Data science Training – Day 2

Pandas:

Pandas is a Dataframe. It resembles Excel sheets.

Panel Sheets == Panel Data == Pandas

Popularly used for data cleaning, data exploration and map to most of the data sources like delimited files, JSON files and DB etc.

We have None and Nan – unknown value. There is no "Null" value considered in python. For this we need to perform data cleaning operations.

Statistics:

Understanding the information is essential to get the things better and make right choices. Choose right option and deal with things.

We will be able to describe the information using statistics.

Descriptive Statistics:

The basic descriptive statistics to give us an idea on the variables and their distributions

Permit the analyst to describe many pieces of data with few indices.

Central Tendencies:

Mean/ Average:

- Generalization capabilities – without seeing the datapoints, we will know where the data lies.
- Mean () function is used to calculate average
- Mean is not a good measure in presence of outliers

Median:

- Sort the data values in ascending or descending order. Then take up the mid value.
- True mid-point value

Median

- Mean is not a good measure in presence of outliers
- For example Consider below data vector
  - 1.5,1.7,1.9,0.8,0.8,1.2,1.9,1.4, 9 , 0.7 , 1.1
- 90% of the above values are less than 2, but the mean of above vector is 2
- There is an unusual value in the above data vector i.e 9
- It is also known as outlier.
- Mean is not the true middle value in presence of outliers. Mean is very much effected by the outliers.
- We use median, the true middle value in such cases
- Sort the data either in ascending or descending order

Median

1.5	0.7
1.7	0.8
1.9	0.8
0.8	1.1
0.8	1.2
1.2	**1.4**
1.9	1.5
1.4	1.7
9	1.9
0.7	1.9
1.1	9

- Mean of the data is 2
- Median of the data is 1.4
- Even if we have the outlier as 90, we will have the same median
- Median is a positional measure, it doesn't really depend on outliers
- When there are no outliers then mean and median will be nearly equal
- When mean is not equal to median it gives us an idea on presence of outliers in the data

If mean and median is not close to each other, then it indicates presence of outliers.

Mean and Median

Import “Census Income Data/Income_data.csv”

```
#Mean and Median on python
gain_mean=Income["capital-gain"].mean()
gain_mean

gain_median=Income["capital-gain"].median()
gain_median
```

Mean is far away from median. Looks like there are outliers, we need to look at percentiles and box plot.

Mode:

Most popular value.

Most popularly we use mean and median.

Standard Deviation:

Standard Deviation

Customer ID	Name	Surname	Gender	Age	Age Grou	Height	egion	Job Classification	Tenure Months	Balance	Spend On Groceries
200000262	Zoe	Clarkson	Female	59	5	62	cotland	Other	24	23550.89	70.77
200001214	Carolyn	McDonald	Female	58	5	61.2	cotland	Other	24	69027.62	67.1
400000497	Anna	Chapman	Female	26	2	65.1	orthern Ireland	White Collar	46	5789.63	46.23
400001939	Richard	Dowd	Male	21	2	70.9	orthern Ireland	White Collar	23	10248.59	36.48
300002298	Phil	Arnold	Male	37	3	70.4	Vales	Blue Collar	15	80824.89	36.11

$$\{ 61.2, 62, 65.1, 70.4, 70.9 \}$$

$$\text{Mean} = \frac{61.2 + 62 + 65.1 + 70.4 + 70.9}{5} = 65.92$$

Difference between mean and each value, will give the idea of how similarity they are. Some differences can be positive and negative. You can square and sum it. Divide by the total.

Variance:

- Tells the variation in the dataset.
- No units
- Large the variance = more are the values are different from each other
- Less variance = similar values

Standard deviation:

- Square root of variance. Has units.
- Both variance and standard deviation conveys the same thing.

$$\{ 61.2, 62, 65.1, 70.4, 70.9 \}$$

$$\mu \quad \text{Mean} = \frac{61.2 + 62 + 65.1 + 70.4 + 70.9}{5} = 65.92 \text{ meters}$$

$$\text{Variance} = \frac{(61.2 - 65.92)^2 + (62 - 65.92)^2 + (65.1 - 65.92)^2 + (70.4 - 65.92)^2 + (70.9 - 65.92)^2}{5}$$

$$\text{Variance} = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = 16.64 \quad \text{no units}$$

σ^2

$$\text{Std. Dev.} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} = 4.08$$

σ

Dispersion:

Dispersion

- Just knowing the central tendency is not enough.
- Two variables might have same mean, but they might be very different.
- Look at these two variables. Profit details of two companies A & B for last 14 Quarters in MMs

															Mean
Company A	43	44	0	25	20	35	-8	13	-10	-8	32	11	-8	21	15
Company B	17	15	12	17	15	18	12	15	12	13	18	18	14	14	15

- Though the average profit is 15 in both the cases
- Company B has performed consistently than company A.
- There was even loses for company A
- Measures of dispersion become very vital in such cases

Only looking at the mean and judge the dataset is similar, that is wrong!

You can't narrow down with help of mean and median.

Take variance and standard deviation.

Variance and Standard deviation

- Dispersion is the quantification of deviation of each point from the mean value.
- Variance is average of squared distances of each point from the mean
- Variance is a fairly good measure of dispersion.
- Variance in profit for company A is 352 and Company B is 4.9

Value	Value-Mean	(Value-Mean)^2
43	28	784
44	29	841
0	-15	225
25	10	100
20	5	25
35	20	400
-8	-23	529
13	-2	4
-10	-25	625
-8	-23	529
32	17	289
11	-4	16
-8	-23	529
21	6	36
15.0		352

