

Python Data Analysis Assignment

Each question is given below with clear sub-bullets for what to do, what graphs to plot, what trends to extract, and what actions to recommend.

Attached CSV sheets:

1. users_data.csv : customer data of 25 customers (card users)
2. cards_data.csv: Data of cards used by those 25 customers
3. transactions_data.csv: Transaction details of customers with all their cards.

Data Description:

Users_data.csv

Column Name	Description
id	Unique identifier for each user record.
current_age	Current age of the user.
retirement_age	Expected or planned retirement age of the user.
birth_year	Year of birth of the user.
birth_month	Month of birth of the user.
gender	Gender of the user (e.g., Male, Female, Other).
address	Residential address of the user.
latitude	Latitude coordinate of the user's location.
longitude	Longitude coordinate of the user's location.
per_capita_income	Average income per person in the household/region.
yearly_income	Total yearly income of the user.
total_debt	Total outstanding debt of the user.
credit_score	Credit score indicating the creditworthiness of the user.
num_credit_cards	Number of credit cards held by the user.

cards_data.csv

Column Name	Description
id	Unique identifier for each card record.
client_id	Identifier mapping the card to a specific customer.
card_brand	Brand of the card (e.g., Visa, MasterCard, Amex).
card_type	Type of card (e.g., Credit, Debit, Prepaid).
card_number	The primary card number (PAN) associated with the account.
expires	Expiry date of the card (month/year).
cvv	Card Verification Value, used for transaction authentication.
has_chip	Indicates if the card is chip-enabled (Yes/No or Boolean).
num_cards_issued	Number of cards issued for this client/account.
credit_limit	Credit limit assigned to the cardholder.
acct_open_date	Date when the account/card was originally opened.
year_pin_last_changed	Year when the cardholder last updated their PIN.
card_on_dark_web	Flag showing whether the card has been detected in dark web databases (Yes/No).

Transactions_data.csv

Column Name	Description
id	Unique identifier for each transaction.
date	Date when the transaction occurred.
client_id	Identifier linking the transaction to a specific customer.
card_id	Identifier of the card used for the transaction.
amount	Transaction amount in monetary value.
use_chip	Indicates if the transaction was done via chip (Yes/No).

merchant_id	Unique identifier of the merchant where the transaction took place.
merchant_city	City where the merchant is located.
merchant_state	State where the merchant is located.
zip	ZIP/postal code of the merchant's location.
mcc	Merchant Category Code – industry classification of the merchant.
errors	Error flags if the transaction had issues (e.g., declined, incorrect PIN, etc.).

Hero

Task A - Data Pre-Processing

1. Ensure Numeric values for core continuous columns. If not numeric then convert them into appropriate int and float datatypes.
2. Check Data Hygiene throughout the data e.g. Turn “amount” column (with \$ signs) into a clean number column. “\$46.26” --- 46.26
3. Check how many duplicate rows are there, handle them.
4. Display number and columns having missing values. Also visualize and then finally handle them with appropriate action.
5. Convert the use_chip column to have only three values: swipe, chip, or online.
6. How can you clean up city names, so they don't have extra spaces and always start with capital letters?
7. Make sure state codes are always two capital letters (like CA, NY)?
8. How do you turn ZIP codes into a proper 5-digit string, keeping leading zeros?

Task B - Data Analysis

1. Customer Card Profile:
 - Explore the user matrix across card_brand, card_type and credit_limit.
 - Visualize using appropriate graphs for them.
 - Extract insights and recommend actions.

2. Explore Transactions:

- What is the time frame of the data collected?
- Which client has spent the most amount.
- Highest and lowest transactions amounts spent
- Transactions amounts within use_chip category
- Use appropriate Graphs and extract insights.

3. Customer Spend Profile vs. Credit Health

- For each client_id, compute: total spend, average ticket size, transaction count, and monthly frequency.
- Create a dataframe for credit_score, yearly_income, total_debt, num_credit_cards.
- Visualize their co-operation.
- Plot Scatter of average amount spent vs yearly_income.

4. Age Portfolio:

- Create appropriate age bands. What share of customers falls in each age band? How does the average credit_limit vary by band?
- Visualize age vs chip_usage, income and credit score.
- Extract insights and recommend actions.
- Are young customers online heavy?

5. Gender Analysis:

- Avg transaction amount by gender.
- Time-of-day / day-of-week profiles: does one group spend more at night/weekends?
- Visualize with appropriate graphs.
- Suggest marketing campaign ideas.

Artifacts to be generated (For Learners):

- **Jupyter notebook(.ipynb)**
- **All code should be readable and clean.**
- **Questions should be specified in the solution.**
- **Insights and conclusions to be written within the .ipynb file next to code for each question in part B.**
- **Artifacts generated need to be submitted in vLearn on or before the deadline.**
- **File Name:**
 - **File name: firstname_lastname_CPDA_Batch.ipynb**
 - **E.g., Kartik_Mudaliar_CPDA_B1.ipynb**