Data Acquisition & Data Preparation

Kittanut Chungnoi (March)
kittanutc@gmail.com

Topic

- Data Categories
- Data Acquisition
- Data Representation
- Data Measurement Scales
- Data Preparation

Data Categories

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotitaw	28

Structured data

```
<?xml version="1.0"?>
- <birds>
  - <owl id="1201">
      <species>Bubo bubo</species>
      <name>Eagle Owl</name>
      <region>Eurasia</region>
    </owl>
  - <owl id="1202">
      <species>Strix occidentalis</species>
      <name>Spotted Owl</name>
      <region>North America</region>
    </owl>
</birds>
```

Semi-structured data

Data literacy is the ability to read, understand, create, and communicate data as information. Much like literacy as a general concept, data literacy focuses on the competencies involved in working with data.

Unstructured data

What is Structured Data

- Structured data is data that has a standardized format for efficient access by software and humans alike.
- It is typically tabular with rows and columns that clearly define data attributes.
- Example: Google Sheet or Excel

What is Semi-Structured Data?

- Semi-structured data is a type of data that is **not purely structured**, but also **not completely unstructured**.
- It contains some level of organization or structure, but does not conform to a rigid schema or data model, and may contain elements that are not easily categorized or classified.
- For example, an **XML document** might contain tags that indicate the structure of the document, but may also contain additional tags that provide metadata about the content, such as author, date, or keywords.

What is Unstructured Data?

- Unstructured data is the data which does not conforms to a data model and has **no easily identifiable structure** such that it can not be used by a computer program easily.
- Unstructured data is not organized in a pre-defined manner or does not have a pre-defined data model, thus **it is not a good fit for a mainstream relational database**.
- Example: images (JPEG, GIF, PNG, etc.), videos, and surveys.

Data Acquisition: Primary Data Collection

- Surveys and Questionnaires
  - Researchers design structured questionnaires or surveys to collect data from individuals or groups. These can be conducted through face-to-face interviews, telephone calls, mail, or online platforms.
- Interviews
  - Interviews involve direct interaction between the researcher and the respondent. They can be conducted in person, over the phone, or through video conferencing. Interviews can be structured (with predefined questions), semi-structured (allowing flexibility), or unstructured (more conversational).

Data Acquisition: Primary Data Collection

- Observations
  - Researchers observe and record behaviors, actions, or events in their natural setting. This method is useful for gathering data on human behavior, interactions, or phenomena without direct intervention.
- Experiments
  - Experimental studies involve the manipulation of variables to observe their impact on the outcome. Researchers control the conditions and collect data to draw conclusions about cause-and-effect relationships.

Data Acquisition: Secondary Data Collection

- Published Sources
  - Researchers refer to books, academic journals, magazines, newspapers, government reports, and other published materials that contain relevant data.
- Online Databases
  - Numerous online databases provide access to a wide range of secondary data, such as research articles, statistical information, economic data, and social surveys.

Data Acquisition: Suggested Data Sources

- UCI Machine Learning Repository
  - https://archive.ics.uci.edu/ml/index.php
- Kaggle
  - https://www.kaggle.com/datasets
- Open Government Data of Thailand
  - https://data.go.th/

Data Representation

- Qualitative Data
  - This data cannot be described using numbers and basic mathematics.
  - This data is generally described using natural categories and language.
- Quantitative Data
  - This data can be described using numbers.
  - Basic mathematical procedures are possible on the set.

Data Measurement Scales

- Categorical Attribute
  - Nominal
  - Ordinal
- Numeric Attributes
  - Interval
  - Ratio

Categorical Attributes

- One that has a set-valued domain composed of a set of symbols.
- Such as Gender = {M,F}, Education = {High School, BS, MS, PhD}, etc.
- **Nominal**
  - Attribute values in the domain are unordered.
  - Can only equality (=) compare.
  - Such as gender, type of hair, etc.
- **Ordinal**
  - Attribute values are ordered.
  - • Can both equality (=) and inequality () compare.
  - Such as education, feel (unhappy, OK, happy), etc.

