Case Study: CoffeeConnect – Data-Driven Insights for Café Chain Expansion

CoffeeConnect is a rapidly growing café chain that specializes in premium coffee offerings across urban centers in India. The company has installed a modern POS (Point-of-Sale) system in all its outlets to track customer purchases, payment methods, and sales patterns.

To make strategic decisions around:

- **inventory,**
- **marketing campaigns,**
- **staffing shifts,**
- **and potential city-wise expansion,**

CoffeeConnect has asked you, a Data Analyst Intern, to analyze two months of transaction-level data collected across their outlets.

You've been given two CSV files:

1. **`index_1.csv` – Transactions from February**
2. **`index_2.csv` – Transactions from March**

Each file includes:

- **`datetime` – Exact timestamp of the transaction**
- **`card` – Anonymized card ID of the customer**

- **`coffee_name` – Name of the coffee purchased**
- **`cash_type` – Payment method (cash/card/UPI)**
- **`money` – Amount spent**
- **`date` – Derived from `datetime`**

You are expected to answer several key business questions through data exploration, filtering, grouping, merging, and visualizations:

1. **Customer Behavior:**
    - **Which coffees are most popular?**
    - **Do people prefer card, cash, or UPI?**
    - **At what time of the day are people most likely to buy coffee?**
2. **Sales Analysis:**
    - **Which coffee earns the most revenue?**
    - **What is the average ticket size for different coffees?**
    - **How do daily sales vary over time?**
3. **Operational Insights:**
    - **Are there peak hours for transactions?**
    - **Which outlets or cards have frequent repeat customers?**
    - **Is there a noticeable difference in customer behavior between February and March?**

4. **Cross-Month Comparison:**
    - **Combine both months and analyze growth trends.**
    - **Track changes in payment preferences.**
    - **Compare sales of specific coffees month-over-month.**

Exercise:

Basic Slicing & Filtering

1. Display all rows where `money > 30` in `index_1.csv`.
2. Filter transactions from `index_1.csv` made with `cash_type` = `'card'`.
3. Find rows in `index_2.csv` where `coffee_name` is either `'Latte'` or `'Americano'`.
4. Retrieve all records from `index_1.csv` where `datetime` is after `"2024-03-01 12:00:00"`.
5. Find the first 5 transactions of `card = 'ANON-0000-0000-0002'` from `index_1.csv`.

Grouping & Aggregations

6. Group both datasets by `coffee_name` and find the **total `money` collected per item**.
7. Count the number of transactions per `cash_type` in both datasets.
8. Find the **average `money` spent per coffee_name** in `index_1.csv`.
9. Which coffee earned the **highest total revenue** in `index_2.csv`?

Merging / Joining

10. Create a new DataFrame with only `card` and `money` columns from both files and **merge** them using `outer` join on `card`.

11. Merge `index_1.csv` and `index_2.csv` using `inner` join on `coffee_name`. How many records do you get?

12. After merging on `coffee_name`, calculate total `money` per `coffee_name`.

Analysis-Based

13. Combine both datasets and find the **total revenue per `date`**.

14. From the combined data, identify the **most frequently purchased coffee**.

15. How many **unique cards** made more than **one purchase**?

16. From the combined data, add a column for `hour` and count transactions by **hour of day**.

17. Plot a **bar chart** of total revenue per `coffee_name` using Seaborn.

18. Plot a **line chart** showing total daily sales from combined dataset.