MODULE 04

DATA SCIENCE BOOTCAMP

Data Cleaning, Visualization and EDA

Vertical Institute

Reminder

1. Please turn on Zoom camera the whole duration of classes.

2. At the start of all classes, please rename yourselves to: Name + Last 3 digits and letter of your NRIC. Example: John Tan (123A)

Agenda

- Data Visualization Fundamentals
- Python Visualization Packages
- Data Cleaning & Exploratory Data Analysis (EDA)
- Common Data Cleaning Steps

Attendance Photo Taking

MODULE 4: DATA CLEANING, VISUALIZATION & EXPLORATORY DATA ANALYSIS

Data Visualization

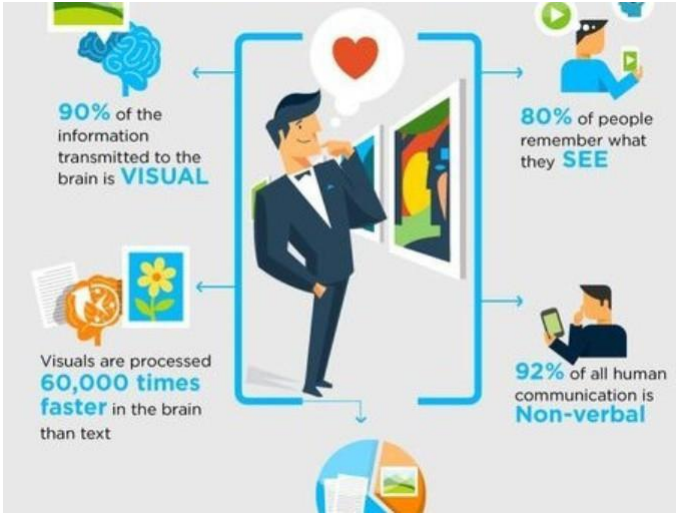
What is Data Visualization?

Data visualization is the graphical representation of information and data. It is the practice of translating information into visual context such as graph or map, to make data easier for human brain to understand and pull insight from.

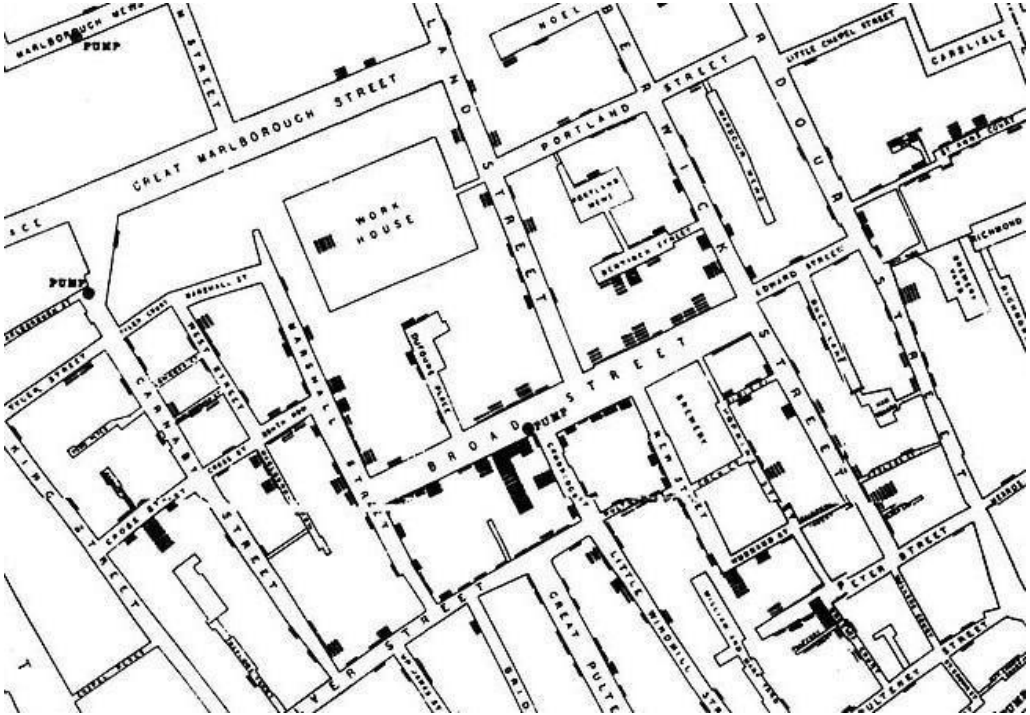
Why Data Visualization?

