BIG DATA & BUSINESS INTELLIGENCE WEEK 2

Winter Semester 2025-2026

Lecturer: Narges Chinichian
SRH University of Applied Science

RECAP FROM LAST WEEK

Short quiz:

1. What is the difference between a data lake and a warehouse?
2. Give one example of structured and one unstructured data.
3. What is a KPI?

RECAP OF LAST WEEK

- We got to know each other and set up our GitHub repos for course work.
- Introduced the course structure (6 weeks) and goals.
- Discussed what data is and the 3 Vs of Big Data: Volume, Velocity, Variety.
- Compared types of data: structured, semi-structured, and unstructured
- Walked through the data lifecycle:

collection → storage → processing → analysis → decision → archival

- Looked at databases, data warehouses, and data lakes.
- Defined Business Intelligence (BI) and Key Performance Indicators (KPIs) as bridges from raw data to actionable insight.
- Ended with hands-on Python & SQL refreshers and a homework notebook using the Harvard Dataverse electrification dataset.

OUTLINES OF SESSION 2

1. Present your findings from the electrification “PeopleSuN” dataset.
2. Choose and justify a dataset relevant to your interests.
3. Formulate 2–3 measurable KPIs.
4. Conduct basic exploratory data analysis to understand data structure, quality, and first insights.

PRESENTATIONS

- We have 10 students.
- Each student has 4 minutes presenting their findings.
- We have ~2 minutes asking questions.

We have 1 hour starting now!

WHAT IS THE PLAN?

We would like to create a **professional dashboard** for users to access a certain **dataset** and be able to interact with it to answer **at least 3 KPIs.**

WHERE CAN WE FIND QUALITY DATASETS?

Popular open data sources:

Harvard Dataverse → academic, social, and development research data

Kaggle or Zindi Datasets → mostly clean, curated data from competitions & community

World Bank Open Data → global indicators on health, energy, education, economy

OpenStreetMap (OSM) → geographic and infrastructure data (roads, buildings, utilities)

Government Open Data Portals → e.g. data.gov, data.europa.eu, data.gov.in

A FEW EXAMPLES OF DASHBOARDS AND KPIS

US POPULATION DASHBOARD

Source: US Census Data (2010–2019)
Goal: Explore how population and migration vary across U.S. states and over time.

Key KPIs

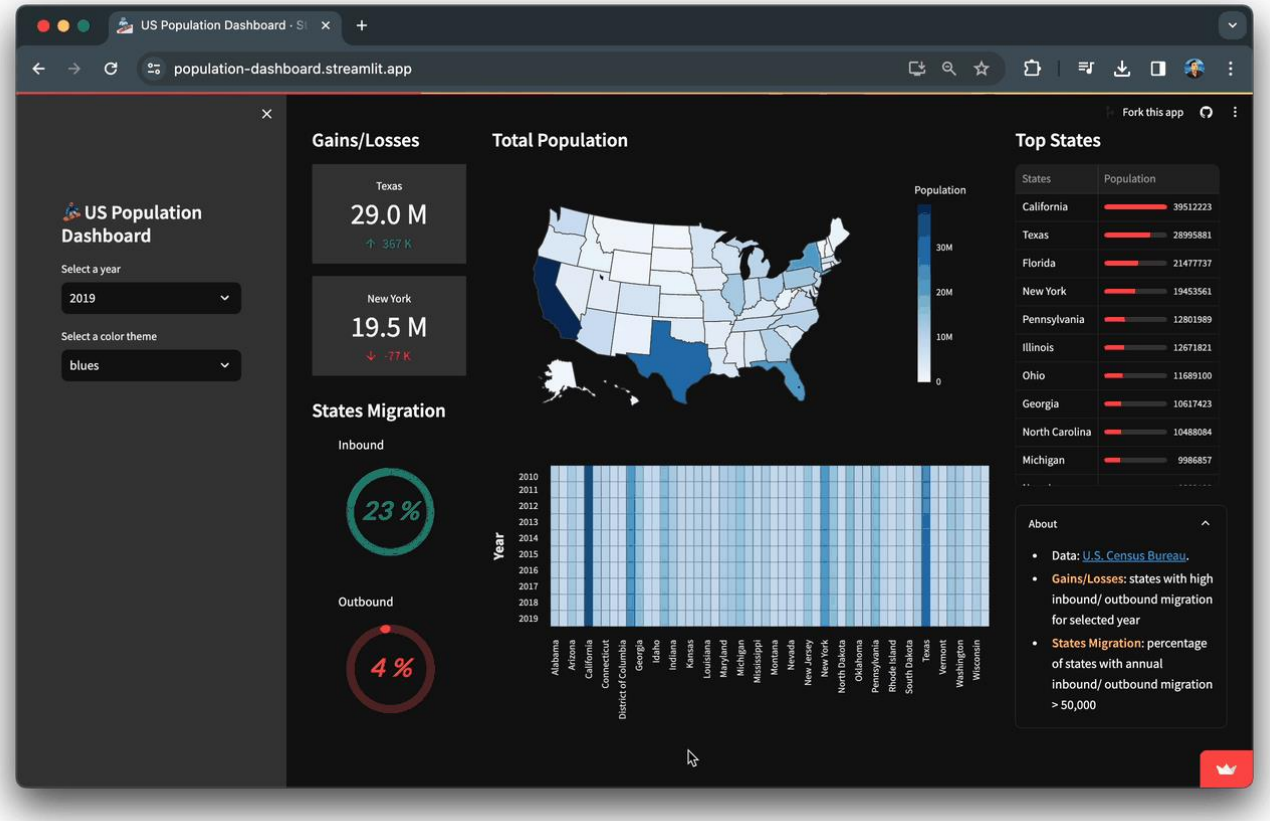
- Total population by selected year
- Year-over-year % population change
- Net migration (in – out) per state

