Right Business Decisions

* High Revenue
* Reduce expenses
* Customer experience

Walmart :

Beer and diaper case study

Stages of data science project :

* Understand the problem statement .
* Data gathering .
* Data Engineering :
* EDA : Exploratory data analysis .
* Data Cleaning :
* Data preparation :
* Data models .
* Test the model created above .
* Deploy the model in production .

https://cran.r-project.org/bin/windows/base/

<https://rstudio.com/products/rstudio/>

SAS

Dplyr : handling data .

Ggplot2 : visualization .

Links to learn R

<https://www.tutorialspoint.com/r/index.htm>

<https://www.rdocumentation.org/packages/base/versions/3.6.2>

<https://www.datacamp.com/>

<https://towardsdatascience.com/>

<https://statistics.berkeley.edu/computing/r-dates-times>

Link for data sets and explore further in the world of data science

<https://www.kaggle.com/>

<http://www.z-table.com/>

<http://www.ttable.org/>

Statistics :

H0 -> Null Hypothesis

H1 -> Alternate Hypothesis

Population data -> total data

Sample data -> subset of population data

µ - Mean of population data

σ – Standard deviation.

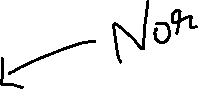
– Mean of the sample data.

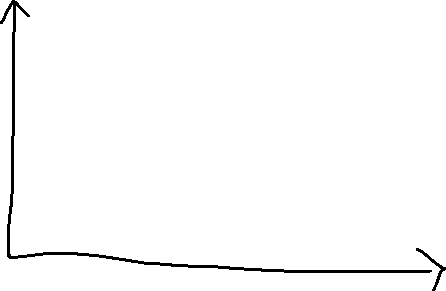
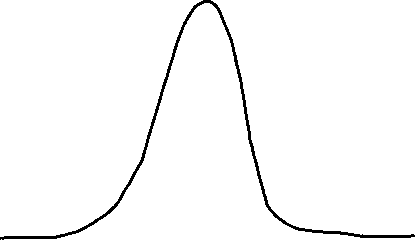
n – number of records in sample data.

N – number of records in population data.

**Central Limit Theorems :**





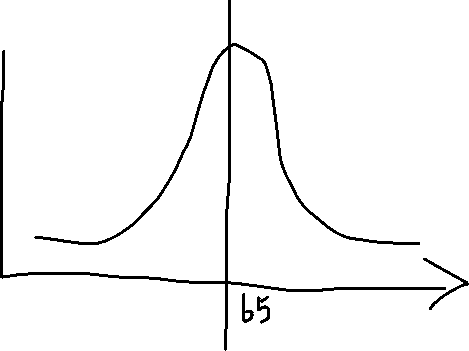






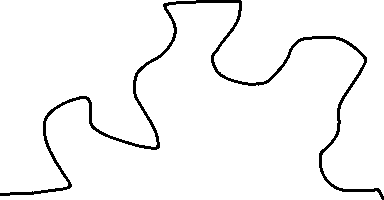
1 - When the population data is normally distributed then the sample means also follow a normal distribution.

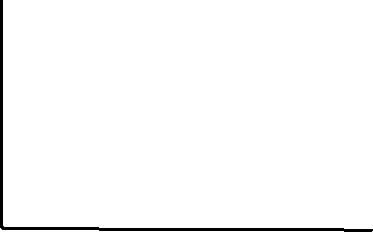




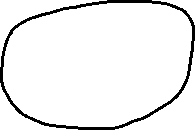


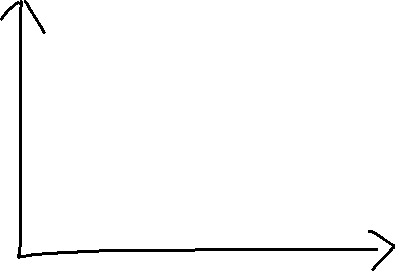
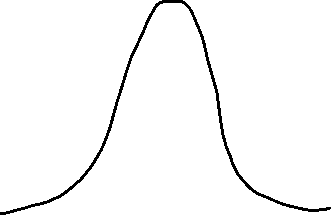
2 – When the sample size is large , the sample mean follows a normal distribution irrespective of the distribution of the population data









