3/2 start

01.07.19

Data Science

Data Science Process -

Ask an interesting question
↓
Get the data
↓
Explore the data
↓
Model ~~the~~ ~~to~~ data
↓
Communicate and visualize the result

Book: An Introduction to Statistical Learning

by
Gareth James
Daniella Witter
Trevor Hastie
Robert Tibshirni

7th or
8th
print

19.08.19

Data Science

The Process of Data Science -

Ask an interesting question

Get the data

Explore the data

Model the data

communicate and visualize the results.

Modules

What is in this course ?

1. Data Collection - data wrangling / munging, cleaning and sampling to get a suitable dataset.

2. Data Management - accessing data quickly and reliably.

3. Exploratory Data Analysis (EDA), generating Hypothesis and buiding intuition.

4. Prediction or statistical learning

5. Communication

Ask a question :

Original Question :

" What does the data tell us about the ride share program ? "

who → ☒ More male user or female user ?

☒ More registered user or onetime user ?

☒ Older or younger people ??

Where

☐→ More in city center or in remote area

→ More in commercial areas of residentia areas

→ More in tourist areas !

When → ☐ More during weekends or week day?

☐ More during rush hour?

☐ More in summer?

Get the data

A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data are multiple of such measurements.

Sources of data -

Internal ~~Sou~~ Source -

Existing external source -

External sources requiring collection effort

09.09.19

Data Science

Data Science Process.

→ Asking Question.

→ Data collection and preprossing

→ Exploratory data Analysis

→

→

→

Data Sources

Internal Sources

Existing External sources

External Sources Requiring Collection Effort

How to get data generated online

→ API

→ RSS (Rich site summary)

→ web scrapping

Types of data -

→ Numeric
→ Boolean
→ String

→ Date & Time
→ Lists
→ Dictionaries

How is your data represented and stored ?

→ Tobular data - csv, tsp, xlsx
→ Structured data - json, xml
→ semi structured data

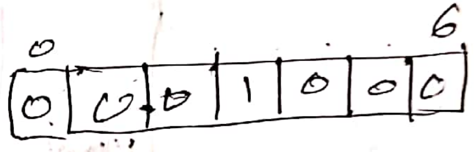
→ Texual Format
→ Temporal format
→ Geolocation data

