Computer Science Department
CS667 – Practical Data Science (CRN: 72872)
Fall 2025

Project #1 / Due 07-Oct-2025

Performing Exploratory Data Analysis (**EDA**) on data is of paramount importance for every Data Scientist / Data Analyst. Exploratory Data Analysis is often used to uncover various patterns, trends, and insights present in the data and to draw conclusions from it. EDA is the core part when it comes to developing a Machine Learning model. This takes place through analysis and visualization of the data which will be fed to the Machine Learning Model. A Machine Learning Model is as good as the training data - you must understand it if you want to understand your model.

Project Overview
You will carry out a full Exploratory Data Analysis (EDA) on the Retail Sales Dataset from Kaggle. Your aim is to understand the dataset's structure, uncover patterns, generate insights, and communicate findings clearly. This will give you experience with data cleaning, visualization, statistical summarization, and deriving business-relevant insights.

Retail datasets are excellent for working with time-series data, customer behavior, and sales forecasting. The data is often dirty and contains a mix of categorical and numerical features.

- **Area of interest**: Sales forecasting, customer segmentation, and supply chain management.
- **Tasks to perform**:
    - **Advanced EDA**: Analyze sales trends over time, identify seasonality and holiday effects, and detect outliers like unusual spikes in purchases.
    - **Feature engineering**:
        - Extract time-based features (e.g., day of week, month, quarter) from timestamps.
        - Create rolling averages and time-decaying features to model recent customer behavior.
        - Use advanced encoding for categorical data like customer IDs or product categories, such as target encoding or embeddings.
    - **Insights & Business Recommendations**:
        - Analyze sales trends over time, identify seasonality and holiday effects, and detect outliers like unusual spikes in purchases.
        - What patterns did you discover? (E.g. busiest months / days, product lines with highest margin/sales, regional differences etc.)
        - Are there customer segments worth focusing on?
        - Suppose you are the retail business: what operational or strategic actions might you take based on the data?
        - What further data would you want, if you were advising the business, to improve decision-making?

Objectives
By the end of this project, you should be able to:

- Clean and preprocess real retail sales data
- Produce descriptive statistics and visual summaries for both categorical and numeric features
- Identify trends, outliers, and anomalies
- Explore relationships between multiple variables
- Formulate insights that could support business decisions
- Communicate your findings clearly, both in visual form and as a written summary

Write **Python** scripts to complete the above tasks along with their output. All work should be done and submitted in a single (**Colab, Jupyter**) **Notebook.**

How to approach the project

1. **Select a dataset**: you have been given a retail sales customer shopping dataset.
2. **Start with basic EDA**: Understand the data's shape, content, and quality. Check for missing values, duplicates, and general statistics.
3. **Perform advanced EDA**: Dive deeper by visualizing relationships, identifying correlations, and detecting anomalies. This may involve plotting time-series data or analyzing feature distributions.
4. **Engineer new features**: Based on your EDA, create meaningful new features. For example, convert timestamps into useful seasonal features or aggregate transactions into customer-level metrics.
5. **Clean and transform the data**: Use advanced techniques to handle missing data (e.g., model-based imputation), outliers, and skewness (e.g., log or Box-Cox transformations).
6. **Evaluate features**: Use feature importance estimation or selection techniques to identify the most relevant features for a potential model.
7. **Document your process**: Keep detailed notes in a (Jupyter, Colab) Notebook, documenting your steps, findings, and the reasoning behind your feature engineering choices.

Dataset's (retail_sales.xlsx) Metadata Info
Transaction ID: A unique identifier for each transaction, allowing tracking and reference.
Date: The date when the transaction occurred, providing insights into sales trends over time.
Customer ID: A unique identifier for each customer, enabling customer-centric analysis.
Gender: The gender of the customer (Male/Female), offering insights into gender-based purchasing patterns.
Age: The age of the customer, facilitating segmentation and exploration of age-related influences.
Product Category: The category of the purchased product (e.g., Electronics, Clothing, Beauty), helping understand product preferences.
Quantity: The number of units of the product purchased, contributing to insights on purchase volumes.
Price per Unit: The price of one unit of the product, aiding in calculations related to total spending.
Total Amount: The total monetary value of the transaction, showcasing the financial impact of each purchase.