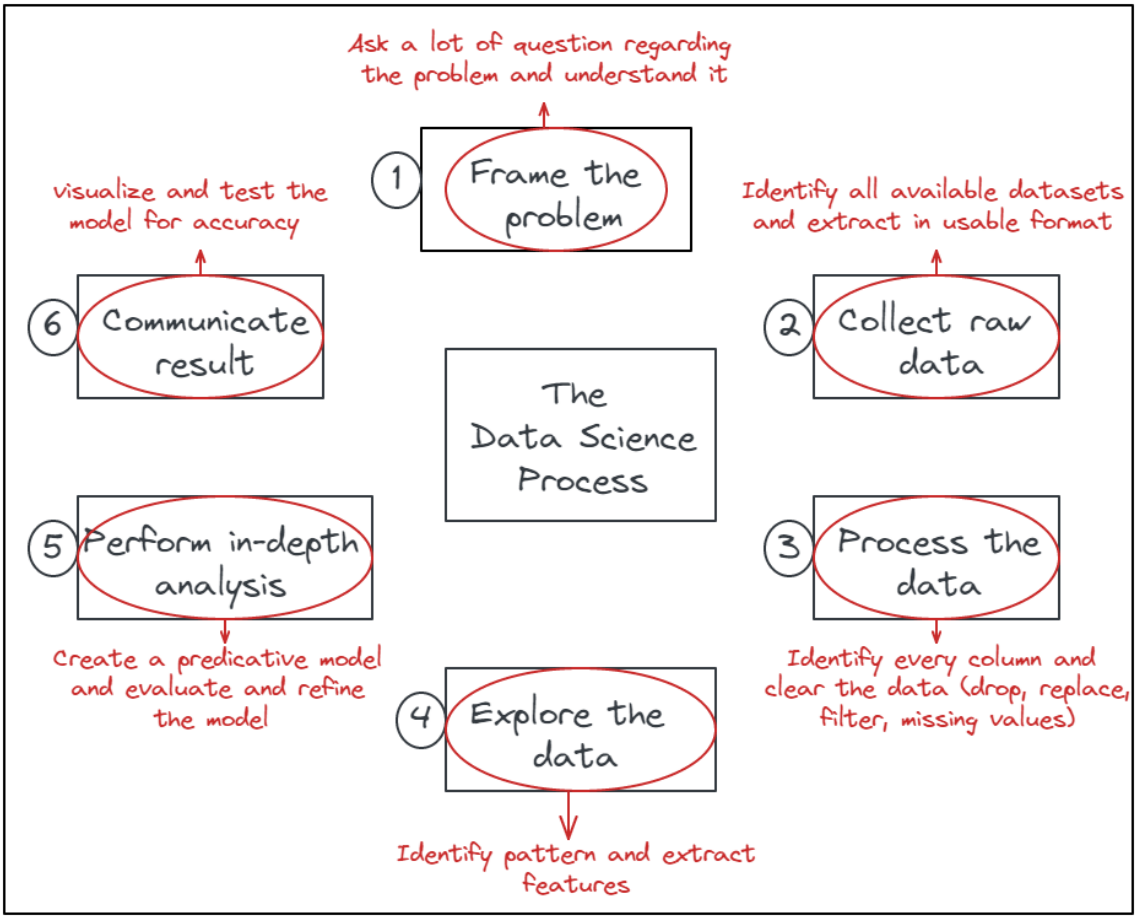
Data Science:

Lecture 1: Introduction

It is blend of various tools, algorithm and machine learning principles with the goal to discover/find hidden patterns from raw data(dataset).



Lecture 2: Types of Data and Levels of Measurement

Lecture 3: Datasets

- **Google Dataset Search:** Dataset is available in different formats and visualization of this dataset is also available.
- **Kaggle:** Dataset and code implementation are both available.
- **Data.gov:** US based datasets (Free and no registration required)
- **Datahub.io:** Financial datasets
- **UCI Machine Learning Repository:** Labeled datasets

Lecture 4: Data Acquisition

Scraping:

To analyze data, we typically need a structure. E.g. in the form of a table (rows and columns).

Found data is often in human readable structure but the idea is to automate this the collect data.