Why It's a Good Example

Simple, clean dataset with temporal + geographic dimensions

Uses maps and heatmaps effectively

Shows how KPIs can summarize broad social trends



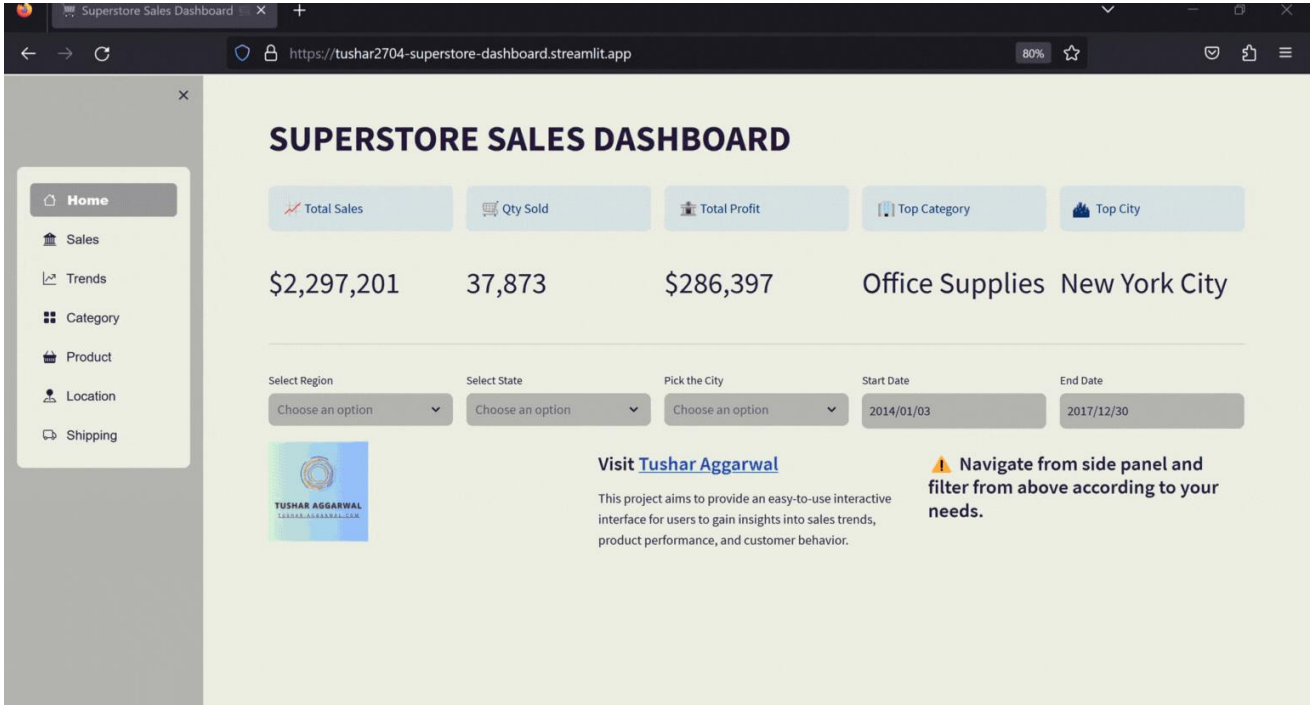
SUPERSTORE SALES DASHBOARD

Source: Sales transaction data (region, product category, customer segment, date, sales, etc.).
Goal: Explore trends, region or product-category performance, customer demographics, etc.

Key KPIs:

- Total sales over time
- Sales by region or category
- Customer segments performance
- Top products by sales

Why it's a good example: More complex filters (region, category, segment) deeper filtering and interactive controls.



IRELAND GENDERGAP DASHBOARD

Source: Irish gender gap reports.
Goal: Visualize and compare pay gaps between male and female employees across Irish entities.

Key KPIs:

- Pay Gap (%) between men and women
- Proportion of men and women in each pay quartile

Why it's a good example:
Important social and ethical issue. Combines categorical (sector, region) and quantitative (pay gap) features

KPIs are relevant to policy discussions

Demonstrates comparison across entities and over time

HANDS-ON: BRAINSTORM YOUR DATASET TOPIC

Preferably you work solo.

If you are interested you can form groups of up to two people.

Pick 2–3 themes you find exciting:

Example: Energy · Health · Education · Mobility · Environment · Economy · Culture · Technology etc.

For each theme, ask:

1. What question would I like to answer?
2. What data might exist for that?

Check available datasets (Dataverse, Kaggle, World Bank, OSM).

Goal: shortlist one dataset per person/team.

Enter it here:

EXPLORATORY DATA ANALYSIS (EDA)

WHAT IS EDA?

Exploratory Data Analysis (EDA) = the process of getting to know your data before modeling or visualization.

Goals

- Understand the **structure** and **content** of the dataset
- Detect **data quality issues** (missing, duplicates, outliers)
- Spot **patterns, trends, and anomalies**
- Generate **hypotheses** for further analysis or KPIs

THE EDA MINDSET

Don’t jump into modeling — explore first

Move between overview and detail

Ask questions like:

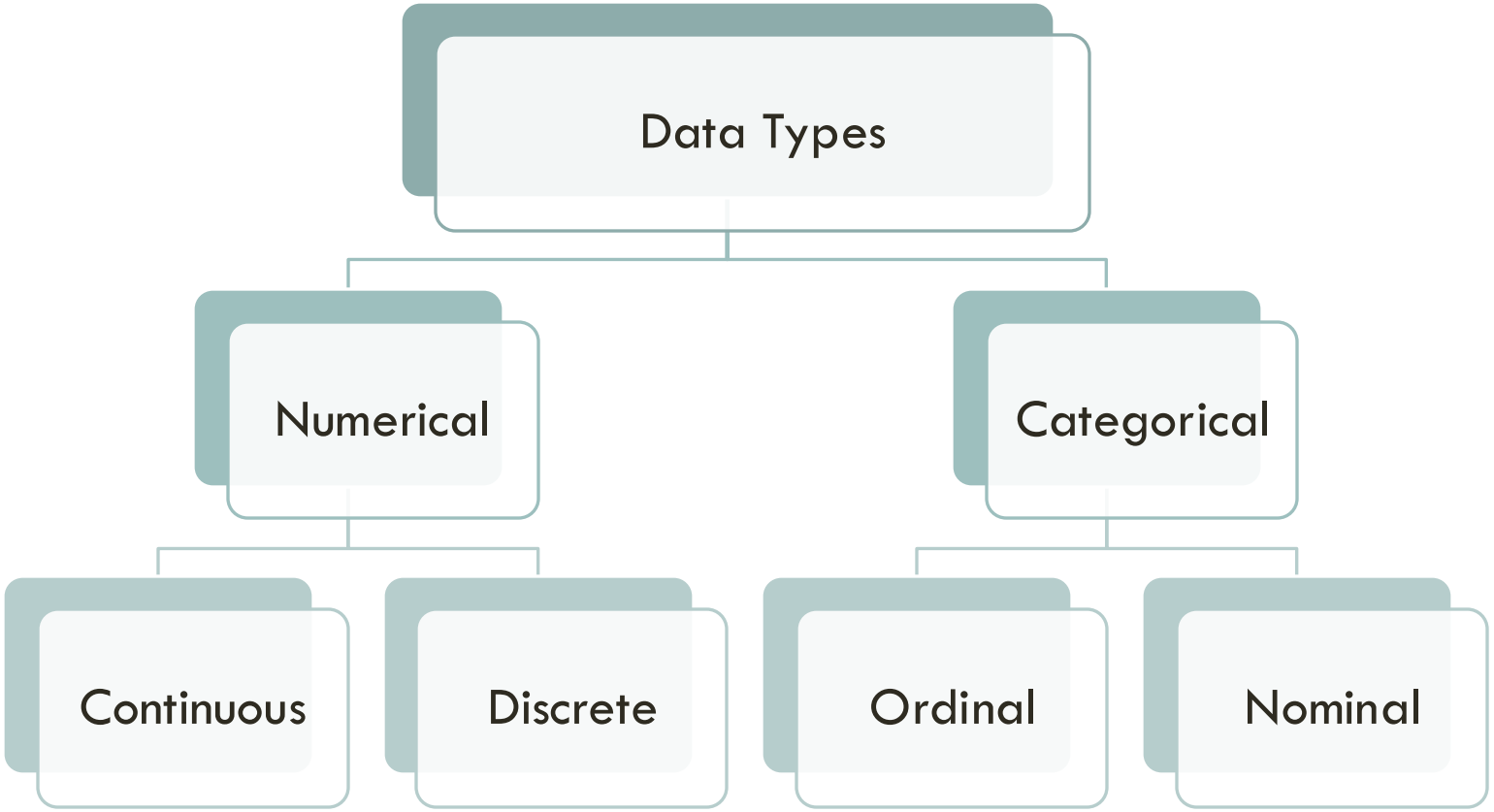
- What does each column represent?
- What are typical and extreme values?
- Are variables correlated?

Combine statistics + visualization + intuition

THE EDA PROCESS

Step	Typical Actions	Python Tools
1. Inspect	load data, view shape, columns, dtypes	`df.head()`, `df.info()`
2. Summarize	descriptive stats, unique counts	`df.describe()`, `value_counts()`
3. Clean	handle missing, duplicates, typos	`df.isna()`, `dropna()`, `fillna()`
4. Visualize	distributions, correlations, trends	`matplotlib`, `seaborn`, `plotly`