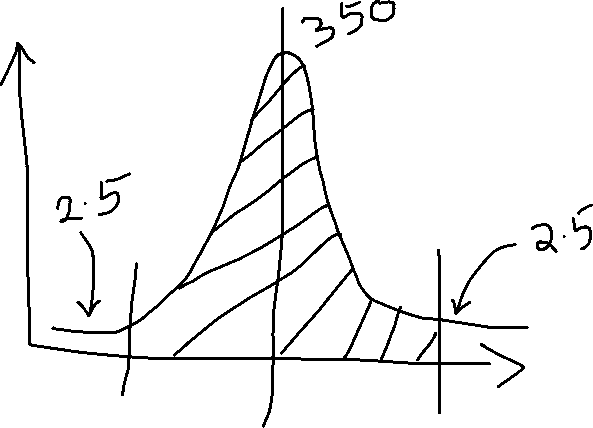
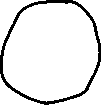


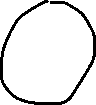
3 –

X ~ N (µ , σ ) - > Population data

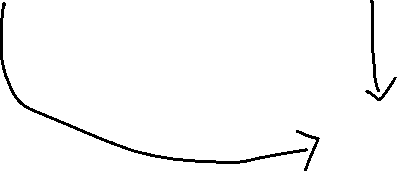
~ N ( µ , ) -> Sample data

H0 - > My friend sells 350 toffees everyday .









Critical region

n = 100

standard deviation of the population = 18

Z test :

Z =

= Z ()

= 350 1.96 ( 18 /100 )

= 350 1.96 \* 1.8

UL : 350 + 3.5 , LL : 350 – 3.5

T test :

t =

School of 1000 students .

Average IQ : 100

N = 1000

= 100

H0 -> IQ of the students will remain the same.

n = 25

= 140

S = 18 -> std dev of sample data .

Level of significance = 0.05

t =

= 11.11

Degree of Freedom: n -1 = 24

-2.064 < t < 2.064 -> accept null hypothesis.

Which means I will reject the null hypothesis and the medicine has a positive impact .

ANOVA TEST :

Analysis of Variance

5 Ponds . polluted .

20 samples from each pond . 

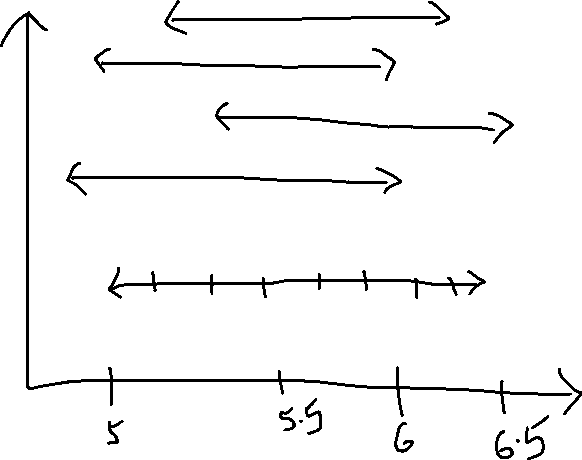
H0 -> P1 = P2 = P3 = P4 = P5 ( all ponds have same pollution)





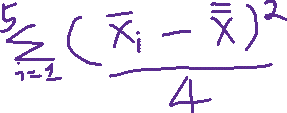




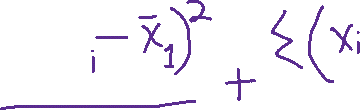
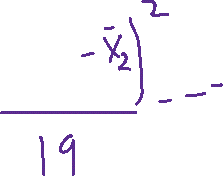




Fstats = 









Parametric tests

Non parametric tests :

Chi square

CHI SQUARE:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Salary | | | |
| Service | Low | Medium | High | Total |
| Excellent | 9 E11 | 10 E12 | 7 E13 | 26 |
| Good | 11 E21 | 9 E22 | 31 E23 | 51 |
| Poor | 12 E31 | 8 E32 | 3 E33 | 23 |
| Total | 32 | 27 | 41 | 100 |

E11 = (26 \* 32) / 100

E12 = (26 \* 27) / 100

E13 = (26 \* 41 )/ 100

.