Ways to collect digital data:

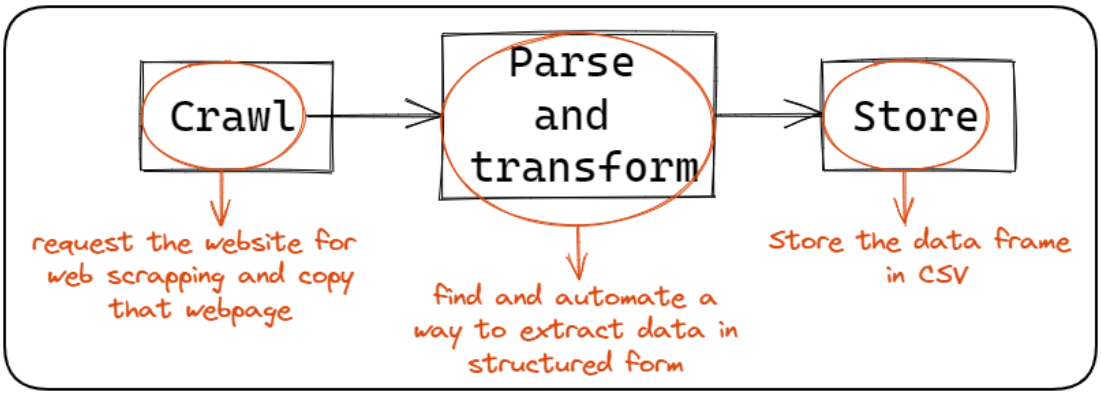
- Languages: R, Python, Ruby, Perl, Java etc.
- Software and APIs

Paper:

The following are the steps:

1. Create digital images
2. Identify colored pixels as characters (OCR)
3. Select the process/software/language
    - Adobe Pro., etc.
    - Best in class commercial: Abby FineReader
    - Now has an API
    - Best in class open-source: Tesseract
    - Python library: pyPdf2 etc.
4. Post-processing

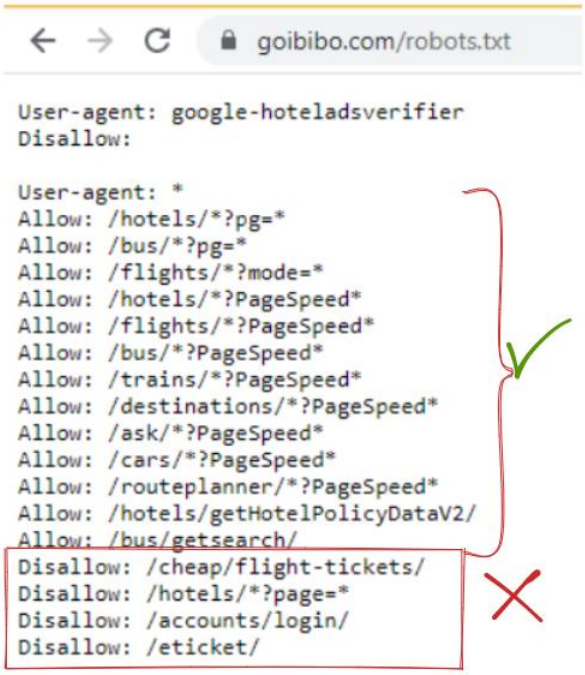
Web Scraping:



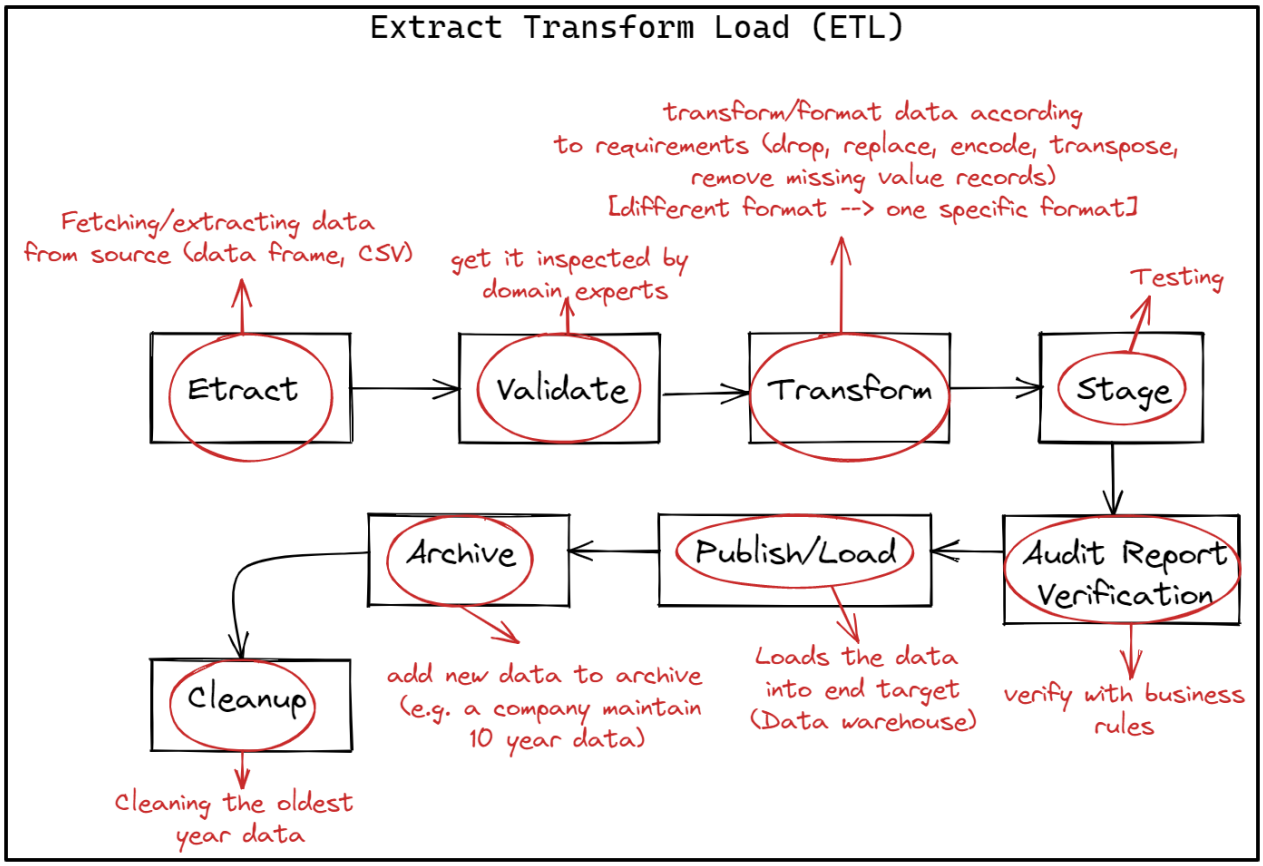
Web scraping Ethics:

Time: Use web scraping at the time when there is less traffic or non-working hours.

robots.txt: Checking which pages are allowed and disallow of the website that you are trying to scrap.

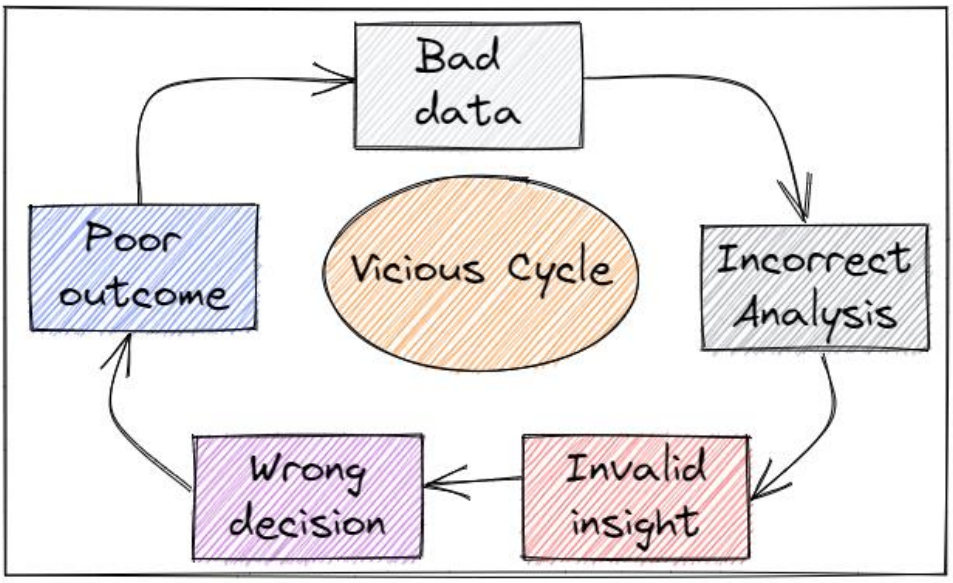


Lecture 7: Extract Transform Load (ETL)



Lecture 8: Data Wrangling

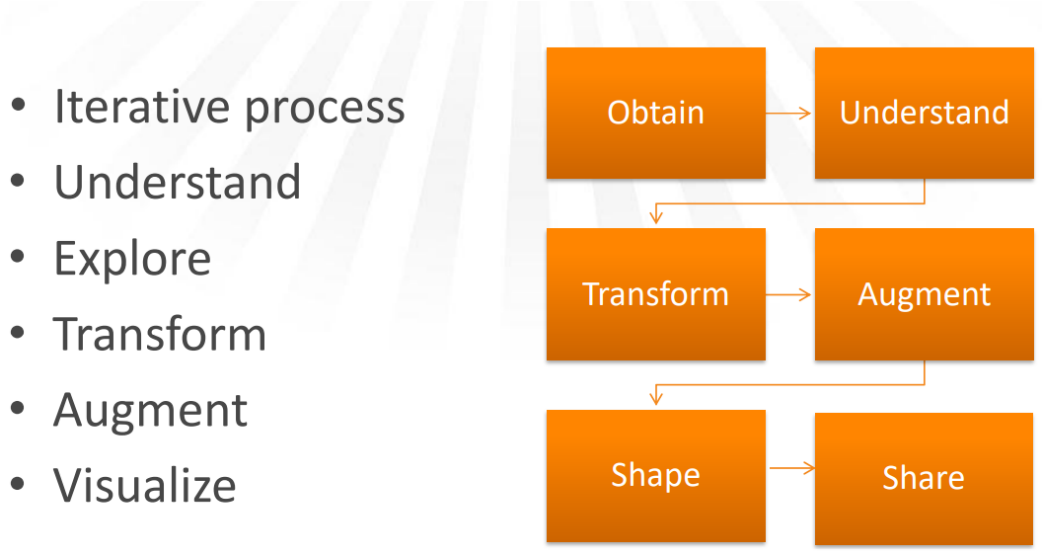
Vicious Cycle:



Data wrangling:

Also known so:

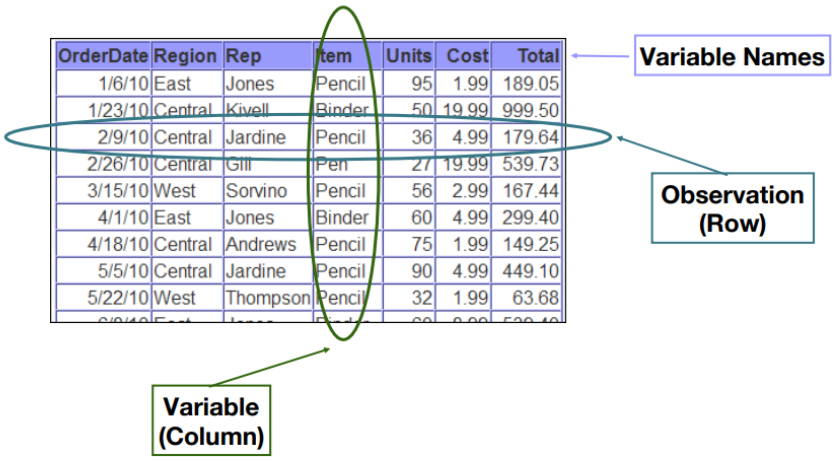
- Data Pre-processing
- Data Preparation
- Data Cleansing
- Data Scrubbing
- Data Munging
- Data Transformation
- Data Fold, Spindle, Mutilate



- Data Ingestion
  - CSV
  - PDF
  - API/JSON
  - HTML Web Scraping
- Data Exploration
  - Visual inspection
  - Graphing
- Data Shaping
  - Tidying Data
- Data Cleansing
  - Missing values
  - Format
  - Outliers
  - Fat Fingered Data
- Data Augmenting
  - Aggregate data sources

Lecture 10: Data Organization

Data is stored in the form of a *Data Matrix*

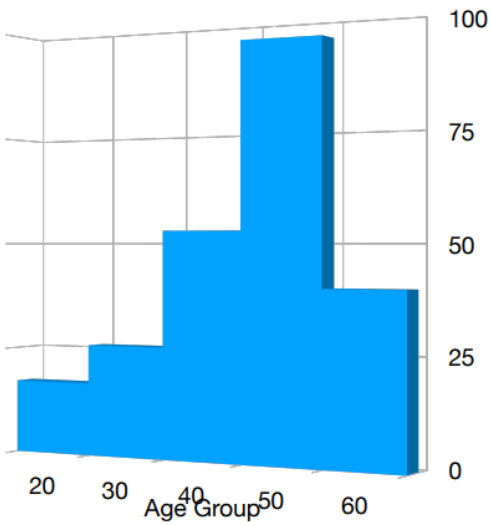


OrderDate	Region	Rep	Item	Units	Cost	Total
1/6/10	East	Jones	Pencil	95	1.99	189.05
1/23/10	Central	Kivell	Binder	50	19.99	999.50
2/9/10	Central	Jardine	Pencil	36	4.99	179.64
2/26/10	Central	Gill	Pen	27	19.99	539.73
3/15/10	West	Sorvino	Pencil	56	2.99	167.44
4/1/10	East	Jones	Binder	60	4.99	299.40
4/18/10	Central	Andrews	Pencil	75	1.99	149.25
5/5/10	Central	Jardine	Pencil	90	4.99	449.10
5/22/10	West	Thompson	Pencil	32	1.99	63.68

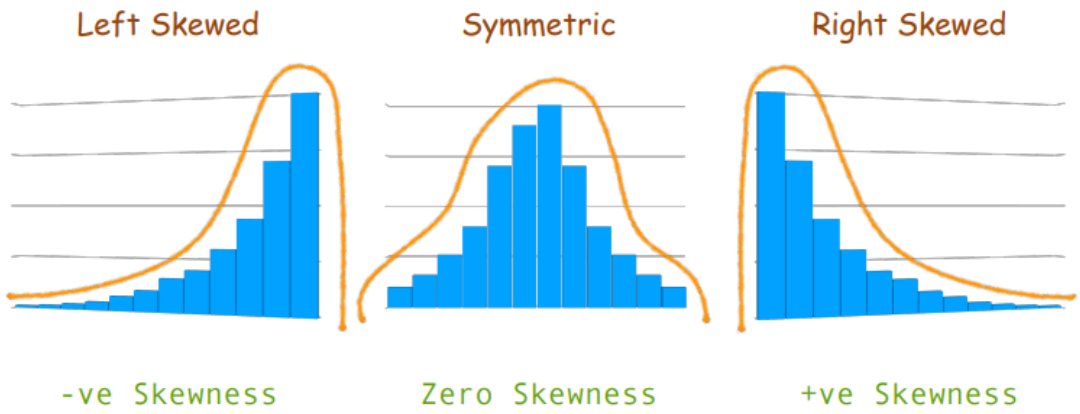
Histograms

- Help to view *data density*
- Help to see *shape of distribution*

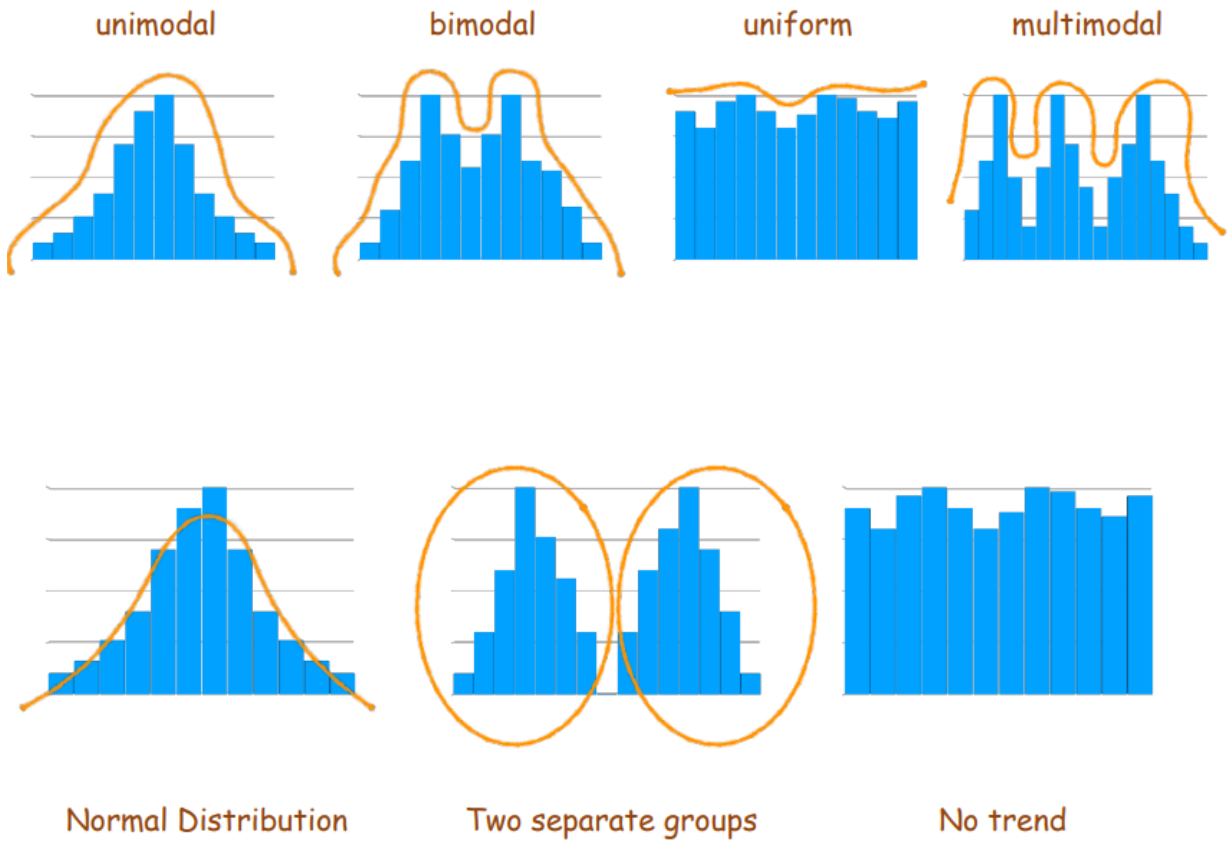
1) Skewness
2) Modality



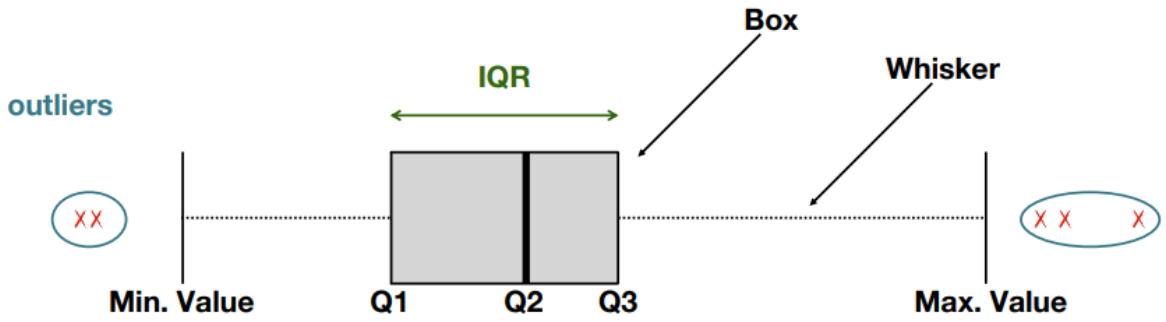
Skewness



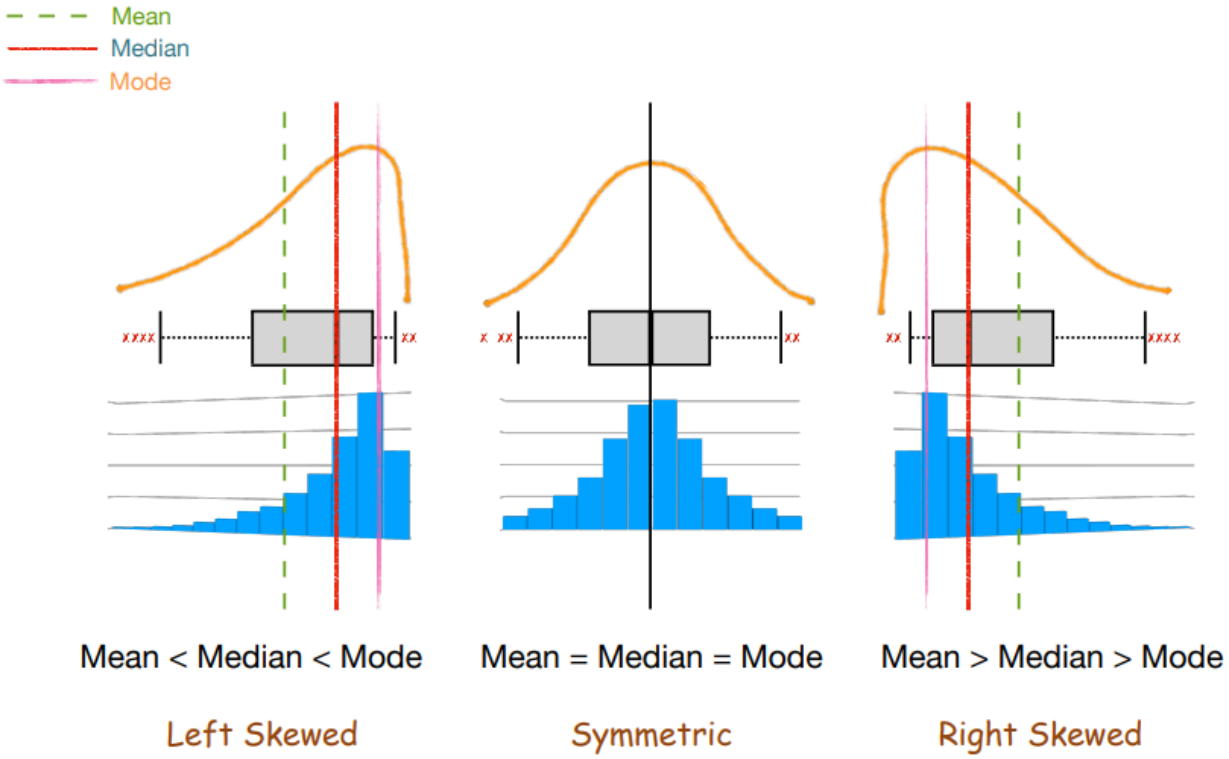
Modality



Box Plots



Min. Value :Lower Extreme (that's not an outlier)
Q1 :Lower Quartile (25% of observations)
Q2 :Median (50% of observations)
Q3 :Upper Quartile (75% of observations)
Max. Value :Upper Extreme (that's not an outlier)
IQR :Inter-Quartile Range = Q3 - Q1 (middle 50% of observations)



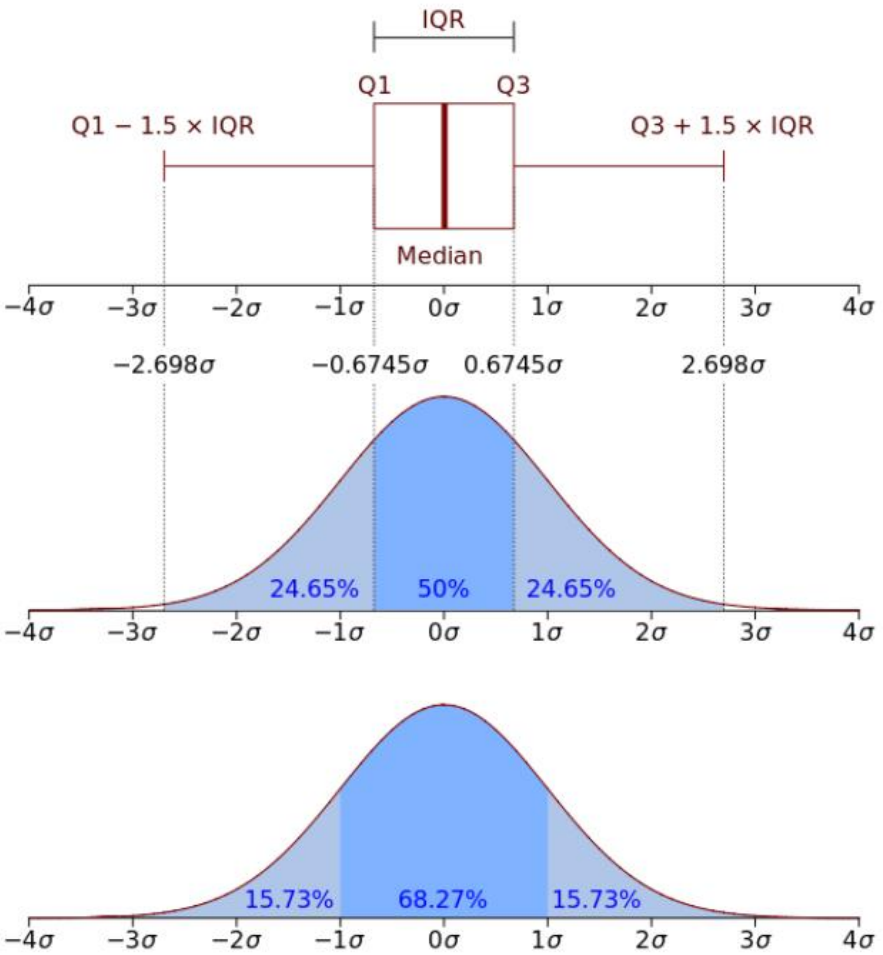
Robust Statistics

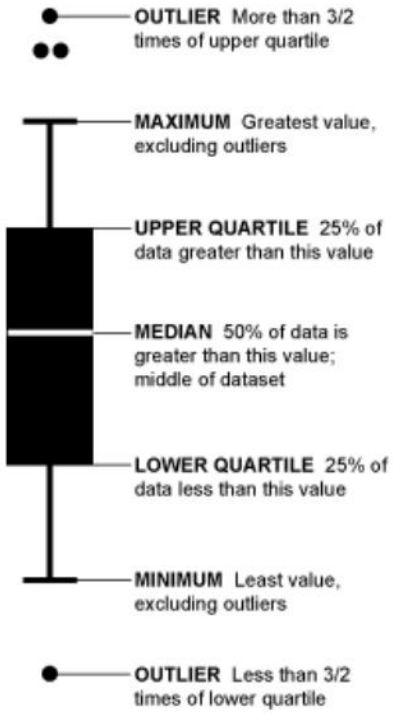
- Measures on which extreme observations or outliers have little effect

	Robust	Non-Robust
Spread	IQR	SD, Range
Center	Median	Mean

Skewed **Symmetric**

Outliers





```
Box Plot Calculation:

                                   n
10, 20, 30, 40, 50, 80, 81, 83, 85, 88, 90, 91, 92, 92, 94
            Q1                  Median                Q2

Q2(Median) = 83

Q1 = 1/4(n+1) = 40

Q3 = 3/4(n+1) = 91

Iqr = Q3 - Q1
    = 91 - 40
    = 51

LQ = Q1 - 1.5 x Iqr
   = 40 - 1.5 x 51
   = -36.5
UQ = Q1 + 1.5 x Iqr
   = 91 + 1.5 x 51
   = 167.5
```

Why do EDA:

- To understand data properties
- To find patterns in data
- To suggest modelling strategies
- To "debug" analyses
- To communicate results