Types of data -

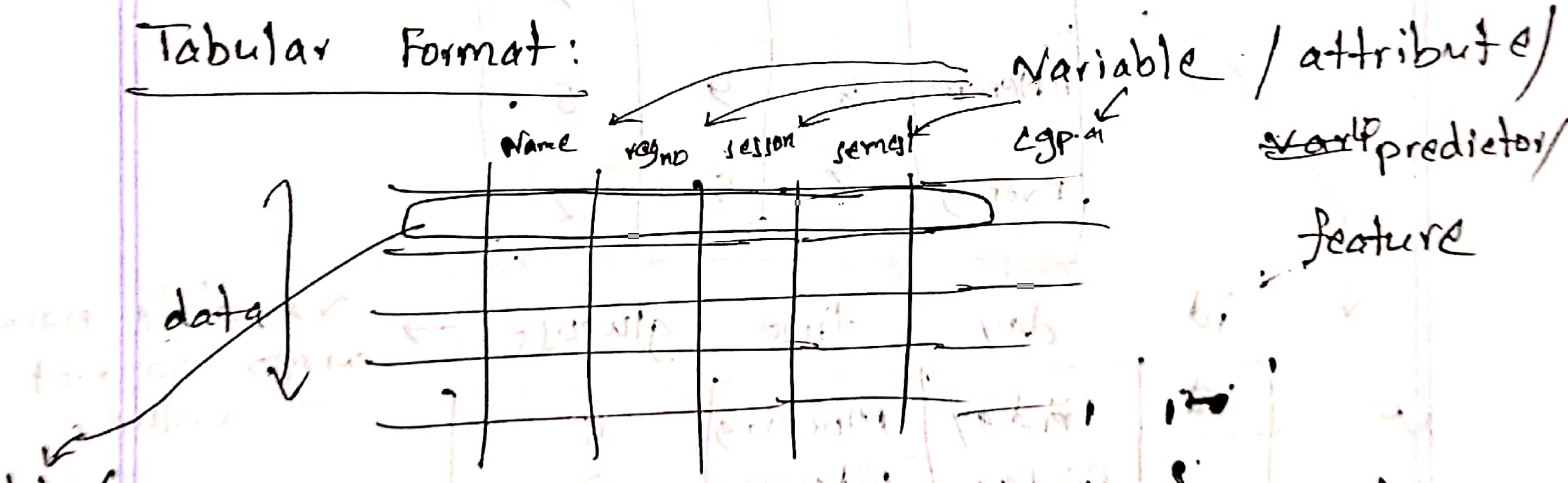
→ Qualitative Variable

- discrete
- continuous

→ Categorical variable.



Tabular Format:



Data Dimension - number of variable

Data Preprocessing :

Common issues with data

→ Missing value

→ Wrong value

→ Messy format

→ Not usable

Data Preprocessing :

✗

	Friday	Saturday	Sunday
Morning	5	4.5	7
Afternoon	8	9	8
Evening	7	5	7

id day time glucose → Variable name থাকতে হবে, not value

✓

id	day	time	glucose
1	friday	Morning	5
2	friday	Afternoon	8
3	friday	Evening	7
4	Saturday	M	4.5
5	"	A	9
6	"	E	5
7	sunday	M	7
8	"	A	8
9	"	E	7

Common Causes of Messiness -

→ Column headers are values, not variable names.

→ Variables are stored in both rows and columns

→ Multiple variables are stored in one column

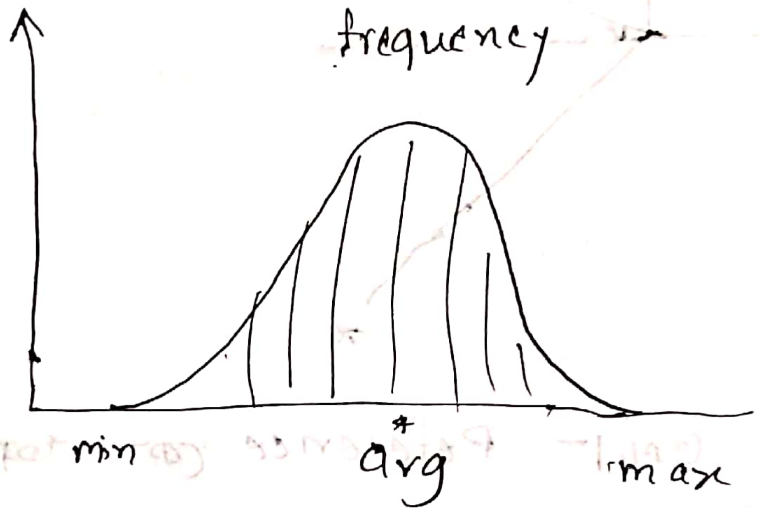
→ Multiple types of experimental units are stored in same table.

12.09.19

Data Science

Exploratory Data Analysis (EDA)

Normal Distribution:



Distribution of data:

is a listing or function showing all the possible values (or intervals) of the data and how often they occur.

Normal Distribution: 2 parameters: (mean, σ^2)

A Population is the entire set of objects or events under study

A sample is a representative subset of the objects or events under study. It is needed because it is impossible to obtain/or intractable to compute with ~~entit~~ entire population data.

Biases in Sample :-

→ Selection Bias - some objects or ~~recoe~~ records. are more likely to be selected.

→ Volunteer / Non - responsive bias - some subjects may not be easily available or represented.