.

.

.

.

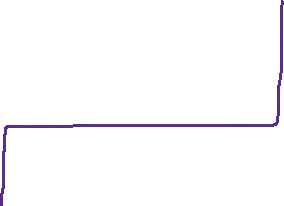
.

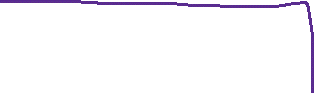
E33 = 23 \* 41 /100

Xsq = 18.658



Machine Learning :

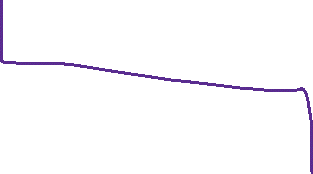






Supervised Learning Unsupervised Learning

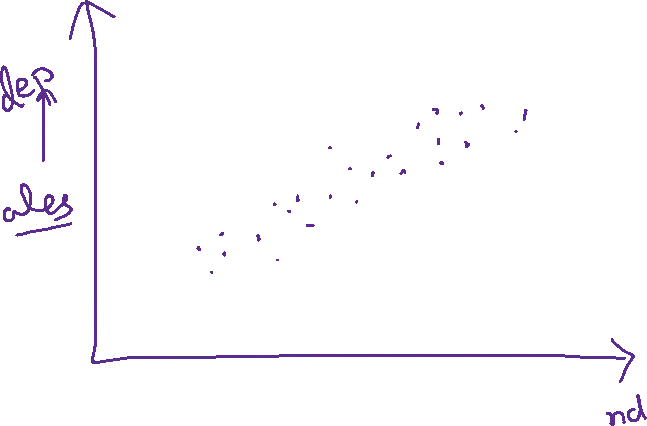
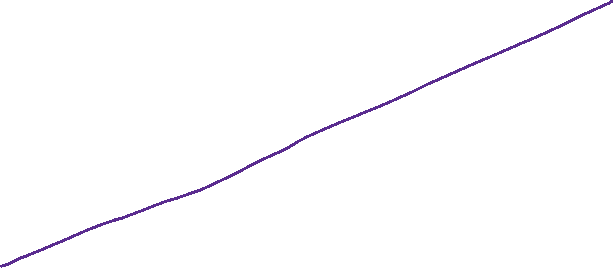
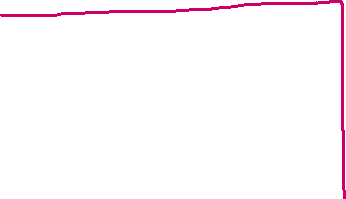
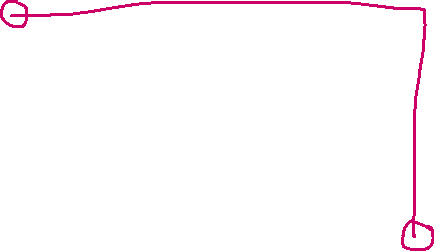
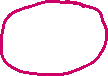
( dependent column) ( no dependent columns)



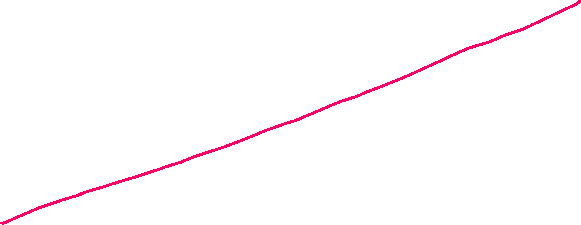
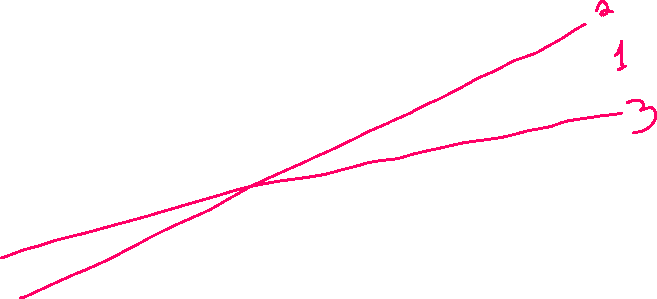
Regression Classification