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

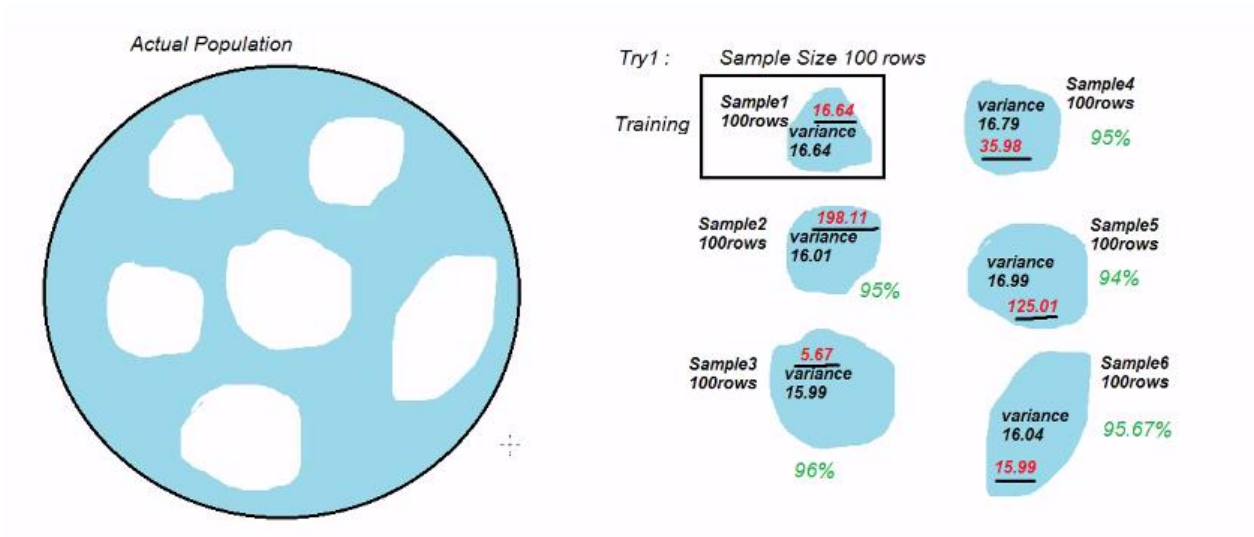
Value	Value-Mean	(Value-Mean)^2
17	2	4
15	0	0
12	-3	9
17	2	4
15	0	0
18	3	9
12	-3	9
15	0	0
12	-3	9
13	-2	4
18	3	9
18	3	9
14	-1	1
14	-1	1
15.0		4.9

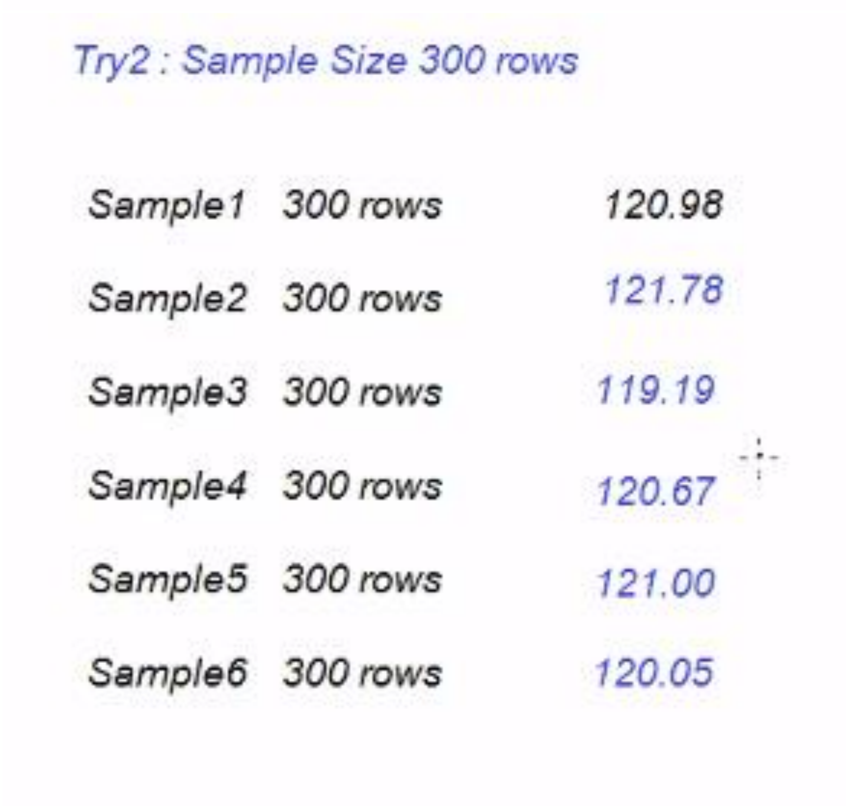
When we are going to give the data into the machine, we need to train it and then test it.

Training set	10	20	30	40	50	60
Testing set	11.23	21.34	35.67	42.11	56.77	59.99

You can see Testing set will be close to the training set.

To build generalization, we will pick up the samples and perform the operations.





Try2 : Sample Size 300 rows

Sample1	300 rows	120.98
Sample2	300 rows	121.78
Sample3	300 rows	119.19
Sample4	300 rows	120.67
Sample5	300 rows	121.00
Sample6	300 rows	120.05

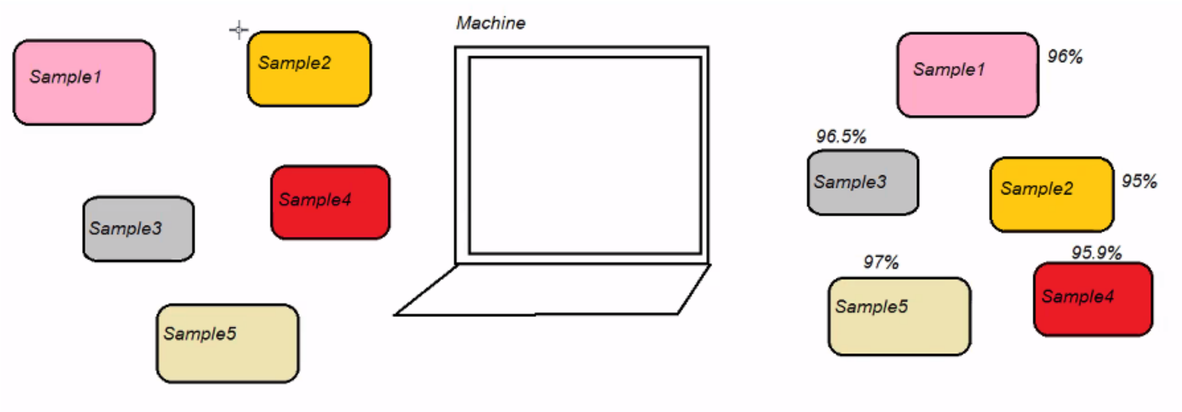
If there is a stable variations, then we consider it has good sample.

Generalization capability is the main thing we focus from variation and standard deviation

Similar variations:

Good fitting!

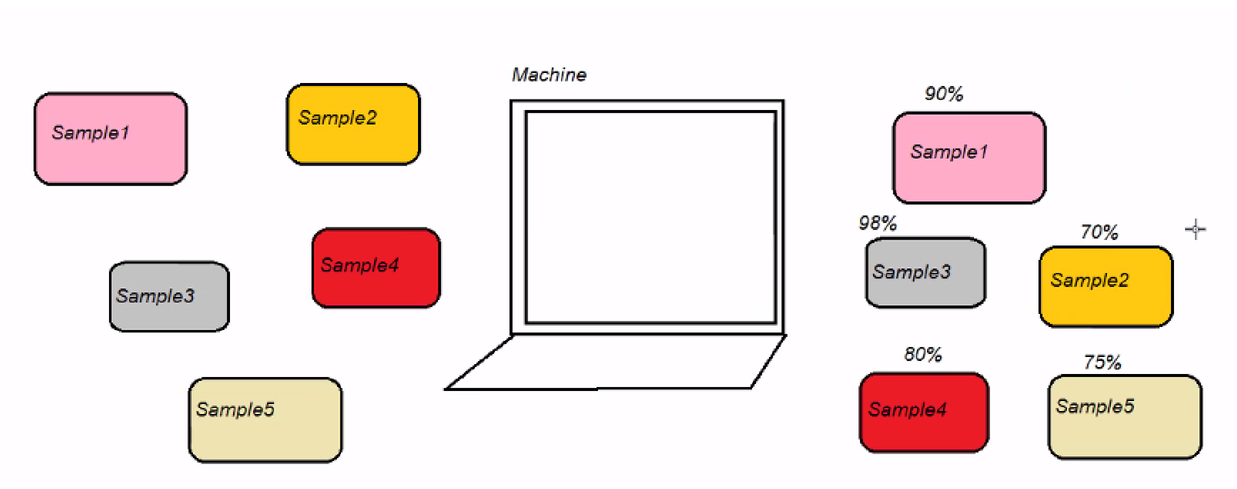
If the samples, we pick up is good and gives a better generalization capability.



It has got enough visibility to all the samples.

Huge variations:

Overfitting!



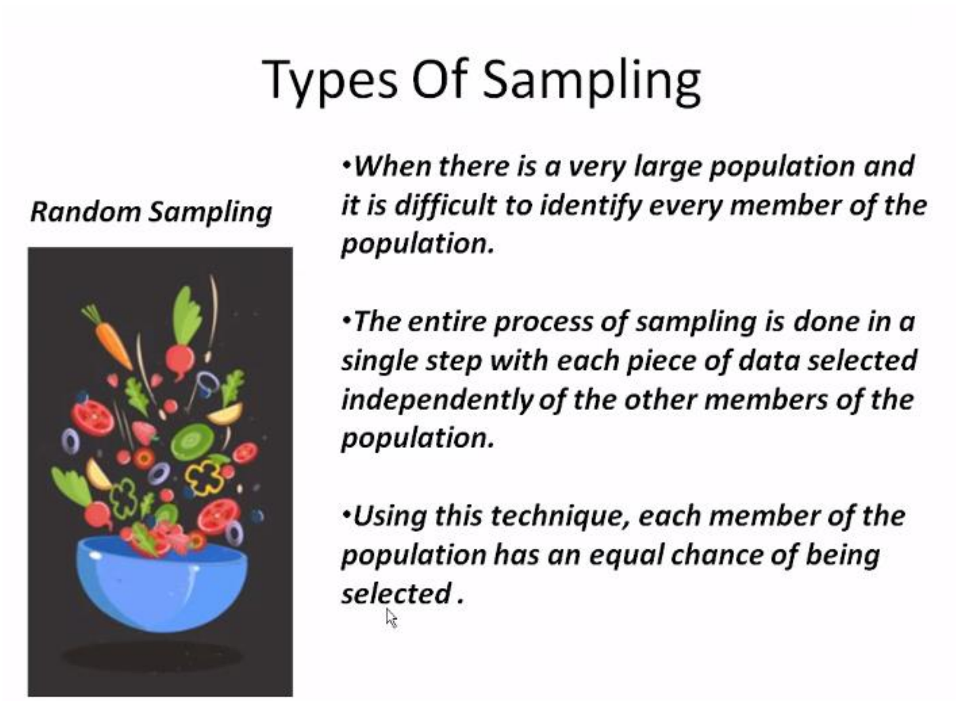
Training Set	10k rows from actual population if 100k rows
Testing Set1	1000 rows from actual population if 100k rows
Testing Set2	1000 rows from actual population if 100k rows
Testing Set3	1000 rows from actual population if 100k rows
Testing Set4	1000 rows from actual population if 100k rows

Sampling can be with or without replacement.

Deciding how much we must give as samples – it's an art. It's all about trial and error!

Types of Sampling:

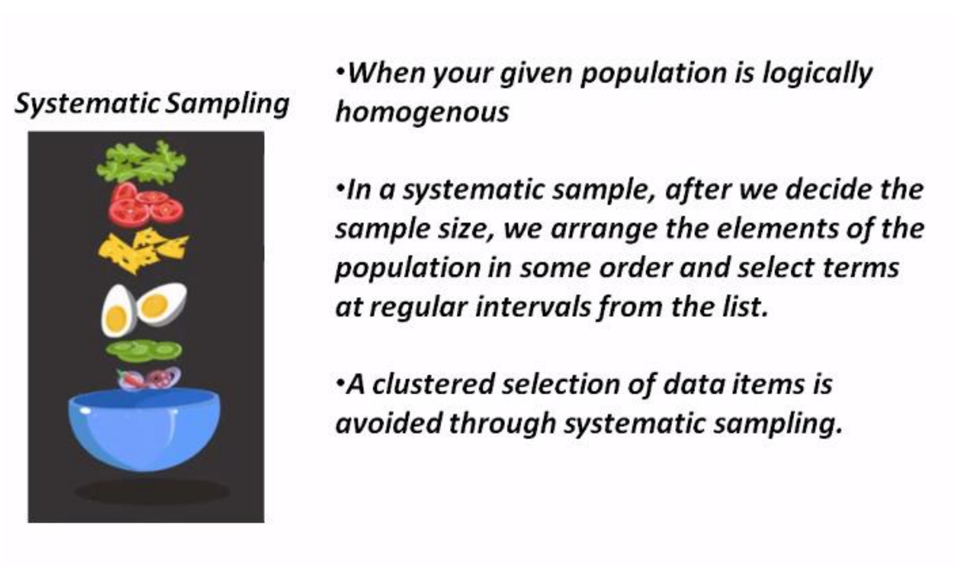
Random sampling:



Values are similar inside the dataset and variation is too similar, then random sampling is good.

