Data cleaning, analysis

Step 0

- .csv
- .xls
- .db

Step 1

- Normalization
- Correction
    - USP zip code in specific form – characters in their normal form
    - Post addresses have a lot of forms
        - 9553 NE 96 Blvd., #102
        - 1210 W Daton St, but 1210 E Dayton St  doesn't exist
            - We put all the data for consistency, so the "E" will be ignored, basically change the small changes to correct it

Step 3

- Scaling
    - Eg. Currency
        - International data
            - Comparing EU and US dollar

Step 4

- Entity matching
    - Spelling in different address, it is referring the same place
    - Last name – Groefe
    - Dr. …..
    - Mr. ….. Göetz
    - ……. Gräfe
        - All the string referring the same person but in difference spelling
        - Eg. Different county names will be Americanized …etc
        - Eg. John Smith or John Smyth or J. Smith → match the same person

^

|

---

Modernized business

Step 5

- random and anonymous
- age of people in classroom
    - distribution
    - correlations
        - not lose info and also randomize

implement in separate products

- eg. ETL  extract transform load