Data Due Diligence Report

Hibah Masoom

Online Master of Science in Data Science, Merrimack College

DSE5004: Visual Data Exploration

Dr. Michael Dupin

11th August 2024

1.0 Introduction

This report provides an in-depth analysis of the customer dataset, including the categorization of all variables, the creation of new features, and a comprehensive summary of the data due diligence and feature engineering processes. The goal is to enhance the dataset's utility for predictive modeling and customer segmentation by ensuring data quality, adding meaningful features, and identifying key relationships within the data.

2.0 Variable Categorization

Each variable in the dataset has been categorized by its meaning, objective, and type. Below is a summary of the categorization:

2.1 Demographics

Categorical	Real-Valued	Ordinal
1. Region 2. Gender 3. MartialStatus 4. AgeGroup 5. Retired 6. PoliticalPartyMem 7. Votes 8. EmploymentLengthCategory	1. Age 2. EmploymentLength	1. TownSize

2.2 Financial

Categorical	Real-Valued
1. LoanDefault	1. HHIncome 2. DebtToIncomeRatio 3. CreditDebt 4. OtherDebt 5. TotalDebt 6. AnnualCardSpend 7. DebtToAgeRatio 8. RetirementSavingsPotential

2.3 Household

Categorical	Real-Valued
1. HomeOwner	1. HouseholdSize 2. NumberPets 3. CarsOwned 4. PetOwnership 5. PetDensity

2.4 Product Ownership and Usage

Categorical	Real-Valued
1. CarOwnership 2. CarBrand 3. CreditCard	1. CarValue 2. CarTenure 3. CardTenurePerAge

2.5 Telecom

Categorical	Real-Valued
1. EquipmentRental 2. WirelessDate 3. Internet 4. CallWait	1. PhoneCoTenure 2. VoiceLastMonth 3. VoiceOverTenure

3.0 Data Due Diligence and Feature Engineering Summary

The data due diligence process involved cleaning, transforming, and augmenting the dataset to ensure its readiness for analysis. Below is a summary of the key steps taken:

3.1 Data Cleaning:

- **Handling Missing Values**: Missing values in categorical variables were imputed using the mode, while continuous variables were imputed using the median to preserve distribution integrity. This approach ensured that the dataset remained complete and ready for analysis without introducing biases from missing data.
- **Normalization of Financial Data**: Variables like HHIncome and CarValue were normalized to create HHIncomeNormalized and CarValueNormalized. This standardization allows for better comparisons across customers with different income levels and car values.