Sample mean (average) of ~~a~~ n observations of a ~~xi~~ variable is defined as-

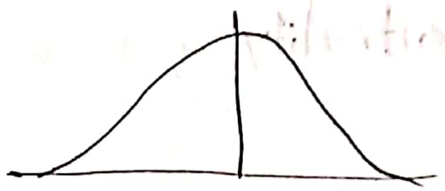
$$\bar{x} = \frac{x_1 + x_2 + \cdots\cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

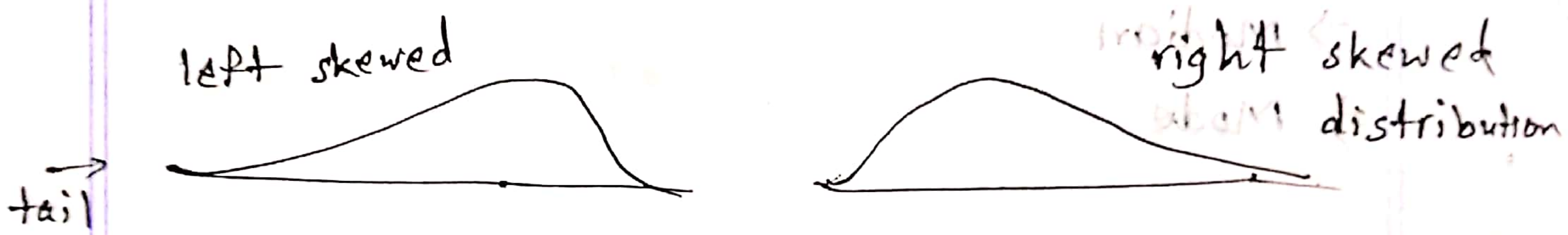
- The mean describes what a typical sample looks like
- It describes the center of the distibution

Sample <u>Median</u> of a set of n observations in a sample, ordered by value is defined by -

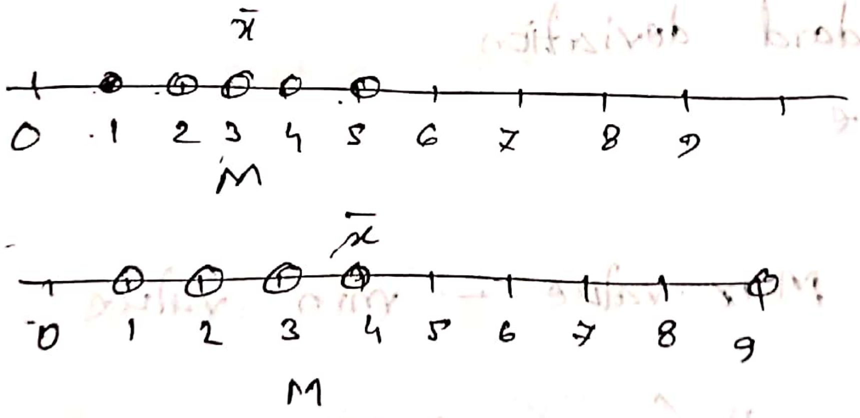
$$\text{Median} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{(n+1)/2}}{2}, & \text{if } n \text{ is even} \end{cases}$$

17 19 21 $\dfrac{22 + 23}{2}$ 23 38 38





Mean vs median -



* Mean is more sensitive to outlies

* Mean → O(n)
Median → O (n logn)

For categorical data -
we calculate Mode of the sample.
→ Most frequently occured data/ objects/ events

σ^2 σ^2 σ^2 σ

Measures of centrality :

- ⇒ Mean
- ⇒ Median
- ⇒ Mode

Measures of spread -

- ⇒ Variance
- ⇒ Standard deviation
- ⇒ range

Range : Max value - min value

Variance : ~~$\frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2 = \sigma^2$~~

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

population variance

estimation of σ^2 is s^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2$$

The sample variance s^2 ~~is~~ the measures how much on average the sample values deviate from the mean.

$$\text{standard deviation} = \sqrt{s^2} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2}$$

Networking:

Subnet mask:

~~sub~~

255.255.254.0

172.16.28.0/22

Subnet 0 0000000/23

Visualizations help us to analyze and explore data.