5. Interpret	note insights & questions	Markdown cells / slides

ROUGH CLASSIFICATION OF DATA TYPE



QUANTITATIVE EXPLORATION

Use statistics to capture first impressions:

Shape of data: df.shape, df.columns

Data types: numerical, categorical, datetime

Central tendency & spread: mean, median, std

Missing values: percentage per column

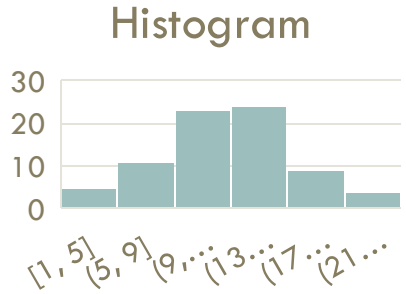
Outliers: use .describe() and quantiles

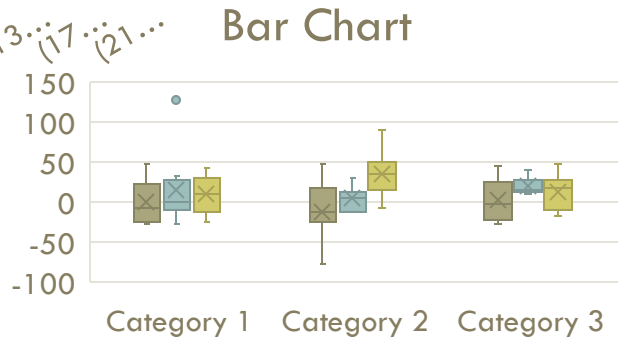
Univariate, Bivariate and Multivariate Analysis: df.corr()

