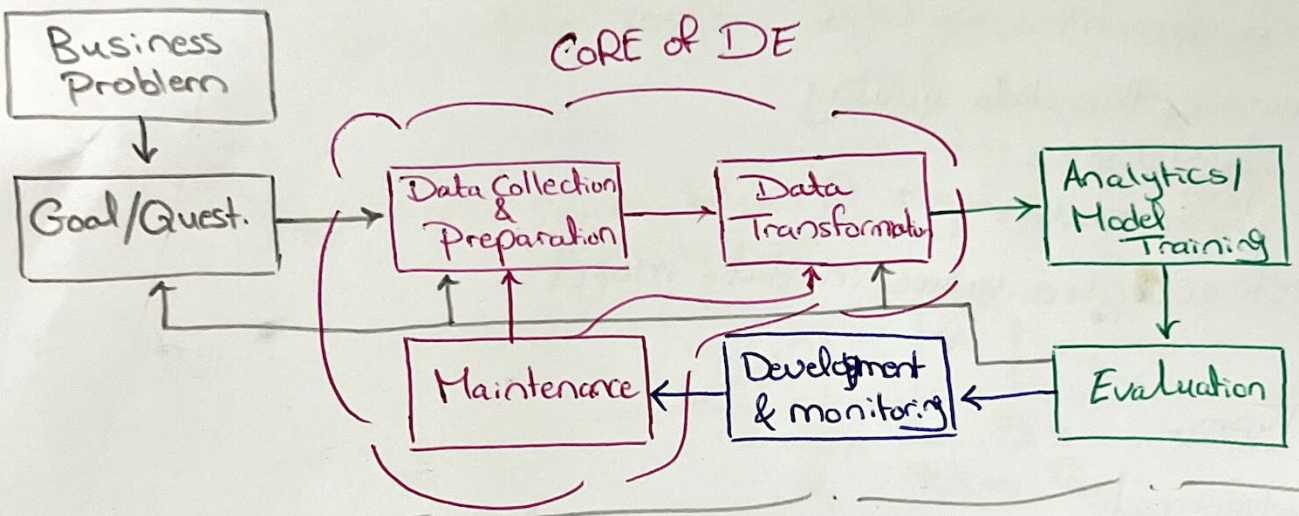
CSEN 1095 - Data Engineering

What is data Engineering?!

Processes that collect & integrate raw data from multiple diverse resources into a unified accessible repository to be analyzed & used in other application (ML Models, etc.)

Challenges:-

* Raw data can suffer from incompleteness, inaccuracy, inconsistency, etc.
* Applications may need data to be processed in batches or streams
* Continuous serving for Analytics



Jobs:-

↳ Data Scientist:- will be able to take data science projects from end to end, they can help store large amounts of data, create predictive modelling & present findings

↳ Data Engineers:- versatile generalists who use computer science to process the data focusing on coding, cleaning up the set & implement requests that come from scientists

↳ Data Analysts:- help people from across the company to understand queries with charts

To build a successful model:-

* Exploratory Analysis (10%)
* Data cleaning (20%)
* Feature Engineering (25%)
* Algorithm selection (10%)
* Model Training (15%)
* Others (20%)

Course Assessments:

- Project 30%
- Midterm 20%
- Final Exam 40%
- Lab Assignments 10%

Building a Data Profile to Assess Data Quality

1. Understand the data
2. Identify its issues
   - Completeness ⇒ Missing values, nulls, ...
   - Conformance ⇒ format, value
   - Correctness ⇒ logic
3. Measure the data quality
   - Statistics
   - Thresholds
4. Get accepted values in data model

Data Types:

- Categorical
  - * Nominal ⇒ each value represent category, can be represented in numbers (discrete ⇒ gender, hair color) not ordered
  - * Binary ⇒ nominal but only with two values
    - Symmetric ⇒ gender (have the same weight)
    - Asymmetric ⇒ States are not equally important
  - * Ordinal ⇒ values have a meaningful order or ranking grade, size, satisfaction (ordered values)
- Continuous
  - * Interval-scaled ⇒ scale of equal size units (temp, year) IQ
  - * Ratio-scaled ⇒ have a true zero point, can be represented as multiples (%, years of experience ...)

(Many methods that are used to analyze continuous data are not the same as those used for categorical data)

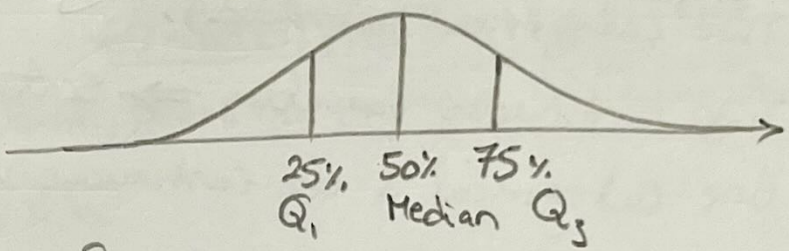
Basic Statistical Descriptions of Data.

↳ Central tendency (local of distribution) (not accurately represent the value of distribut)

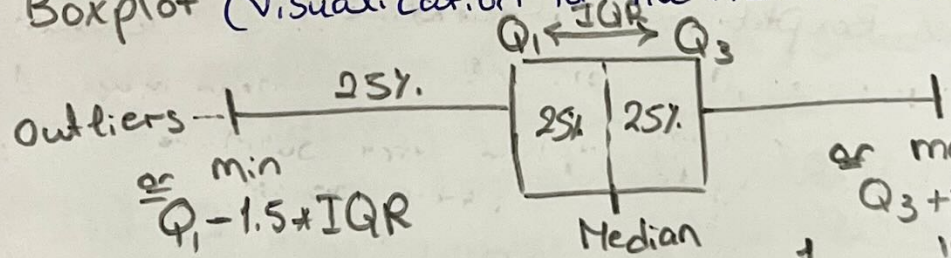
* Mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ Sensitive to outliers
  weighted average
  → sol ⇒ trimmed mean = chop off extreme values at both ends of observations
* Median 50% P_{50} 2nd Quatile Q_2
  Expensive to compute for large # of observations
* Mode → the most frequent value (can be more than 1)
  What if there is no mode?!

↳ Variability (Scale) @ 1st we have to order the observation the compute

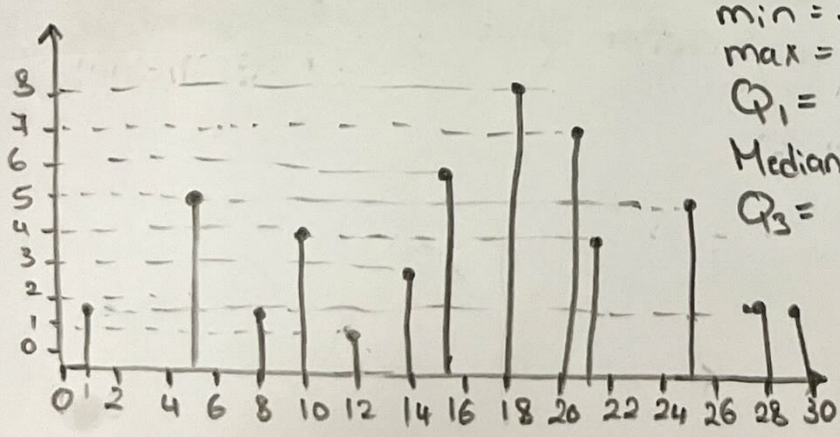
* Range, max − min
* Standard deviation $s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$ how spread out a data distribution
  Low SD → close to mean
  High SD → large range of values
* Interquartile range, $IQR = Q_3 - Q_1$
  3^{rd} Quantile
  ↓?!
* Quantiles, points taken @ regular intervals of data distribution



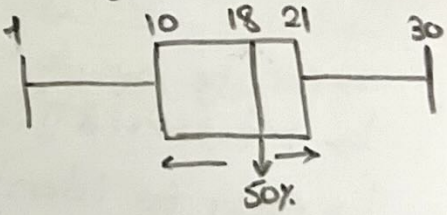
* Five-Number Summary
  Min, Q_1, Median(Q_2), Q_3, Max
* Boxplot (visualization for five number Summary)



Ex:



min = 1
max = 30
Q_1 = 10
Median = 18
Q_3 = 21



Nominal Data:

- Frequency f_i
- Proportion $\frac{f_i}{N}$
- Percentage $\frac{f_i}{N} \times 100$
- Illustrated using bar chart or Pie chart
  - bars don't touch eachother for categorial and non continuous data

Ordinal Data:

- Frequency, Proportion, percentage
- Percentiles
- Mode
- Median
- Interquartile Range
- Illustrated using barchart or Pie chart

Continuous Data:

- Percentiles, median, interquartile range
- Mean, Median, Mode
- SD, range, IQR
- illustrated using Histogram or Boxplot
  - Histogram: can describe continuous data; not a good way to represent outliers; define number of modes
  - Boxplot: more robust; not able to define no. of modes & distribution

Relationship between two variables is complicated than previously mentioned (each participant must have the two variables)

Number of computers in home by country ⇒ tabular form (bar chart)
(*Two categorical variables)

* Two continuous variables ⇒ scatter plot
* One categorical & one continuous variable ⇒ side by side box plot

Common Questions:

- When to use Median instead Mean?!
- When to use IQR instead of SD?!
- When to use Boxplot instead of bar chart?!

→ Depends on type of data

Outliers:

- small 1.5IQR
- Large 3IQR
- can be due to human error when entering the data
- Median & IQR are robust to outliers ⇒ Use them in presence of outliers
- easy to identify in Scatter plot

Data Challenges:-

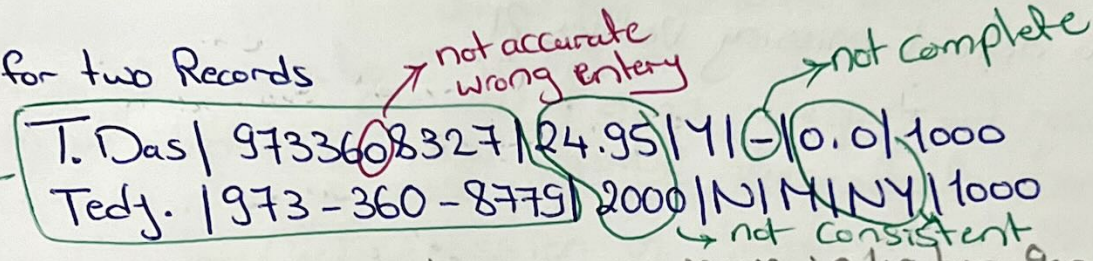
1. Massive Data
2. Dimensionality ⇒ high dimensional problem in terms of features
3. Missing data values
4. Wrong data values
5. data not factual

To increase the quality of the data, we have to prepare & preprocess the data

Factors of data quality:-

1. accuracy (no errors)
2. Completeness
3. Consistency (different datatypes)
4. Timeliness (old)
5. Integrity (poor relationship)
6. Interpretability (how easy it can be understood)

Ex:- Data for two Records



* Interpretability ⇒ Name, phone no., revenue, indicator, gender, State, Usage

To enhance the quality of data
- ↳ extensive domain knowledge is needed
- ↳ Solutions should be chosen case by case

To improve the quality of data:-

* Data cleaning ⇒ filling in missing values, Smoothing noisy data, identify & remove outliers
* Data transformation ⇒ normalization, discretization
* Data reduction ⇒ produce the same analytic results with smaller data
* Data integration ⇒ include data from multiple sources

Data Cleaning

- ↳ incomplete
- ↳ noisy (errors & outliers)
- ↳ inconsistent (age = 42, birthday = 3/7/2010)

How to clean the data?!

impose semantic & syntactic validity of raw data

1. fill missing value (missing value imputation)
2. smoothout noise (binning, regression, outlier detection, clustering, discretization)
3. Correcting inconsistencies (fuzzy joins, regular expressions, database profiling, metadata)

Data cleaning method must be reproducible & data must be preserved in its original form

How to detect missing values?!

- ↳ Scan for gaps in rows & columns
- ↳ Cross check data schema with actual data for missing attributes
- ↳ check data loss during communication
- ↳ check history of data source
- ↳ keep track of estimated values & error bounds

Missing Values

Missing Completely @ random (MCAR)

* probability of being missing is the same for all values
* no relationship between missing data & any value

Missing @ random (MAR)

* Systematic relationship between the propensity of missing value & observed data

ignorable if enough data is available to impute missing data

Missing @ not random (MANR)

* relationship between the propensity of value missing & its values

⇓ non ignorable

w to handle missing values?!

↳ Fill in missing values randomly ⇒ time consuming
↳ use global constant ⇒ null
↳ Listwise deletion
] ⇒ simplistic Methods

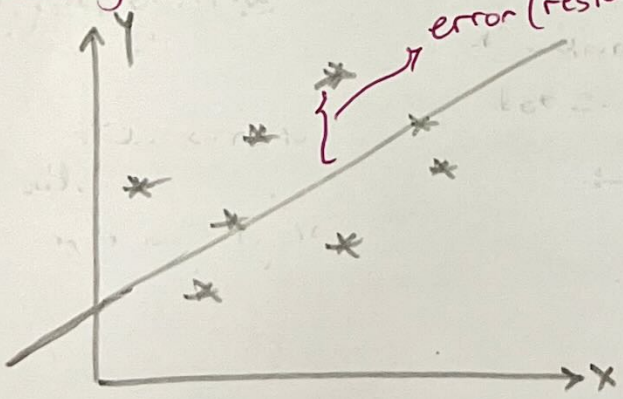
↳ replace missing values with plausible values (uncertain)] ⇒ imputation

↳ Univariate Imputation
↳ Mean/Average imputation ⇒ replace missing data with mean of non-missing data
SD will be underestimated
↳ use value drawn from distribution of non-missing data (interpolation ex:)
(can be used if missing data 2-3%, this imputation is a poor choice for some ML algorithms

↳ Multivariate Imputation
↳ use mean or median
↳ use most probable (estimated) value
"Regression or Bayesian"

↳ Linear Regression → dependent variable is continuous
↳ Multiple Regression → multi independent variables
↳ Logistic Regression → dependent variable is binary
↳ Multinomial Logistic Regression → dependent variable is categorical
] ⇒ Regress

Linear Regression :-



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

ε_i ⇒ random error term

to find the best regression model ⇒ Gradient Descent