Gift This Course

A-Z Machine Learning using Azure Machine Learning (AzureML)

Hands on AzureML: From Azure Machine Learning Introduction to Advance Machine Learning Algorithms. No Coding Required.

BEST SELLER ★★★★★ 4.3 (215 ratings) 1,597 students enrolled

Created by Jitesh Khurkhuriya Last updated 3/2018 English CC English

Summarize Data

Summarize Data Module

- Generates a basic descriptive statistics for the columns in a dataset
- All Columns with Missing Values
- Get a count of categorical values for a column
- Numerical statistics such as mean and standard deviation of the column

Some Additional Terms

Mean Deviation

$$\text{Mean} = \frac{\text{Sum of Salary}}{\text{Number of Observations}}$$

$$= \frac{\$\ 103{,}723}{15}$$

$$= \$\ 6{,}915$$

Row Number	Salary
1	$ 3,725
2	$ 4,155
3	$ 4,627
4	$ 5,147
5	$ 5,718
6	$ 6,347
7	$ 7,039
8	$ 7,210
9	$ 7,423
10	$ 7,556
11	$ 8,369
12	$ 8,810
13	$ 8,940
14	$ 9,200
15	$ 9,458

Mean Deviation

Row Number	Salary	Distance from Mean
1	$ 3,725	$3,190
2	$ 4,155	$2,760
3	$ 4,627	$2,288
4	$ 5,147	$1,768
5	$ 5,718	$1,197
6	$ 6,347	$568
7	$ 7,039	$124
8	$ 7,210	$295
9	$ 7,423	$508
10	$ 7,556	$641
11	$ 8,369	$1,454
12	$ 8,810	$1,895
13	$ 8,940	$2,025
14	$ 9,200	$2,285
15	$ 9,458	$2,543

Mean = $ 6,915

Mean Deviation = $ 1,569

Sample Variance & Standard Deviation

Salary X	Distance from Mean	Square of the distance
$ 3,725	$3,190	$1,01,76,100
$ 4,155	$2,760	$76,17,600
$ 4,627	$2,288	$52,34,944
$ 5,147	$1,768	$31,25,824
$ 5,718	$1,197	$14,32,809
$ 6,347	$568	$3,22,624
$ 7,039	$124	$15,376
$ 7,210	$295	$87,025
$ 7,423	$508	$2,58,064
$ 7,556	$641	$4,10,881
$ 8,369	$1,454	$21,14,116
$ 8,810	$1,895	$35,91,025
$ 8,940	$2,025	$41,00,625
$ 9,200	$2,285	$52,21,225
$ 9,458	$2,543	$64,66,849

Mean = $ 6,915

$$\text{Variance } (S^2) = \frac{\text{Sum of Squared distances}}{N-1}$$

$$\text{Sample Standard Deviation} = \sqrt{Variance}$$

Quartile

1st Quartile → row 4

Median → row 8

3rd Quartile → row 12

Row Number	Salary
1	$ 3,725
2	$ 4,155
3	$ 4,627
4	$ 5,147
5	$ 5,718
6	$ 6,347
7	$ 7,039
8	$ 7,210
9	$ 7,423
10	$ 7,556
11	$ 8,369
12	$ 8,810
13	$ 8,940
14	$ 9,200
15	$ 9,458

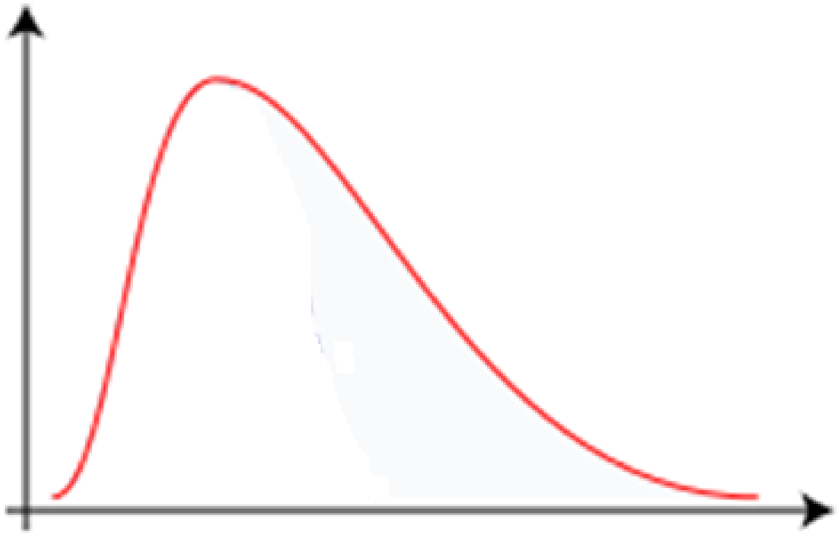
Q3 – Q1

Inter Quartile Range

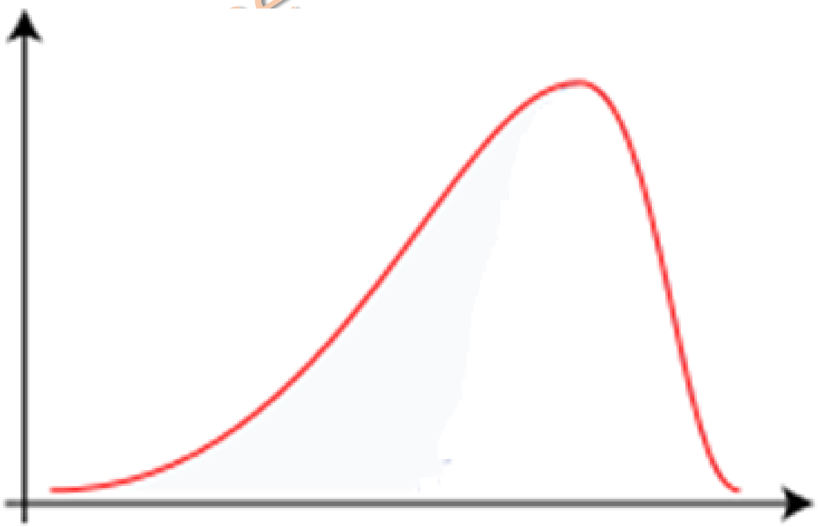
IQR

Skewness

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean – Wikipedia



Positive Skew



Negative Skew

Outliers

Outliers

- Observation that is distant from other observations
- Impacts the predictions or estimates

Mean = $107,600 / 12 = $8,967

Mean = $ 62,600 / 10 = $ 6,260

Salary
$ 4,000
$ 4,500
$ 8,000
$ 5,300
$ 5,700
$ 7,200
$ 7,400
$ 7,900
$ 6,400
$ 21,000
$ 24,000
$ 6,200

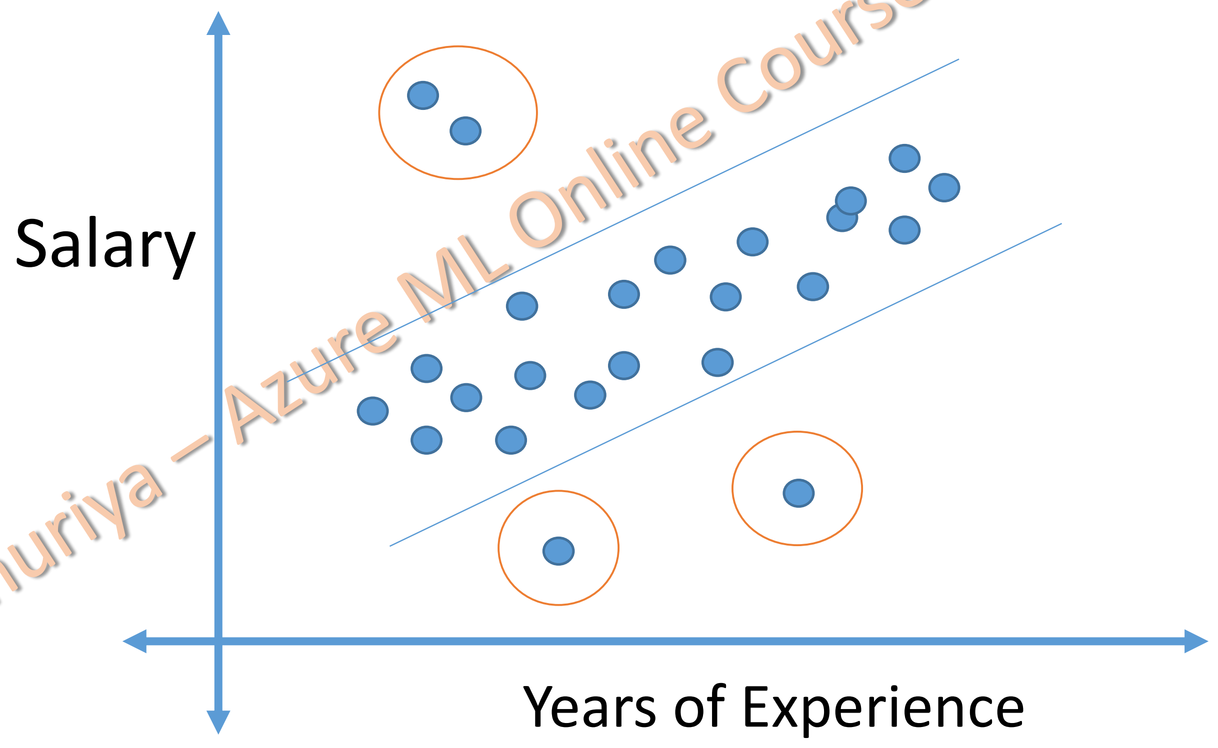
Outliers – Occurrences and Causes

- Human error
- Malfunction of the measurement equipment
- Data transmission or transcription error
- System Behaviour
- Fraudulent behaviour
- Natural Outliers
- Sampling error