3.2 Feature Engineering:

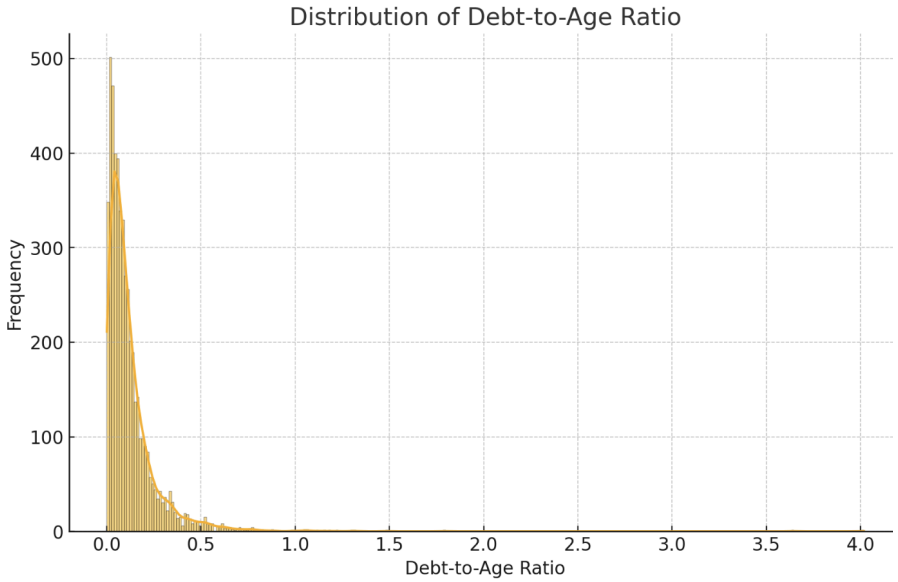
1. **TotalDebt**: Created by summing CreditDebt and OtherDebt. This feature provides a consolidated view of a customer's financial obligations, which is crucial for understanding their financial stability.
2. **DebtToAgeRatio**: Created to assess the relationship between a customer's age and their debt levels. This feature helps identify potential financial stress relative to the customer's life stage.
3. **CardTenurePerAge**: This ratio was created to understand how long a customer has held a credit card relative to their age, providing insight into financial maturity and behavior.
4. **PetOwnership**: Combined NumberPets, NumberCats, NumberDogs, and NumberBirds to create a single measure of total pet ownership, simplifying the analysis of customer pet-related behaviors.
5. **PetDensity**: Created by dividing PetOwnership by HouseholdSize to understand the concentration of pets within households, which can be used for targeted marketing of pet-related products.
6. **AnnualCardSpend**: An estimated annual expenditure based on monthly credit card spending, providing insight into customer spending habits and identifying potential high-value customers.
7. **AgeGroup**: Categorized age into groups to simplify demographic analysis and segmentation. This feature helps in tailoring marketing strategies based on the age demographic.
8. **RetirementSavingsPotential**: A simplified estimate of potential savings for retirement, calculated based on the customer's current income and age. This feature is useful for financial planning and advisory services.

9. **EmploymentLengthCategory**: Categorized employment length into bins to help understand the stability and career progression of customers, providing insight into their long-term financial planning.
10. **DebtToCarValue**: Created by dividing TotalDebt by CarValue, this feature helps assess the financial burden relative to an owned asset, such as a car.

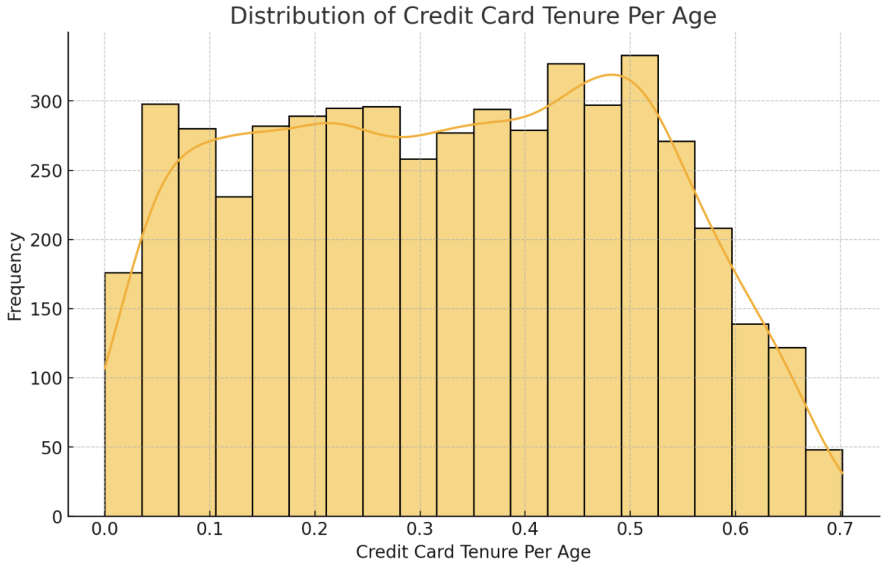
4.0 Visualizations/Graphs and Explanations

4.1 Single-Variable Plots:

- **Debt-to-Age Ratio (DebtToAgeRatio)**:
  - The histogram below illustrates the distribution of the DebtToAgeRatio feature, showing how debt levels vary relative to age among the customer base. A higher ratio indicates that a customer has more debt relative to their age, which could suggest financial stress, especially for younger individuals. The distribution highlights that debt is more concentrated among certain age groups, indicating potential financial vulnerability in those demographics.