~~It~~ They help to -

⇒ Identify hidden patterns and trends

⇒ Formulate / test hypothesis

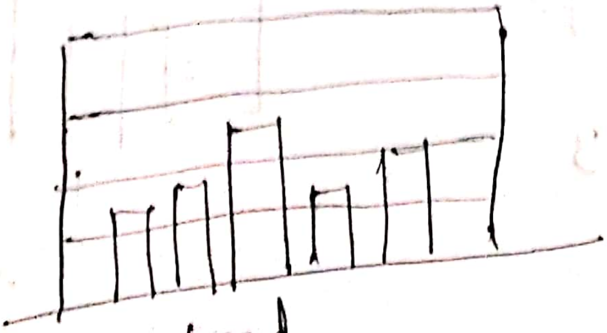
⇒ Communicate any modelling results

- Present information and ideas
- Provide evidense and support
- Influence and persuade

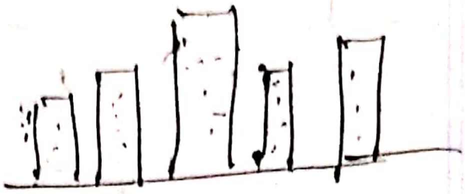
⇒ Determine next step in analysis/modeling

Principles / Good practices of Visualization :

⇒ Maximize data to ink ration



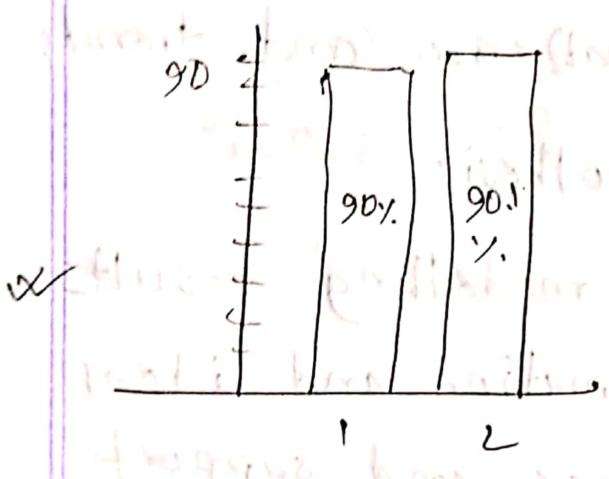
Good

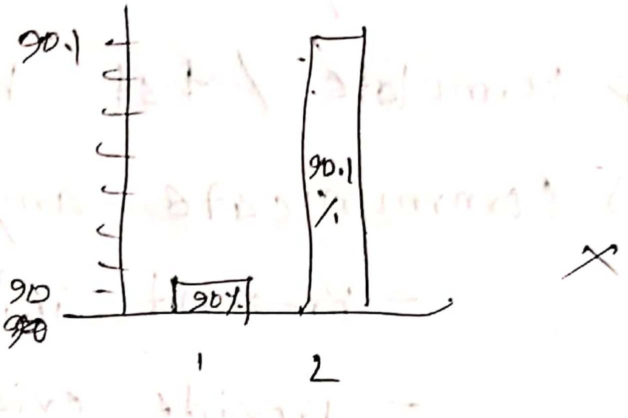


Better

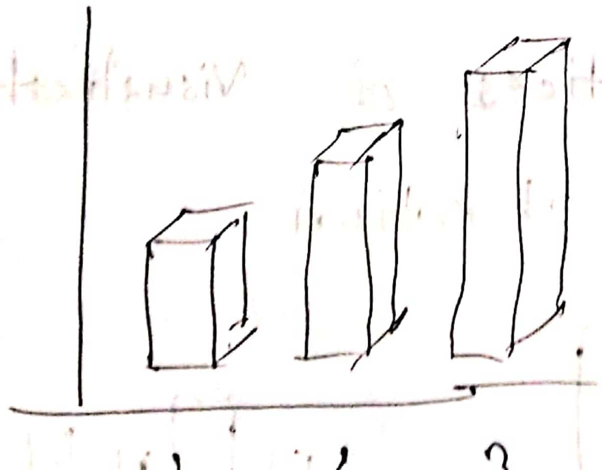
⇒ Don't lie with scale -

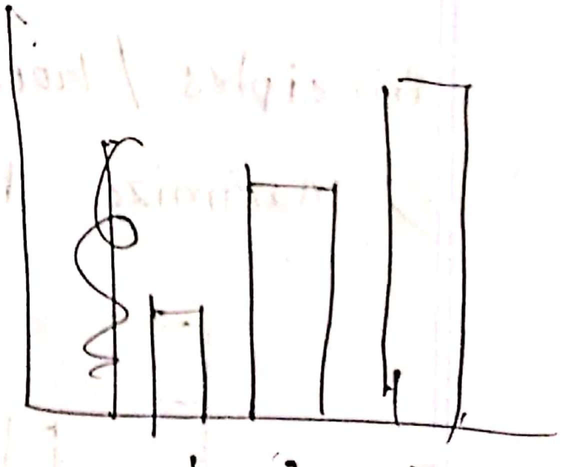
