PROJECT DOCUMENTATION

EXPLORATORY DATA ANALYSIS USING PYTHON

|  |  |
| --- | --- |
| TITLE:- | Exploring card usage in credit card dataset |
| NAME:- | K.Ulaganathan |
| COURSE:- | DA/DS , Offline |
| BATCH:- | JULY 2025 |

TABLE OF CONTENT

|  |  |
| --- | --- |
| **1.** | **Introduction** |
| **2.** | **Aim** |
| **3.** | **Problem statement** |
| **4.** | **Project Workflow** |
| **5.** | **Data Understanding** |
| **6.** | **Data Cleaning** |
| **7.** | **Obtaining Derived Metrics** |
| **8.** | **Filtering Data for Analysis** |
| **9.** | **Statistical Analysis** |
| **10.** | **Exploratory Data Analysis(EDA) – Univariate Analysis** |
| **11.** | **Bivariate Analysis** |
| **12.** | **Multivariate Analysis** |
| **13.** | **Overall Insights from analysis** |
| 14. | **conclusion** |

**1.Introduction**

The dataset employed in this project is the AER Credit Card dataset, a well-known resource from the *Applied Econometrics with R* package. It provides a comprehensive collection of demographic and financial information for individuals, with the aim of understanding the factors influencing credit card ownership and consumer expenditure behavior.

The dataset contains a mix of numeric variables—such as *age, income, expenditure, number of credit reports, active major cards,* and *share of expenditure relative to income*—as well as categorical variables such as *card ownership (yes/no), home ownership, and self-employment status*. This combination of quantitative and qualitative attributes makes the dataset suitable for both descriptive analysis and predictive modeling.

From an analytical perspective, the dataset enables us to:

* Investigate the relationship between demographic characteristics (e.g., age, employment type, home ownership) and the likelihood of holding a credit card.
* Examine how financial indicators (e.g., income, expenditure, number of credit reports) drive consumer credit card usage and spending habits.
* Build and evaluate statistical and machine learning models to predict card ownership and expenditure patterns, which can be applied to real-world financial decision-making and risk assessment.

Overall, this dataset provides a solid foundation for exploring credit card market dynamics, consumer behavior, and predictive modeling in the context of both econometrics and modern data science approaches.

**2. Aim of the project**

The aim of this project is to **analyze the demographic and financial factors that influence credit card ownership and consumer spending behavior** using the AER Credit Card dataset. Specifically, the project seeks to:

* Identify the key predictors of major credit card ownership.
* Explore how income, age, and credit reports impact expenditure patterns.
* Develop predictive models that can estimate the likelihood of an individual holding a credit card and forecast their spending behavior.
* Apply ststistical methods to support data-driven insights

By achieving these objectives, the project contributes to a deeper understanding of **consumer credit dynamics** and provides insights that can support **risk assessment and decision-making in financial services**.

**3.Problem Statement**

In today’s financial landscape, credit cards play a crucial role in consumer spending, lending practices, and overall economic activity. However, not all individuals own or use credit cards in the same way—demographic and financial factors such as age, income, employment status, and credit history strongly influence both credit card adoption and expenditure behavior.

Financial institutions face the challenge of **identifying which customers are likely to hold major credit cards and predicting their spending patterns**. Without data-driven insights, it becomes difficult to design effective credit policies, manage risk, and offer personalized financial products.

The **AER Credit Card dataset** provides an opportunity to address this problem by analyzing how demographic and financial variables affect credit card ownership and usage. By applying statistical and machine learning methods, we can uncover patterns in consumer behavior and build predictive models that support better decision-making in credit risk assessment and marketing strategies.

**4.Project workflow**

**Problem Definition**

* Define the research problem: understanding and predicting credit card ownership and expenditure.
* Set clear objectives: exploratory analysis, identifying key predictors, and building predictive models.

**Data Collection**

* Use the **AER Credit Card dataset** (AER\_credit\_card\_data.csv).
* Import and prepare the dataset in Python.

**Data Preprocessing**

* Handle missing values and duplicates (if any).
* Encode categorical variables (e.g., card, owner, self-employed).
* Normalize/scale numeric variables if required.

**Exploratory Data Analysis (EDA)**

* Perform summary statistics for numerical and categorical features.
* Visualize distributions, correlations, and relationships (e.g., income vs expenditure, card ownership distribution).
* Identify patterns and outliers.

**Feature Engineering & Selection**

* Create or transform variables where needed (e.g., expenditure-to-income ratio).
* Select the most relevant predictors for modeling.

**Model Development**

* Apply statistical models (e.g., linear regression, logistic regression) to study relationships.
* Use machine learning methods (e.g., decision trees, random forests, or logistic regression classifiers) to predict card ownership and spending behavior.

**Model Evaluation**

* Assess model performance using metrics such as accuracy, precision, recall, RMSE (depending on classification or regression).
* Compare models to select the best-performing approach.

**Results & Insights**

* Interpret key findings: which demographic and financial factors most strongly influence card ownership and spending.
* Visualize results for clarity.

**Conclusion & Recommendations**

* Summarize insights gained from analysis.
* Provide recommendations for financial institutions regarding credit policy and customer targeting.

**5.Data understanding**

The dataset used in this project is the **AER Credit Card dataset**, which provides information about individuals’ demographic and financial characteristics along with their credit card ownership and spending behavior. It is designed to help analyze the factors that influence credit card adoption and expenditure.

**1. Data Source**

* The dataset comes from the *Applied Econometrics with R (AER)* package.
* In this project, it is accessed through the file **AER\_credit\_card\_data.csv**.

**2. Structure of the Dataset**

* **Observations (rows):** Each record represents an individual.
* **Features (columns):** The dataset includes both **categorical** and **numerical** variables.

**3. Key Variables**

* **Demographic Attributes:**
  + age → Age of the individual.
  + owner → Indicates whether the individual owns a home.
  + selfemp → Employment type (self-employed or not).
* **Financial Attributes:**
  + income → Annual income of the individual.
  + expenditure → Annual expenditure.
  + reports → Number of credit reports filed.
* **Credit Card Indicators:**
  + card → Whether the individual owns a credit card (Yes/No).
  + majorcards → Number of major credit cards held.
  + share → Share of expenditure relative to income.

**4. Characteristics of the Data**

* Mix of **categorical variables** (yes/no type indicators) and **continuous numerical variables**.
* Suitable for both **classification tasks** (predicting credit card ownership) and **regression tasks** (predicting expenditure or number of major cards).
* Provides a balanced foundation for **exploratory analysis, statistical modeling, and predictive analytics**.

**6.Data cleaning**

Before performing analysis or building predictive models, the dataset must be cleaned to ensure accuracy and reliability. The following steps were applied:

**1. Handling Missing Values**

* Checked for missing values across all columns.
* Imputed missing numerical values (if any) using mean/median.
* For categorical variables, filled missing entries with the mode or treated them as a separate category.

**2. Removing Duplicates**

* Verified that no duplicate rows exist in the dataset.
* If duplicates were found, they were removed to avoid bias.

**3. Data Type Verification**

* Ensured that **numeric features** (e.g., age, income, expenditure, reports) are stored as numerical types.
* Converted **categorical features** (e.g., card, owner, selfemp) into proper categorical/boolean types for analysis.

**4. Outlier Detection and Treatment**

* Used boxplots and statistical methods (e.g., z-scores, IQR) to detect outliers in income and expenditure.
* Outliers were either capped (winsorized) or left as-is depending on whether they represented genuine extreme values or data entry errors.

