-20	2
20-30	1
30-40	3
40-50	1



KDE = Kernal Density function estimator

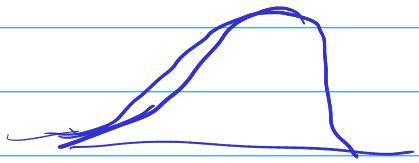


Bell curve
symmetric curve
Normal dist
or Gaussian dist

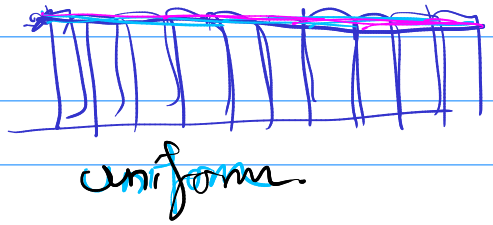


Bimodal

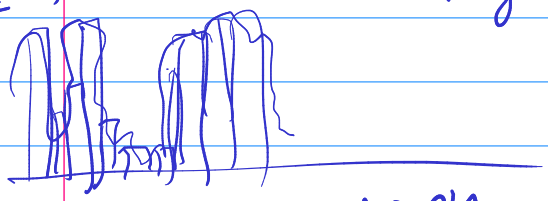
Right Skew



Left Skew



uniform



No-Pattern graph

① → Bar
② - pie
③ countplot
④ distplot → warning
⑤ kdeplot (new)