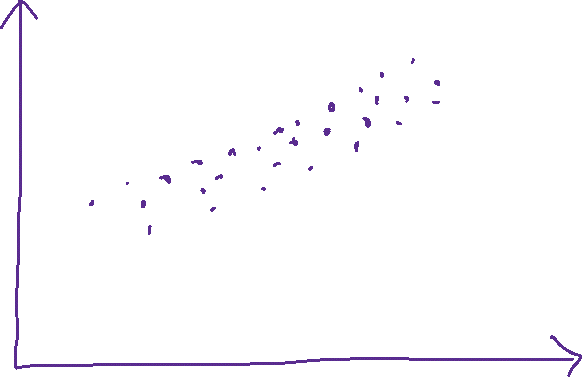
( continuous) ( categorical ) 

**Linear Regression :**

Simple Linear Regression : 





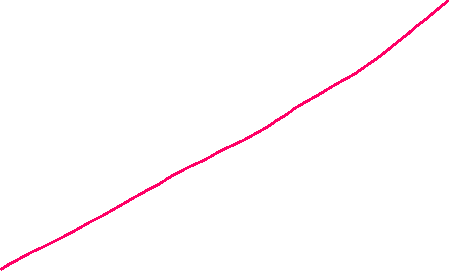
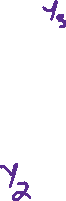




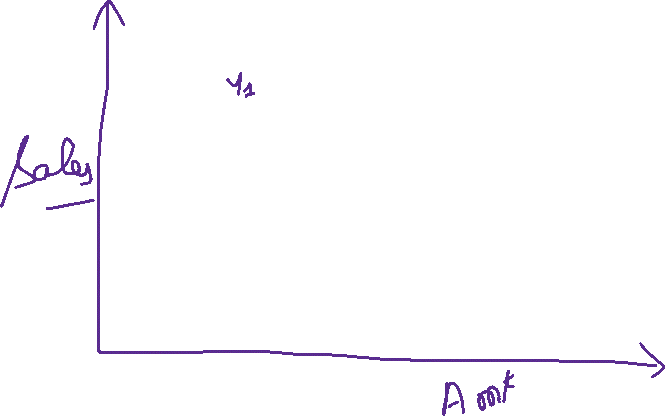
Best fit line :



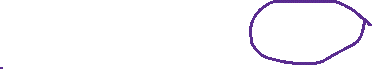


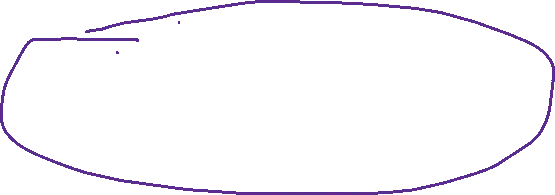








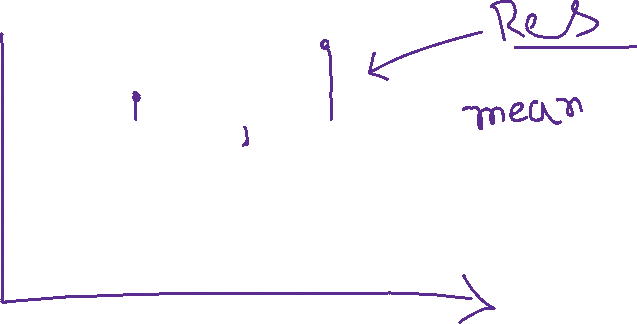


RSS = (Y1 – y1)2 + (Y2 – y2)2 + (Y3 – y3)2 

Residual sum of squares .









