# Personal Loan Classification Using Machine Learning Models: Decision Tree

Yen Yung Geszvain

## Supervised Learning Problem Description

# Prediction

- Example Bank is a US bank with a growing customer base primarily composed of depositors (liability customers). The bank aims to expand its small number of borrowers (asset customers) to increase its loan business and enhance revenue through loan interest. Specifically, management seeks strategies to convert liability customers into personal loan customers while maintaining their deposit relationships.

- A successful campaign last year achieved a conversion rate of over 9% for liability customers, prompting the retail marketing department to develop more targeted marketing initiatives to improve this ratio. As a data scientist at Example Bank, your task is to build a model to help identify potential customers who are more likely to take out loans.

# Preview Data

| ID | Age | Experience | Income | ZIP Code | Family | CCAvg | Education | Mortgage | Personal Loan | Securities Account | CD Account | Online | CreditCard |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 25 | 1 | 49 | 91107 | 4 | 1.6 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2 | 45 | 19 | 34 | 90089 | 3 | 1.5 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | 39 | 15 | 11 | 94720 | 1 | 1.0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 35 | 9 | 100 | 94112 | 1 | 2.7 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 35 | 8 | 45 | 91330 | 4 | 1.0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 |

| ID | Age | Experience | Income | ZIP Code | Family | CCAvg | Education | Mortgage | Personal Loan | Securities Account | CD Account | Online | CreditCard |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4996 | 29 | 3 | 40 | 92697 | 1 | 1.9 | 3 | 0 | 0 | 0 | 0 | 1 | 0 |
| 4997 | 30 | 4 | 15 | 92037 | 4 | 0.4 | 1 | 85 | 0 | 0 | 0 | 1 | 0 |
| 4998 | 63 | 39 | 24 | 93023 | 2 | 0.3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4999 | 65 | 40 | 49 | 90034 | 3 | 0.5 | 2 | 0 | 0 | 0 | 0 | 1 | 0 |
| 5000 | 28 | 4 | 83 | 92612 | 3 | 0.8 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |

# Basic Data Elements

- Rows: 5000
- Columns: 14
- Features:
- [‘ID,' 'Age,' 'Experience,' 'Income,' 'ZIP Code,' 'Family,' 'CCAvg,' 'Education,' 'Mortgage,' 'Personal Loan,' ‘Securities Account,’ 'CD Account,' 'Online,' ‘CreditCard']
- No missing value

# Basic Data Elements

Column Details:

1. **ID**: Unique identifier for each record (int64).
2. **Age**: Age of the individual (int64).
3. **Experience**: Years of professional experience (int64).
4. **Income**: Annual income in monetary units (int64).
5. **ZIP Code**: ZIP code of the individual's residence (int64).
6. **Family**: Size of the family (int64).
7. **CCAvg**: Average monthly credit card spending (float64).
8. **Education**: Level of education (1 = Undergrad, 2 = Graduate, 3 = Advanced/Professional) (int64).
9. **Mortgage**: Mortgage value (int64).
10. **Personal Loan**: Indicator if a personal loan is taken (1 = Yes, 0 = No) (int64).
11. **Securities Account**: Indicator for having a securities account (1 = Yes, 0 = No) (int64).
12. **CD Account**: Indicator for having a certificate of deposit account (1 = Yes, 0 = No) (int64).
13. **Online**: Indicator if the customer uses online banking services (1 = Yes, 0 = No) (int64).
14. **CreditCard**: Indicator if the customer has a credit card (1 = Yes, 0 = No) (int64).

- The dataset is complete, with no missing values across any columns. All features are numerical in nature. The target variable for this analysis is Personal Loan, which indicates whether a personal loan has been taken (1 = Yes, 0 = No).

# Data Preprocessing

- Remove the ID column, as it is not needed
- Fix data types; convert these columns to category type: 'Personal_Loan', ‘Securities_Account',' Family', ‘CD_Account,' ‘Online,' ‘CreditCard,' ‘ZIP_Code,' 'Education'

Column Details:

1. **Age** *(int64)*: Age of the individual.
2. **Experience** *(int64)*: Years of professional experience.
3. **Income** *(int64)*: Annual income in monetary units.
4. **ZIP_Code** *(category)*: Categorical variable indicating the geographical region.
5. **Family** *(category)*: Categorical variable representing family size.
6. **CCAvg** *(float64)*: Average monthly credit card spending.
7. **Education** *(category)*: Categorical variable indicating education level (1 = Undergraduate, 2 = Graduate, 3 = Advanced/Professional).
8. **Mortgage** *(int64)*: Mortgage value in monetary units.
9. **Personal_Loan** *(category)*: Target variable indicating whether a personal loan has been taken (1 = Yes, 0 = No).
10. **Securities_Account** *(category)*: Indicates the presence of a securities account (1 = Yes, 0 = No).
11. **CD_Account** *(category)*: Indicates the presence of a certificate of deposit account (1 = Yes, 0 = No).
12. **Online** *(category)*: Indicates whether the customer uses online banking services (1 = Yes, 0 = No).
13. **CreditCard** *(category)*: Indicates whether the customer owns a credit card (1 = Yes, 0 = No).