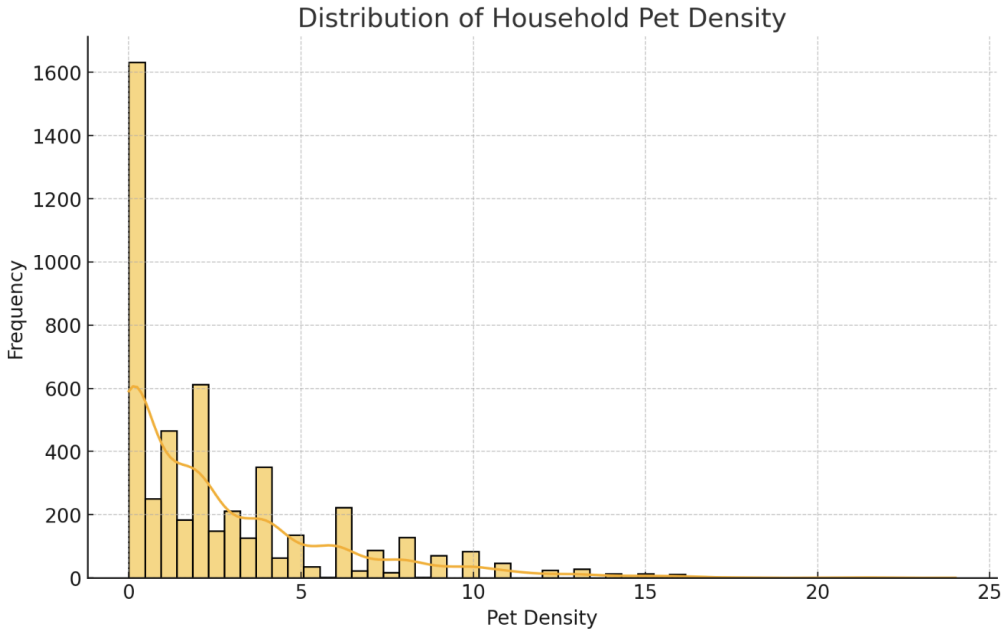


- **Credit Card Tenure Per Age (CardTenurePerAge)**:
  - The histogram shows the distribution of the CardTenurePerAge feature, providing insight into how long customers have held credit cards relative to their age. The distribution may reveal whether younger customers tend to have shorter credit card tenures, while older customers have longer tenures, providing insights into financial maturity and credit behavior across different age groups



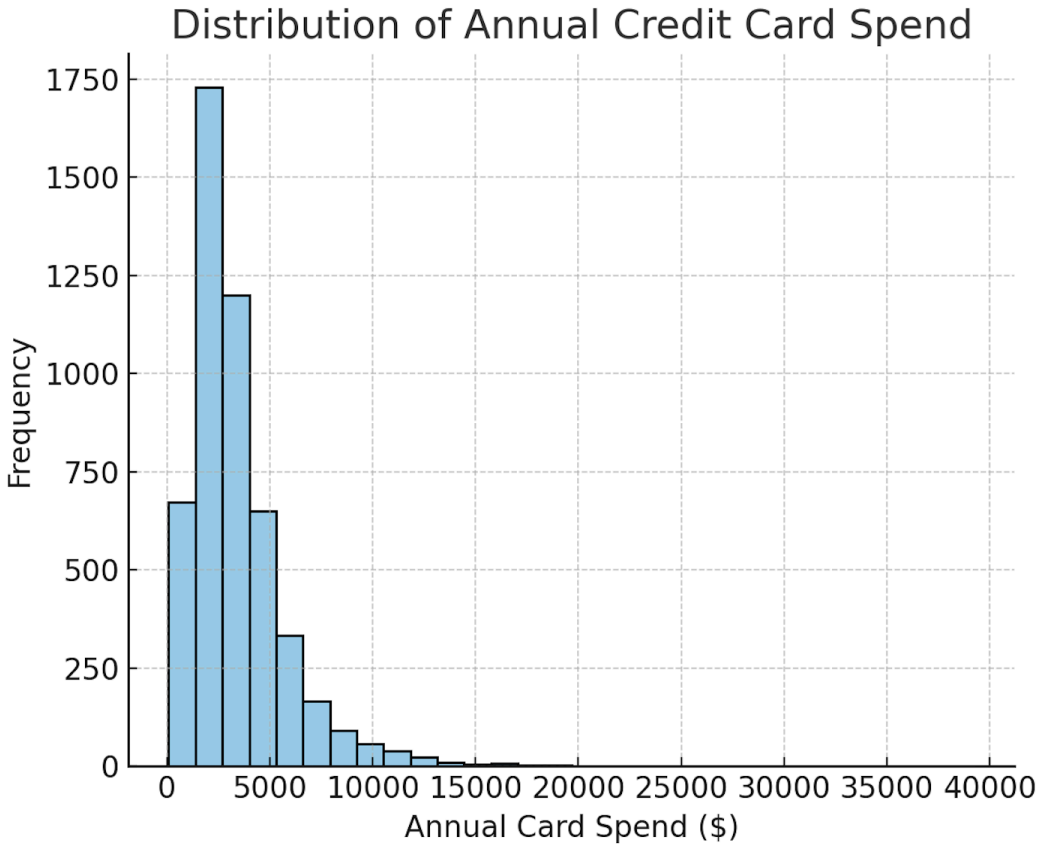
➢ **Household Pet Density (PetDensity)**:

- This histogram visualizes the distribution of PetDensity, which represents the ratio of pets to household size, highlighting households with varying levels of pet ownership. The distribution can help identify target markets for pet-related products, especially in areas where pet density is high, suggesting a higher potential for sales of such products.



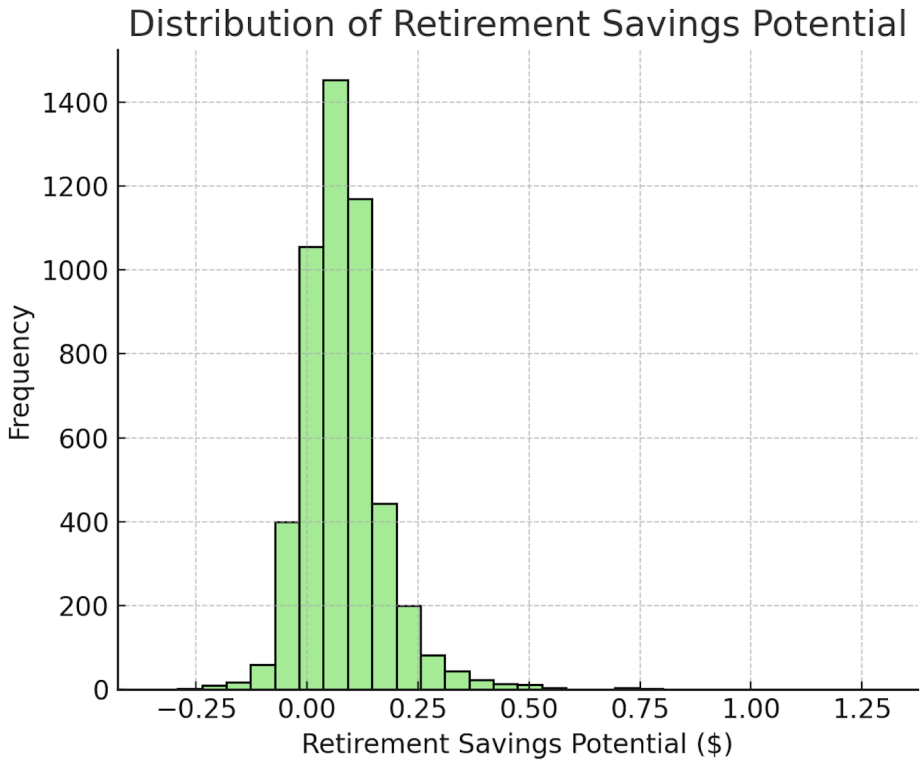
➢ **Annual Credit Card Spend (AnnualCardSpend)**:

- The histogram below illustrates the distribution of estimated annual credit card spending, identifying spending behaviors across the customer base. It shows the range of spending behaviors, identifying potential high-value customers who spend significantly more on their credit cards. This information is crucial for targeted marketing strategies.



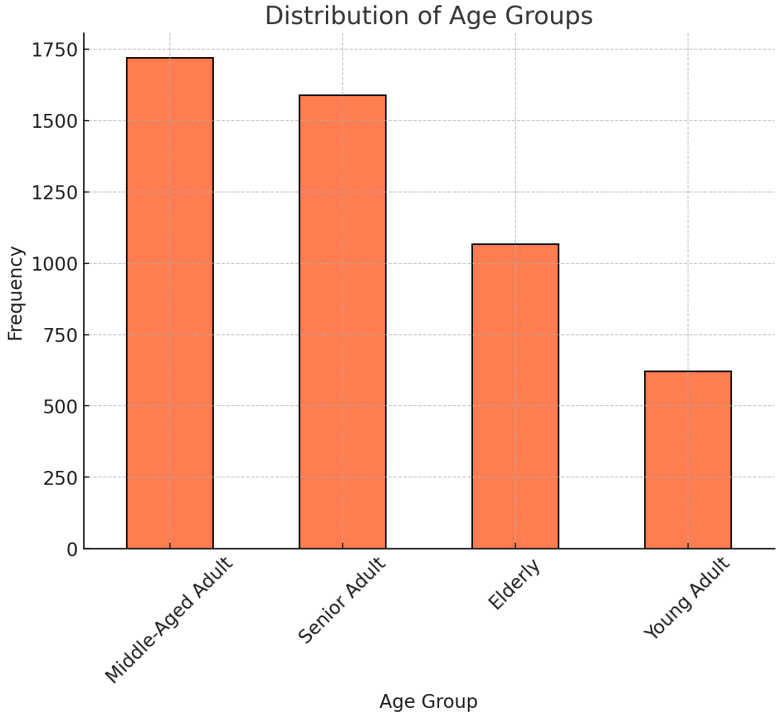
➢ **Retirement Savings Potential (RetirementSavingsPotential)**:

- The histogram illustrates the distribution of estimated retirement savings potential, showing how customers vary in their financial preparedness for retirement. It highlights how prepared different customers are for retirement, with some having higher potential savings and others lower.