Data visualization provides a quick effective way to communicate information in a universal manner using visual information.

It helps business identify which factors affect customer behaviour and pinpoint areas that need to be improved or need more attention.

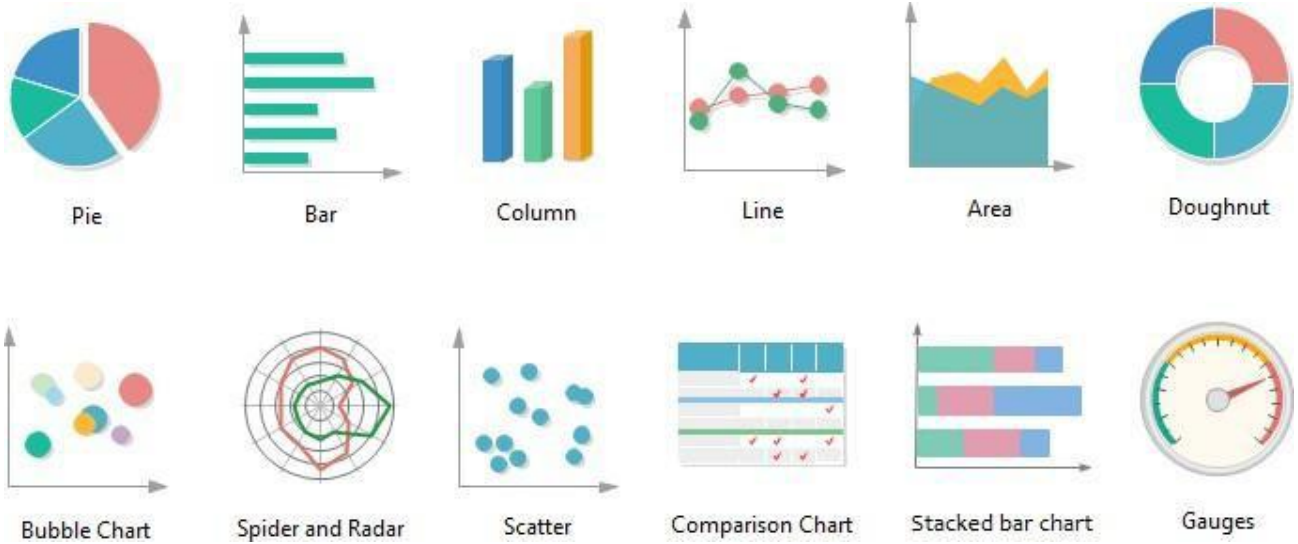


John Snow's Map of The 1854 London Epidemic



Types of Data Charts

Data visualization is the graphical representation of information and data.



Key Principles of Data Visualization

Simple

- Maximize **impact**, minimize **noise**.
- If it doesn't **add value** or **serve a purpose**, get rid of it.