Memory Usage:

The dataset occupies approximately **260.1 KB** in memory, optimized due to the presence of categorical data types.

# Exploratory Data Analysis

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Age | 5000.0 | 45.338400 | 11.463166 | 23.0 | 35.0 | 45.0 | 55.0 | 67.0 |
| Experience | 5000.0 | 20.134600 | 11.415189 | 0.0 | 10.0 | 20.0 | 30.0 | 43.0 |
| Income | 5000.0 | 73.774200 | 46.033729 | 8.0 | 39.0 | 64.0 | 98.0 | 224.0 |
| CCAvg | 5000.0 | 1.937938 | 1.747659 | 0.0 | 0.7 | 1.5 | 2.5 | 10.0 |
| Mortgage | 5000.0 | 56.498800 | 101.713802 | 0.0 | 0.0 | 0.0 | 101.0 | 635.0 |

**Observations**

**Age:** Customer ages range from 23 to 67 years, with both the mean and median age approximately 45 years.

**Experience:** Professional experience varies from 0 to 43 years, with a mean and median of approximately 20 years. The maximum value of 43 years may warrant verification for accuracy.

**Income:** Annual incomes span from $8,000 to $224,000, with a mean of $73,000 and a median of $64,000. The maximum income value of $224,000 should be reviewed to ensure validity.

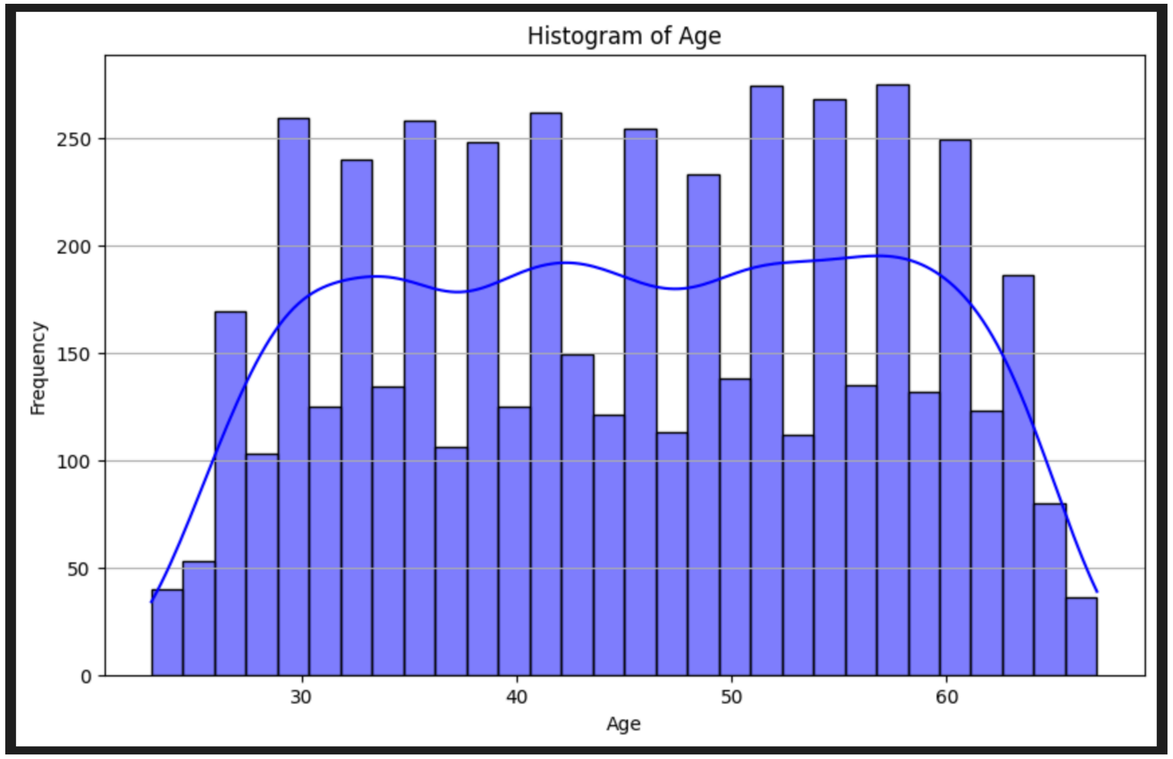
**Mortgage:** The maximum recorded mortgage value is $635,000, which may need further validation. The dataset does not provide summary statistics for average or median mortgage values.