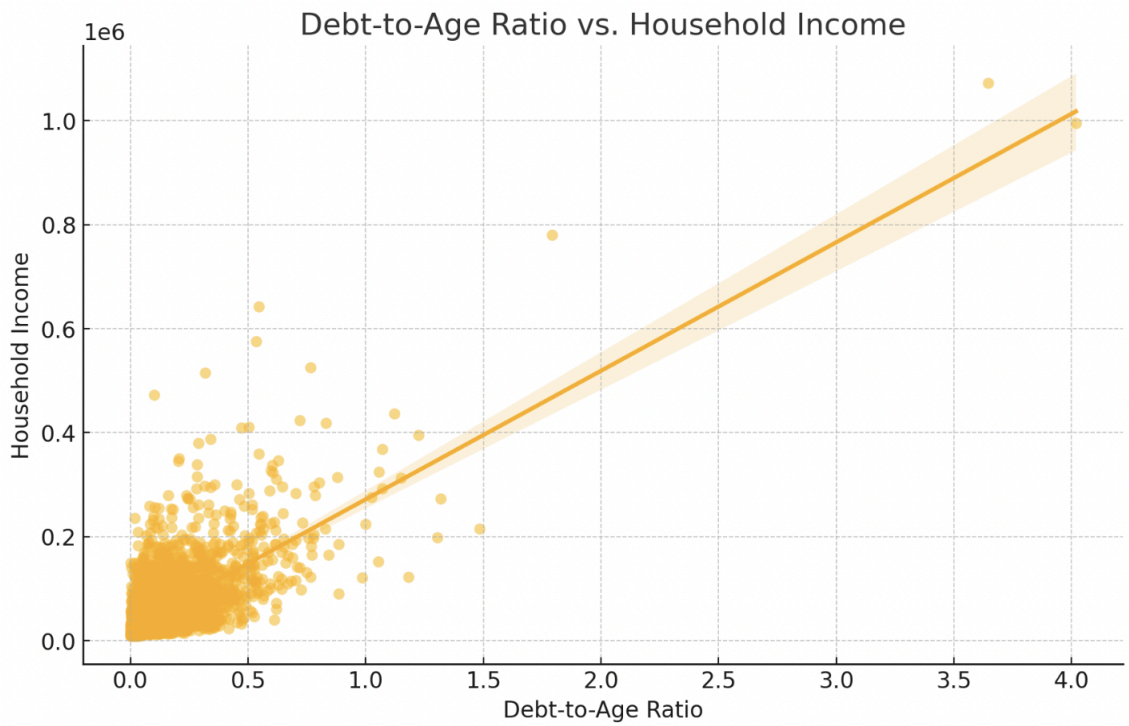
➢ **Age Group (AgeGroup)**:

- The bar plot shows the count of customers in each AgeGroup, providing a clear view of the age demographics within the dataset. This information is crucial for targeting young adults with student loan offers or elderly customers with retirement planning services.



4.2 Two-Column Analyses:

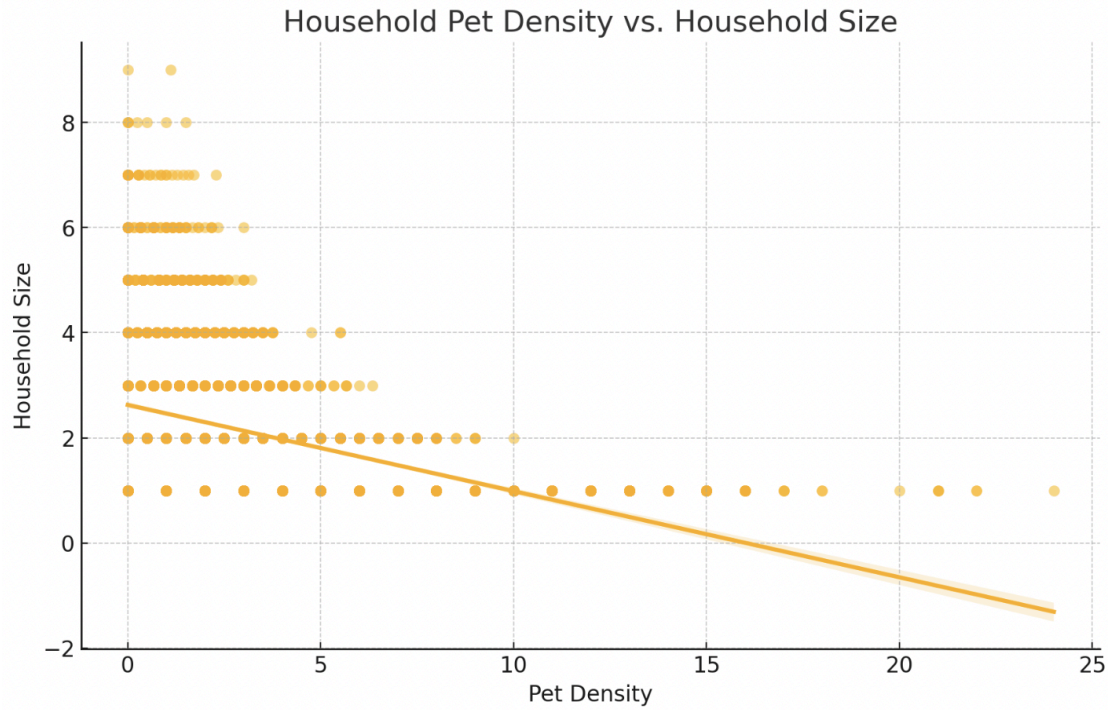
- **Debt-to-Age Ratio vs. Household Income**:
  - The scatterplot below illustrates the relationship between DebtToAgeRatio and HHIncome, showing how debt relative to age correlates with household income. Customers with lower incomes might have a higher DebtToAgeRatio, indicating that they are carrying more debt relative to their age.