Boxplot

Graph for Bivariate Analysis (2 feature) → category & Numerical

- Categorical & Categorical
- numerical & numerical
- numerical & Categorical

① Categorical & Categorical

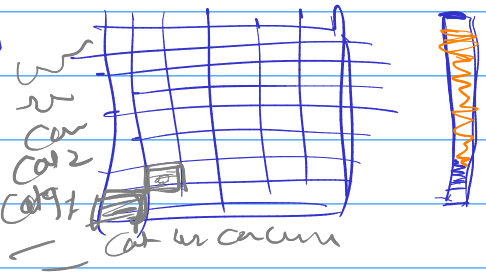
1) Contingency Table / Crosstab

Survived	p-class
0	1
1	2
	3

Survive	Pclass 1	2	3
0	42	31	63
1	71	118	13

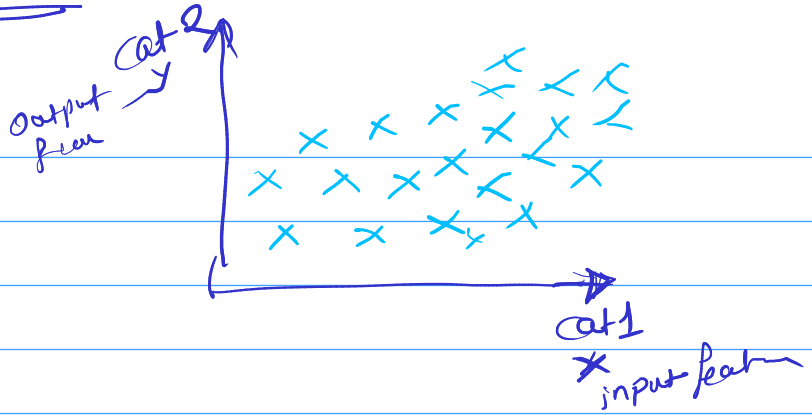
② Heatmap

Categoric



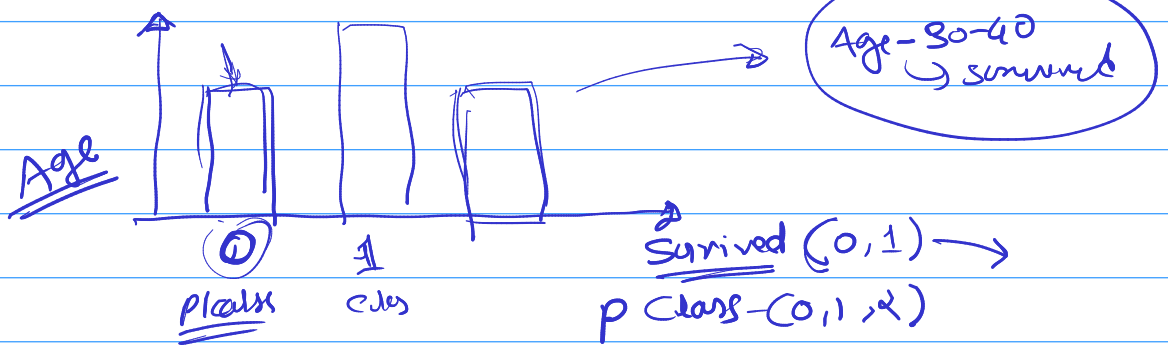
2) Numerical – Numerical

i) Scatter plot

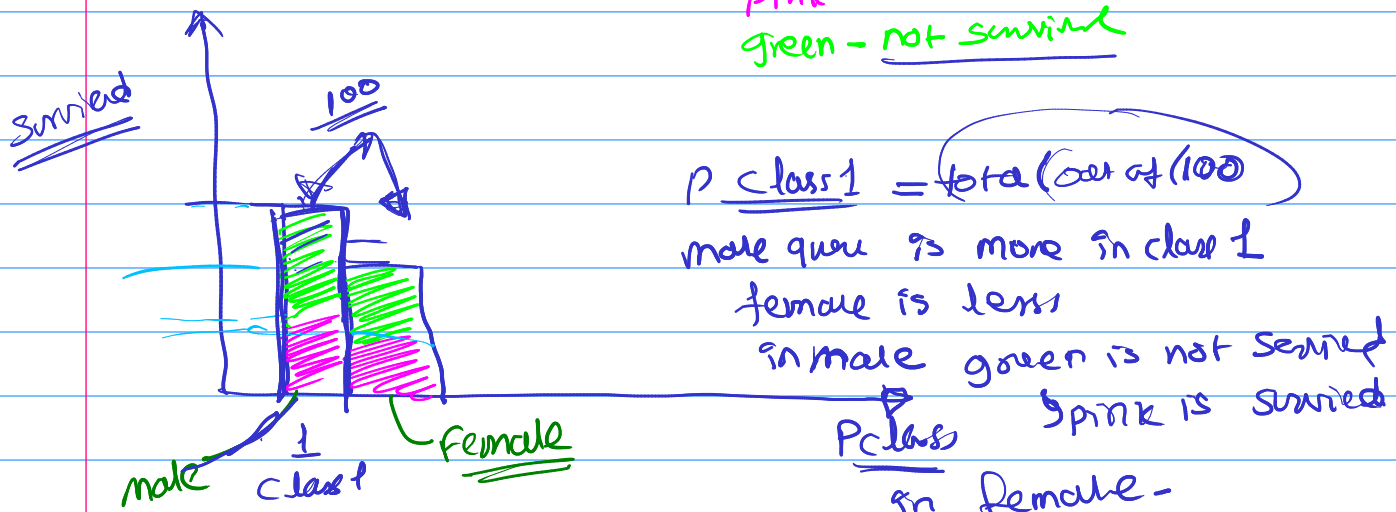


3) Categorical – Numerical / Numerical & Categorical.

① Bar



② Hue (Pre-Bar)
& (multivariate Analysis)



multivariate analysis

['sex'] ['Pclass'] ['survived']

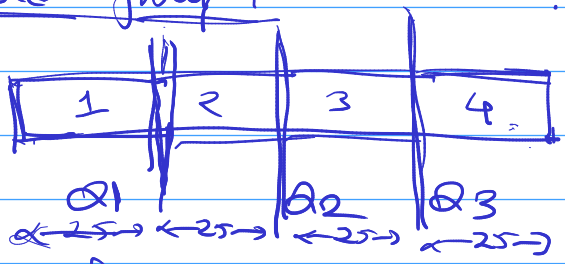
Hue

x axis

y axis

Quantile and percentile

— Quantile are Statastical measure use to divide a set of numerical data into equal sized group



Type of Quantile

① Quartile → Divide the data into four equal Part
Q_1 - 25th percent Q_2 = 50th Q_3 = 75th

② Deciles — we divide data into 10 equal Part.

③ Percentiles — Divide the data into 100 equal Part

④ Quintiles — Divide into 5 equal Part

↳ ① Soot the ~~data~~ —
② You get an index location

① Percentile i→

percentage = $\frac{\text{The no. we got}}{\text{total number}} \times 100$

$= \frac{560}{800} \times 100 = 70.75\%$

A percentile is a statastical measure that represent the percentage of observations in datatoset that fall below a particular value.

