Inspire…Educate…Transform.

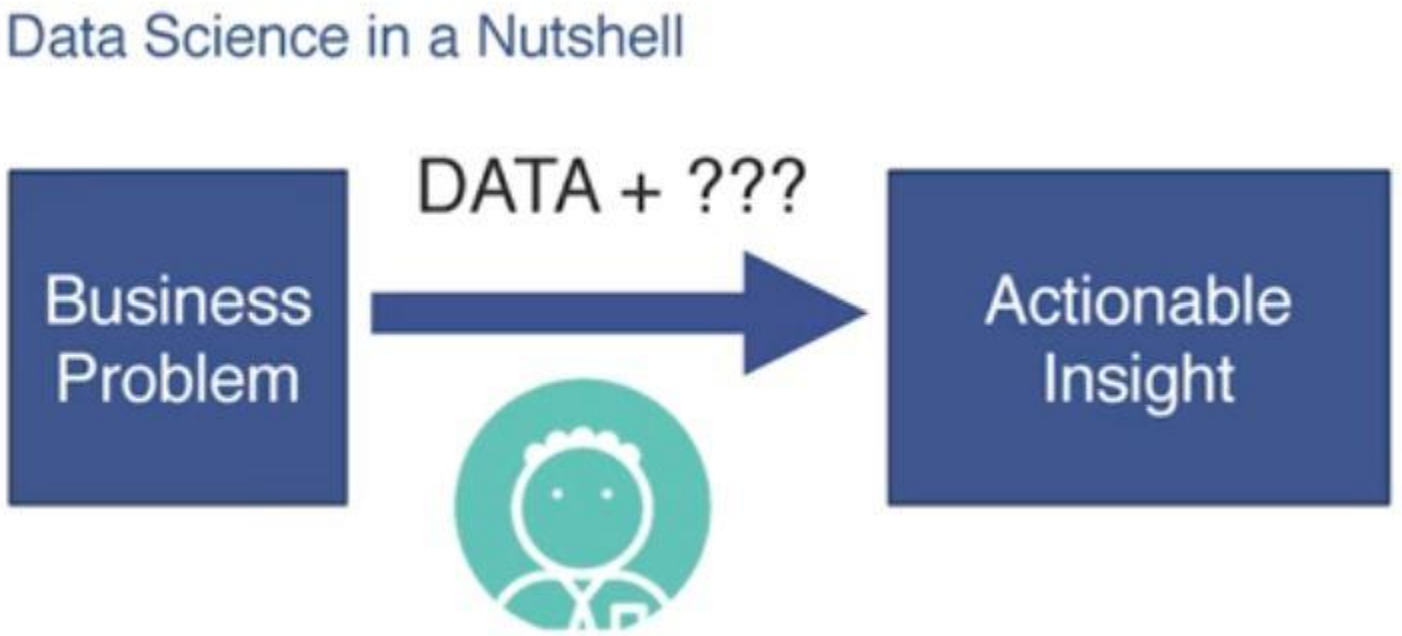
Introduction to Data processing in R and Python