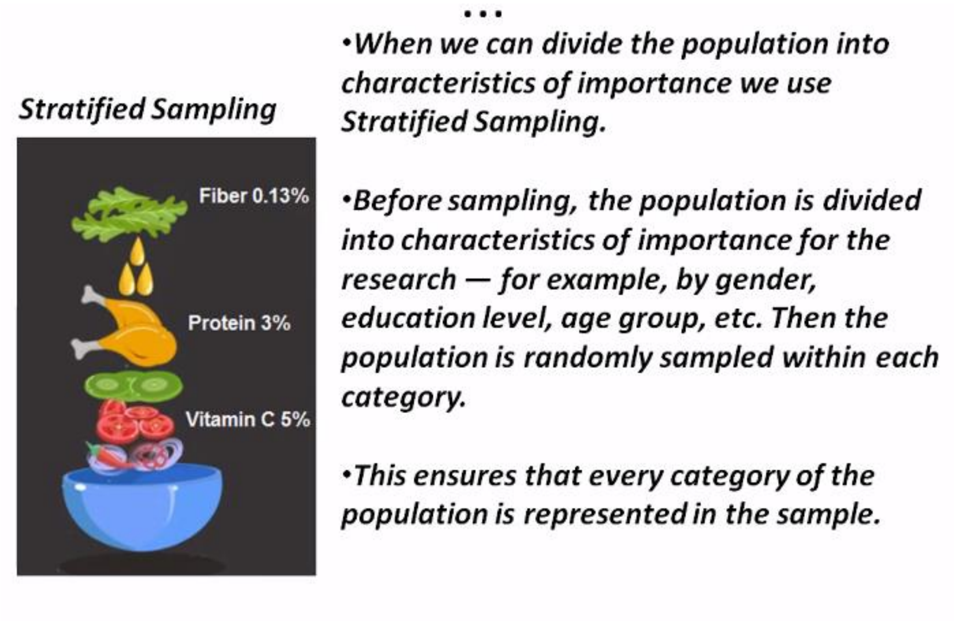
In random sampling, we face clustered selection.

Systematic Sampling:



When the values and variation is not similar, then systematic sampling is good.

Stratified Sampling:



Give importance to each category. Ensure all categories are picked up in the samples. More population more probability of the dataset. Less prominent cities, then less probability of the data.

Testing samples are randomized sampling. Training samples can be systematic and stratified sampling.

Sampling of Dataframe implementation:

https://towardsdatascience.com/how-to-sample-a-dataframe-in-python-pandas-d18a3187139b

Types of values/ features:

1) Discrete

2) Continuous

Discrete:

Concrete boundaries – well defined boundaries

Example:

Year of birth – 2000,2001, 2002, 2003 and so on..

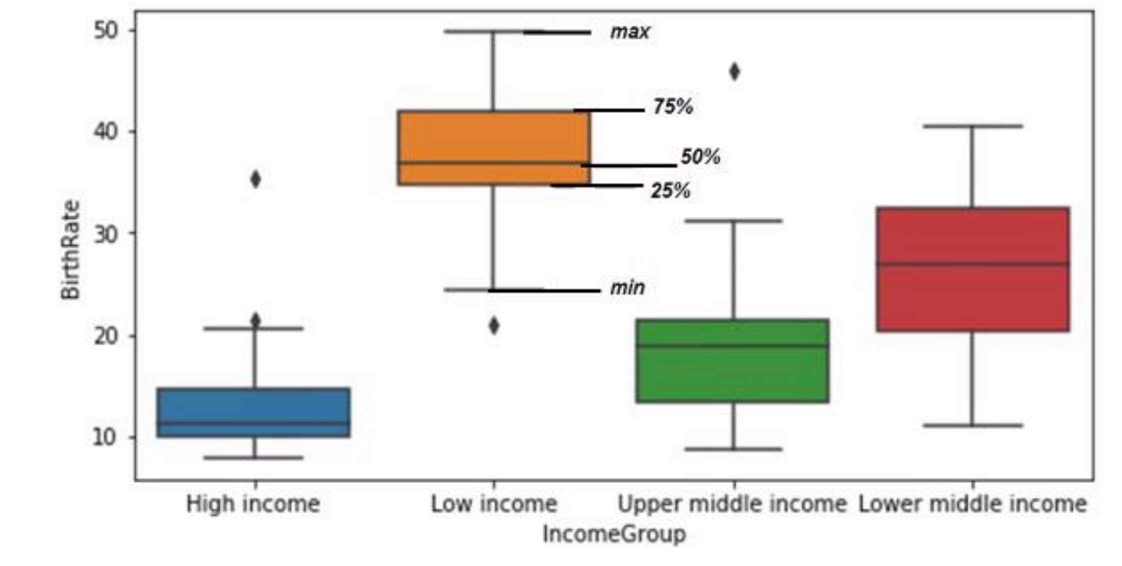
Response – 0 or 1

DeptID - 10 20 30 40 50

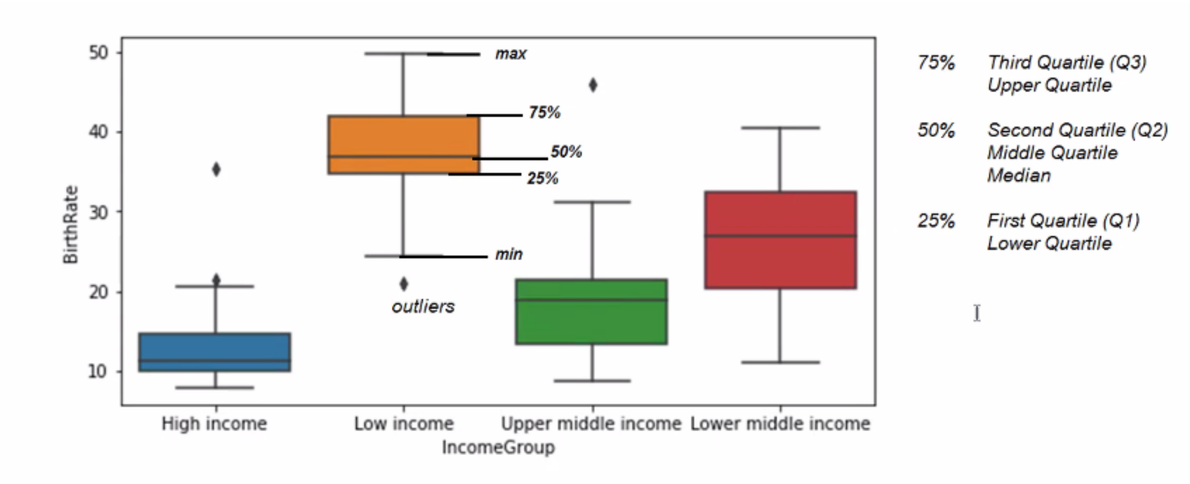
Continuous:

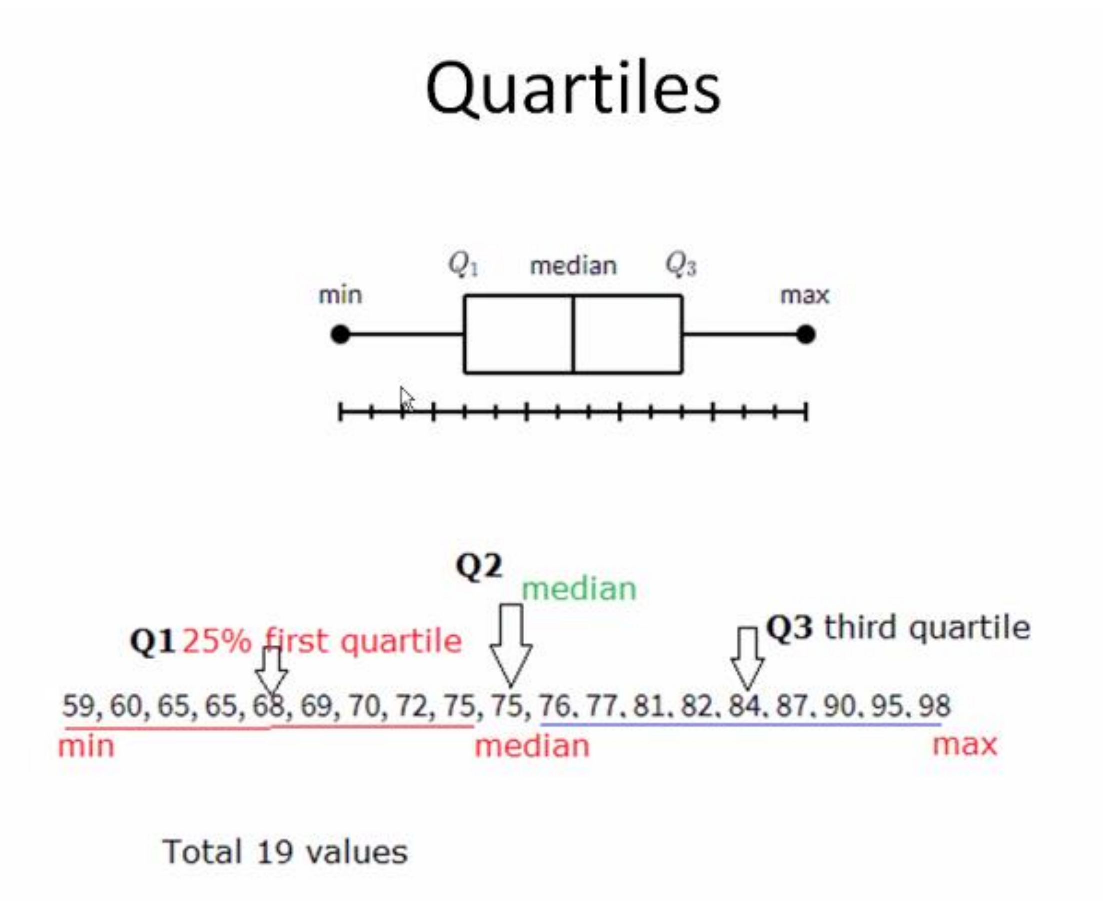
Example:

Exact weight of animal in the jungle?? 1Kg ~ 100 Kg

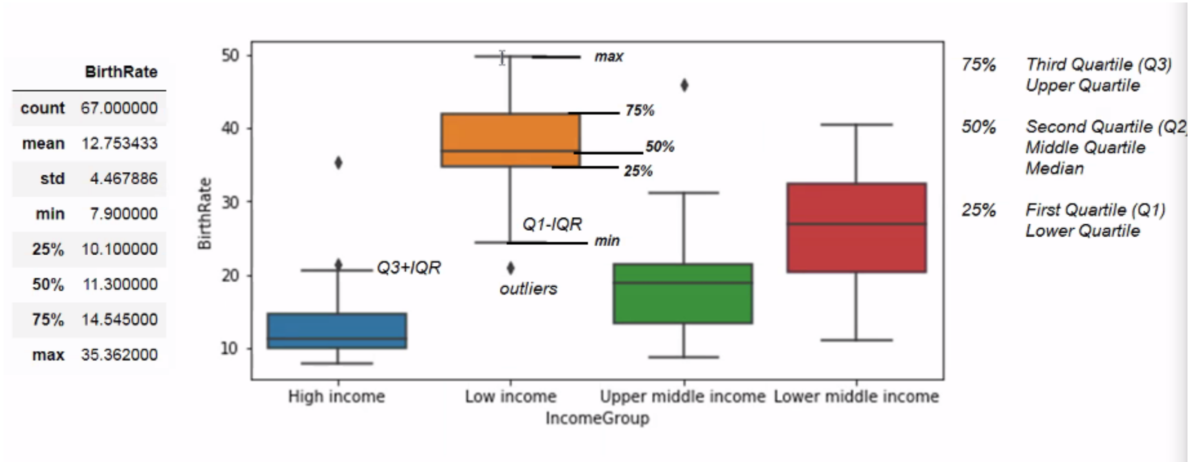


75%	*Third Quartile (Q3)* *Upper Quartile*
50%	*Second Quartile (Q2)* *Middle Quartile* *Median*
25%	*First Quartile (Q1)* *Lower Quartile*





Calculate the outlier:



	BirthRate
count	67.000000
mean	12.753433
std	4.467886
min	7.900000
25%	10.100000
50%	11.300000
75%	14.545000
max	35.362000

Outlier Calculation

Step 1: Calculate IQR (Inter Quartile Range) variance in Quartile

IQR = (Q3 – Q1) *1.5 = (14.545 - 10.10) * 1.5 = 6.667

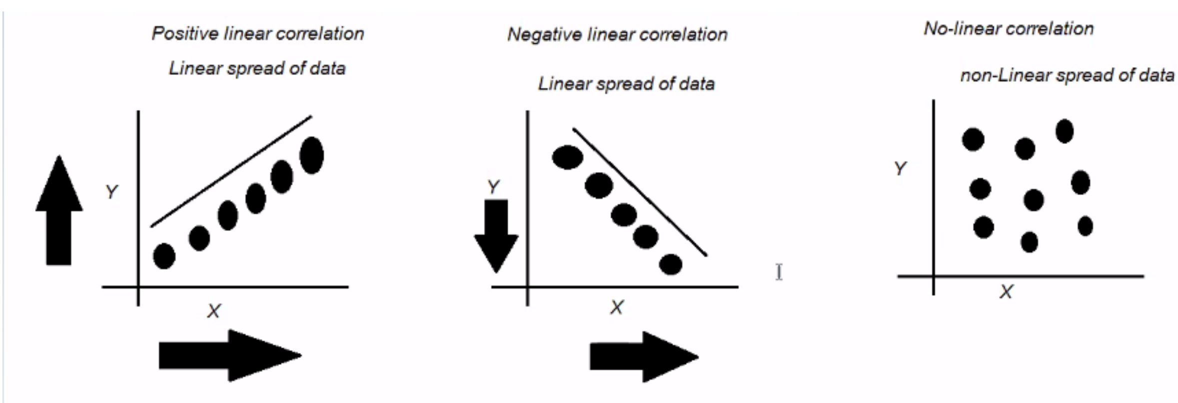
Step 2:

Q3 + IQR = 14.545 + 6.667 = 21.212 # all values above this are outliers

Q1 – IQR = 10.10 - 6.667 = 3.43 # all values above this are outliers

At middle income, you see mean and median are very similar.