Lecture 5: Database

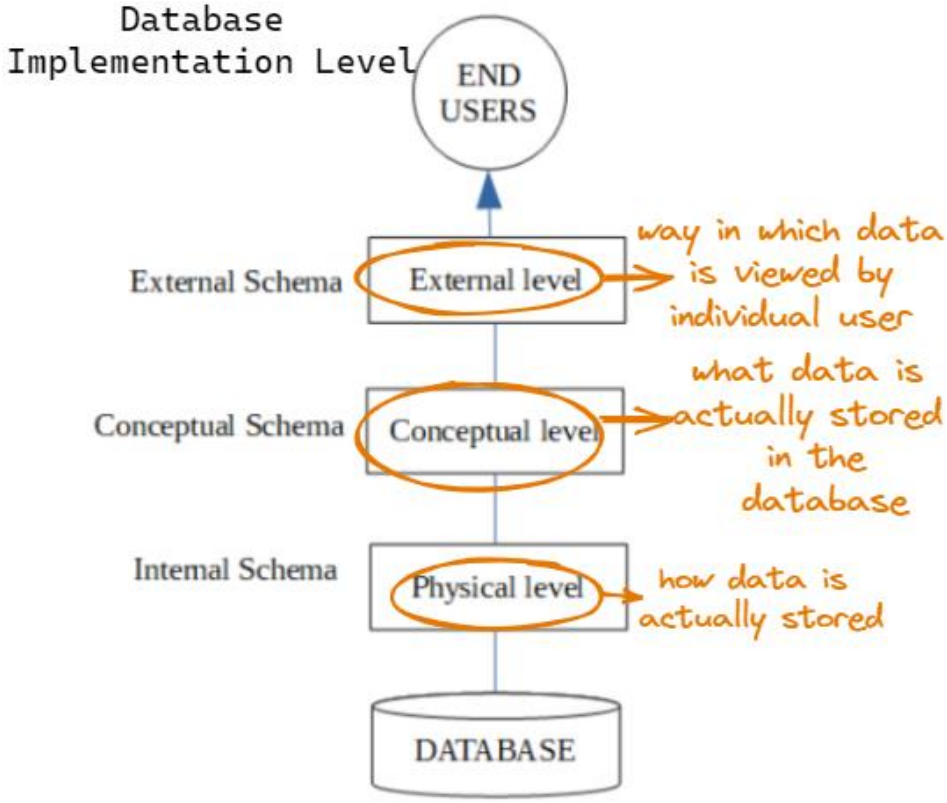
The term database describes a collection of data organized in a manner that allows access, retrieval, and use of that data.

Name	D.O.B	Fees
Harsh	23/01/1993	Not paid
Amar	04/11/1994	Paid
Devendra	14/06/1992	Not paid
Harsh	23/01/1993	Not paid

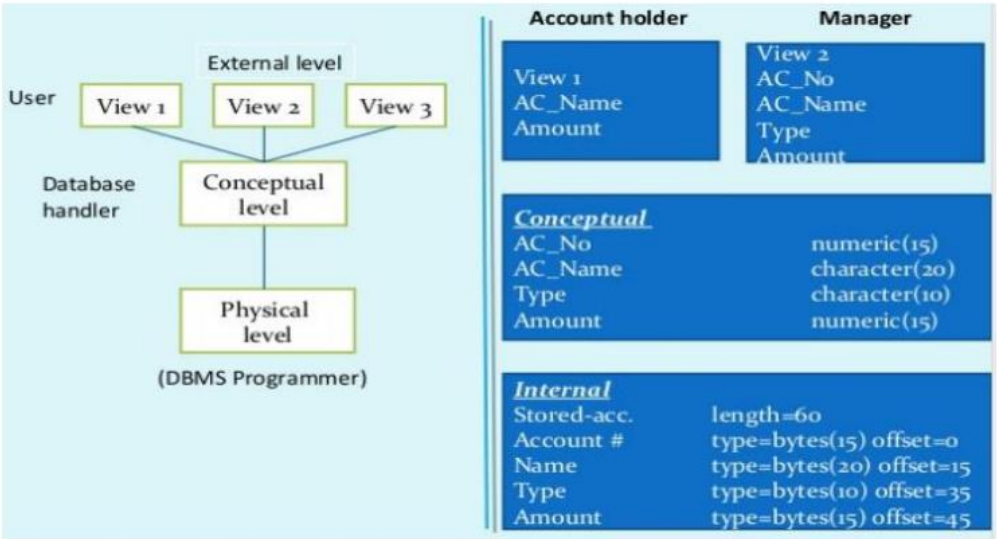
Data Redundancy

Name	D.O.B	Fees
Harsh	23/01/1993	Paid
Amar	04/11/1994	Paid
Devendra	14/06/1992	Paid
Harsh	23/01/1993	Not paid

Data Inconsistency



Levels of Abstraction



Types of Databases:

Single-user database supports only one user at a time

- Desktop database: single-user; runs on PC

Multiuser database supports multiple users at the same time

- Workgroup and enterprise databases

Centralized database

- data located at a single site

Distributed database

- data distributed across several different sites

Operational database

- supports a company's day-to-day operations
- transactional or production database

Data ware house

- stores data used for tactical or strategic decisions

Relational Database:

A database structured to recognize relations between stored items of information.