Jagan Basa
Lead Data Scientist, INSOFE

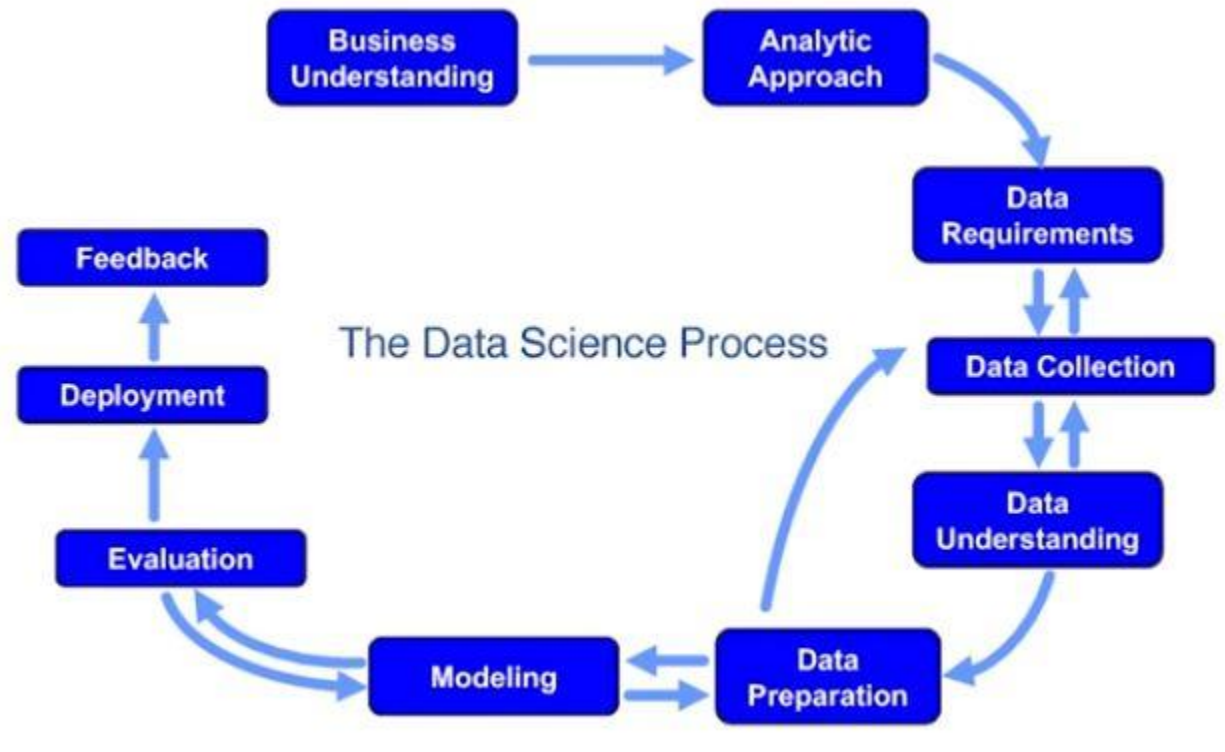
We will learn

- What is Data science & pre-processing
- Data science Process
- Skills required and mapping the skills
- Key steps in the Data processing