Correlations: Measure of linearity between X and Y.



Pearson Coefficients:

Pearson Coefficient

```
              N *sum(xy) - sum(x)*sum(y)
Pearsonr =   ------------------------------------------------------------
              sqrt([N*sum(x^2)- sum(x)^2]*[N*sum(y^2)-sum(y)^2])
```

Using scipy we can find it.

Correlation is calculated using r-value / Pearsonr-value & p-value

Pearsonr-value conveys the percentage of correlation between X & y

p-value conveys the percentage of uncorrelation between X & y

==

Pearsonr-value (-1 to 1)

Pearsonr-value = 0.95	X & y are 95% correlated
Pearsonr-value = 0.08	X & y are 8% correlated

Pearsonr-value

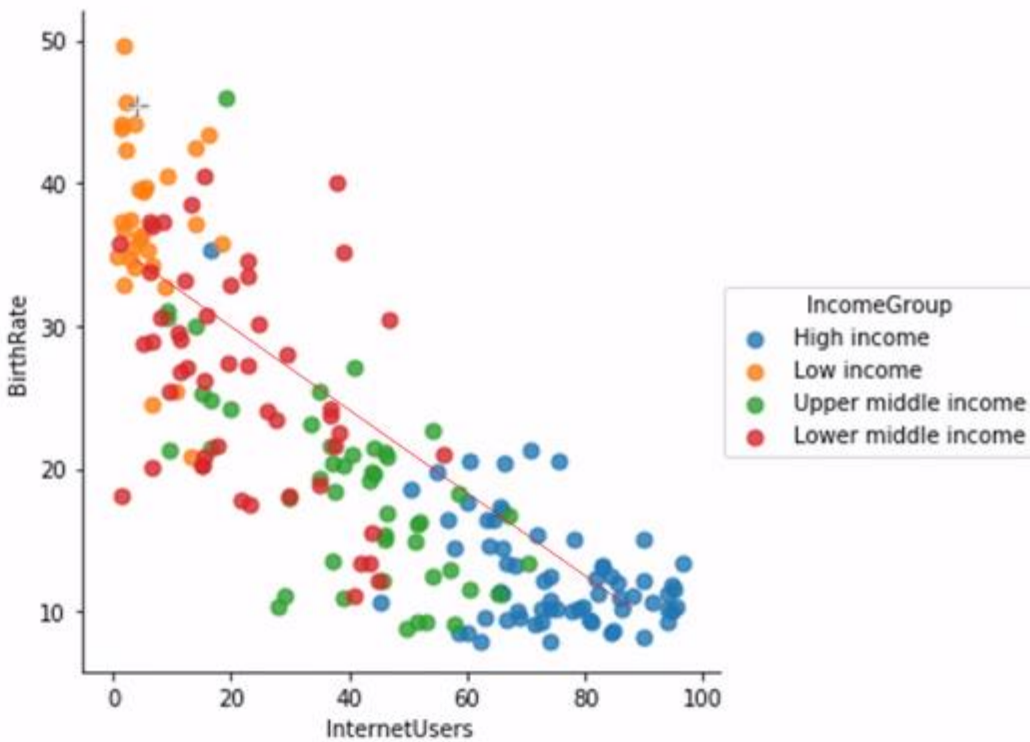
0 to < 0.25	No correlation , No relevance between x & y
0.25 to < 0.50	Negligible correlation / relevance between x & y
0.50 to < 0.75	Moderate correlation / relevance between x & y
> 0.75	Very Strong correlation / relevance between x & y

==
p-value

p-value=0.98 X & y are 98% uncorrelated
p-value=0.02 X & y are 2% uncorrelated

X features with p-value above 0.05 highly uncorrelated , we ignore or avoid choosing those X features

X features with p-value below 0.05 highly correlated , we consider choosing those X features



X features 10 in number to predict Y

8 features have correlation in moderatae to very strong
2 features have neglible to no-correlation
1 x feature pearsonr- 45%
2nd X feature pearsonr- 12%

Flow of data in Data science

Data Collection

Data Cleaning

Data exploration: Calculate mean median, mode, std deviation and variance

Data Preprocessing: Whatever data you are cleaning, you will apply preprocessing techniques on it. Examples: Encoding, dummy variable creation, train-test split, scaling of values.

It is a value-added step. This will improve the quality of the data.

Model implementation:

Supervised learning – Accept both input and output data – Remember & Generalize

Regression – Value to predict is Continuous

Linear Regression

Polynomial Regression

Decision Tree Regression

Random Forest Regression

Classification – Value to predict is Discrete

Logistic Regression

Support Vector Machine

Decision Tree classifier

Random Forest Classifier

Naïve Bayes

Unsupervised learning – Accept only input data – Remember & Generalize

Clustering

K-means clustering

Reinforcement learning – Accept data on the go – online learning – Adaptive

Upper Confidence Bound

Thompsan Sampling

Deep learning – Accept input data and output data – Remember & Generalize

Artificial Neural Network

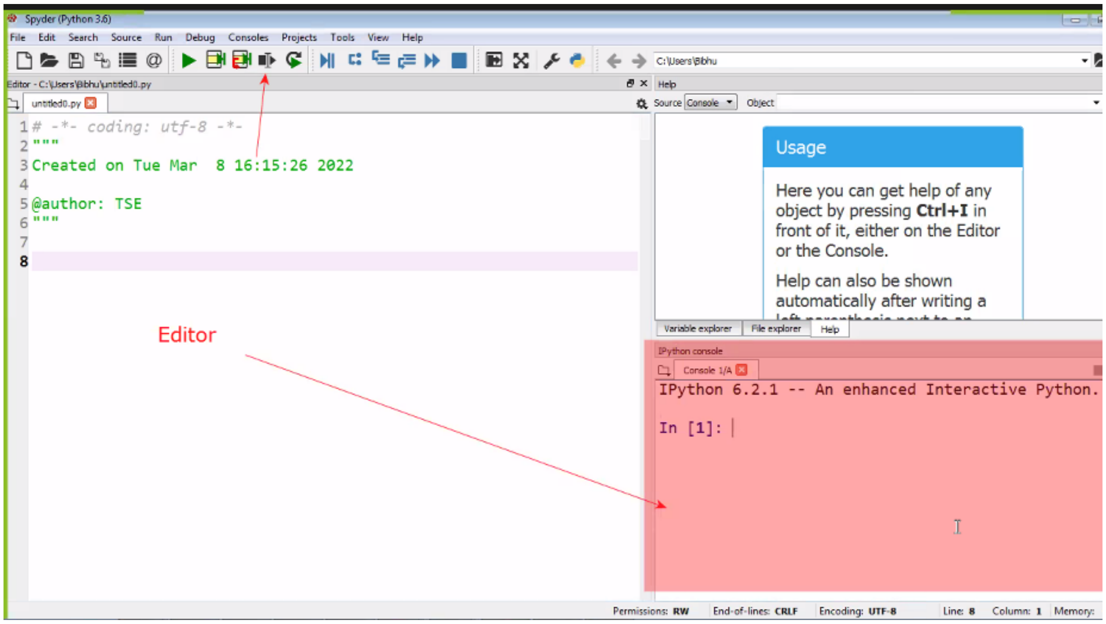
Convolutional Neural Network

Recurrent-Neural Network – LSTM (Long short-term memory)

Natural language processing – Preprocessing text to number

Regression Problem:

Open Spyder:



Continuous value Prediction:

Continuous Value Prediction, measure of model performance is Error

YrsExp	ActualSal	PredictedSal	Error = diff(acutal,predicted) = avg((diff(acutal,predicted) ^2))
1	10000	10500	-500
2.5	12000	12000	0
3	15000	20000	-5000
3.5	17000	11000	6000
4	20000	19000	1000
4.5	25000	27000	-2000
			----------------Error

Less the error => Better is the algorithm => Target

Error is the parameter of judgement.

Performance measure is error – sum of error divided by the total number of values.

Discrete Value prediction:

Discrete Value Prediction, Measure of model performance is Accuracy percentage

Age	ActualResponse	PredictedResponse
20	0	0
30	1	1
21	0	0
35	1	0
40	1	1
45	1	1
50	0	1
18	0	0
19	1	0

6/9*100=66% accuracy score

Here, you must count the total number of correct prediction and divided by the total predictions.

Accuracy is the key measure of judgement.

Linear regression is to be chosen when the data is linear data distribution / moderate or strong correlation based on PSNR value. No need to plot every time and check. Based on Peasonr value also we can decide.

Linear Regression

$$y = b_0 + b_1 * x_1$$

y is the dependent variable
x1 is independent variable
b1 is coefficient of X or the slope of the line
b0 is the constant called intercept

Linear Regression:

Y = b0 + b1 * x1

Salary Prediction

*Salary = base Package + Amount * Total Experience*

Base Package are the people with 0 years of experience.

Example:

B0 = 25000

Total experience = 5000

Amount = 2

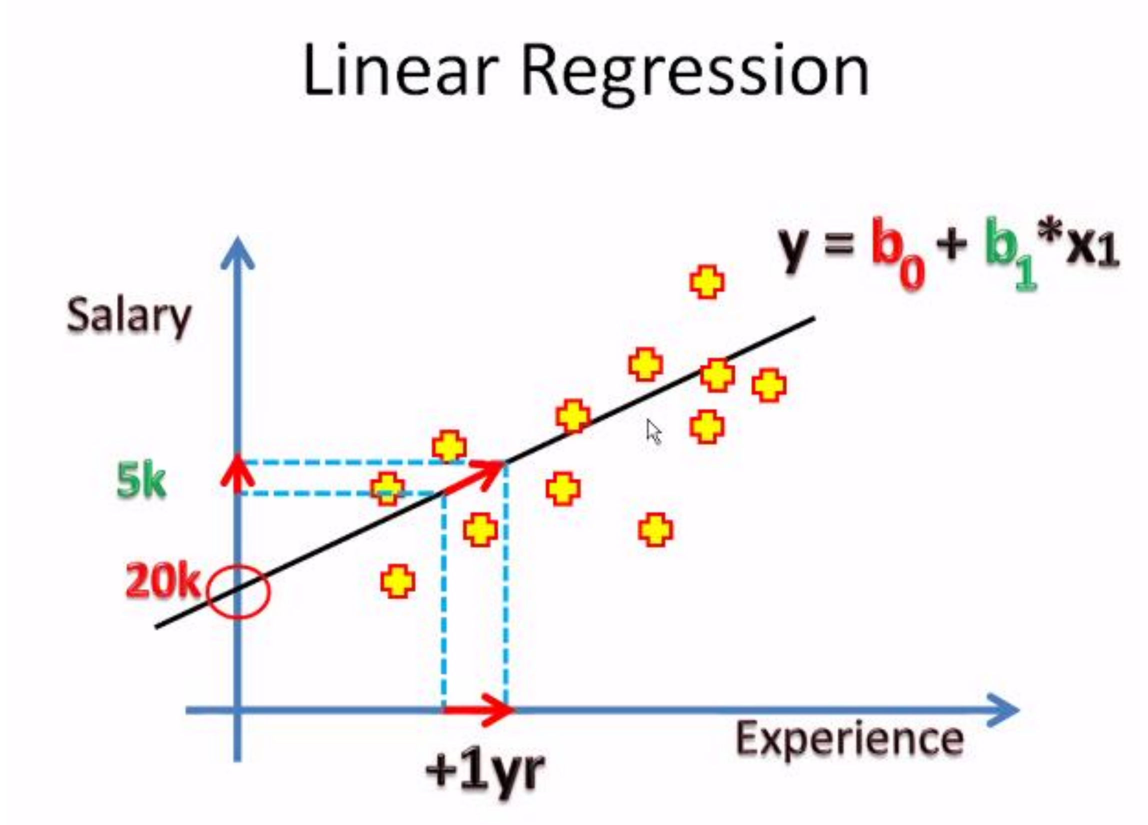
Salary = 35000

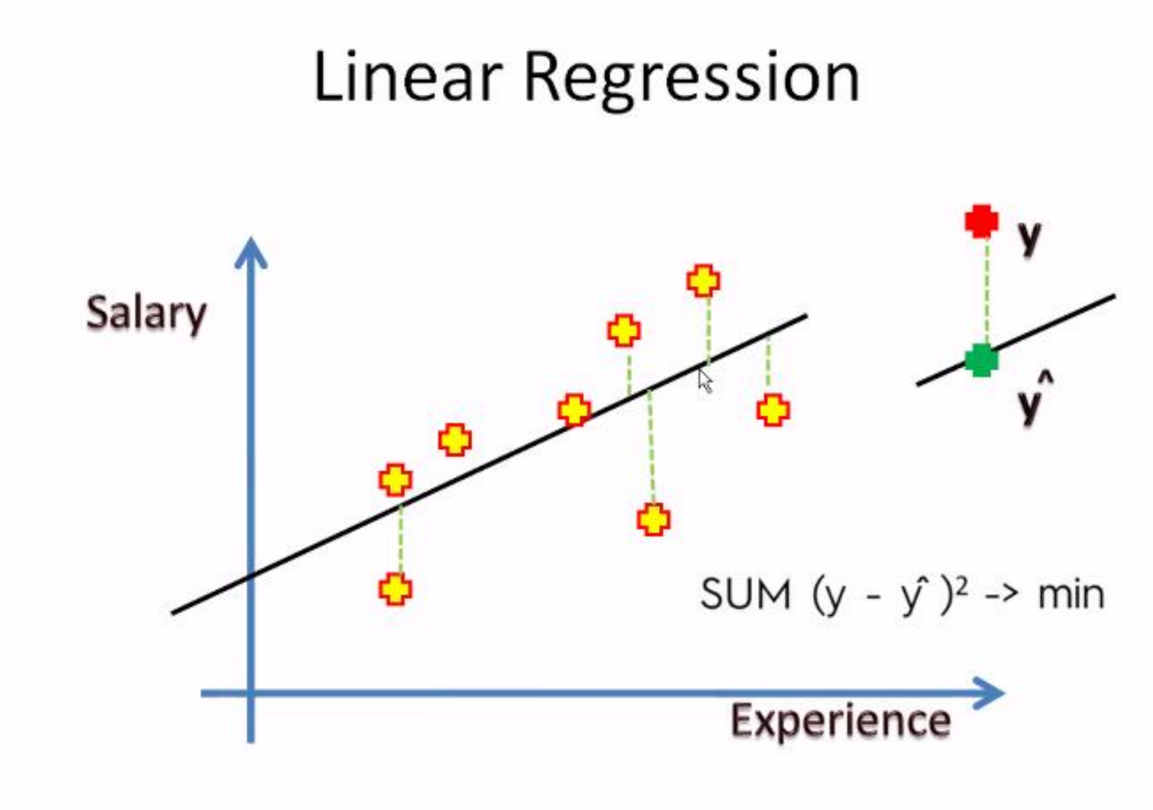
How b0 and b1 is calculated?

```
sum(y)*sum(x^2) - sum(x)sum(xy)
------------------------------------------b0 = intercept
n*sum(x^2) - sum(x)^2



n*sum(xy) - sum(x)*sum(y)
--------------------------------b1 = slope
n*sum(x^2) - sum(x)^2
```



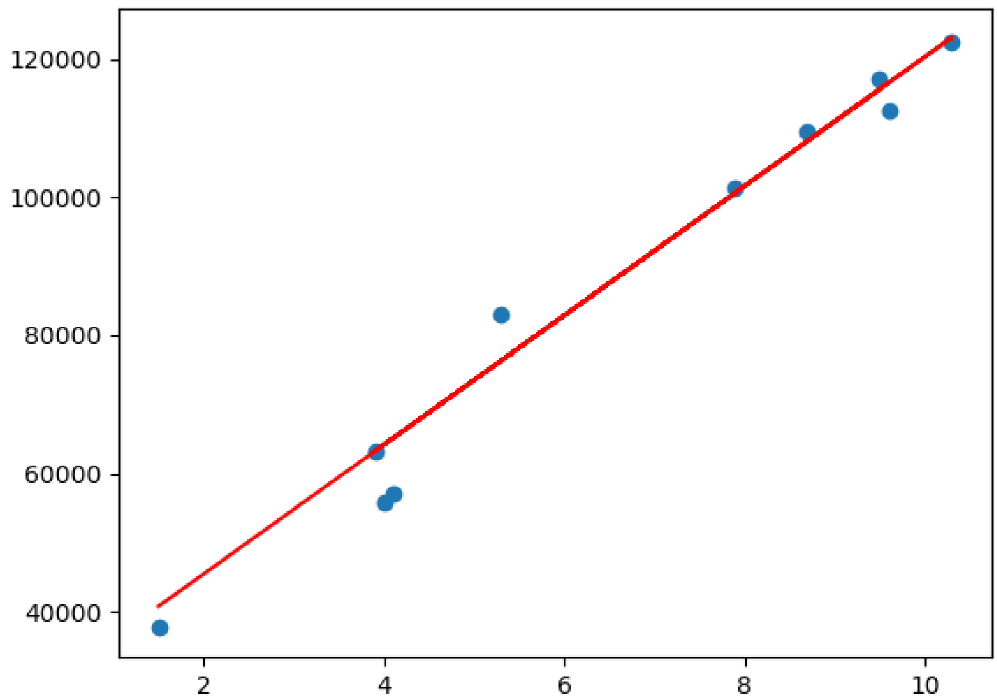


Why is intercept needed?

If there is no intercept, then you won't be able to draw the best fir line. That gets difficult!

Intercept is the grace mark which pushes the slope line to fit the datapoints.

Output of simple regression:



Sample 100 rows

Random Split 80-20 or 70-30 or 75-25

Train Data	80%	80 rows	X_train,y_train	YrsExp,Salary
Test Data	20%	20 rows	X_test,y_test	YrsExp,Salary

Machine Getting trained X_train,y_train YrsExp,Salary

Machine Testing

Predict salary for X_test (YrsExp)
PredSalary = y_pred
Actual Salary = y_test

```
In [10]: np.mean((y_test-y_pred)**2)
Out[10]: 21026037.329511296

In [11]: np.sqrt(np.mean((y_test-y_pred)**2))
Out[11]: 4585.4157204675885
```

Here, we have only one train data and we can have multiple test data.

Test Data1	20%	20 rows	X_test,y_test	YrsExp,Salary	4585
Test Data1	20%	20 rows	X_test,y_test	YrsExp,Salary	4401
Test Data1	20%	20 rows	X_test,y_test	YrsExp,Salary	4500
Test Data1	20%	20 rows	X_test,y_test	YrsExp,Salary	4587
Test Data1	20%	20 rows	X_test,y_test	YrsExp,Salary	4505

Use Case
Client: I want a ML solution for my business

Business Objective / Acceptance criteria
I will accept ML solution if I se improvement in customer satisfaction by 25% and revenue improvement of 2million

Data : Data cleaning , Data Exploration, Data Preprocessing

Train the Machine

Test the Model => 98% accurate results

Situation1: Model1 98% accuracy
improvement in customer satisfaction 5% excepted was 25%
revenue improvement of 50k expected was 2million

Situation2: Model2 80% accuracy
improvement in customer satisfaction 35% excepted was 25%
revenue improvement of 5million expected was 2million