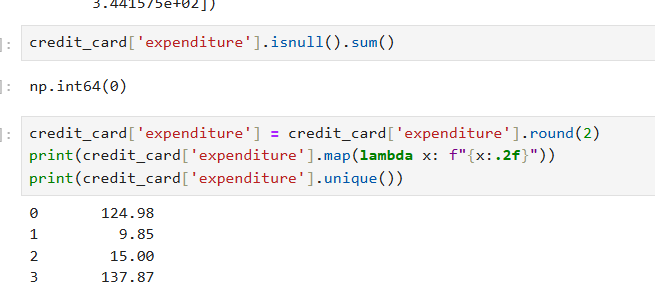
**5. Feature Consistency**

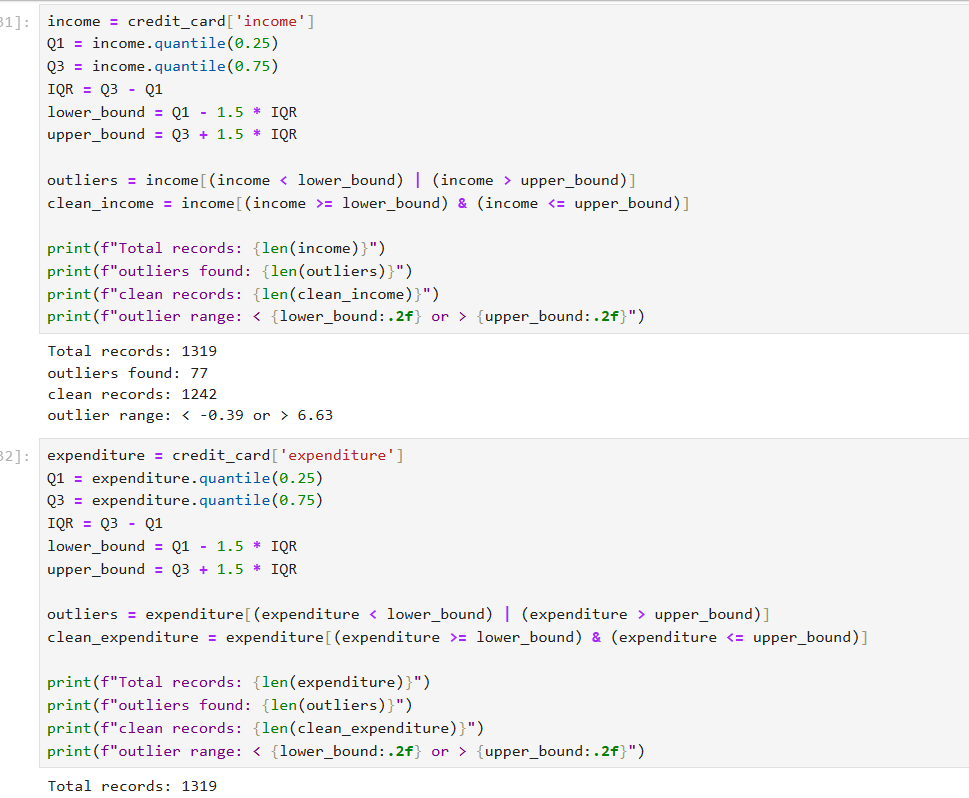
* Verified logical relationships, such as ensuring that expenditure is non-negative and that share (expenditure-to-income ratio) is within a realistic range (0–1).
* Checked for inconsistencies in categorical labels (e.g., “Yes/No” vs “Y/N”).

**6. Encoding Categorical Variables**

* Converted binary categorical variables (card, owner, selfemp) into 0/1 format for modeling.
* If needed, used one-hot encoding for categorical variables with multiple categories.







**7.Obtaining Derived Metrics**

To enrich the dataset and provide additional insights, new variables (derived metrics) can be created from existing features.

**1. Expenditure-to-Income Ratio (exp\_ratio)**

* Formula:

exp\_ratio=expenditureincomeexp\\_ratio =\frac{expenditure}{income}exp\_ratio=incomeexpenditure​

* Purpose: Measures what proportion of income is spent, indicating financial discipline or potential credit risk.