What is Data Science



The Data Science Process



4

The Data Science Skills

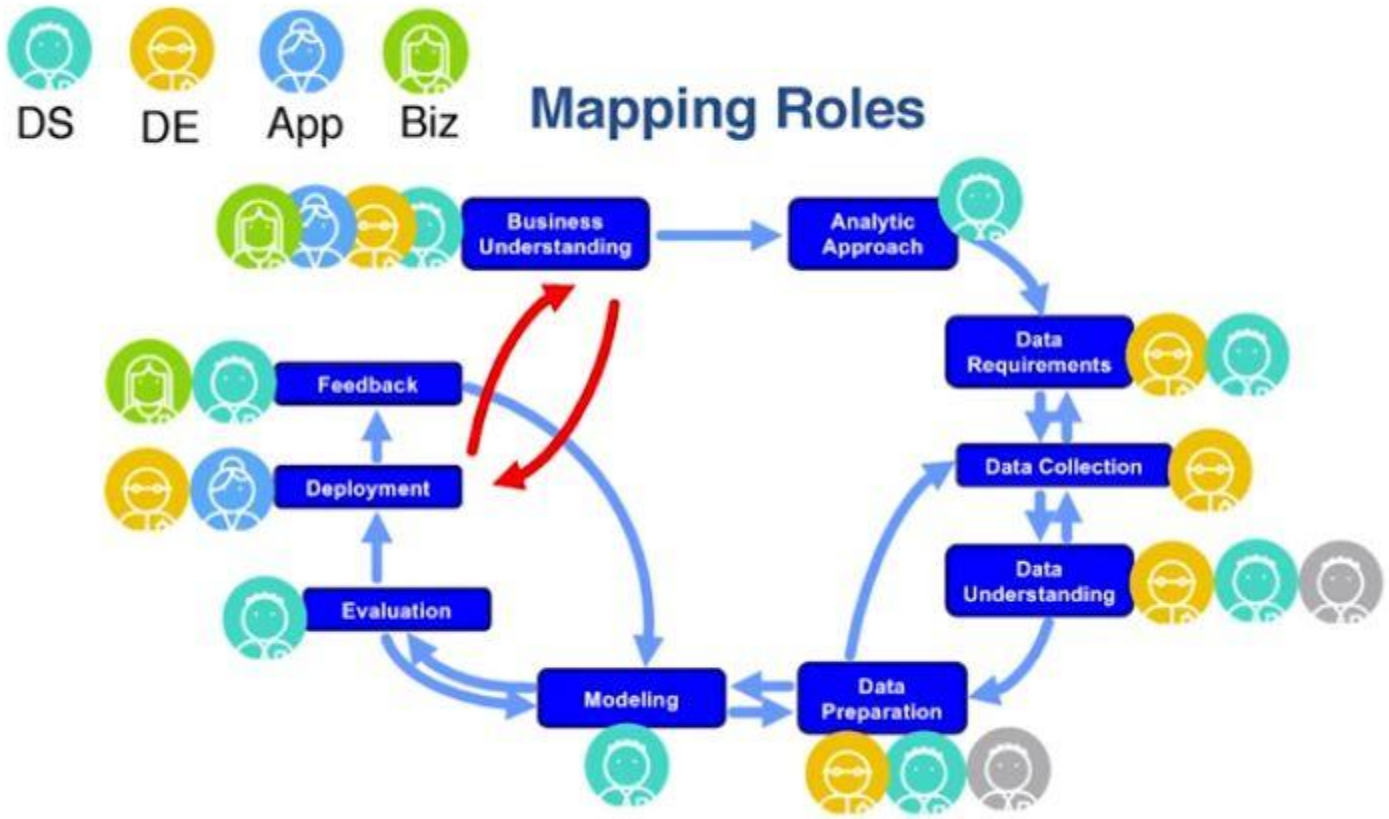
The "Perfect" Hire

Can't hire one person for each role?

- Subject matter expertise & industry experience
- Caught up with the latest tools & technologies
- Experience with big data & big data infrastructure
- Experience with machine learning and building pipelines
- Experience with app development
- Also happy to do simple ad-hoc reports

The Data Science Skill Map