Types of Outliers

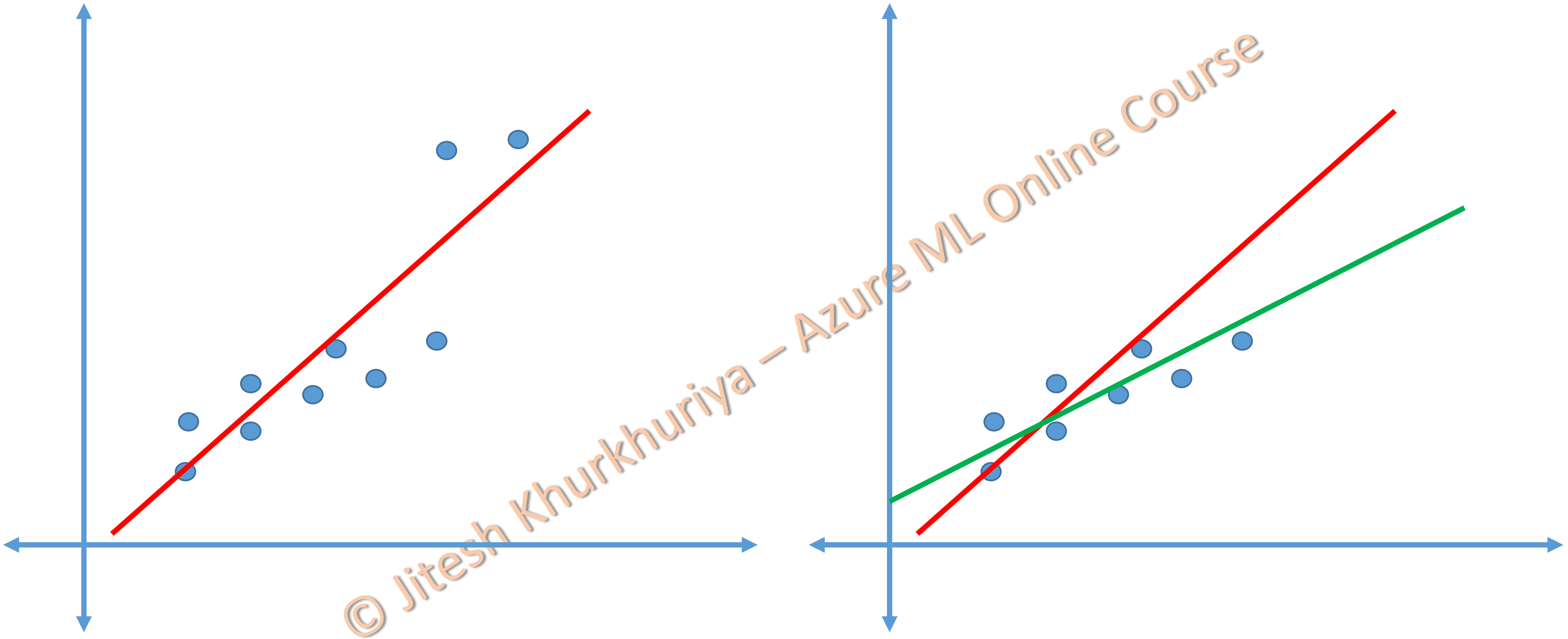
Salary
$ 4,000
$ 4,500
$ 8,000
$ 5,300
$ 5,700
$ 7,200
$ 7,400
$ 7,900
$ 6,400
$ 21,000
$ 24,000
$ 6,200

Univariate



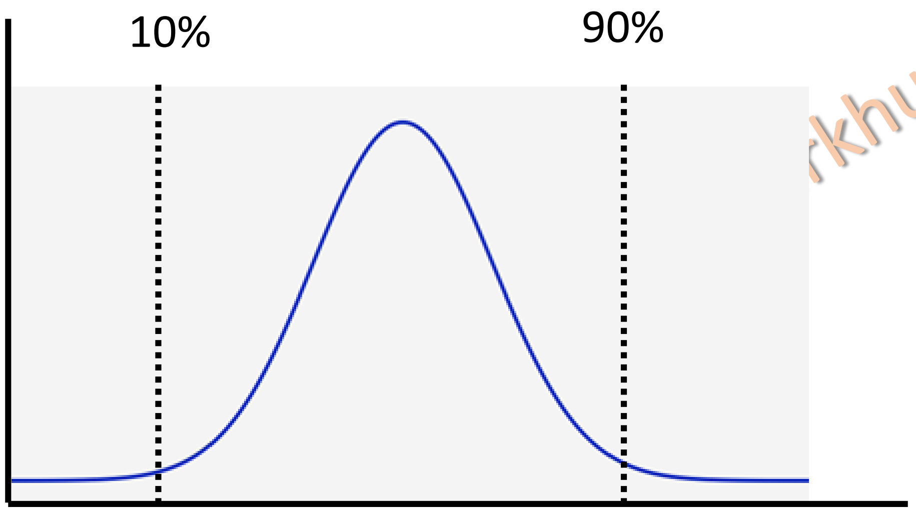
Multivariate

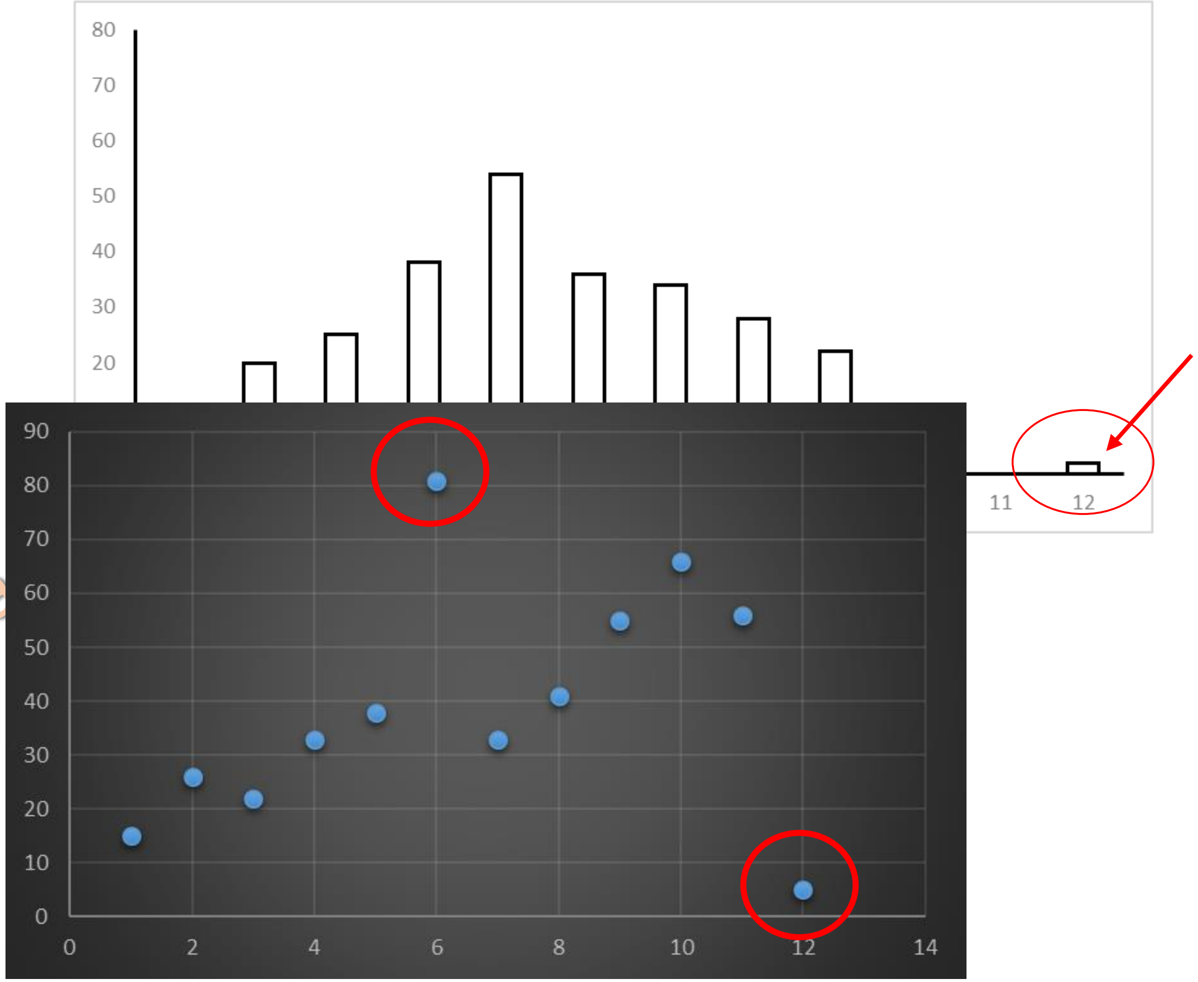
Impact of Outliers



How to Detect Outliers?

- Most common method is visualisation
- Box Plot, Histogram, Scatter plot
- Percentile measures





Normalize Data

What is Normalization?

