Wiley Edge Data Analysis Project

Marco Arnoldi

# Chosen dataset:

I have chosen a public dataset of Brazilian E-Commerce by Olist, a popular department store in Brazil that runs an online marketplace and connects smaller retailers to customers through their site and logistics network.

This dataset was chosen because it contains 9 different entities that can be linked together with a relational database design. Therefore, a relational database can be set up using MySQL, which can then be linked to Python for conducting various data analyses and to PowerBI for reporting data visualizations.

# Objective:

The objective of this project is to demonstrate a strong understanding of data analytics and an aptitude for the various software applications that are required for this type of work. I will import and process the data conduct exploratory data analysis, as well as advanced analysis with predictive models.

Additionally, this document serves as an action plan for the project, outlining and detailing the necessary steps and timeline for completion of the project.

Project Requirements**:**

Various software applications and programming libraries will be needed for this project:

* PowerBI Desktop
* MySQL
* Python 3.10.7
  + (list of packages to be updated)

# Timeline

**Day 1: Monday (Full Day) – Dataset Inspection**

* **Objective**: Create ER diagram and develop plan for project
* **Tasks**:
  + Create a relational schema diagram.
  + Import the data into SQL and set up the database.
  + Verify data consistency and integrity.
* **Outcome**
  + Dataset store in SQL, database structured and populated.

**Day 2: Tuesday (Half Day) - Data Setup**

* **Objective**: Link the SQL database to Python.
* **Tasks**:
  + Link the SQL database to python and import that data.
  + Perform data cleaning and preprocessing.
  + Begin exploratory data analysis.
* **Outcome**: Dataset imported into Python, clear idea of required analyses and visualizations.

**Day 3: Wednesday (Full Day) - Exploratory Data Analysis (EDA) and Modelling in Python**

* **Objective**: Perform exploratory data analysis to understand the dataset.
* **Tasks**:
  + Finish with the relevant EDA.
  + Create Python classes to handle different parts of the project.
  + Create initial advanced analytics in Python to identify patterns and trends.
* **Outcome**: Cleaned dataset and initial exploratory analyses in Python, project classes structured and mostly coded

**Day 4: Thursday (Half Day) - Advanced Analysis and Modeling in Python**

* **Objective**: Apply statistical analysis and modeling techniques.
* **Tasks**:
  + Perform advanced analytics - regression, clustering, or any relevant analysis.
  + Develop predictive models if applicable.
  + Evaluate and refine models.
* **Outcome**: Advanced analysis results, potential predictive models.

**Day 5: Friday (Full Day) - PowerBI Integration and Reporting**

* **Objective**: Generate visual reports for presentation in PowerBI.
* **Tasks**:
  + Integrate the analyzed data with PowerBI.
  + Design and create visual dashboards in PowerBI.
  + Generate reports with key insights and findings.
* **Outcome**: Interactive visual reports and analysis available on PowerBI.

**Python Classes to Create:**

1. **DatabaseConnector:**
   * **Responsibility:** Connects and interacts with the relational database.
   * **Methods:**
     + **connect\_db():** Establishes a connection to the database.
     + **fetch\_data(query):** Retrieves data based on SQL queries.
     + **insert\_data(data):** Inserts data into the database.
     + **close\_connection():** Closes the database connection.
2. **DataProcessor:**
   * **Responsibility:** Handles data cleaning, preprocessing, and EDA.
   * **Methods:**
     + **clean\_data():** Cleans the raw data obtained from the database.
     + **explore\_data():** Performs EDA using statistical methods and visualizations.
     + **preprocess\_data():** Preprocesses the data for modeling.
3. **ModelTrainer:**
   * **Responsibility:** Trains and evaluates machine learning or statistical models.
   * **Methods:**
     + **train\_model():** Trains a model using prepared data.
     + **evaluate\_model():** Evaluates the model's performance using metrics.
     + **tune\_model():** Hyperparameter tuning and optimization.
4. **PowerBIReporter:**
   * **Responsibility:** Prepares data for PowerBI and manages visualization-related tasks.
   * **Methods:**
     + **prepare\_data\_for\_powerbi():** Structures data for efficient PowerBI usage.
     + **create\_visuals():** Generates Python-based visualizations for PowerBI integration.
     + **export\_data\_to\_powerbi():** Pushes the analyzed and processed data to PowerBI.

**Optional additions (if there is enough time)**

* Flask integration
  + Create a basic page where users can input new data into the database using an interface that communicates with our program through flask.

**Potential Roadblocks:**

* **Day 1**: Roadblocks might include data integrity issues, requiring extensive cleaning.
* **Day 2**: Potential issues with linking the SQL database to Python effectively.
* **Day 3**: Potential outcome could be discovering key trends or anomalies within the dataset. Roadblock might include troubles with setting up the Python classes effectively
* **Day 4**: Roadblocks might involve complexities in modeling and analysis, requiring iteration.
* **Day 5:** An outcome could be compelling and actionable visual reports for stakeholders.

# Relational Schema Diagram: