**📊 EDA & Dashboard Project: From Database to Insights**

**Project Overview**

In this project, students will work on the **end-to-end data analysis process** starting from pulling data from a **MySQL database** into Python, performing **Exploratory Data Analysis (EDA)** with visualizations, cleaning the data, and finally creating a **Power BI dashboard** for presenting insights.

The goal is to give students hands-on experience with the **data analysis pipeline**:

* Connecting databases to Python
* Cleaning and exploring data
* Visualizing findings
* Publishing results through an interactive dashboard – Power BI

At the end of the project, students will present their work as if they were **Data Analysts** delivering results to business stakeholders.

**📌 Steps & Guidelines**

**Step 1: Data Extraction (MySQL → Python)**

* Connect Python to a MySQL database using libraries such as mysql-connector-python or SQLAlchemy.
* 3 Datasets are provided at the end of this file. Pull any one dataset of interest
* Store the data in a **Pandas DataFrame** for further analysis.

✅ Deliverable: Python code that successfully extracts data from MySQL.

**Step 2: Data Cleaning & Preprocessing**

* Check for **missing values, duplicates, and inconsistent formats**.
* Handle missing data (drop or replace).
* Convert data types (e.g., dates, categorical variables) if pandas recognize incorrectly.
* Create new derived variables if required (e.g., profit = sales – cost).

✅ Deliverable: Cleaned dataset saved as a CSV file (to be later imported into Power BI).

**Step 3: Exploratory Data Analysis (EDA) in Python**

* Perform **summary statistics** (mean, median, mode, standard deviation, percentiles).
* Explore **data distributions** (histograms, boxplots).
* Check **relationships** between variables using scatter plots, heatmaps, and correlation analysis.
* Use visualization libraries like matplotlib, seaborn..

✅ Deliverable: Jupyter Notebook with well-documented EDA and visualizations.

**Step 4: Upload Cleaned Dataset to Power BI**

* Import the cleaned dataset (CSV) into Power BI.
* Perform **basic transformations** in Power Query if needed.
* Create relationships if multiple tables are used.

✅ Deliverable: Data model in Power BI.

**Step 5: Dashboard Creation in Power BI**

* Create **interactive visuals** such as bar charts, line charts, maps, KPIs, slicers, and filters.
* Highlight **key insights** found during EDA (e.g., sales trends, customer demographics, performance comparisons).
* Ensure dashboard is **clean, professional, and tells a story**.

✅ Deliverable: Power BI Dashboard (.pbix file).

**Step 6: Presentation & Submission**

* Students should **present their project** to the class as if presenting to a client/business team.
* The presentation should cover:
  + Problem Statement
  + Steps taken (ETL, EDA, cleaning)
  + Key findings & insights
  + Dashboard walk-through
* Final submission should include:
  + Python code (Jupyter Notebook) - documentation
  + Cleaned dataset (CSV)
  + Power BI dashboard file (.pbix)
  + Presentation slides

**✅ Evaluation Criteria**

1. **Data Extraction & Cleaning (20%)** – Proper extraction, handling of missing/duplicate values.
2. **EDA in Python (25%)** – Depth of analysis, quality of visualizations, clarity of insights.
3. **Dashboard in Power BI (30%)** – Interactivity, design, and storytelling with data.
4. **Presentation (15%)** – Clarity, professionalism, ability to explain insights.
5. **Code & Documentation (10%)** – Well-documented Jupyter Notebook and reproducible steps.

**You can Pick any one dataset for the project**. Click on the dataset name to access the link.

**Dataset 1:** [**Retail / E-commerce Sales Data**](https://www.kaggle.com/datasets/carrie1/ecommerce-data)

Data Description:

InvoiceNo - a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation

StockCode - a 5-digit integral number uniquely assigned to each distinct product

Description**-** product name

Quantity - the quantities of each product (item) per transaction

InvoiceDate - the day and time when each transaction was generated

UnitPrice - product price per unit

CustomerID - a 5-digit integral number uniquely assigned to each customer

Country - the name of the country where each customer resides

**Description:**  
Transaction-level sales data from an online or physical retail store.

* Analyze **customer purchase behaviour, product sales trends, seasonal patterns, and revenue generation**.

**Insights they can derive:**

* Which products are the most/least selling?
* Customer segmentation by spend.
* Revenue contribution by country.

**Dataset 2:** [**Banking: Finance Loan Approval Prediction Data**](https://www.kaggle.com/datasets/krishnaraj30/finance-loan-approval-prediction-data) **Consider train dataset**.

**Description:**  
This dataset includes features such as **Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History**, and other demographics and financial details

Explore loan approval patterns, understand demographic trends, and analyze what factors influence lending decisions. **Insights they can derive:**

* Compare average income or credit history scores between approved vs. rejected loans
* Visualize distributions of LoanAmount, ApplicantIncome, or CoapplicantIncome
* Analyze approval rate by gender, education level, marital status, or number of dependents
* Investigate correlation among numeric features.

**Dataset3:** [**Ride-Hailing: Uber Ride Analytics Dashboard**](https://www.kaggle.com/datasets/yashdevladdha/uber-ride-analytics-dashboard?select=ncr_ride_bookings.csv)

* **Description: The dataset “Uber Ride Analytics Dashboard” centers around ride bookings in the NCR region, with over 148,770 bookings across various vehicle types, offering a comprehensive view of ride-sharing operations.**
* **Possible Analytical Angles:**
  + **Hourly/daily booking trends and peak demand periods**
  + **Ride distribution by vehicle type or status (e.g., completed, canceled)**
  + **Average ride duration or distance (if available)**
  + **Cancellation patterns and their correlations (e.g. by time of day, vehicle)**
  + **Rider patterns across weekdays vs. weekends**