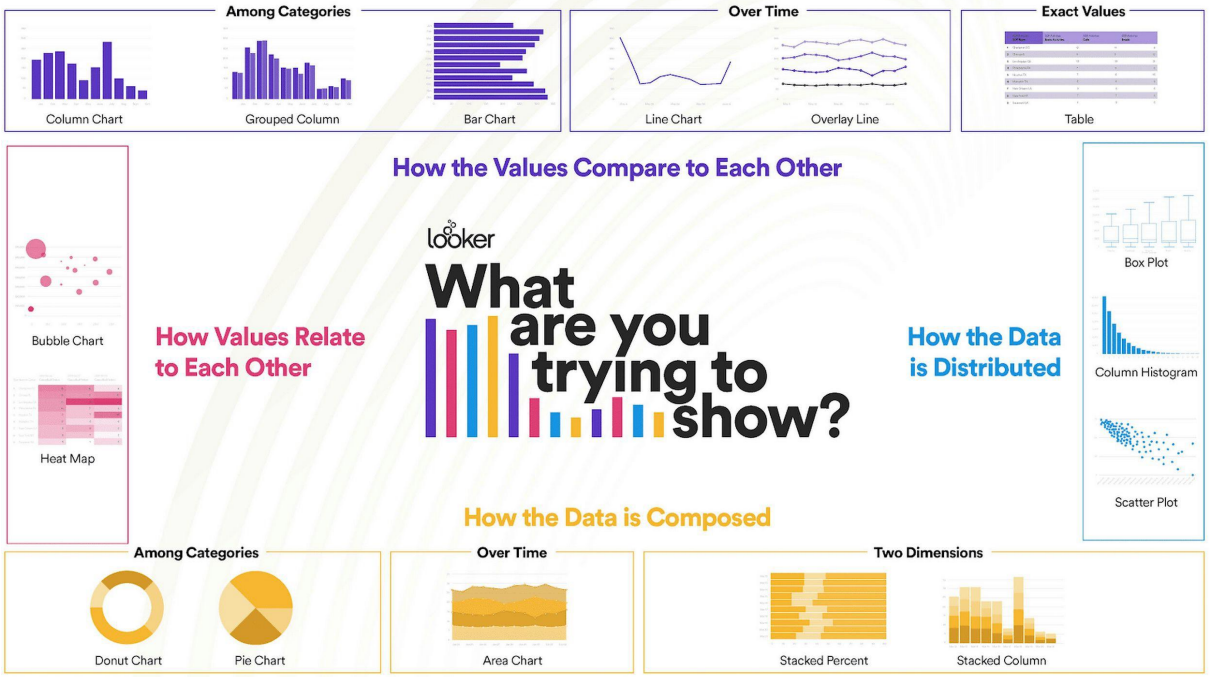
Narrative

- Don't just show data, **tell a story**.
- Communicate key insights **clearly**, **quickly**, and **powerfully**.

Balance between Design and Function

- Selecting the right type of chart is **critical**.
- **Beautiful** is good, **functional** is better, **both** is idea.

When to Use What Type of Visualization?



MODULE 4: DATA CLEANING, VISUALIZATION & EXPLORATORY DATA ANALYSIS

Python Visualization Packages

Python Visualization Packages

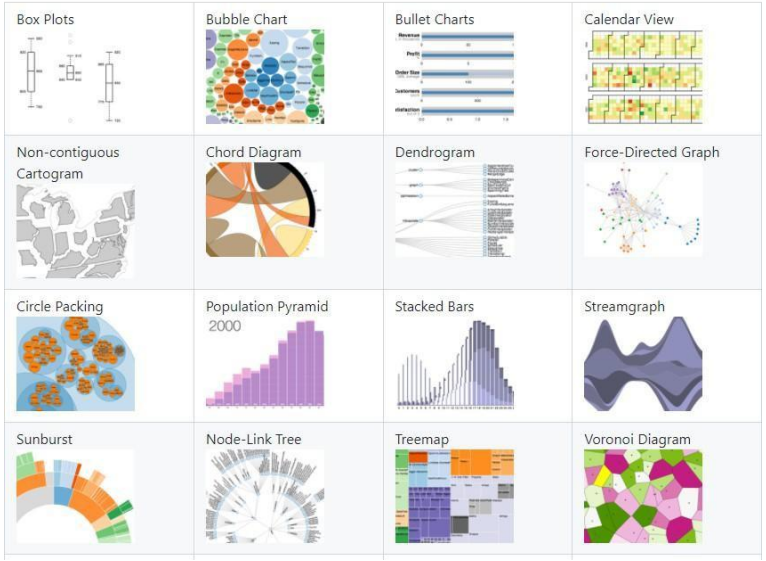
- Matplotlib
- Ggplot2
- Seaborn
- Bokeh
- Plotly