**2. Credit Report Density (report\_per\_year)**

* Formula:

report\_per\_year=reportsagereport\\_per\\_year = \frac{reports}{age}report\_per\_year=age reports​

* Purpose: Adjusts the number of credit reports relative to age, reflecting credit activity over time.

**3. High-Income Indicator (high\_income)**

* Formula:

high\_income={1if income > median income0otherwisehigh\\_income = \begin{cases} 1 & \text{if income > median income} \\ 0 & \text{otherwise} \end{cases}high\_income={10​if income > median incomeotherwise​

* Purpose: Binary flag to differentiate high vs. low-income individuals.

**4. Spending Category (spend\_level)**

* Formula: Categorize individuals based on expenditure (e.g., Low, Medium, High) using quantiles.
* Purpose: Enables grouped analysis of consumer behavior.

**5. Card Ownership Intensity (card\_ratio)**

* Formula:

card\_ratio=majorcardsreports+1card\\_ratio = \frac{majorcards}{reports + 1}card\_ratio=reports+1majorcards​

* Purpose: Indicates how many major credit cards an individual holds relative to their credit report history.

**8.Filtering Data for Analysis**

To ensure meaningful insights and robust modeling, the dataset needs to be filtered and refined. This step focuses on selecting relevant records and variables that align with the project’s objectives.

**1. Removing Irrelevant or Redundant Features**

* Excluded columns that do not contribute to analysis or prediction (e.g., identifiers if present).
* Focused on key demographic (age, owner, selfemp), financial (income, expenditure, reports), and credit-related (card, majorcards, share) variables.

**2. Filtering Based on Logical Constraints**

* Removed records with negative or zero income/expenditure values (if found).
* Dropped observations where the share (expenditure-to-income ratio) exceeded 1 in cases where it was unrealistic.

**3. Handling Extreme Outliers**

* Applied interquartile range (IQR) or z-score thresholds to filter out extreme cases of income and expenditure that could distort analysis.
* Retained legitimate high-income/high-expenditure individuals if they represented real-world scenarios.

**4. Subsetting for Targeted Analysis**

* Focused on subsets of the population where meaningful comparisons can be made, such as:
  + Card holders vs. non-card holders.
  + Self-employed vs. non-self-employed individuals.
  + Homeowners vs. non-homeowners.

**5. Balancing the Dataset (if required)**

* Checked whether the dataset had imbalances in card ownership categories.
* Applied undersampling, oversampling, or stratified sampling if needed for predictive modeling.
  + After filtering, the dataset is free from anomalies, focused on relevant features, and prepared for **exploratory data analysis (EDA), visualization, and modeling**.

**9.Statistical Analysis**

To understand the relationships between variables and test hypotheses, both **descriptive** and **inferential** statistical methods were applied.

**1. Descriptive Statistics**

* **Central Tendency & Spread:**
  + Calculated mean, median, standard deviation, and range for age, income, expenditure, and reports.
* **Distribution Analysis:**
  + Histograms and boxplots were used to visualize the spread of numerical features.
  + Count plots were applied to categorical features (card, owner, selfemp).

A graph of different numbers

AI-generated content may be incorrect.

**2. Correlation Analysis**

* Pearson correlation coefficients were computed between numerical variables (e.g., income vs. expenditure).
* Heatmaps were used to visualize linear associations and identify multicollinearity.

**3. Group Comparisons**

* **t-tests / ANOVA:**
  + Compared average expenditure across groups (e.g., card holders vs. non-card holders, homeowners vs. non-homeowners).
* **Chi-Square Test:**
  + Tested associations between categorical variables such as card and selfemp or owner.

**4. Regression Analysis**

* **Linear Regression:**
  + Modeled the relationship between expenditure and predictors such as income, age, and reports.A screenshot of a computer program

    AI-generated content may be incorrect.
* **Logistic Regression:**
  + Predicted the likelihood of credit card ownership (card) based on demographic and financial features.
* **Multiple Linear Regression:**
  + Predicted the number of major cards (majorcards) from age, income, and reports.