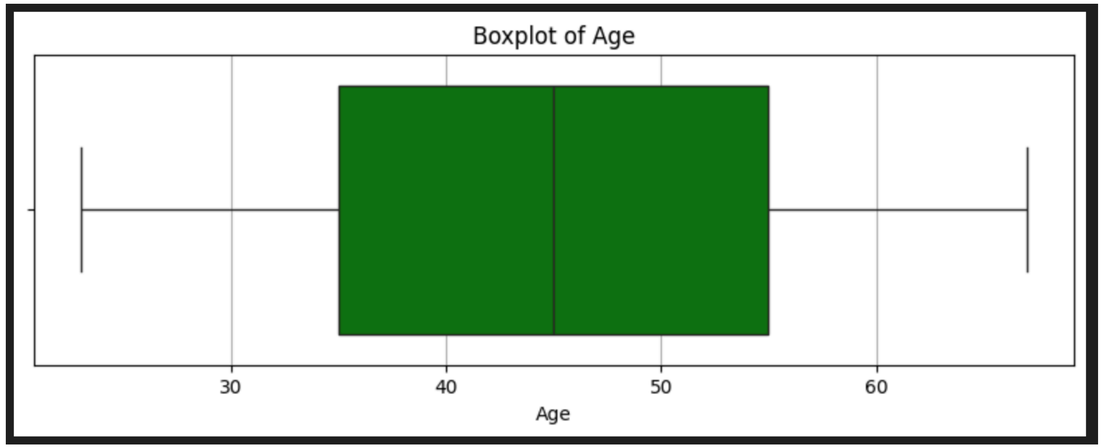
**Credit Card Spending:** Average monthly credit card expenditures range from $1 to $10,000, with a mean of approximately $1,900 and a median of $1,500.

**Geographical Data:** A total of 1,095 customers are located in Los Angeles County. Loan History: 480 customers have previously borrowed a loan.