Numeric Attributes

- One that has a real-valued or integer-valued domain.
- Such as age, height, grade, frequency, etc.
- **Discrete**
  - Take on a finite or countably infinite set
  - Such as integer, grade, number of object, etc.
- **Continuous**
  - Take on any real value
  - Such as height, weight, size, etc.

Numeric Attributes

- **Interval-scaled**
  - Can compute only differences (addition or subtraction)
  - For example, temperature measured in ˚C or ˚F.
    - If it is 20 ˚C on one day and 10 ˚C on previous day
    - We can talk about a temperature drop of 10˚C.
    - We cannot say that it is twice as cold as the previous day.

Numeric Attributes: Data Representation

- **Ratio-scaled**
  - Can compute both differences and ratio between values,
  - For example, age.
    - If Jone is 20 years old and Jim is 10 years old.
    - We can say that Jone older than Jim with 10 years.
    - We can say that Jone is twice as old as Jim.

Summary of Data Measurement Scales

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Beautiful Soup: Build a Web Scraper With Python

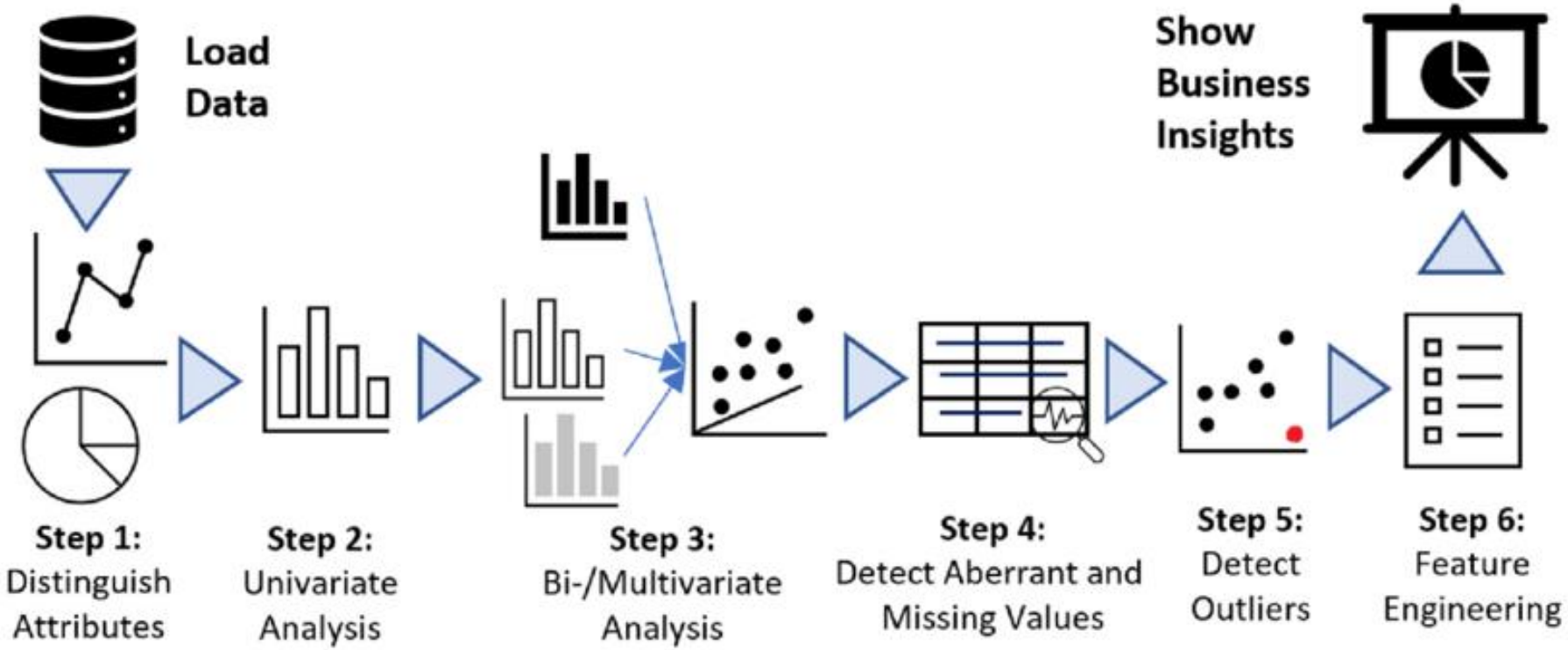
Exploratory Data Analysis

https://tinyurl.com/2wy88ckb

What is Exploratory Data Analysis (EDA)?

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations

Exploratory Data Analysis (EDA) Process



https://www.researchgate.net/profile/Cristina-Sousa-2/publication/342282008/figure/fig2/AS:903839578353665@1592503556728/Exploratory-Data-Analysis-EDA-steps.png

Exploratory Data Analysis (EDA) Process

- Distinguish Attributes : การเลือกเตรียมข้อมูลเพื่อวิเคราะห์ โดยพิจารณาจาก Attribute หรือ คุณลักษณะต่างๆ เช่น เพศ สี อายุ เป็นต้น
- Univariate Analysis : การวิเคราะห์ข้อมูลตัวแปรเดียว เพื่อพิจารณาพฤติกรรมของแต่ละ Attribute เช่น ค่าเฉลี่ย ผลรวม ความแปรปรวน เป็นต้น