The Data Science Skill Map



Why pre-process

- Poor model on good data is likely to be better than great model on poor data

Raw data

- Normally it is available in multiple tables
  - Merge them
  - Fill missing values

More on attributes

- Type
  - Numeric, Categorical and Ordinal
- Actionable
  - Focus, changeable

Process

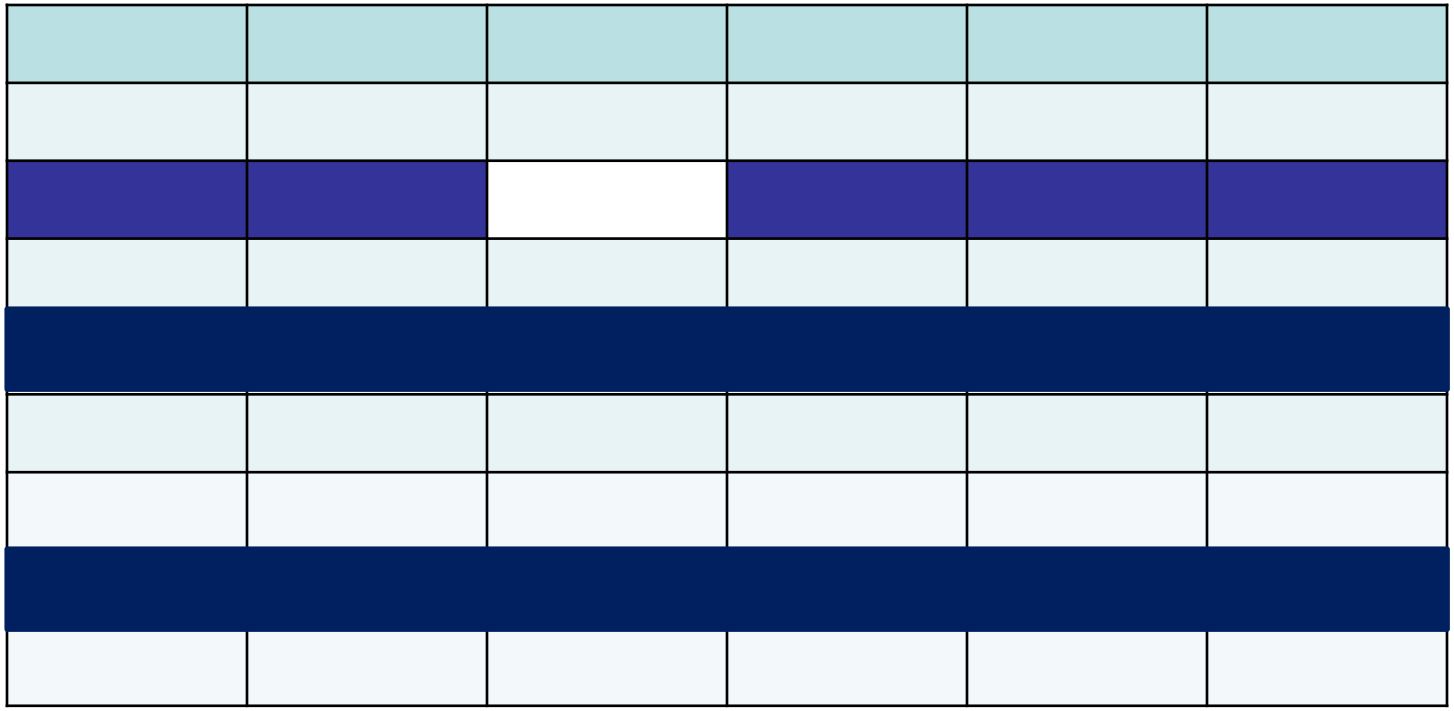
- Take a subset from a table if needed
- Typeset the attributes correctly
- Do descriptive statistics and understand the data