minimize $\dfrac{\text{size of effect in graph}}{\text{size of effect in data}}$ | lie factor





⇒ Minimize chart junk - show data variation, not design variation





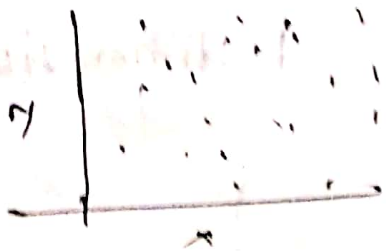
Types of visualization -

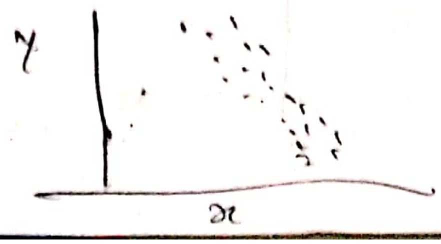
□ Distribution : How the variables In the data set distribute over a range of possible values

□ Relationship : How the values of multiple variables in the dataset relate to each other.

□ Composition : How the data set breaks down into subgroups.

□ Comparison: How trends in multiple variable of data set compare



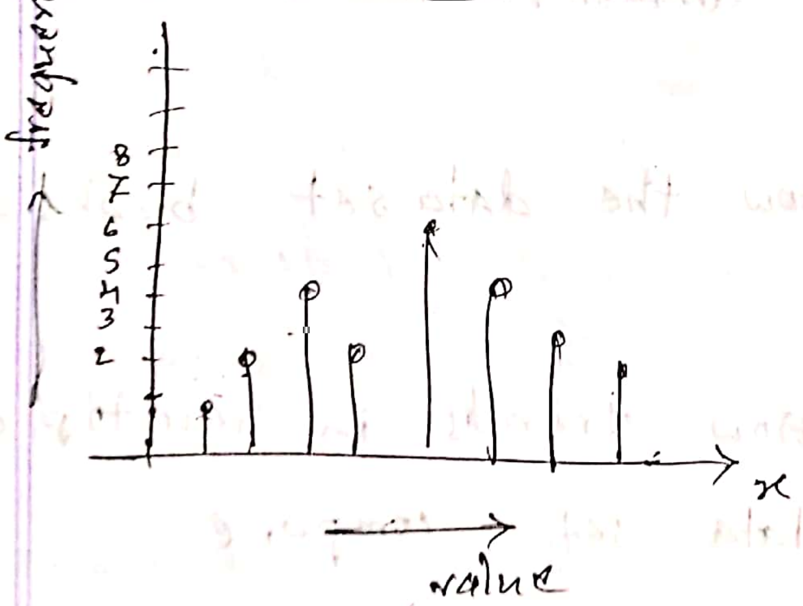


Common used plots / diagrams.