- **Credit Card Tenure Per Age vs. Age**:
  - This scatterplot shows the relationship between CardTenurePerAge and Age, highlighting how credit card tenure relates to the customer's age. This information helps in understanding financial behavior and stability across different age groups.

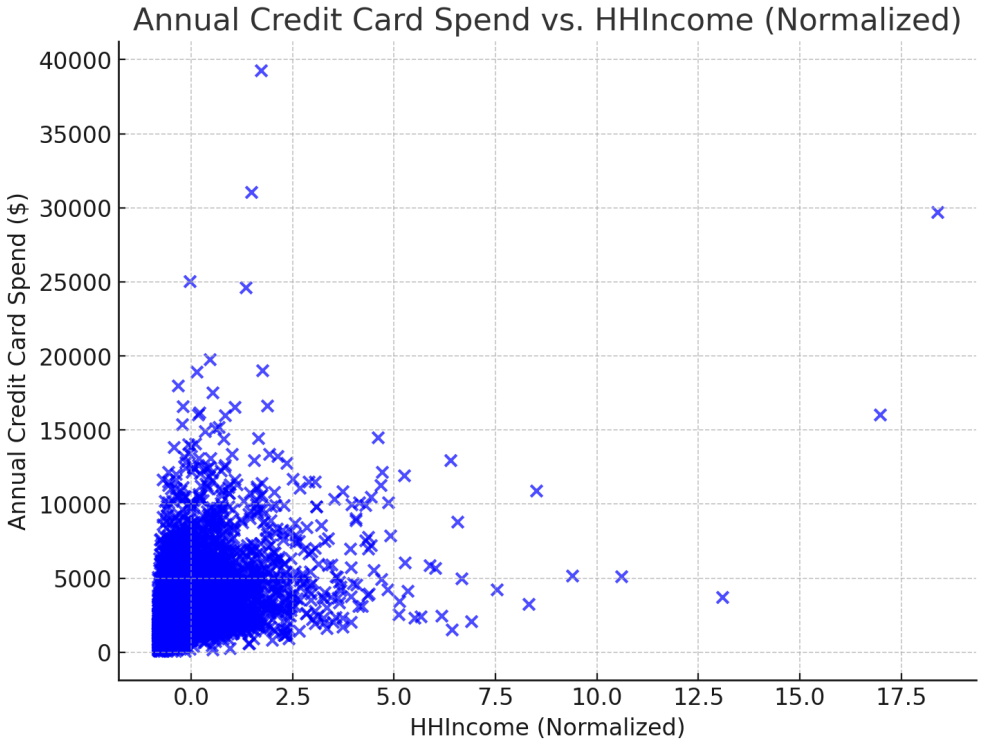
➢ **Household Pet Density vs. Household Size**:

- The scatterplot below illustrates the relationship between PetDensity and HouseholdSize, showing how pet ownership is distributed across different household sizes. This information is valuable for businesses targeting pet owners, as it helps identify the household types most likely to own pets.