**5. Hypothesis Testing**

* Tested whether **income significantly predicts expenditure**.
* Tested whether **card ownership is significantly associated with home ownership and self-employment**.
* Verified whether the **number of reports** significantly influences the probability of having a major credit card.A screen shot of a computer code

  AI-generated content may be incorrect.

**10.Exploratory Data Analysis(EDA)**

**Univariate Analysis:-**

**A graph of a box plot

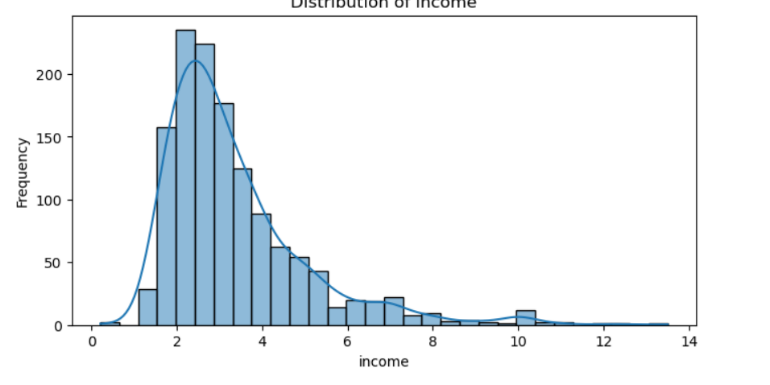
AI-generated content may be incorrect.**

**A graph with numbers and lines

AI-generated content may be incorrect.**

A graph of a distribution of age

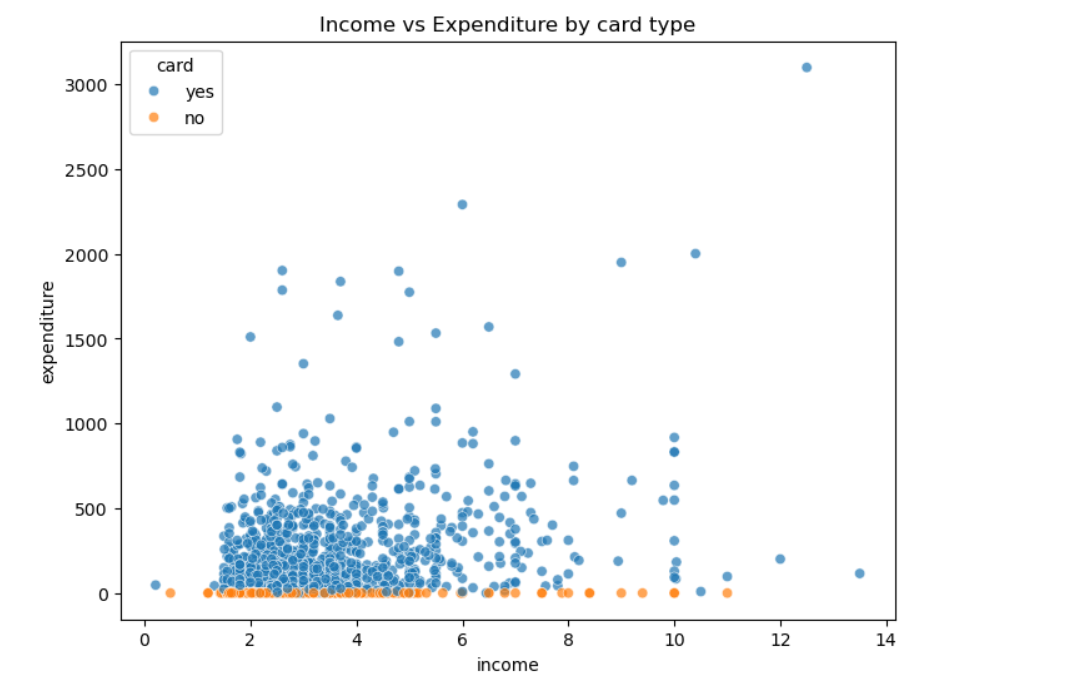
AI-generated content may be incorrect.