# Exploratory Data Analysis

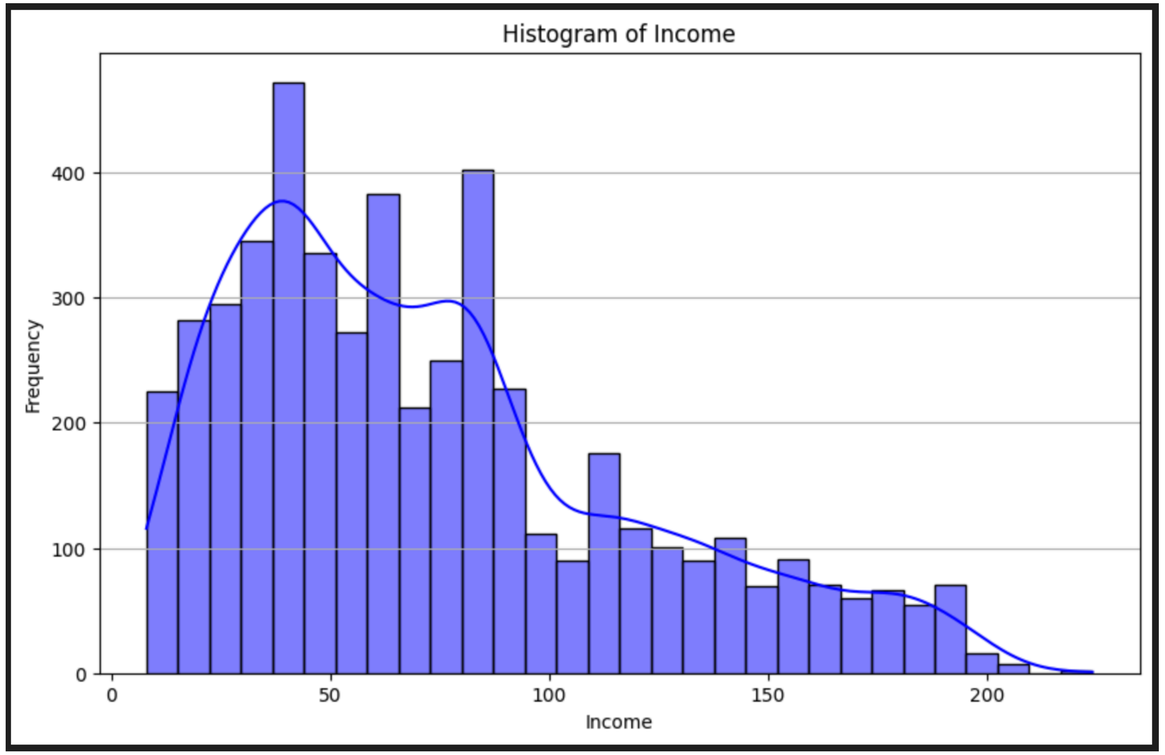
## Univariate Analysis

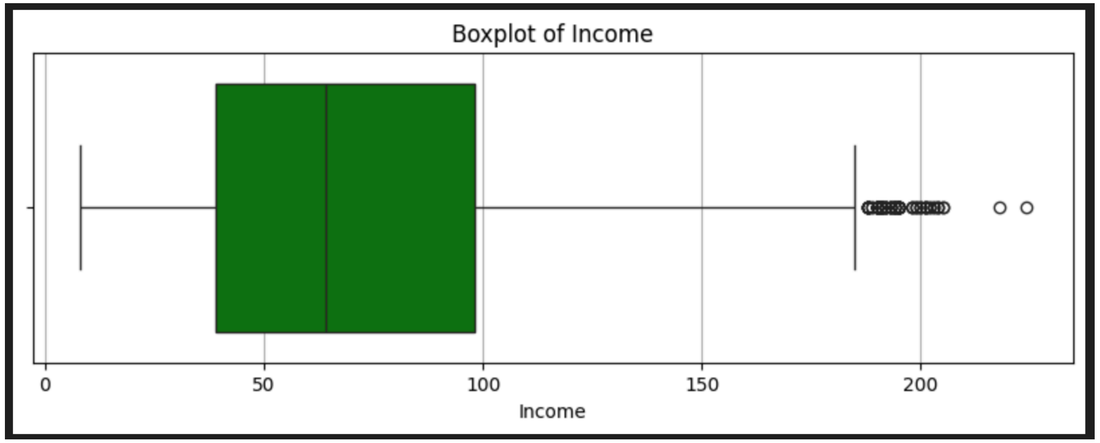


# Exploratory Data Analysis

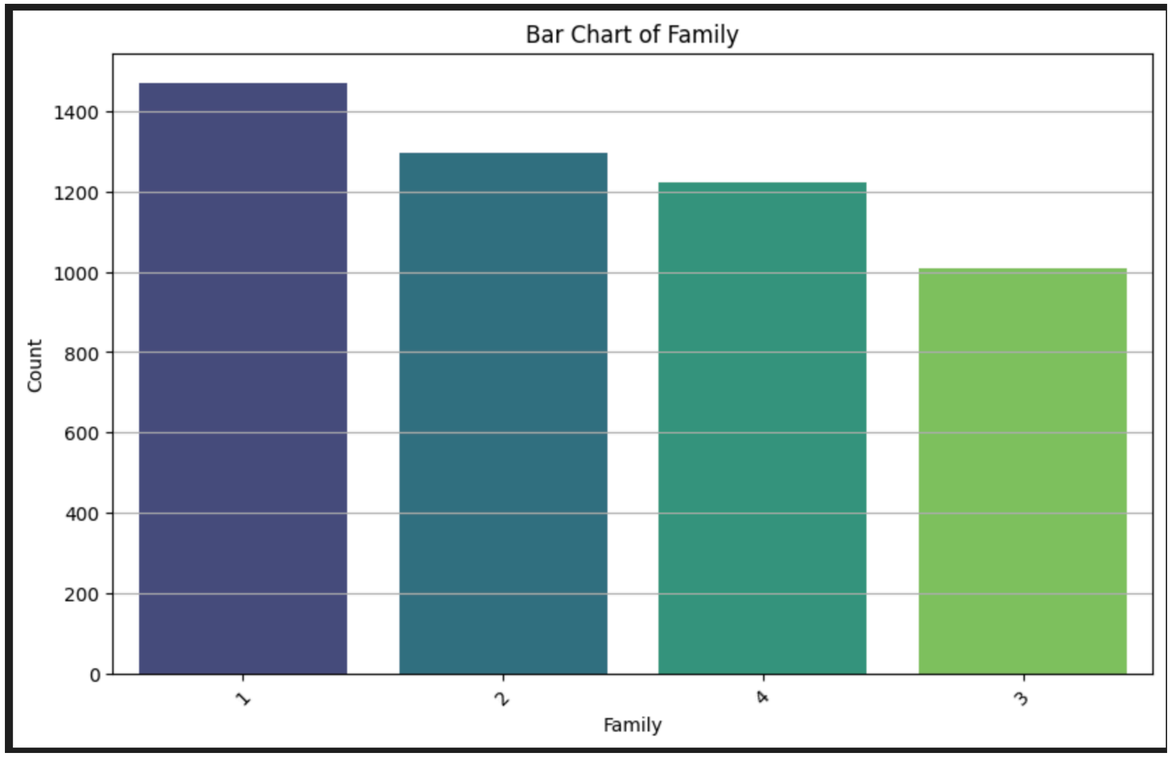
## Univariate Analysis



# Exploratory Data Analysis

## Univariate Analysis



# Exploratory Data Analysis

## Univariate Analysis

**Observations**

**Age and Experience:** Both variables exhibit similar distributions, characterized by spikes at intervals of 5 years, likely due to data rounding or grouping.

