Instacart Basket Analysis - Data Analytics Project

Objective:

This project analysed customer purchasing behaviour using Instacart's open-source grocery dataset to identify trends that could guide marketing and sales strategies. The analysis focused on understanding customer profiles, shopping frequency, spending patterns, and regional differences, ultimately providing insights into how Instacart can better target promotions and improve customer retention.

Project Overview

This project involved working with over 30 million transaction records. The workflow followed a professional data analytics pipeline:

Data Import & Cleaning:

- Imported large datasets (orders, products, order_products_prior, and customers) using Python and Pandas.
- Checked for and treated missing values, duplicates, and inconsistencies.
- Dropped irrelevant PII (e.g., first and last names) to ensure data privacy compliance.

Data Consistency Checks:

- Validated all merges across the datasets.
- Conducted cross-tabulations (pd.crosstab) to confirm consistency between orders, departments, and customer data.
- Logged all checks (missing values, duplicates) in an Excel "Consistency Checks" tab.

Data Merging:

- Combined the four datasets into a master DataFrame (instacart_dataset_complete.pkl).
- Verified total row counts after each merge for accuracy.

Data Derivation:

Created several new columns to improve analysis granularity, including:

- loyalty_flag → classifying customers as New, Regular, or Loyal based on number of orders.

- price_range_loc → grouping products into Low-range, Mid-range, and High-range categories based on price.
- spender_level → separating customers into Low Spender and High Spender groups.
- customer_activity_level → tagging users with High or Low Activity depending on order count.
- customer_profile → defining behavioral profiles such as Young Single Shopper, Budget Family, or Affluent Professional using age, income, and dependents.

Exploratory Data Analysis (EDA):

- Explored numerical and categorical variables to identify key patterns.
- Analyzed which departments received the most orders.
- Investigated spending distribution by time of day, region, and customer loyalty.
- Conducted cross-tab analyses between customer_profile and both region and department to explore behavioral links.

Data Visualization:

Created bar charts and heatmaps in Jupyter using Matplotlib and Seaborn to visualize:

- Top product departments.
- Average spending by customer profile.
- Order distribution by region and loyalty.
- Relationships between age, family status, and order frequency.

Tools & Skills Used

- Languages: Python (Pandas, NumPy, Matplotlib, Seaborn)
- Environment: Jupyter Notebook
- Data Files: .csv, .pkl
- Documentation: Excel for consistency tracking & flow documentation

Concepts Applied:

- Data wrangling and cleaning
- Deriving variables
- Exploratory data analysis
- Visualization and storytelling
- Cross-tabulation and segmentation analysis

Key Deliverables

- Cleaned and merged dataset (instacart_dataset_complete.pkl)
- Full analysis notebook (4-10_ExcelReporting_Elia_Part1.ipynb)
- Visualizations for customer profiles, price ranges, and product departments
- Excel reporting file with population flow and consistency check documentation

Outcome

This project demonstrated the ability to manage and analyse large-scale datasets, build meaningful customer segments, and communicate insights that directly inform marketing decisions. It showcased end-to-end analytical thinking, from raw data cleaning to actionable business recommendations.