TSS = (Y1 – M)2 + (Y2 – M)2 + (Y3 – M)2 

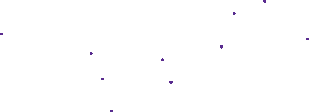
Total sum of squares .

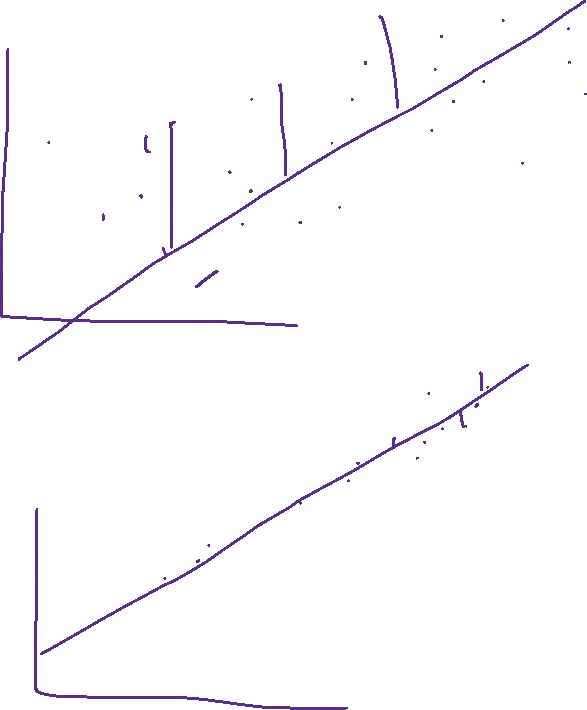
R2  = 1 – ()

0 < R2 < 1

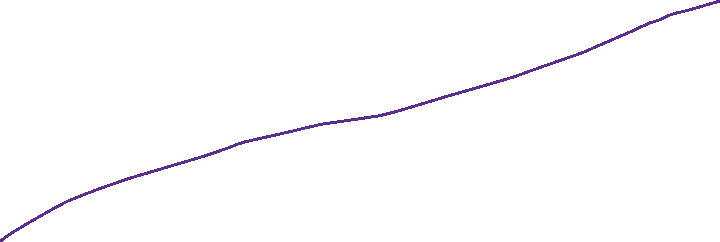
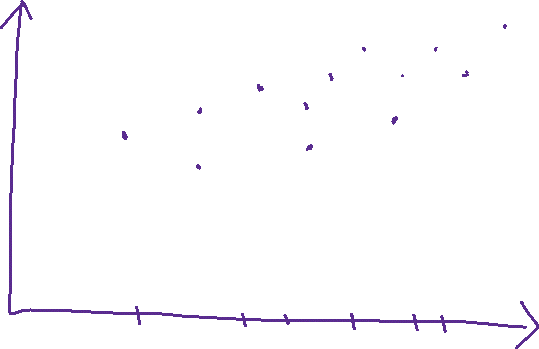
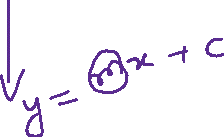
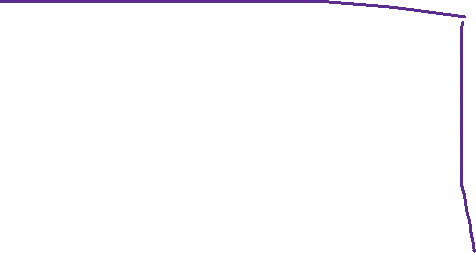
R2  = 0.75 R2  = 0.89

Higher the rsquare , the better my model is .







-1 < correlation +1 

Correlation is near zero then there exists a little relation between the variables.

Y = 1.000656 x - 0.107265

**Multiple Linear regression :**

Two independent columns and one dependent column , my graph is three dimensional.

Equation of multiple linear regression with n independent columns

Y = β0  + β1X1  + β2 X2  + β3 X3 + β4 X4 + ……βn Xn