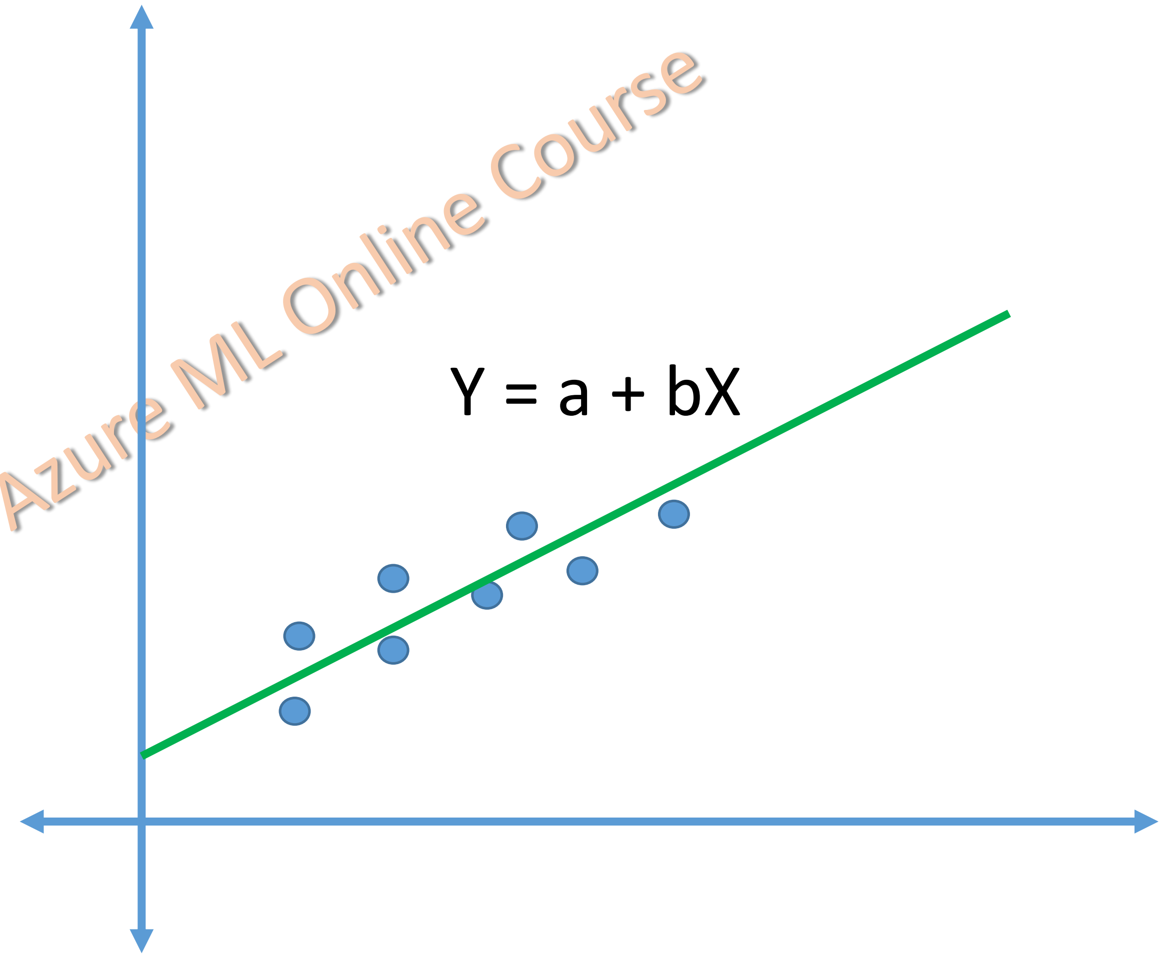
- A method to standardise the range of independent variables or features of data
- Variables are fitted within a certain range (Generally between 0 and 1)
- Applied on numeric columns

Why to Normalise the data?

$Y = a + b_1X_1 + b_2X_2^2$

$X_1 = 1,2,3.....20$

$X_2 = 1000,2000,.....20000$



Normalize data – Transformation Methods

ZScore

$$Z = \frac{X - mean(x)}{stdev\ (x)}$$

MinMax

$$Z = \frac{X - min(x)}{Max\ (x) - min(x)}$$

Logistic

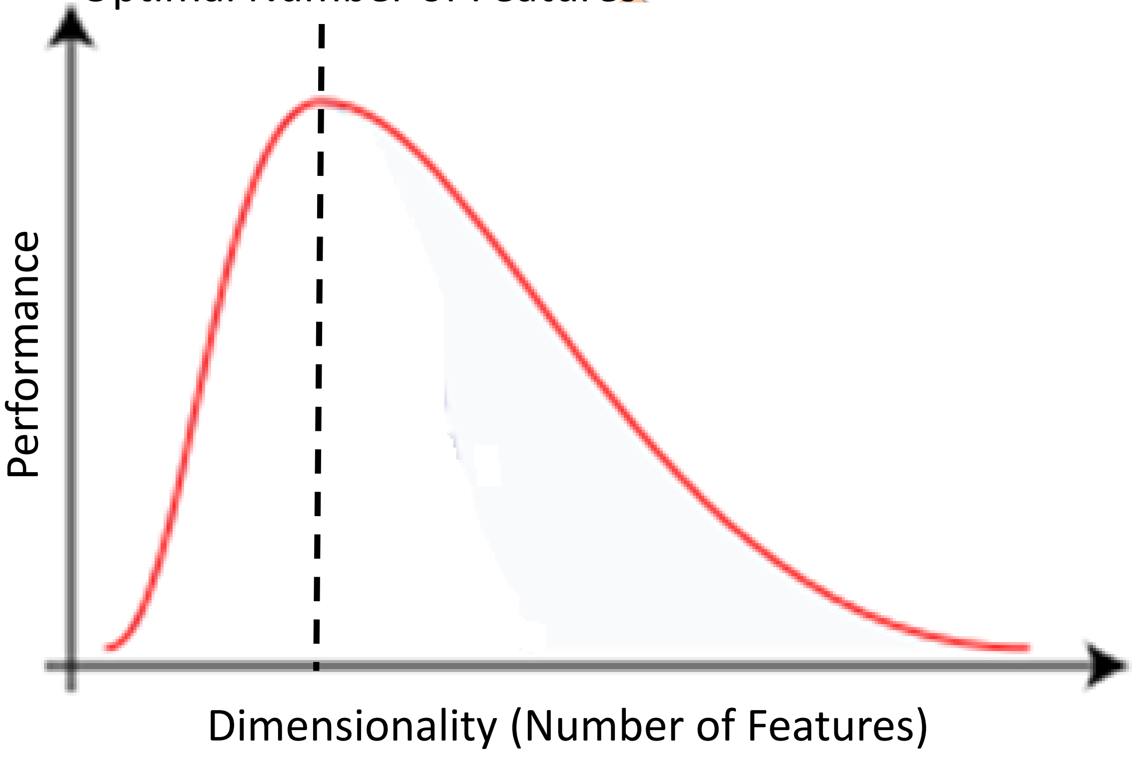
$$Z = \frac{1}{1 + exp(-x)}$$

Most commonly used transformation methods

Principal Component Analysis

Curse of dimensionality

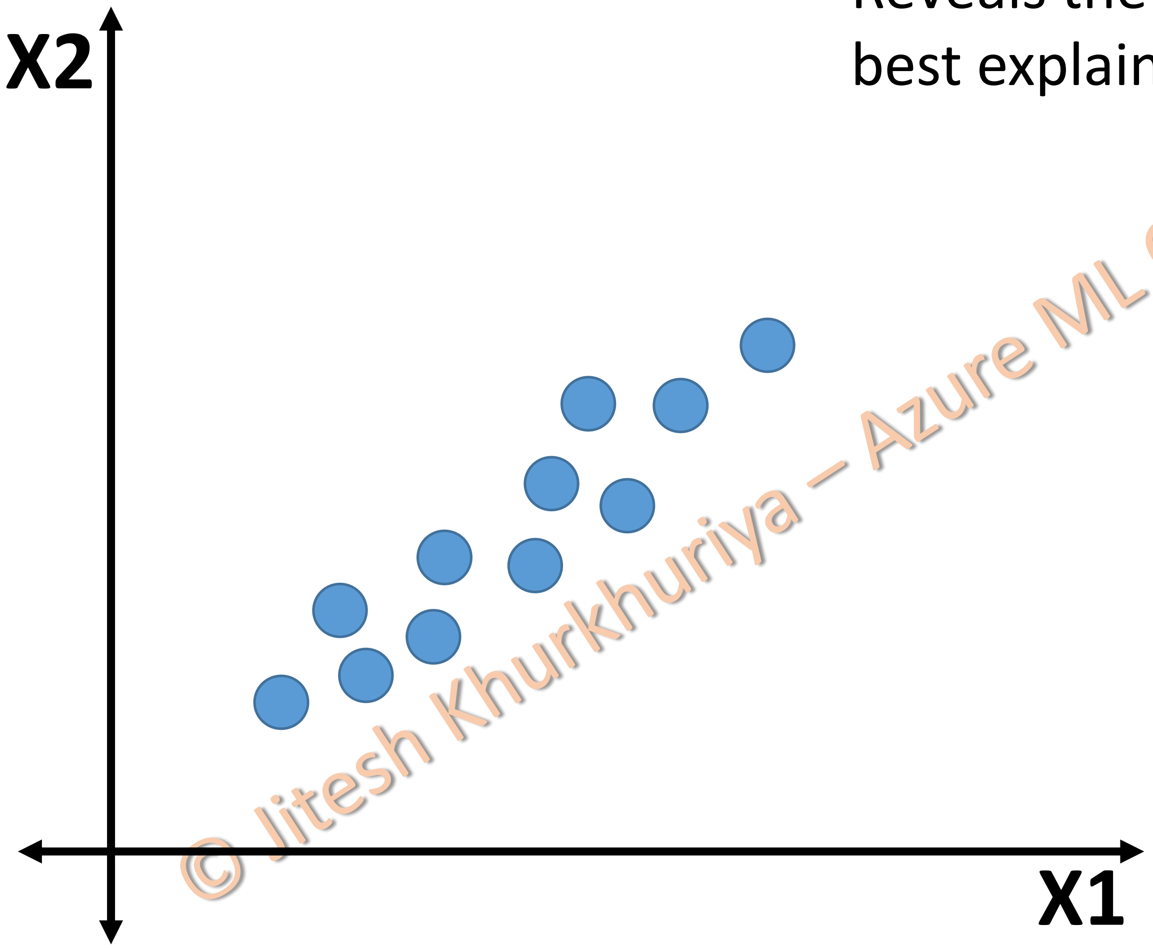
- 100s or 1000s of variables in a dataset
- Data becomes sparse as the available space increase multi-fold
- Sparse data can result in lesser accuracy
- Requires higher run-time
- May Lead to overfitting



What is a Principal Component?