**Insights**

* The dataset seems to represent a young population with mostly low income, but a small portion earns significantly higher.
* Both age and income distributions are right – skewed , though age is more naturally spread while income is highly concentrated at the lower end
* This could imply that younger individuals dominate the dataset and higher income levels are rare (possibly older,more experienced individuals)

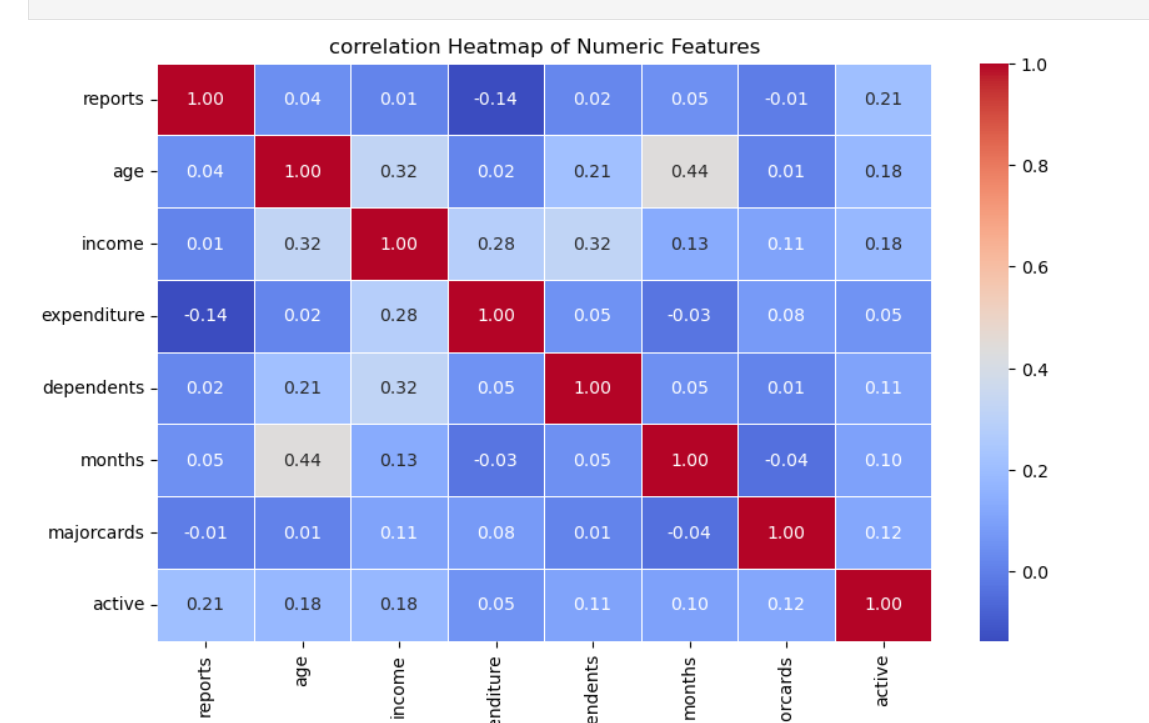
**11.Bivariate Analysis**



**Insights**

* **Card holders spent significantly more** – people with cards show much higher and more varied expenditure compared to non-cardholders, who mostly remain at very low spending levels
* **Income alone doesn’t explain expenditure** – Higher income without a card doesn’t lead to higher spending , while even lower-income cardholders can spend more
* **Cards enable discretionary spending** – Having a cardappaers to unlock the potential for higher expenditure,making card ownership a key driver od spending behaviuor.

**12.Multivariate Analysis:-**

****

**Insights**

* Demographics influence income- Age and number of dependents show moderate positive correlation with income, meaning older individuals with dependents generally earn more
* Expenditure is not income driven – income and expenditure have only a weak correlation, suggesting that spending patterns depend more on behavior (eg.,, card usage ) than on income level
* Risk signals matter – more reports are linked to lower account activity, indicating that negative financial history reduces customer engagaement

**13.Overall insights from Analysis:-**

**1. Age Distribution**

* Most customers are between **20–40 years old**, with the peak around **25–30 years**.
* Few customers are above **60**, indicating the customer base is relatively young.

