Inspire…Educate…Transform.

Structured Data Processing

R Commands Practised

- Read and Write Data
- Difference between Vector, Matrix and Data Frame
- Sub-setting data
- Merge two data files
- In-built function
  - seq, length, print, sample, mean, sd, is.na, na.rm, str, matrix, colnames, data.frame, which.max, ifelse
- For Loop, Apply
- User defined functions
- Visualizations
  - hist, plot, boxplot

- DPLYR
  - select, filter, arrange, mutate, summarize, group_by

High-level Steps in a Project

- Understanding the problem
- Data Preparation
  - Data Exploration
  - Data Pre-processing
- Model Implementation
- Results Interpretation
- Delivering/Deploying the solution

Currently Focusing

- Data Preparation
  - Data Exploration
  - Data Pre-processing

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

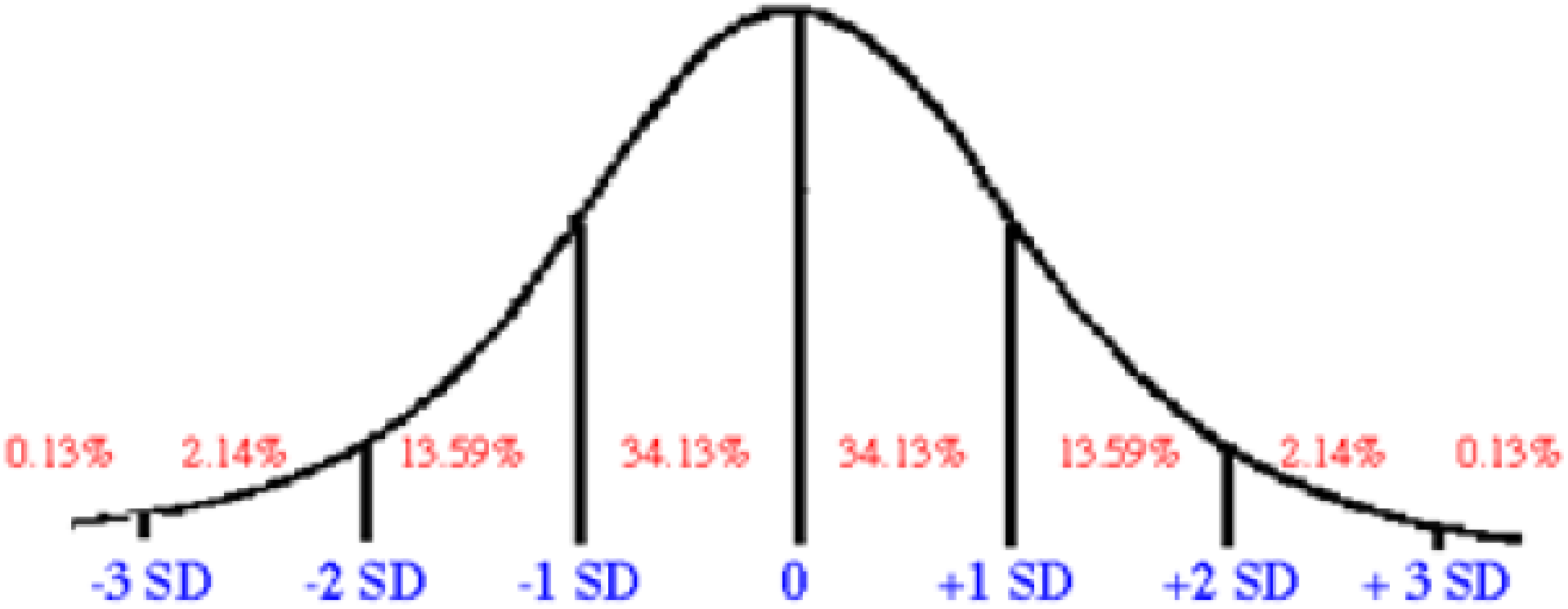
Data standardization

- 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

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


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

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)

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

Ordered to numeric

- If all divisions are important
  - Identify a range
  - Split it uniformly
  - 1, 2,3,4 gets changed to 0, 0.25, 0.5, 1

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,…