- Creates a new set of coordinates for the data
- Reveals the internal structure of the data that best explains the variance in data
- Reduces the dimensionality of the multivariate dataset

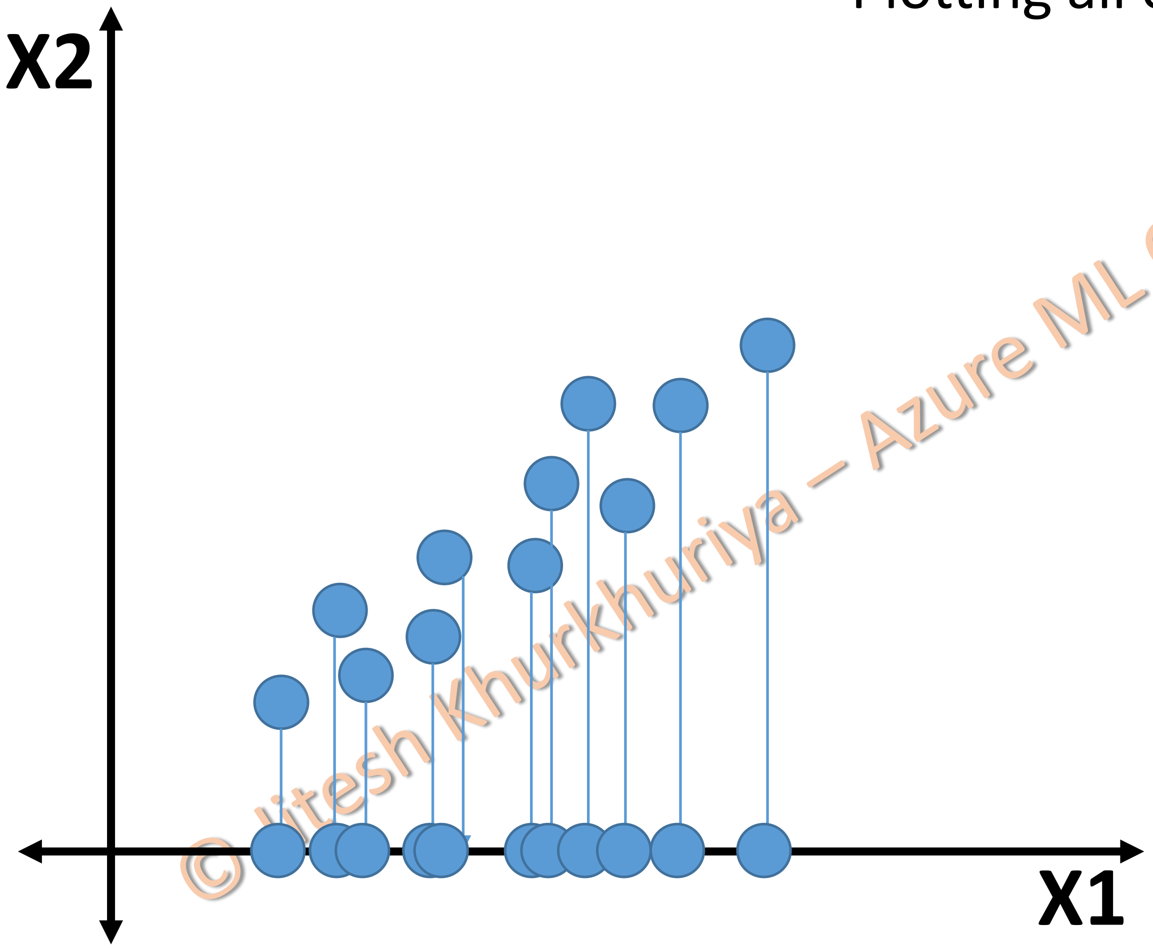
What is PCA?



Reveals the internal structure of the data that best explains the variance in data

What is PCA?

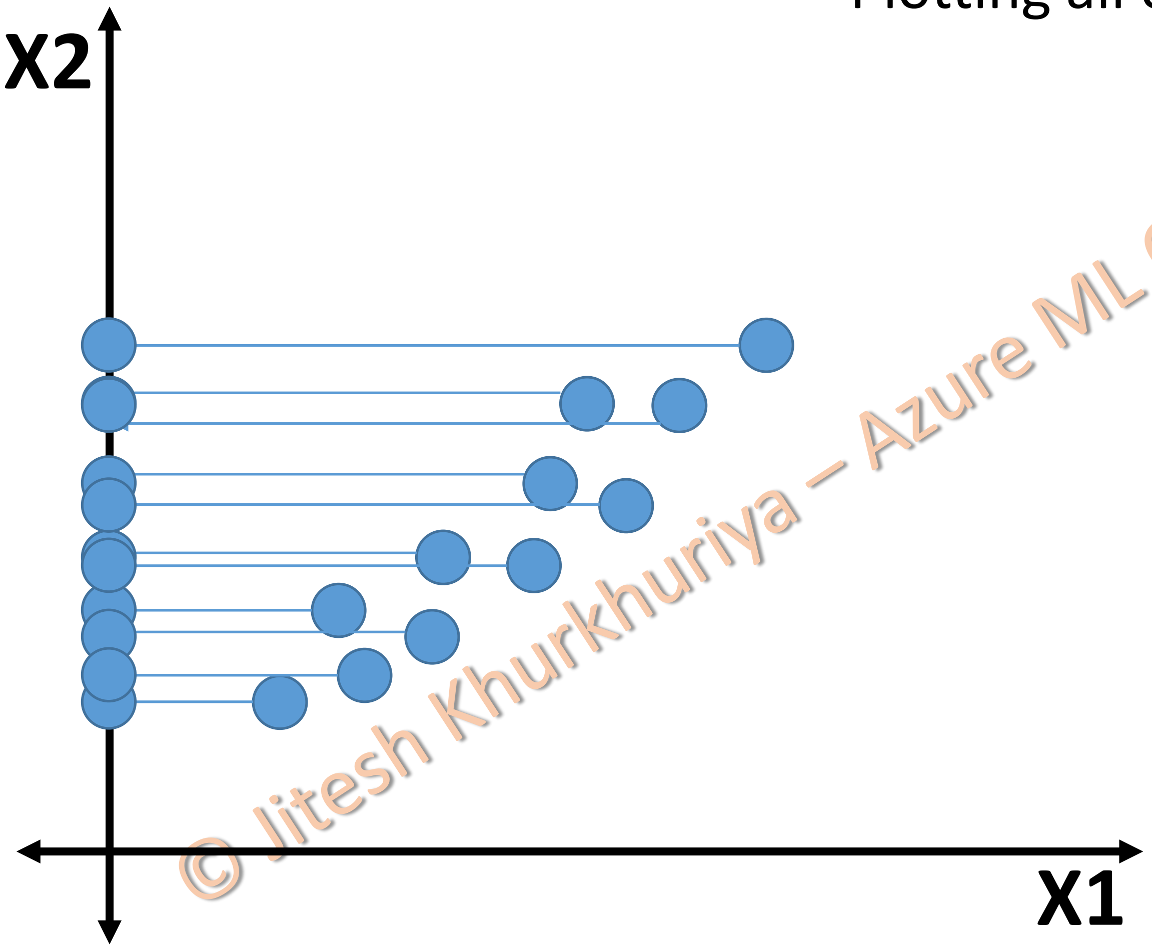
Plotting all observations on X1



© Jitesh Khurkhuriya – Azure ML Online Course

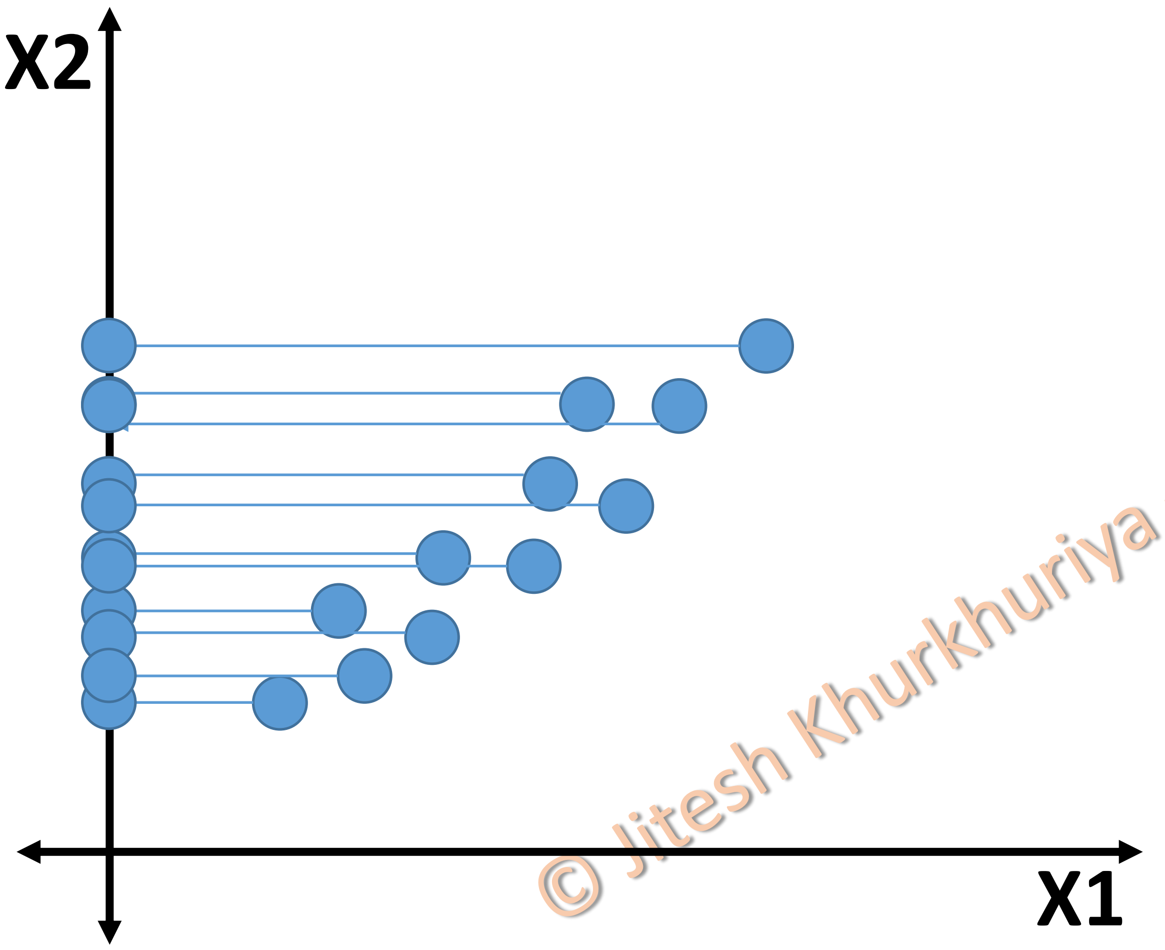
What is PCA?