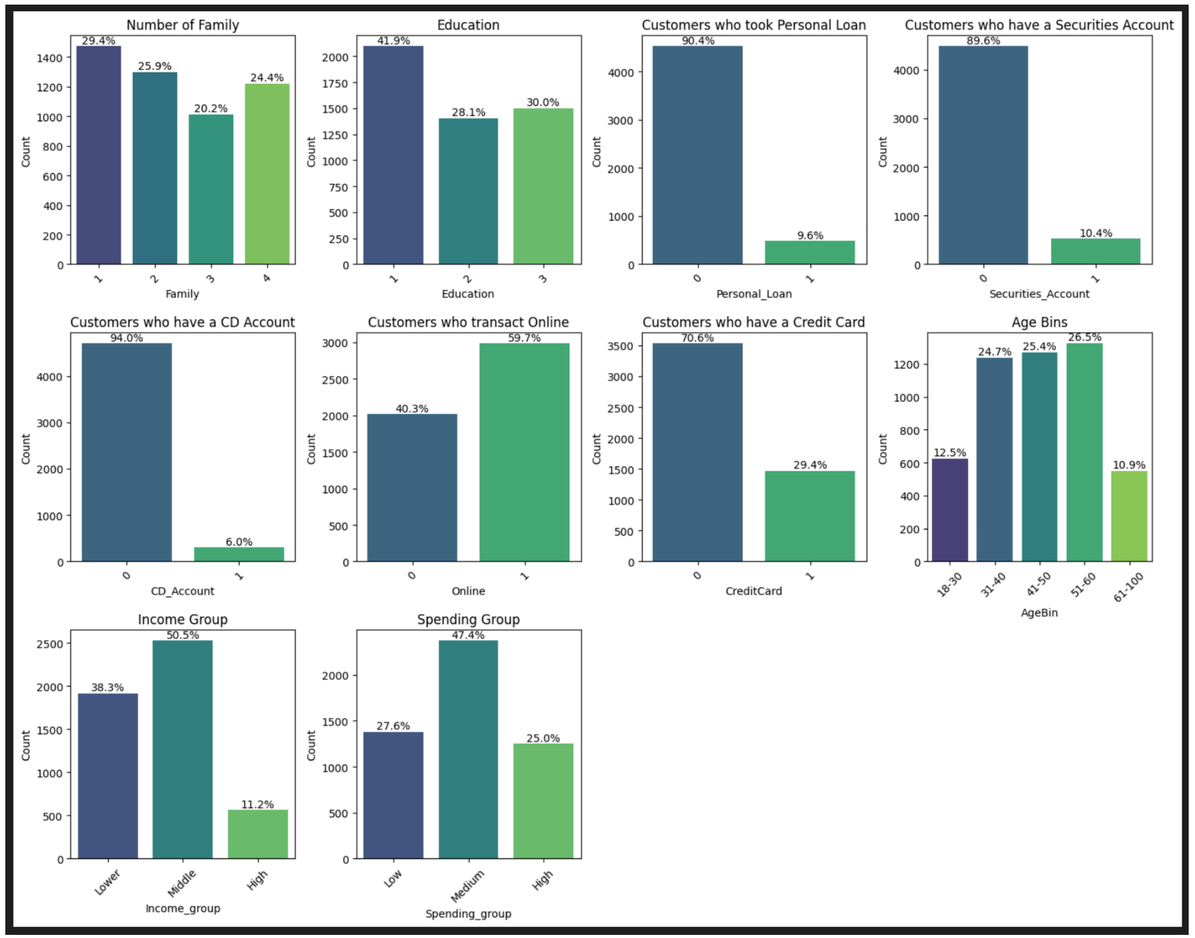
**Income:** The distribution of income is right-skewed, with a few high-value outliers that could distort the analysis. These outliers may be clipped or capped to reduce their impact.

**Average Monthly Credit Card Spending:** This variable is also right-skewed, with numerous high-value outliers. Clipping or capping these extreme values can help achieve a more balanced representation.

**Mortgage:** The majority of the values are zero, indicating that many customers do not have a mortgage. However, for non-zero values, the distribution is right-skewed with significant outliers on the higher side. These outliers could also benefit from clipping to enhance data reliability.

# Exploratory Data Analysis

## Univariate Analysis



# Exploratory Data Analysis

## Univariate Analysis

**Observations**

**Family:** The majority of customers have two members in their family (29.4%), followed by three members (25.9%) and one member (24.4%).

**Education:** A significant portion of customers have a bachelor's degree (41.9%), while 30.0% have a master's degree.

**Personal Loan:** A substantial 90.4% of customers have taken a personal loan.

**Securities Account:** A high percentage (89.6%) of customers hold a securities account.

**CD Account:** A large majority (94.0%) of customers have a CD account.

**Online Transactions:** A considerable number of customers transact online (59.7%).

**Credit Card:** A significant proportion (70.6%) of customers have a credit card.

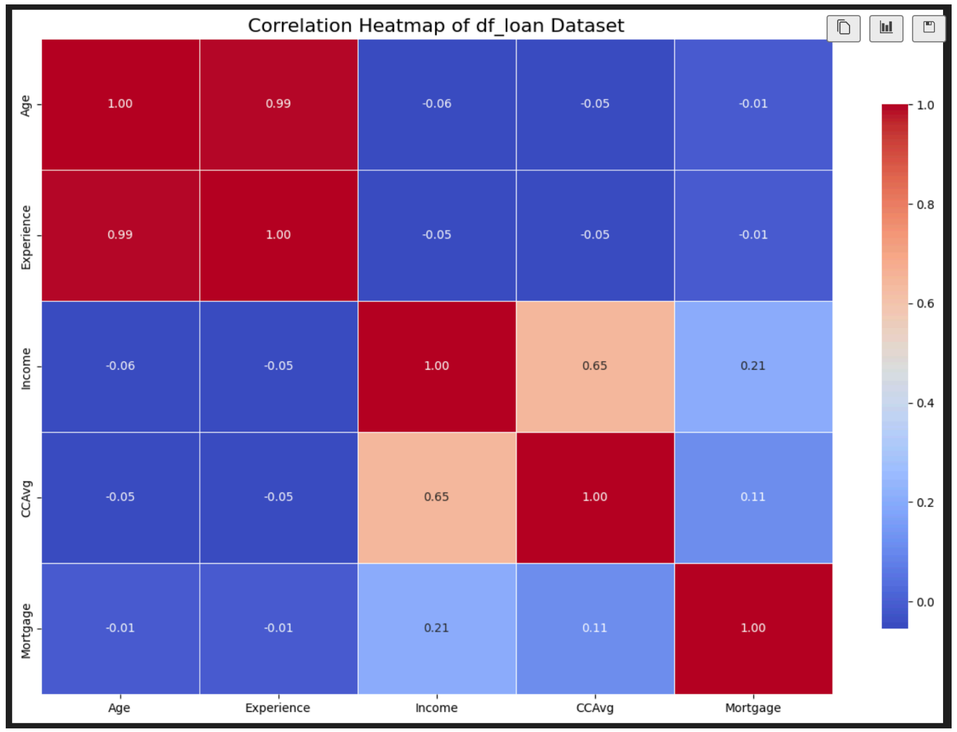
**Age Bins:** The largest age group is 41-50 (29.4%), followed by 31-40 (26.5%) and 51-60 (24.7%).

**Income Group:** The majority of customers fall in the middle income group (50.5%), followed by the high income group (38.3%) and the low income group (11.2%).

**Spending Group:** The highest spending group is the medium group (47.4%), followed by the high group (27.6%) and the low group (25.0%).

# Exploratory Data Analysis

## Bivariate and Multivariate Analysis




**Observations**

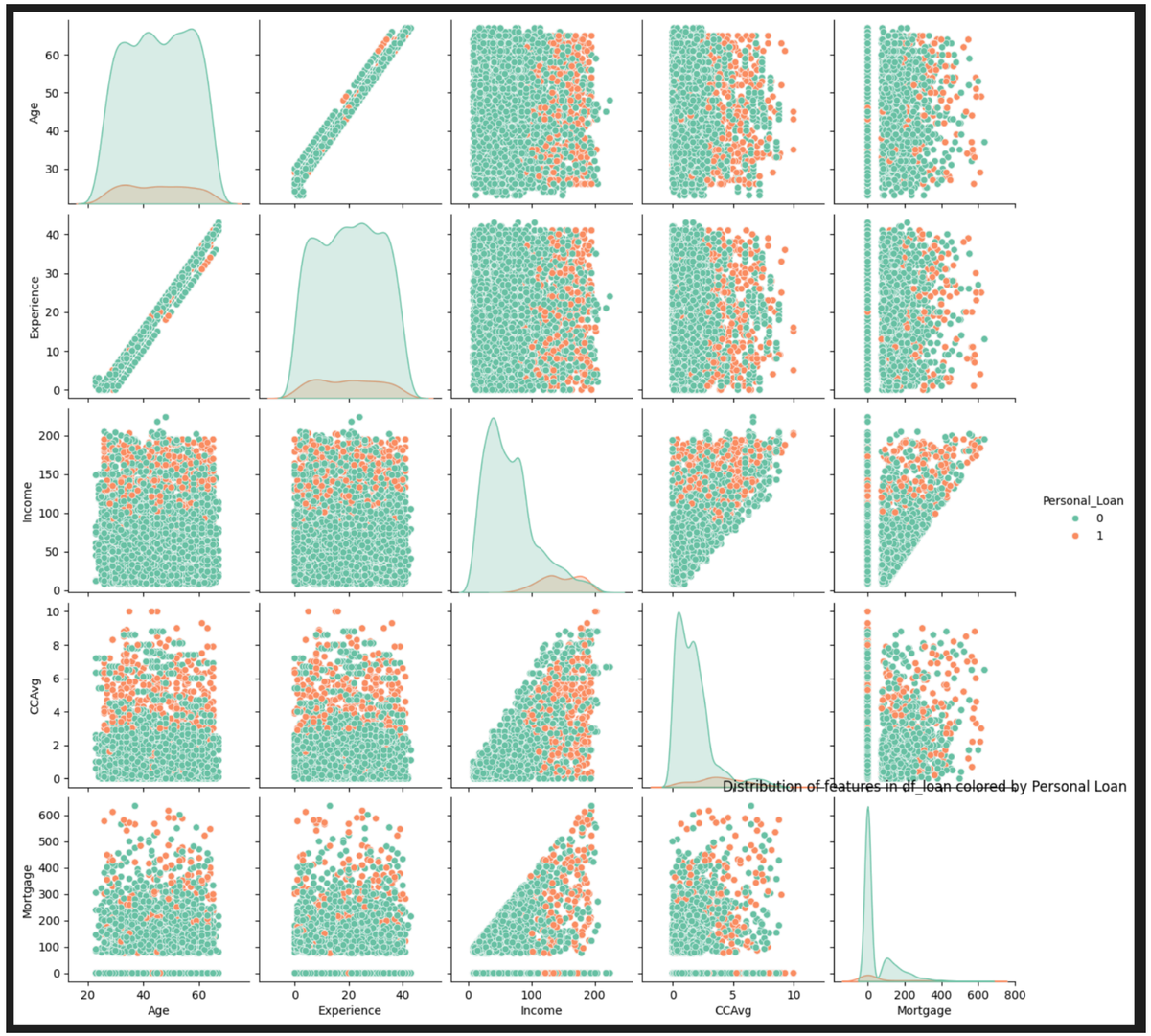
**Strong Positive Correlation:** Age and Experience exhibit a very strong positive correlation (0.99), which is expected as experience generally increases with age.

**Moderate Positive Correlation:** Income and CCAvg (average credit card spending) have a moderate positive correlation (0.65), suggesting that individuals with higher incomes tend to spend more on credit cards.

**Weak Correlations:** Other pairs of variables show weak or negligible correlations, indicating limited linear relationships between them.

# Exploratory Data Analysis

## Bivariate and Multivariate Analysis



# Exploratory Data Analysis

## Bivariate and Multivariate Analysis

**Observations**

**Age and Experience:** There's a strong positive correlation between age and experience, as expected. The distribution of both variables appears to be similar for customers with and without personal loans.

**Income and CCAvg:** A positive correlation is observed between income and average credit card spending (CCAvg). Customers with higher incomes tend to have higher credit card spending. There seems to be a slight separation between customers with and without loans based on these variables, with customers with loans generally having higher income and CCAvg values.

**Mortgage:** The distribution of mortgage amounts appears to be similar for both groups. However, there's a subtle trend suggesting that customers with loans might have slightly lower mortgage amounts.

# Decision Tree (model building and training)

## Data Preparation

- Drop columns not needed: “AgeBin," ‘Experience,' ‘Income_group,' ‘Spending_group,' and ‘ZIP_Code.'

# Decision Tree (model building and training)

## Split data into train and test set

```
Training set features shape: (3500, 13)
Testing set features shape: (1500, 13)
Training set target shape: (3500,)
Testing set target shape: (1500,)
```