- ⇒ histogram
- ⇒ Bar diagram
- ⇒ Pie chart
- ⇒ Scatter plot

- ⇒ stacked area graph
- ⇒ multiple histogram
- ⇒ box plot.

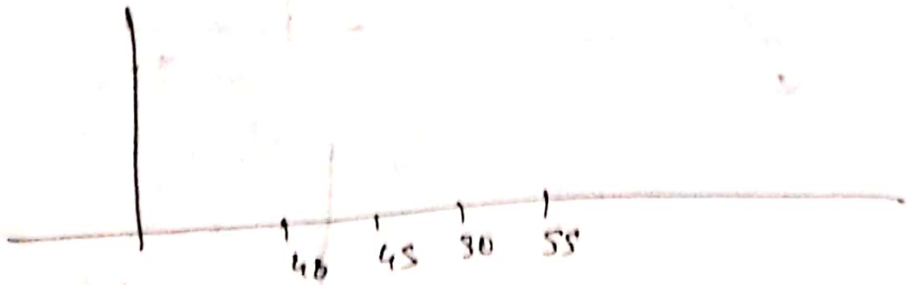
histogram.



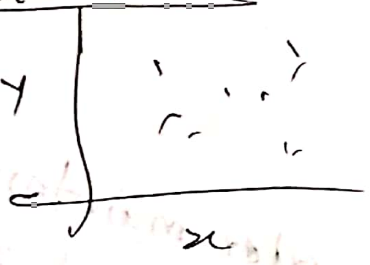
1 dimensional data

bin.

- কত রেঞ্জের- value গুলোকে- একটা হিসেবে count করছি।

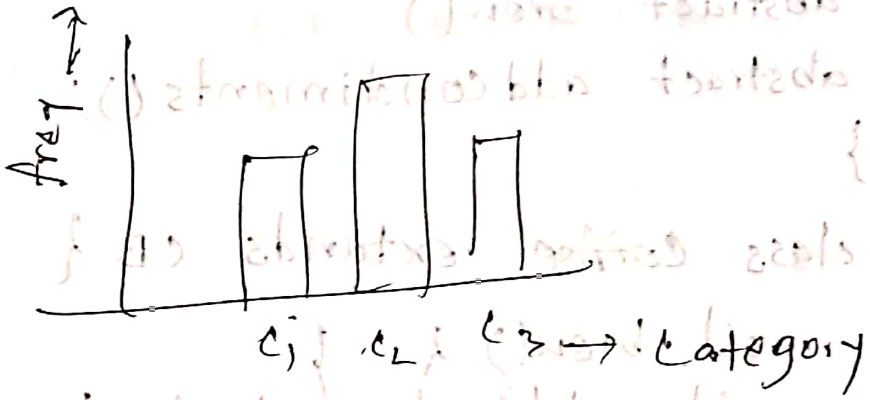


Scatter plot:

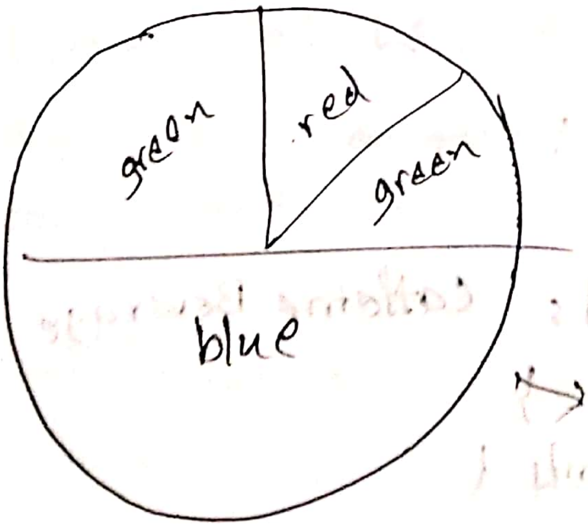


For categorical value:

Bar diagram



Pie chart:



black	:	12%
red	:	13%
green	:	25%
blue	:	50%

→ total : 100%