Plotting all observations on X2



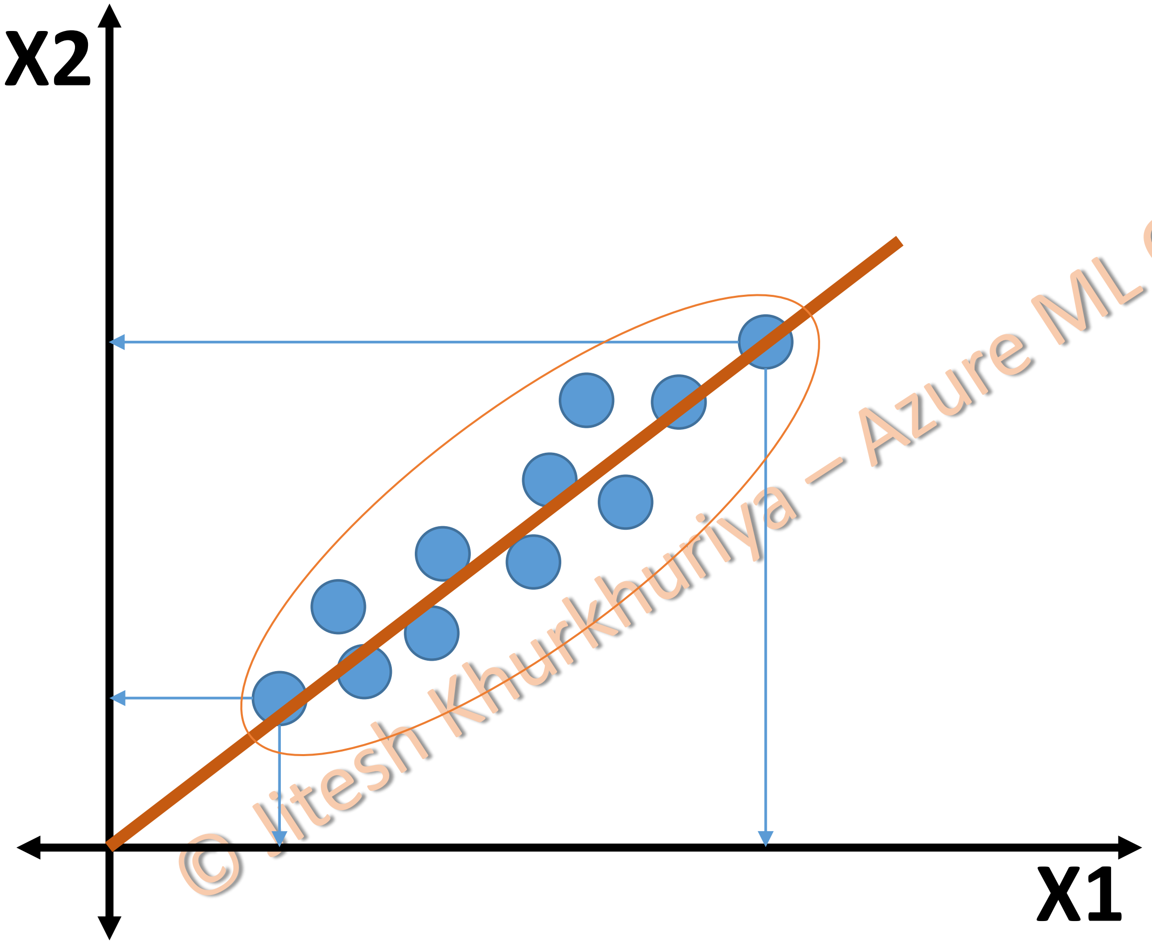
© Jitesh Khurkhuriya – Azure ML Online Course

What is PCA?

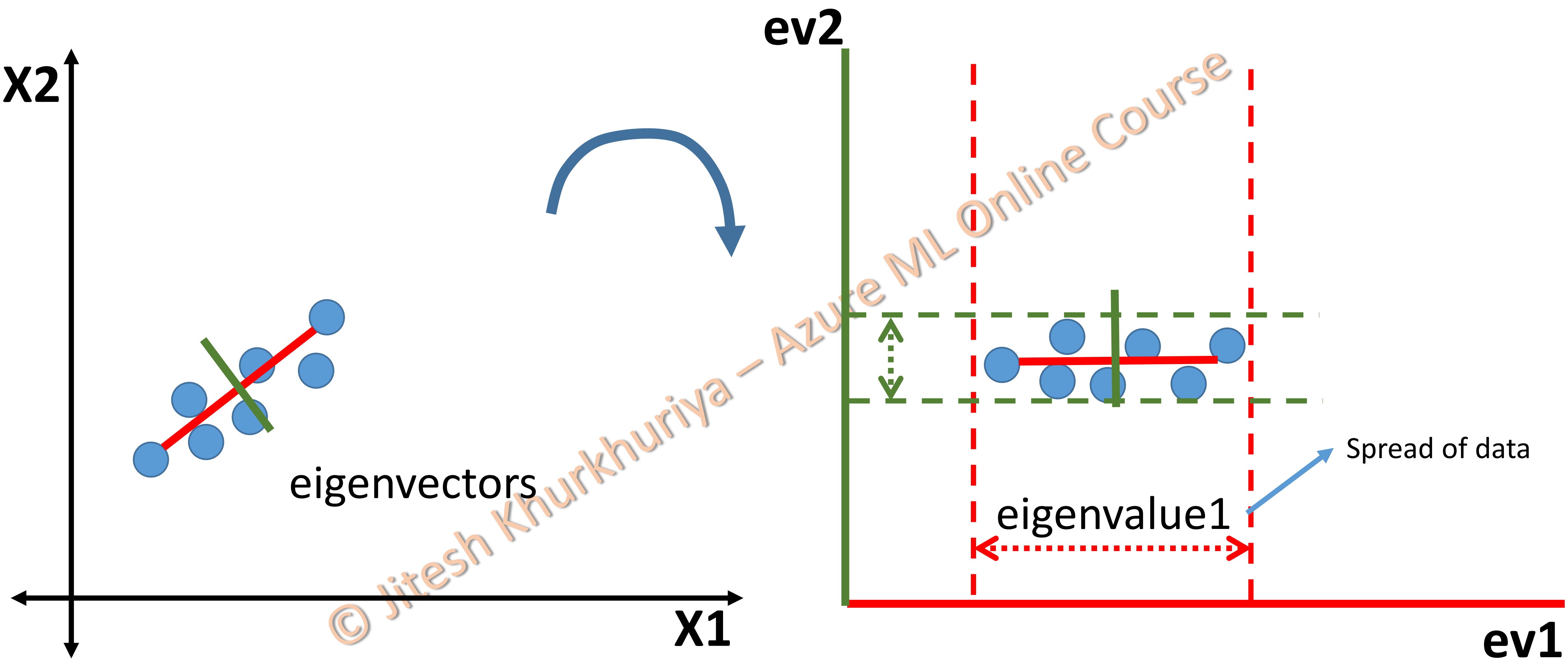


© Jitesh Khurkhuriya – Azure ML Online Course

What is PCA?