- Bi-/Multivariate Analysis : การวิเคราะห์มากกว่า 1 ตัวแปร เพื่อพิจารณาถึงความสัมพันธ์ขั้นต้น เช่นการหา Correlation และการเขียนกราฟ Scatterplot
- Detect Aberrant and Missing Values : การพิจารณาสิ่งผิดปกติที่เกิดขึ้นในชุดข้อมูลและกลุ่มข้อมูลที่หายไป
- Detect Outlier : การวิเคราะห์หาข้อมูลที่ผิดปกติไปจากค่ากลาง หรือ Outlier
- Feature Engineering : การสร้าง feature หรือตัวแปรที่จะนำไปวิเคาะห์เชิงลึก

Data Exploratory Analysis using DataPrep

- Data Preparation : Collect, clean, and visualize your data in python with a few lines of code

Analyze Distributions

- We start by calling **plot(df)** which computes dataset-level statistics, a histogram for each numerical column, and a bar chart for each categorical column.

```
from dataprep.eda import plot
plot(df)
```

Computing compute_hll_array-chunk-estimate_count-agg-af1f00907198a5831056fbbe19aafb49: 99% 603/612 [00:00<00:00, 1226.99it/s]

Hide Stats and Insights

Dataset Statistics	
Number of Variables	15
Number of Rows	48842
Missing Cells	6465
Missing Cells (%)	0.9%
Duplicate Rows	52
Duplicate Rows (%)	0.1%
Total Size in Memory	30.1 MB
Average Row Size in Memory	645.7 B
Variable Types	Numerical: 6 Categorical: 9

Dataset Insights	
workclass has 2799 (5.73%) missing values	Missing
occupation has 2809 (5.75%) missing values	Missing
native-count… has 857 (1.75%) missing values	Missing
fnlwgt is skewed	Skewed
education-num is skewed	Skewed
capital-gain is skewed	Skewed
capital-loss is skewed	Skewed
hours-per-we… is skewed	Skewed
capital-gain has 44807 (91.74%) zeros	Zeros
capital-loss has 46560 (95.33%) zeros	Zeros

Understand a column

- We can thoroughly investigate a column of interest col1 using **plot(df, col1)**.
- The output is of **plot(df, col1)** is different for numerical and categorical columns.

```
plot(df, "age")
```