VISUAL EXPLORATION

Key plot types:

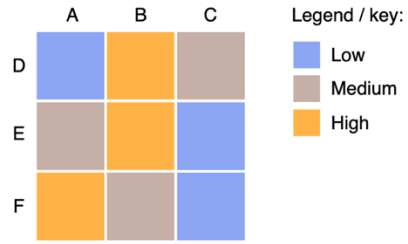
Histograms & Boxplots → distribution and outliers

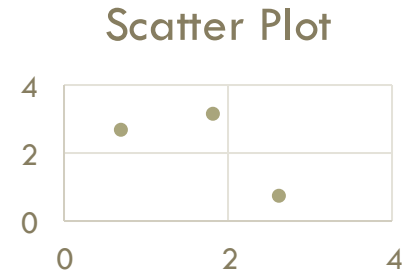




Scatter plots → relationships between two variables

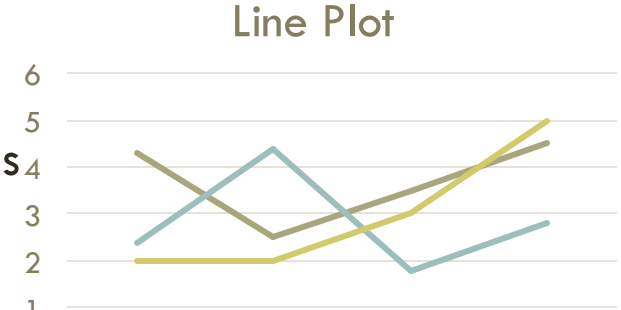
Heatmaps → correlations

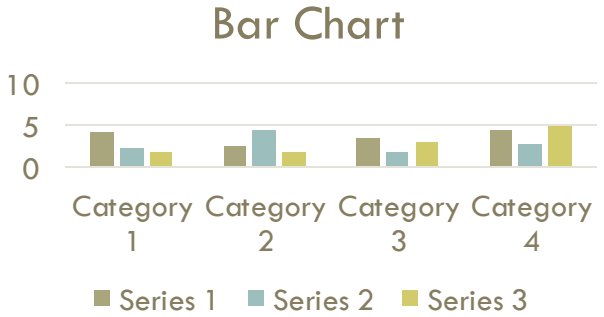




Bar charts / Count plots → categorical frequencies

Line plots → time trends





Heatmap image from: https://datavizcatalogue.com/methods/heatmap.html

FROM EDA TO KPI DESIGN

EDA reveals what is measurable and relevant

Example: if electrification varies by income → KPI: “Electrification rate by income group.”

Each KPI should be supported by a small EDA summary:

- Variable description
- Visualization
- Short interpretation

HANDS-ON ACTIVITY/ NOTEBOOK 1

- Setup + data loading (uses your CSV; falls back to a synthetic demo if the file isn't found).
- Data overview (shape, columns, dtypes, quick head).
- Datetime coercion.
- Descriptive stats for numeric and categorical columns.
- Missingness and duplicates awareness.
- Univariate distributions (histograms with optional log view).
- Categorical frequencies (bar charts of top categories).
- Bivariate relationships (scatter plots) and grouped summaries (numeric vs categorical).
- Correlation matrix (matplotlib imshow).
- Time-series trends (if a datetime column exists).
- Optional geo-style summaries (per region/country).
- Markdown sections for insights and KPI drafts.

HOMEWORK UNTIL NEXT WEEK

Perform an apropriate level of EDA on your own data.

Make a few slides for a ~5min presentation about your data.

See you next week