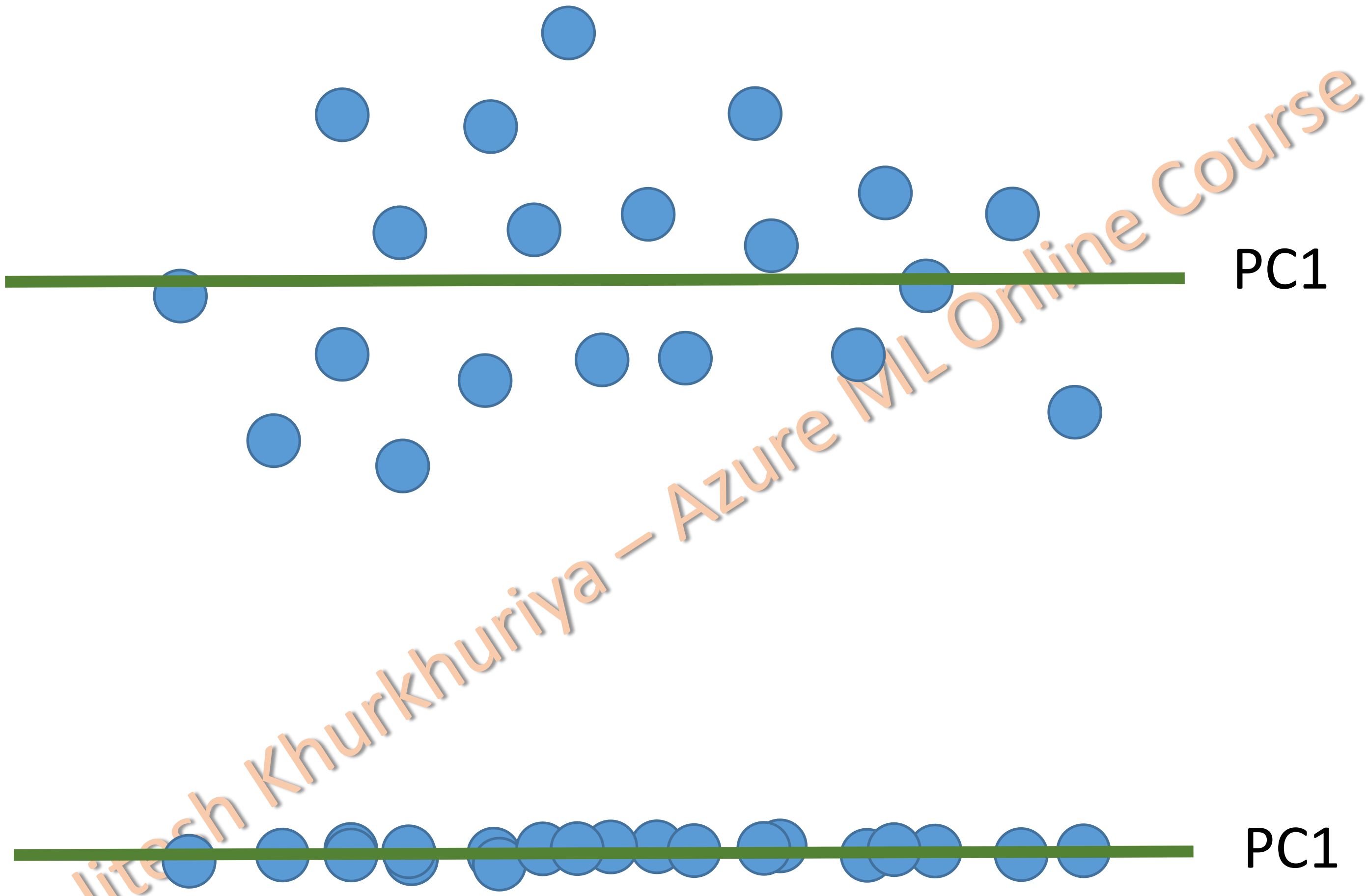


Understanding the PCA



ev1 has higher eigenvalue. Hence drop **ev2** as it explains much lesser variation compared to **ev1**

PCA



Clean Missing Data with MICE

MICE

- Replace with mean, mode or custom value – Single Imputation Method
- Multivariate Imputation using Chained Equation or Multiple Imputation by Chained Equations
- Each variable with missing data is modelled conditionally using the other variables in the data
- Data is Missing at Random
- Regression for predicting continuous variables and classification for categorical missing values

Simple example

Original Dataset

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	$ 5,500
38	$ 7,000
42	$ 7,500
33	$ 6,200
46	$ 7,800
48	$ 8,000
51	$ 8,500
43	$ 7,600
55	$ 8,500

Simple example

Missing Values

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	
38	$ 7,000
42	
33	$ 6,200
	$ 7,800
48	$ 8,000
	$ 8,500
43	$ 7,600
55	$ 8,500

MICE Steps

Step 1 – Calculate the Mean based on the available values

Step 2 – Replace all missing values with mean

Step 3 – Choose Dependent column and restore original

Step 4 – Apply transformation and create prediction model

Step 5 – Predict Missing values and repeat steps 3 to 5

Step 1 – Calculate the Mean based on the available values

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	
38	$ 7,000
42	
33	$ 6,200
	$ 7,800
48	$ 8,000
	$ 8,500
43	$ 7,600
55	$ 8,500

Age Mean = 38.1

Salary Mean = $ 7,080

Step 2 – Replace all missing values with mean

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	**$ 7,080**
38	$ 7,000
42	**$ 7,080**
33	$ 6,200
38.1	$ 7,800
48	$ 8,000
38.1	$ 8,500
43	$ 7,600
55	$ 8,500

Age Mean = 38.1

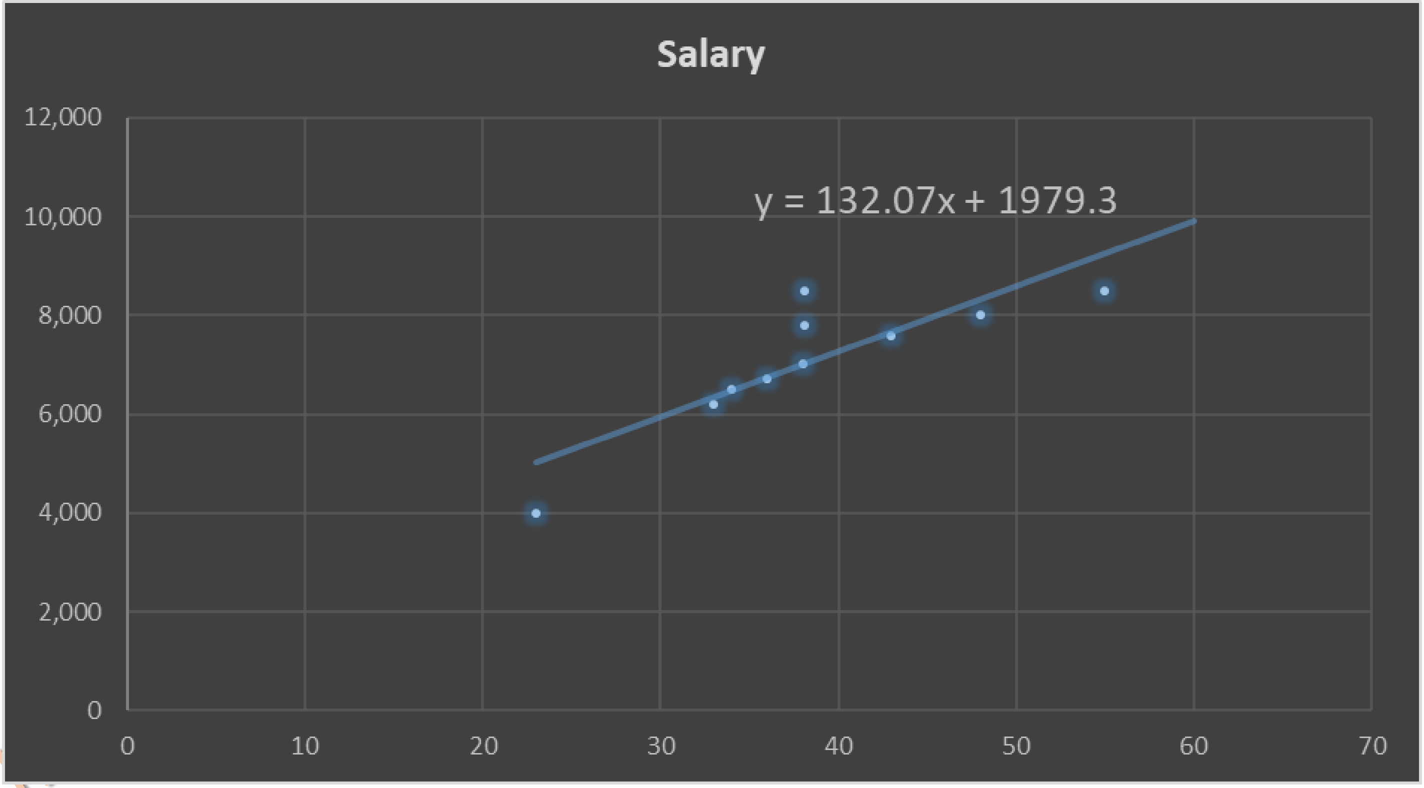
Salary Mean = $ 7,080

Step 3 – Choose Dependent column and restore original

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	
38	$ 7,000
42	
33	$ 6,200
38.1	$ 7,800
48	$ 8,000
38.1	$ 8,500
43	$ 7,600
55	$ 8,500

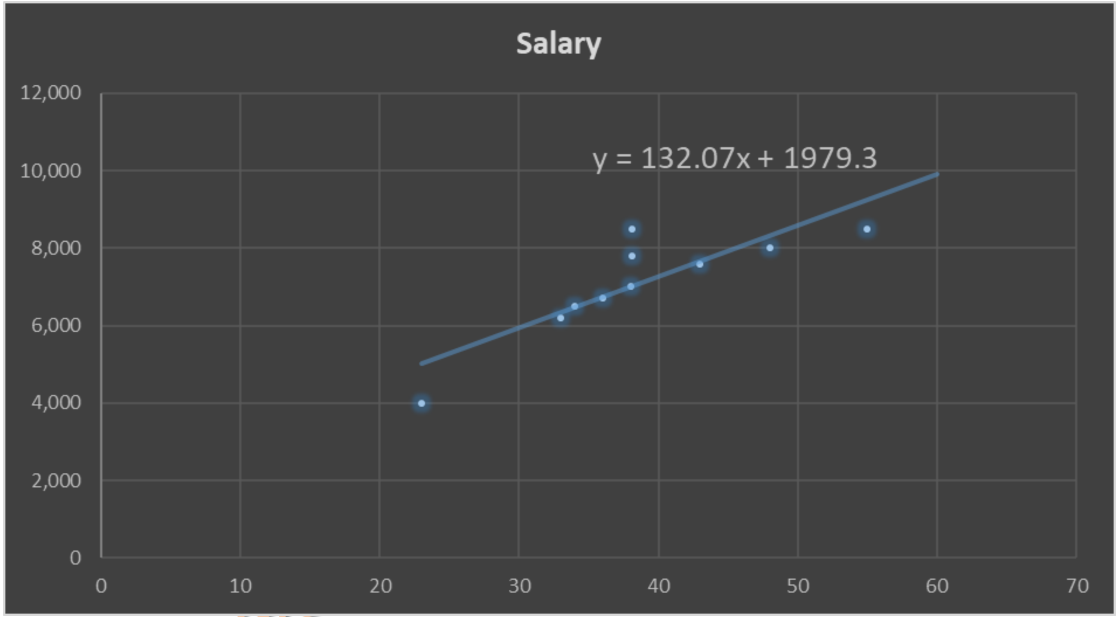
Step 4 – Apply transformation and create prediction model