Computing series-max-agg-134a2c7c45840eebf012c54a9e47b8dc: 0% 0/110 [00:00<?, ?it/s]

Stats | Histogram | KDE Plot | Normal Q-Q Plot | Box Plot | Value Table

Overview

Approximate Distinct Count	74
Approximate Unique (%)	0.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	763.2 KB
Mean	38.6436
Minimum	17

Descriptive Statistics

Mean	38.6436
Standard Deviation	13.7105
Variance	187.9781
Sum	1.8874×10^{06}
Skewness	0.5576
Kurtosis	-0.1844
Coefficient of Variation	0.3548

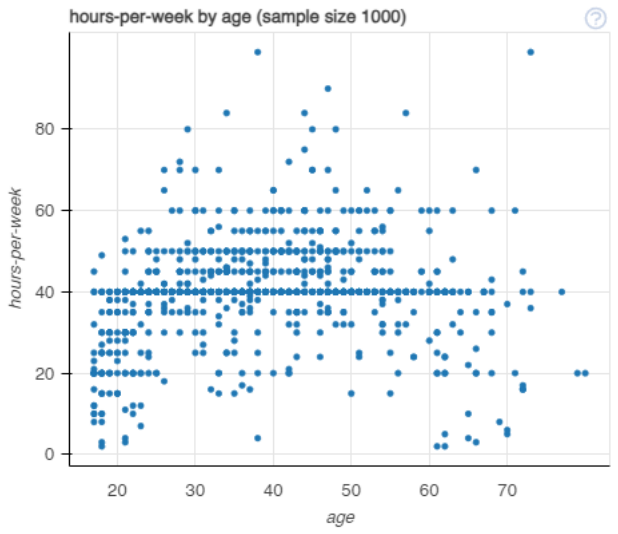
Understand the relationship between two columns

- we can explore the relationship between columns col1 and col2 using **plot(df, col1, col2)**. The output depends on the types of the columns.

```
plot(df, "age", "hours-per-week")
```

Computing cleanup-shuffle-collect-_box_comps-fd4da9ea7ea6eebbb25be1c6eb34b96e: 90% 9/10 [00:00<00:00, 54.10it/s]

Scatter Plot Hexbin Plot Box Plot



Analyze Correlations

- The function **plot_correlation()** explores the correlation between columns in various ways and using multiple correlation metrics.
  - **plot_correlation(df)**: plots correlation matrices (correlations between all pairs of columns)
  - **plot_correlation(df, col1)**: plots the most correlated columns to column col1
  - **plot_correlation(df, col1, col2)**: plots the joint distribution of column col1 and column col2 and computes a regression line

```
from dataprep.eda import plot_correlation
plot_correlation(df)
```

Stats | Pearson | Spearman | KendallTau

	Pearson	Spearman	KendallTau
Highest Positive Correlation	0.672	0.79	0.607
Highest Negative Correlation	-0.683	-0.707	-0.528
Lowest Correlation	0.002	0.001	0.0
Mean Correlation	0.019	0.028	0.021

Analyze Missing Values

- The function **plot_missing()** enables thorough analysis of the missing values and their impact on the dataset.
  - **plot_missing(df)**: plots the amount and position of missing values, and their relationship between columns
  - **plot_missing(df, col1)**: plots the impact of the missing values in column col1 on all other columns
  - **plot_missing(df, col1, col2)**: plots the impact of the missing values from column col1 on column col2 in various ways.

```
from dataprep.eda.missing import plot_missing
plot_missing(df)
```