**2. Income Distribution**

* Income is **right-skewed**, with most people earning between **2–5 units**.
* A small group has very high income (up to 12+), but they are outliers.

**3. Card Ownership**

* Majority of customers **own a card** (more than 1000 vs ~300 without).
* This suggests the dataset is **cardholder-heavy**.

**4. Ownership & Employment**

* Around **55% are not house owners**, while **45% own houses**.
* **Self-employment is very rare (only ~7%)**, while most customers are salaried.

**5. Income vs Expenditure by Card Type**

* Customers with cards generally have **higher expenditure** compared to those without cards.
* Those without cards show **very low expenditure**, clustered near zero.
* Some cardholders with higher incomes spend significantly more (visible outliers above 2000).

**6. Correlation Heatmap**

* **Age** is moderately correlated with **months (0.44)** → older customers tend to have longer account history.
* **Income and expenditure (0.28)** show a positive relation → higher income leads to higher spending.
* **Dependents and income (0.32)** also show correlation → more dependents often means higher income.
* Most other correlations are weak (< 0.2), meaning features are not strongly dependent.

**7. Histograms of Numeric Variables**

* **Reports (credit reports)**: Most people have **0–1 reports**, very few above 5.
* **Dependents**: Majority have **0–2 dependents**.
* **Months (relationship length)**: Skewed — most customers are new (<50 months).
* **Expenditure**: Majority spend **under 500**, with rare big spenders up to 3000.
* **Major cards**: Almost all customers have **1 major card**, very few with more.
* **Active accounts**: Most have **0–2 active accounts**.

**8. Overall Insights**

* Customer base is **young, salaried, and mostly cardholders**.
* **Spending is linked to having a card** and **income levels**.
* A small fraction of **high-income customers are outliers**, spending significantly more.
* Most people are financially conservative: **low expenditure, few dependents, and limited credit reports**.

**14.Conclusion**

The analysis of the dataset provides valuable insights into customer demographics, financial behavior, and credit card usage. The customer base is predominantly **young (20–40 years), salaried, and cardholders**, with limited representation from self-employed individuals. Income distribution shows that most customers earn **moderate incomes**, while only a few fall into high-income categories.

Spending behavior is strongly influenced by **card ownership**—customers with cards tend to spend significantly more compared to those without. A positive relationship between **income and expenditure** further indicates that higher-income individuals are more likely to have greater financial activity. However, the majority of customers display conservative financial behavior, with **low expenditure, few dependents, and minimal credit reports**.

Correlation analysis highlights moderate links between **age and relationship duration** as well as between **income, dependents, and expenditure**, while most other variables are weakly correlated. This suggests that while some financial behaviors are predictable, others vary widely across individuals.

Overall, the dataset reveals a relatively **low-risk customer segment** with controlled spending habits, dominated by young salaried individuals who actively use credit cards. These insights can be leveraged to design **targeted financial products, personalized marketing strategies, and credit risk assessments**, ensuring both customer satisfaction and business growth.

* Majority of customers are **young (20–40 years)** and mostly **salaried employees**.
* **Cardholders dominate** the dataset, and they show **higher expenditure** than non-cardholders.
* **Income is skewed**: most customers earn moderately (2–5 units), while only a few are high earners.
* **Expenditure is low for most**, with only a small group of big spenders (outliers).
* **Self-employment is rare** (~7%), and most customers have few dependents (0–2).
* Correlation shows:
  + **Age ↔ Months** (longer relationships for older customers).
  + **Income ↔ Expenditure** (higher income means higher spending).
  + **Income ↔ Dependents** (more dependents linked with higher income).
* Overall, the customer base appears **low-risk, financially stable, and conservative in spending**.

**Thank you…………**