Inspiration for Charts

https://github.com/d3/d3/wiki/Gallery

https://www.python-graph-gallery.com/

Visual Index

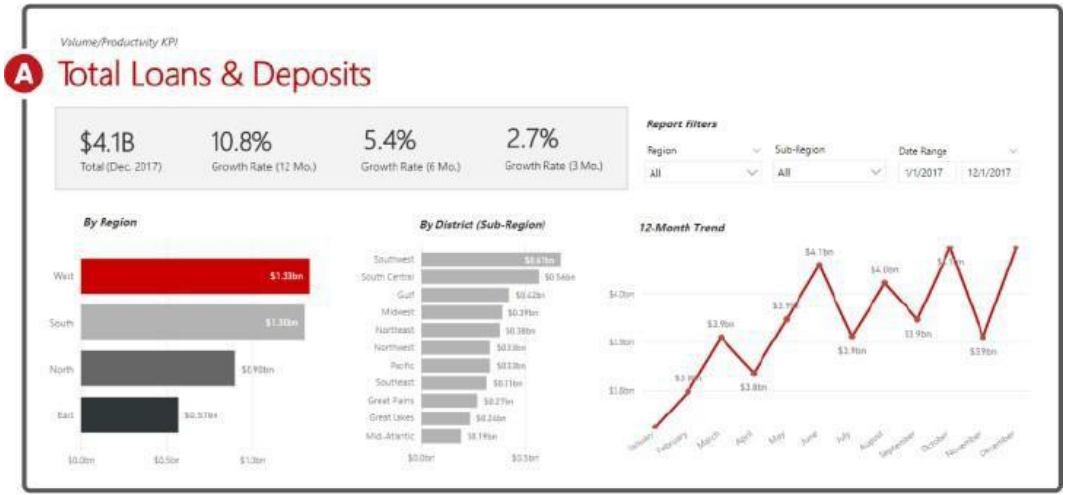


The Python Graph Gallery

Welcome to the Python Graph Gallery, a collection of hundreds of charts made with `Python`. Charts are organized in about 40 sections and always come with their associated reproducible code. They are mostly made with `Matplotlib` and `Seaborn` but other library like `Plotly` are sometimes used. If you're new to python, this online course can be a good starting point.

Consumer Banking Dashboard

Financial institutions actively use data visualization to aid in their understanding of their business and decision-making. Below is a dashboard visualization of consumer banking KPIs.



MODULE 4: DATA CLEANING, VISUALIZATION & EXPLORATORY DATA ANALYSIS

Data Cleaning & Exploratory Data Analysis (EDA)

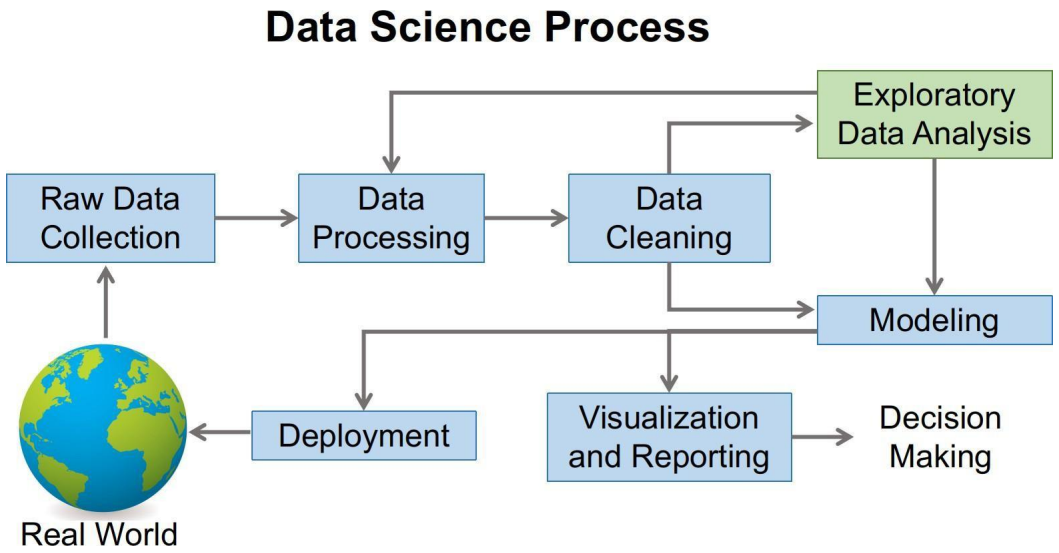
Exploratory Data Analysis (EDA) Framework

EDA is **the process of investigating the dataset to discover patterns and anomalies (outliers)**, and form hypotheses based on our understanding of the dataset.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better.