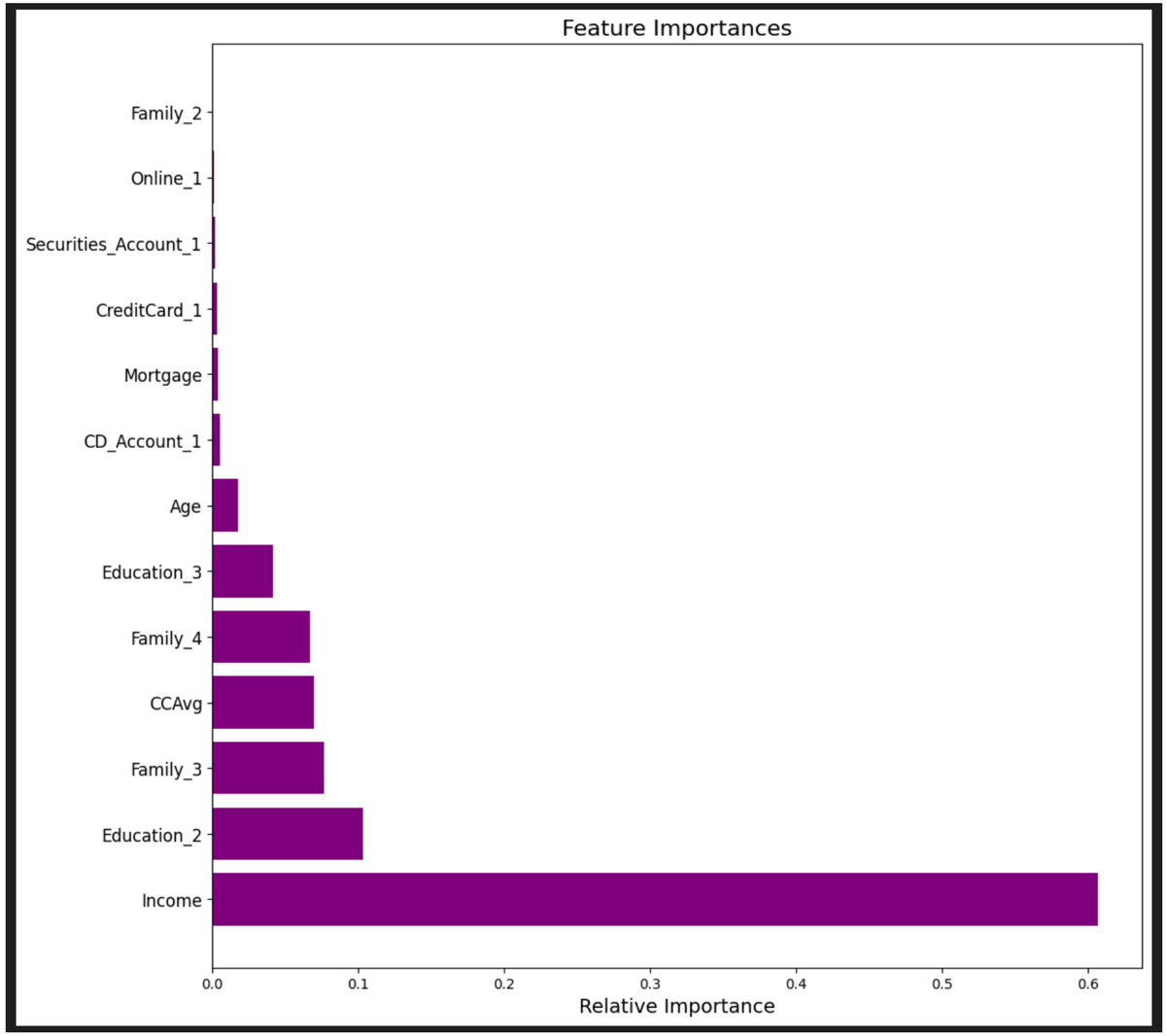
# Decision Tree (model building and training)

## Built a model on the train data

```
from sklearn.tree import DecisionTreeClassifier

# Initialize the DecisionTreeClassifier with Gini impurity, class weights, and a fixed random state
model = DecisionTreeClassifier(
    criterion='gini',                # Gini impurity criterion for splitting nodes
    class_weight={0: 0.15, 1: 0.85}, # Class weights to handle imbalanced classes
    random_state=1                   # Ensure reproducibility of results
)

# Fit the model to the training data
model.fit(X_train_dt, y_train_dt)

# Evaluate the model performance using recall score and confusion matrix
get_recall_score(model)
```

# Decision Tree (model building and training)

## Feature Importance



# Discussion and Conclusion

**1. Accuracy:**

- **Train Accuracy: 1.0000**: This means that the decision tree model is perfectly fitting the training data, correctly classifying all training instances. This is often an indicator of overfitting, where the model has learned the specific details of the training data too well, possibly at the expense of its ability to generalize.
- **Test Accuracy: 0.9787**: The model performs well on the unseen test data, correctly classifying approximately 97.87% of instances. This indicates that despite overfitting on the training data, the model still generalizes well to new data, although a slight drop from the training accuracy is observed, which is expected.

**2. Recall:**

- **Train Recall: 1.0000**: The recall on the training data is perfect, meaning the model is correctly identifying all positive cases (true positives) from the training set. This suggests that the model is very sensitive to detecting the target class on the training data.
- **Test Recall: 0.8750**: On the test data, the recall is 87.5%. This is still quite good, meaning that the model is identifying 87.5% of all actual positive instances in the test set. However, the decrease from 100% to 87.5% suggests that the model may be less sensitive in generalizing to new, unseen examples. The drop in recall indicates that some positive instances in the test set are being missed (false negatives).

## Conclusion:

- **Overfitting:** The model is likely overfitting the training data, given the perfect training accuracy and recall. This means it has learned the noise and details of the training set too well, which may lead to poor generalization to unseen data in some cases.
- **Model Generalization:** Despite overfitting, the model performs very well on the test set, with a high accuracy and a good recall rate, though the drop from training recall to test recall suggests room for improvement.
- **Actionable Insights:**
  - **Tuning:** Consider pruning the decision tree or using cross-validation to optimize the tree's complexity and reduce overfitting.
  - **Evaluation:** It might be helpful to look at other evaluation metrics like **precision**, **F1-score**, or the **confusion matrix** to get a more comprehensive understanding of the model's performance, especially if you are concerned about the recall drop.
  - **Class Imbalance:** If the dataset is imbalanced, the recall score for the positive class might be impacted, so evaluating with techniques like **class weights** or **sampling methods** could also improve performance.