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	
38	$ 7,000
42	
33	$ 6,200
38.1	$ 7,800
48	$ 8,000
38.1	$ 8,500
43	$ 7,600
55	$ 8,500



Step 5 – Predict Missing values

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	
38	$ 7,000
42	
33	$ 6,200
38.1	$ 7,800
48	$ 8,000
38.1	$ 8,500
43	$ 7,600
55	$ 8,500



For Age = 29

Salary = 132.07 (29) + 1979.3

= $ 5,809.33

Original salary $ 5,500

For Age = 42

Salary = 132.07 (42) + 1979.3

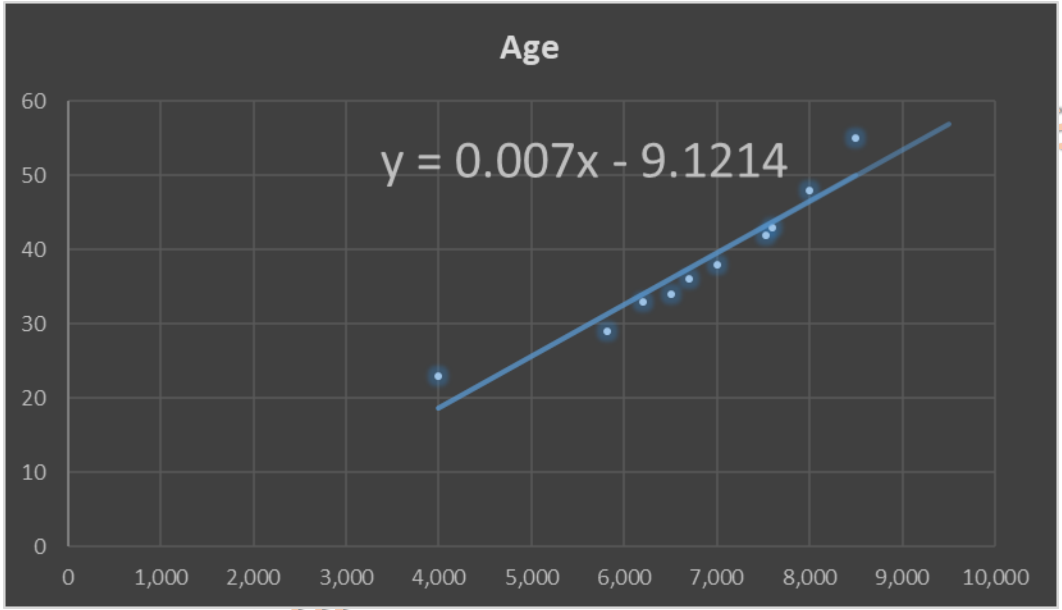
= $ 7,526.24

Original salary $ 7,500

Repeat for Age with new values of Salary

New Prediction Model

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	$ 5,809.33
38	$ 7,000
42	$ 7,526.24
33	$ 6,200
	$ 7,800
48	$ 8,000
	$ 8,500
43	$ 7,600
55	$ 8,500



For Salary = $ 7,800

Age = 0.007(7800) – 9.1214

= 45.48

Original Age 46

For Salary = $ 8,500

Age = 0.007(8500) – 9.1214

= 50.38

Original Age 51

Replace with MICE Result – 2 iterations

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	$ 5,500
38	$ 7,000
42	$ 7,500
33	$ 6,200
46	$ 7,800
48	$ 8,000
51	$ 8,500
43	$ 7,600
55	$ 8,500

Replace with MICE

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	***$ 5,809.33***
38	$ 7,000
42	***$ 7,526.24***
33	$ 6,200
45.48	$ 7,800
48	$ 8,000
50.38	$ 8,500
43	$ 7,600
55	$ 8,500

Replace with Mean

Age	Salary
23	$ 4,000
34	$ 6,500
36	$ 6,700
29	**$ 7,080**
38	$ 7,000
42	**$ 7,080**
33	$ 6,200
38.1	$ 7,800
48	$ 8,000
38.1	$ 8,500
43	$ 7,600
55	$ 8,500

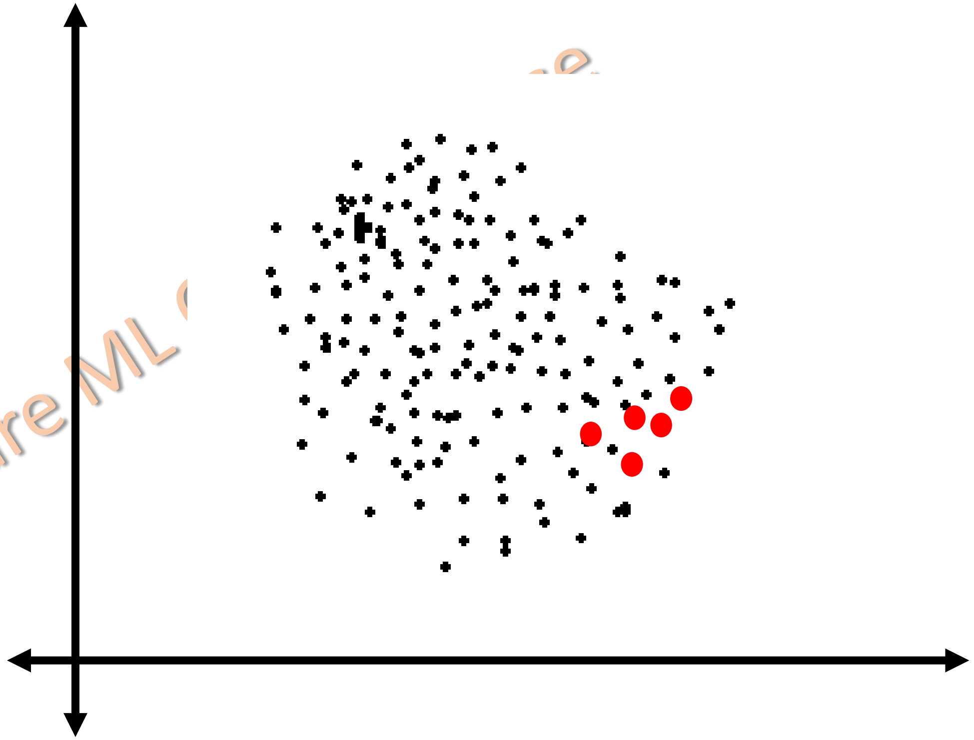
SMOTE

Dealing with Imbalanced Dataset

- Presence of minority class in the dataset

- Challenges related Imbalanced Dataset
  - Biased predictions
  - Misleading accuracy

- Some Examples
  - Credit card frauds
  - Manufacturing defects
  - Rare diseases diagnosis
  - Natural disasters
  - Enrolment to premier institutes



Two Class Classification

No-Fraud → 99.5%
Fraud → 0.5%

Re-Sample the Dataset

- Balance the classes by Increasing minority or decreasing majority

- Random Under-Sampling
  - Randomly remove majority class observations
  - Helps balance the dataset
  - Discarded observations could have important information
  - May lead to bias

Total Observations = 1,000
Fraudulent = 10 or 1%
Normal = 990 or 99%

Reduce normal to 90
Fraudulent = 10 or 10%

- Random Over-Sampling
  - Randomly add more minority observations by replication
  - No information loss
  - Prone to overfitting due to copying same information

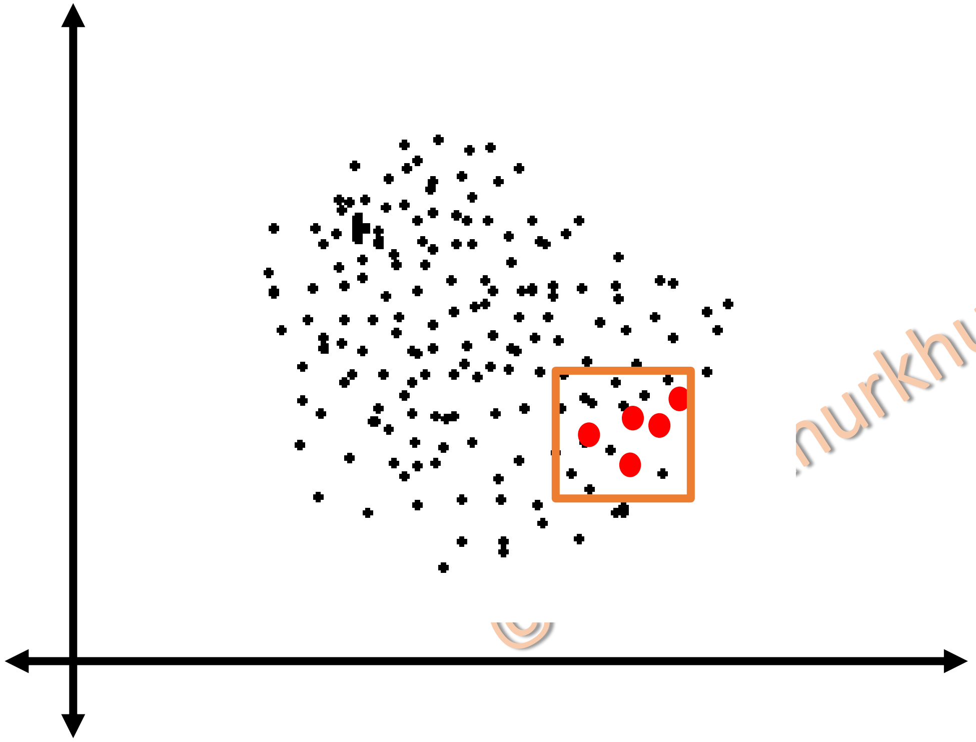
Total Observations = 1,000
Fraudulent = 10 or 1%
Normal = 990 or 99%