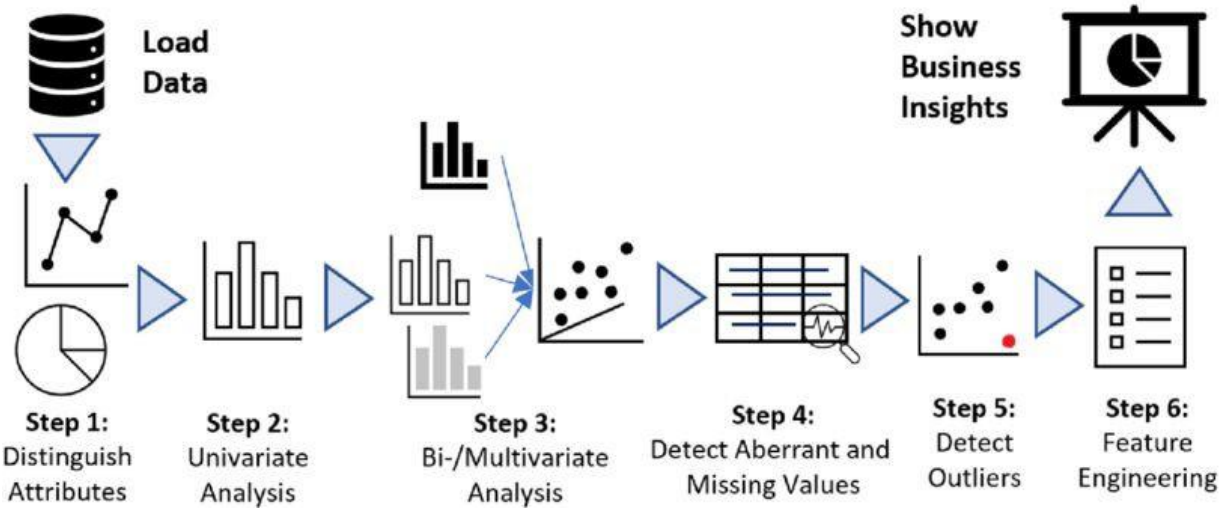
Exploratory Data Analysis (EDA) Framework

Within the EDA process:



Exploratory Data Analysis (EDA) Framework

On a higher level perspective:



Types of Data

Qualitative Data

Qualitative or Categorical Data describes the object under consideration using a finite set or discrete classes. It means that this type of data can’t be counted or measured easily using numbers and therefore divided into categories.

Examples:

Gender: Male or Female

Car Brands: Honda, Toyota, BMW

Types of Data

Quantitative Data

This data type tries to quantify things and it does by considering numerical values that make it countable in nature.

Examples:

Price of smart phone

Height of a person

Qualitative Data

Nominal

These are the set of values that don’t possess a natural ordering.

Examples: Red, Blue, Orange

Ordinal

These types of values have a natural ordering while maintaining their class of values.

Examples:

Sizes - small < medium < large

Grades — A, B, C

Quantitative Data

Discrete

The numerical values which fall under are integers or whole numbers are placed under this category.

Examples: Number of speakers, number of cameras, number of people

Ordinal

The fractional numbers are considered as continuous values

Examples:

Height — 1.71m

Temperature — 36.9 Degrees Celcius

What is Data Cleaning?

Data cleaning is a critical aspect of the domain of data management. The cleaning process involves reviewing all the data present within a database to either remove or update information that is incomplete, incorrect, or duplicated and irrelevant.

This is to ensure data quality as it plays an important part in deriving reliable answers during the data analysis.

Common Strategies in Cleaning Data

- Remove missing values
- Remove incorrect values
- Update incorrect values
- Clean the data formats
- Impute missing or invalid data
- Backfill or forward fill
- Deal with Outliers
- Extracting important variables

Recap time!

What are your favorite takeaways on Data Cleaning, Visualization and EDA today?

Let's share with each other!

Some things to take note...

Link and resource could be accessed in the Learning Portal.

https://elearn.verticalinstitute.com/users/sign_in

Attendance Photo Taking

MODULE 04

DATA SCIENCE BOOTCAMP

Thank you!

Vertical Institute