Computing isnull-961119a5c5bd20cb8ab259d5fc522be8: 0%

Stats | Bar Chart | Spectrum | Heat Map | Dendrogram

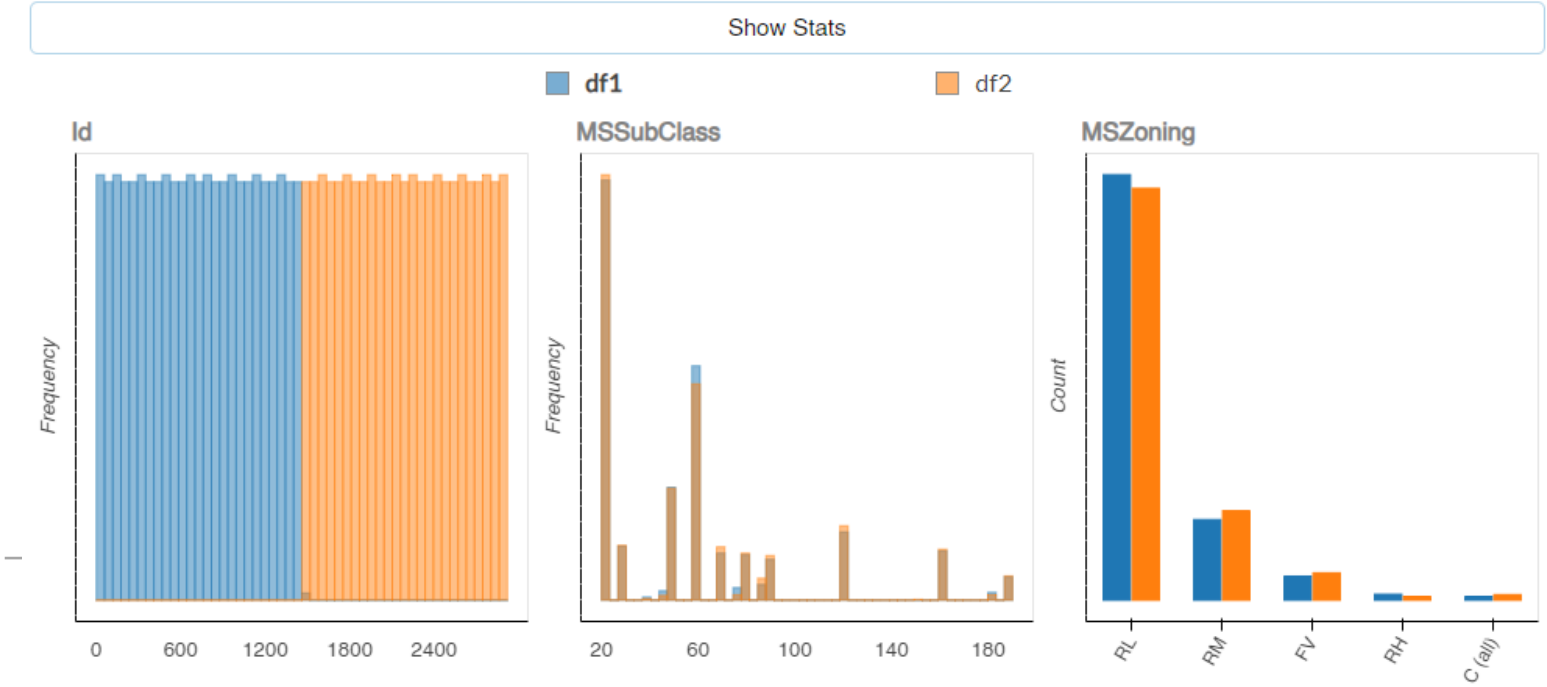
Missing Statistics

Missing Cells	866
Missing Cells (%)	8.1%
Missing Columns	3
Missing Rows	708
Avg Missing Cells per Column	72.17
Avg Missing Cells per Row	0.97

Analyze Differences

- The function **plot_diff()** explores the difference of column distributions and statistics across multiple datasets.

```
from dataprep.eda import plot_diff
plot_diff([df1, df2])
```



Example DataPrep

- https://tinyurl.com/33rzpmfx

Thank You