Missing values

- Ignore
- Fill with a central statistic

KNN Imputation

- Find nearest neighbors based on existing attributes



Euclidian for numeric and 1 or zero for categorical

- Take an average of only the nearest neighbors
  - Mean for numeric
  - Mode for categorical

Data standardization

- Let us say, we are measuring distance between records for some purpose

Employee	Age	Income
1	24	50000
2	25	55000
3	60	51000

This dominates completely

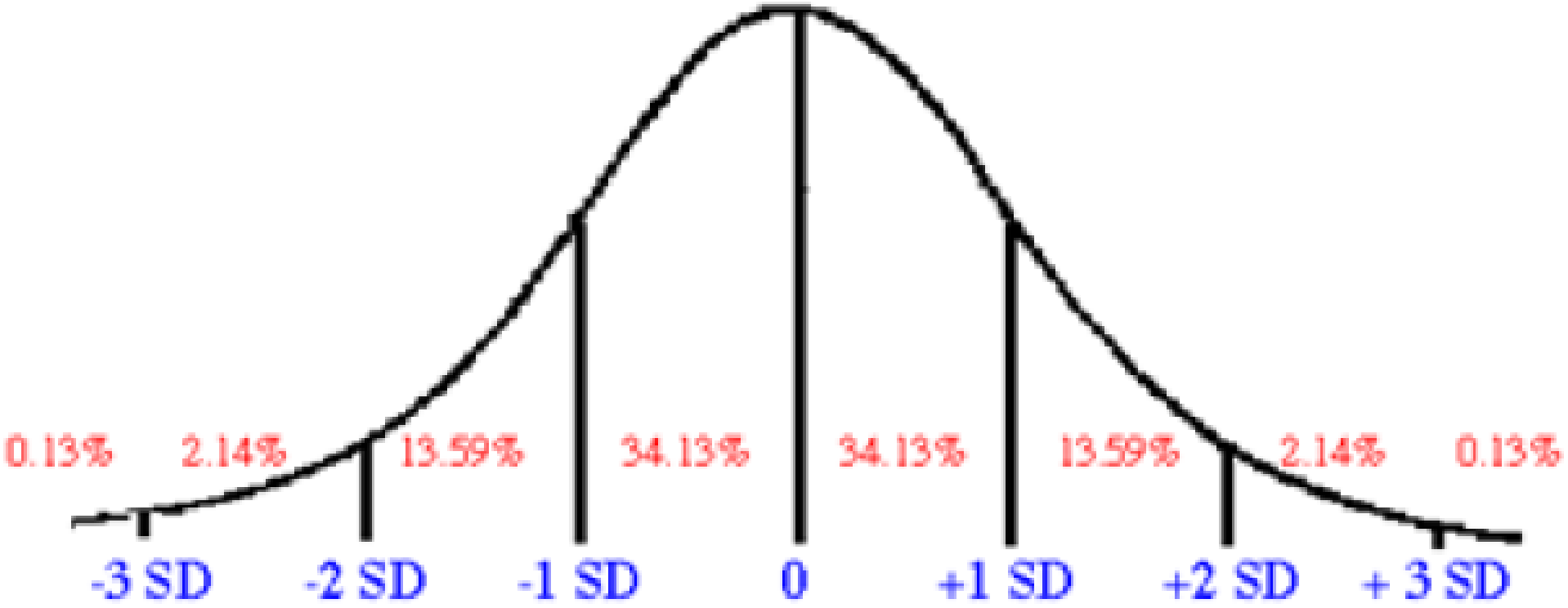
Bring to same range

$$Value_{new} = \frac{Value - minValue}{maxValue - minValue}$$ Range is 0 to 1

$$Min\ \max\ for\ 25: \frac{25 - 24}{60 - 24}$$

- Min-Max is extremely sensitive to the outliers.

- Min-max of : (1, 2, 1001) is (0, 0.001, 1)


0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
-3 SD
-2 SD
-1 SD
0
+1 SD
+2 SD
+ 3 SD

$$newValue = \frac{Value - mean}{standard\ deviation}$$

Changing the type

- Neural networks
  - All attributes must be numeric
- Naive Bayes
  - All attributes must be categorical

Numeric to categorical

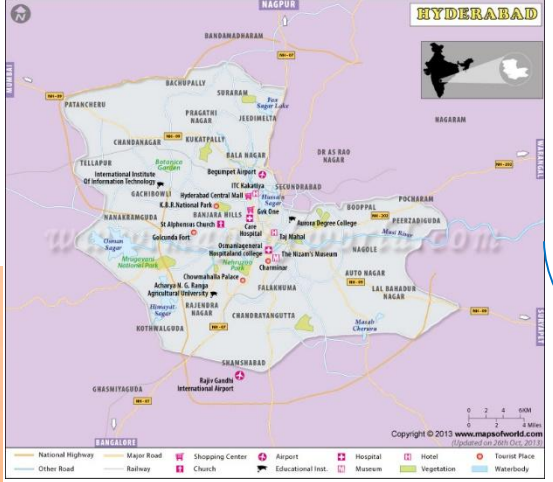
- Manual
- Equal frequency
  - Number of samples in each bin
- Equal width
  - Interval is same (good for uniform distributions)

Ordered and categorical

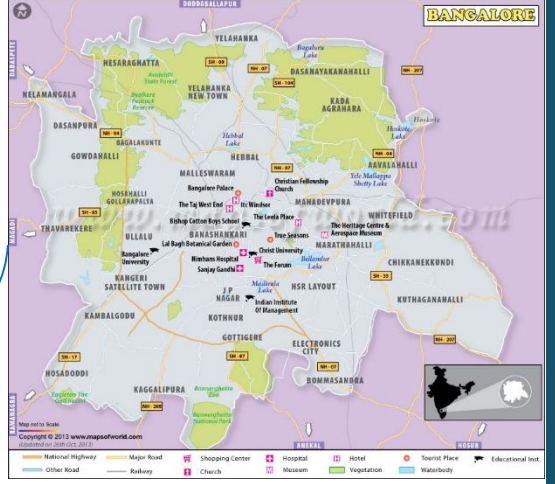
- Merging multiple bins
  - Verify the frequencies
  - Convert them to numeric and recode

Categorical to numeric

- How do we set up categorical variables in distance metrics
  - Create as many dummy variables as there are options
  - Code as 100, 010,…



Inspire...Educate...Transform.



HYDERABAD

Office and Classrooms

Plot 63/A, Floors 1&2, Road # 13, Film Nagar,
Jubilee Hills, Hyderabad - 500 033
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Office

Incubex, #728, Grace Platina, 4th Floor, CMH Road,
Indira Nagar, 1st Stage, Bengaluru – 560038
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Classroom

KnowledgeHut Solutions Pvt. Ltd., Reliable Plaza,
Jakkasandra Main Road, Teacher's Colony, 14th Main
Road, Sector – 5, HSR Layout, Bengaluru - 560102

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.