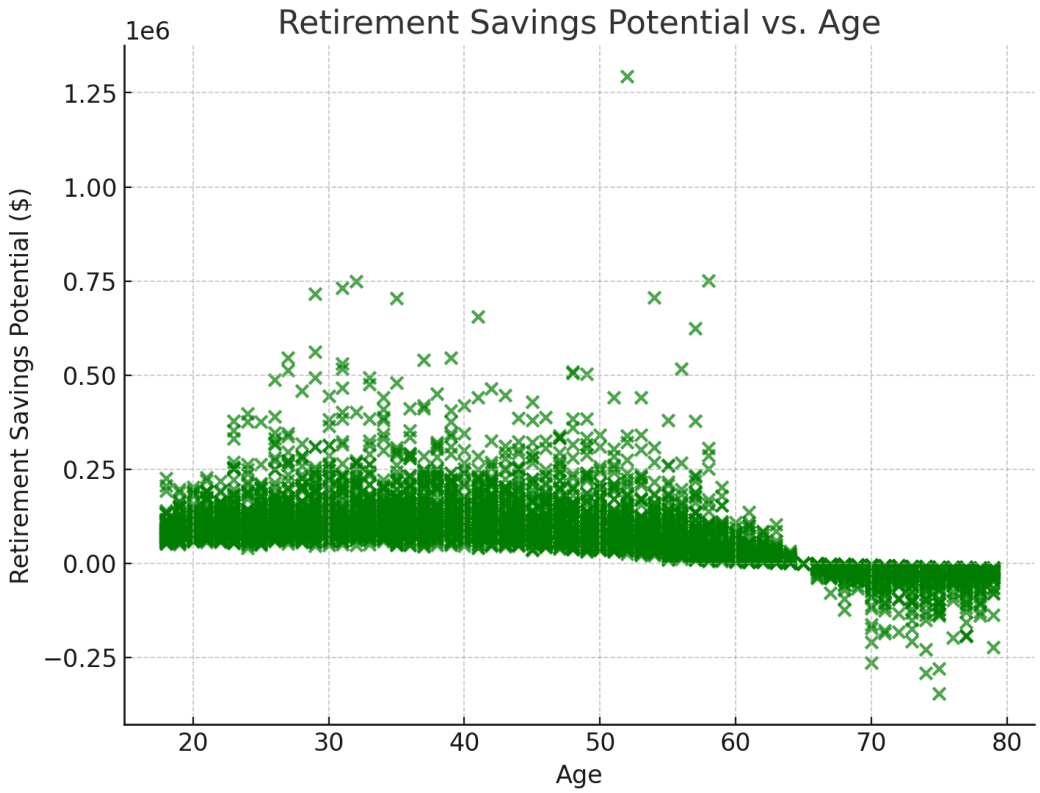
➢ **Annual Credit Card Spend vs. HHIncome (Normalized)**:

- This scatterplot shows the relationship between AnnualCardSpend and HHIncomeNormalized, highlighting how annual spending correlates with household income. This relationship shows us that higher-income customers tend to spend more on their credit cards annually.



➢ **Retirement Savings Potential vs. Age**:

- The scatterplot below illustrates the relationship between RetirementSavingsPotential and Age, providing insights into how age influences potential retirement savings. This information shows us that as customers age, their potential retirement savings generally decrease.



5.0 Conclusion

The steps taken in this analysis have significantly enhanced the dataset's utility for predictive modeling, customer segmentation, and targeted marketing. By categorizing variables, cleaning and normalizing data, and creating new features, we have provided deeper insights into customer behavior, financial stability, and lifestyle factors. The combination of single-variable distributions and two-column analyses enables a comprehensive understanding of the data, setting a solid foundation for further analysis or application in machine learning models.