Class (CAT, SAT, JEE) classroom

75% → 70

Percentile → 74 Student is below you

100 Student

Formula for Percentile $= \frac{P}{100}(N+1)$

P → the percentile rank
N → total no. of observation.

ex = 78, 82, 84, 88, 91, 93, 94, 96, 98, 99 → sorted

(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

0 (1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

if you find 75% percentile

$PL = \frac{P}{100}(N+1) = \frac{75}{100}(10+1) = 0.75(11) = 8.25$

location

8th & 9th average $\frac{96+98}{2} = 97$

75th percentile of data is 97

~~75%~~

rank of 91

percentile rank $= \frac{x + 0.5}{N}$

$x \rightarrow$ numer of value below the given value

$N \rightarrow$ total no. of dataset

rank of 91 $= \frac{4+0.5}{10} = 0.45 = 45\%$

5-Number Summery — Aim → to observe the outlier

(1) minimum value $(Q_1 - 1.5(IQR))$

(2) First Quartile Q_1 (25%)

(3) medial (Q2) (50%)

(4) Third Quartile (Q3) (75%)

(5) maximum value $(Q_3 + 1.5(IQR))$

IQR = Inter Quartile Range $= (Q_3 - Q_1)$

6, 213, 241, 260, 281, 290, 314, 321, 350, 1500

(1) (2) 3 4 5 6 7 8 9 10

$Q_2 = \frac{50}{100}(10+1) = 11 = 5.5$ location $= \frac{5^{th}+6^{th}}{2} = \frac{281+290}{2} = 285.5$ (Q2)

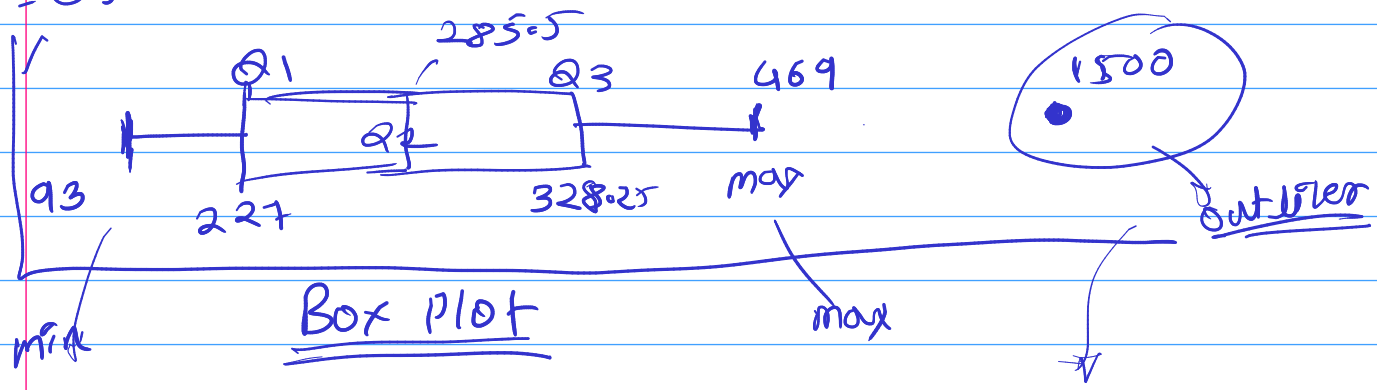
$Q_1 = \frac{25}{100}(11) = 2.75 = \frac{2^{nd}+3^{rd}}{2} = \frac{213+241}{2} = 227$ (Q1)

$Q_3 = 328.25$

$IQR = Q_3 - Q_1 = 94$

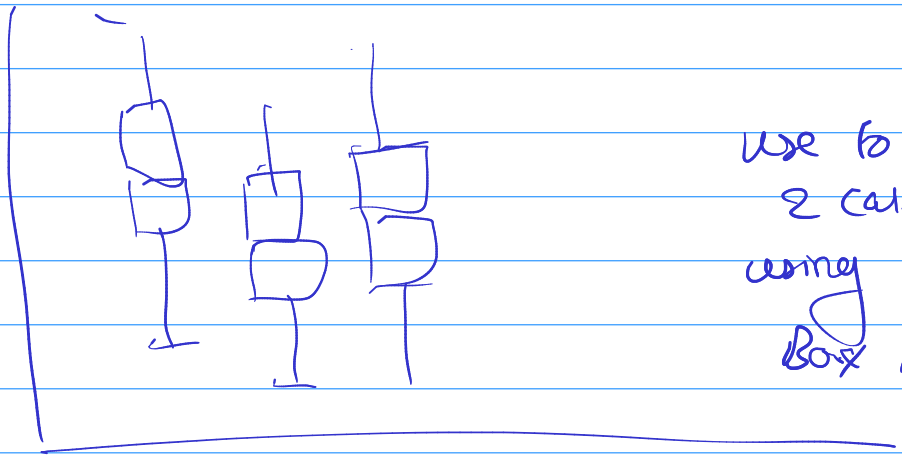
$min = Q_1 - 1.5 \cdot IQR = 93$

$max = Q_3 + 1.5 \cdot IQR = 469$



Box Plot

Multivariate analysis



use to Compare 2 category also using multivariate Box Plot analysis.