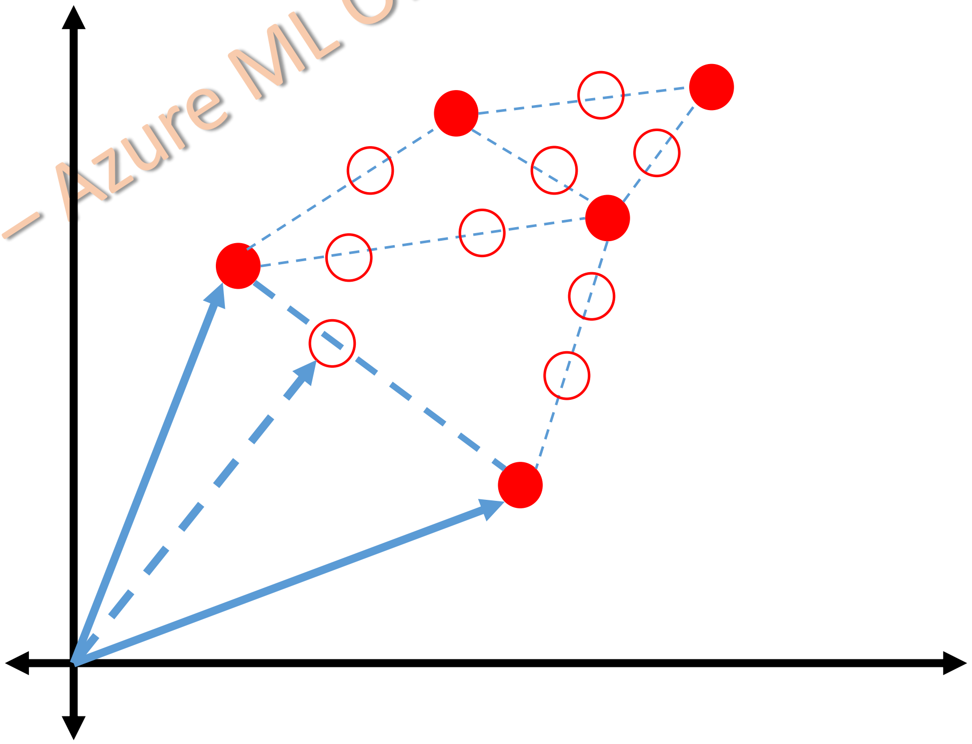
Increase fraudulent by 100
Fraudulent 110 or 10%

SMOTE

- Synthetic Minority Oversampling Technique
- Creates new “Synthetic” observations
- SMOTE Process
  - Identify the feature vector and its nearest neighbour
  - Take the difference between the two
  - Multiply the difference with a random number between 0 and 1
  - Identify a new point on the line segment by adding the random number to feature vector
  - Repeat the process for identified feature vectors

SMOTE





Join Data

What is Join Data?

- Information is provided in two or more datasets
  - Different sources
  - Created at different times

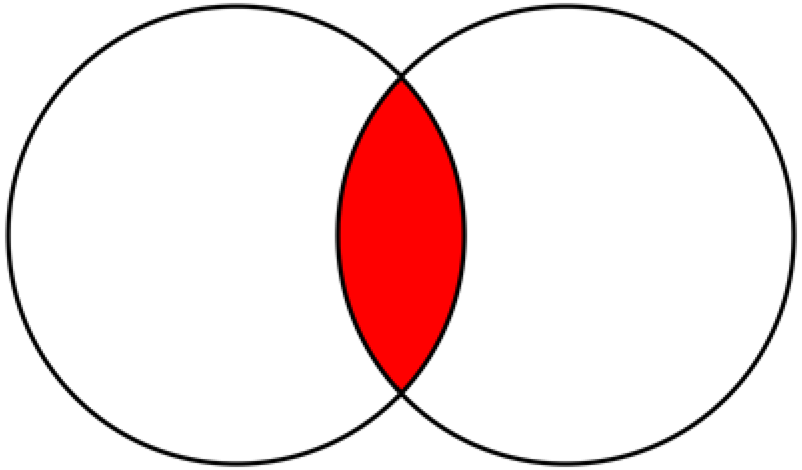
- Datasets are related by key columns

- Different types of Join supported by AzureML
  - Inner Join
  - Left Outer Join
  - Full Outer Join
  - Left Semi-join

Inner Join

EmpID	Salary
EMP001	$ 5,000
EMP002	$ 5,500
EMP003	$ 5,200
EMP004	$ 6,000
EMP007	$ 5,800
EMP008	$ 6,700

EmpID	Department
EMP001	IT
EMP003	IT
EMP004	Marketing
EMP007	Finance
EMP009	Marketing
EMP010	Finance

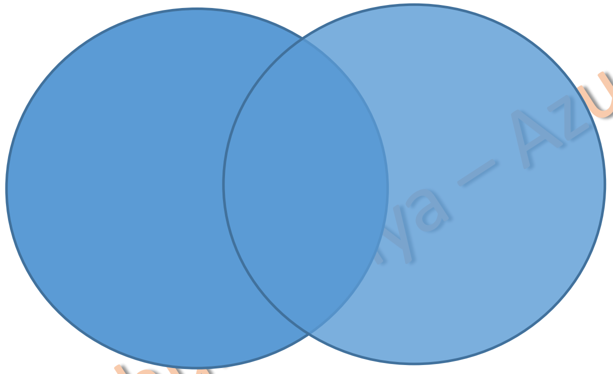


EmpID	Salary	Department
EMP001	$ 5,000	IT
EMP003	$ 5,200	IT
EMP004	$ 6,000	Marketing
EMP007	$ 5,800	Finance

Full Outer Join

EmpID	Salary
EMP001	$ 5,000
EMP002	$ 5,500
EMP003	$ 5,200
EMP004	$ 6,000
EMP007	$ 5,800
EMP008	$ 6,700

EmpID	Department
EMP001	IT
EMP003	IT
EMP004	Marketing
EMP007	Finance
EMP009	Marketing
EMP010	Finance

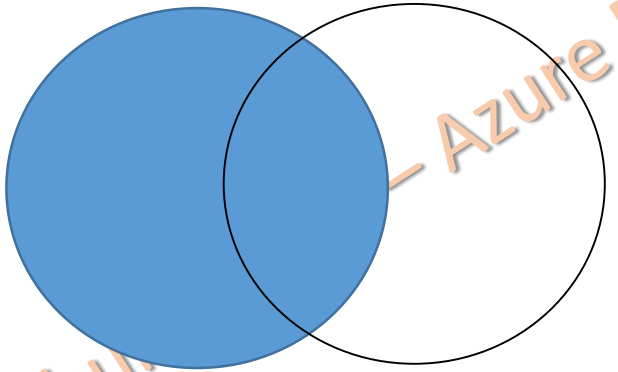


EmpID	Salary	Department
EMP001	$ 5,000	IT
EMP002	$ 5,500	
EMP003	$ 5,200	IT
EMP004	$ 6,000	Marketing
EMP007	$ 5,800	Finance
EMP008	$ 6,700	
EMP009		Marketing
EMP010		Finance

Left Outer Join

EmpID	Salary
EMP001	$ 5,000
EMP002	$ 5,500
EMP003	$ 5,200
EMP004	$ 6,000
EMP007	$ 5,800
EMP008	$ 6,700

EmpID	Department
EMP001	IT
EMP003	IT
EMP004	Marketing
EMP007	Finance
EMP009	Marketing
EMP010	Finance

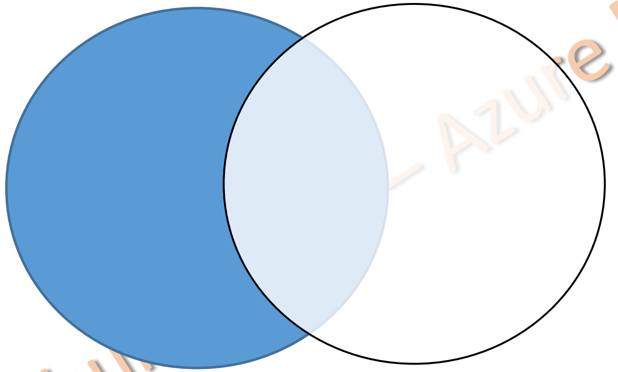


EmpID	Salary	Department
EMP001	$ 5,000	IT
EMP002	$ 5,500	
EMP003	$ 5,200	IT
EMP004	$ 6,000	Marketing
EMP007	$ 5,800	Finance
EMP008	$ 6,700	

Left Semi Join

EmpID	Salary
EMP001	$ 5,000
EMP002	$ 5,500
EMP003	$ 5,200
EMP004	$ 6,000
EMP007	$ 5,800
EMP008	$ 6,700

EmpID	Department
EMP001	IT
EMP003	IT
EMP004	Marketing
EMP007	Finance
EMP009	Marketing
EMP010	Finance



EmpID	Salary
EMP001	$ 5,000
EMP003	$ 5,200
EMP004	$ 6,000
EMP007	$ 5,800

Thank You..!