K-Means Clustering on Survey of Consumer Finances (SCF)

Introduction

Project Goals

The primary goal of this project is to segment households using the Survey of Consumer Finances (SCF) data, focusing specifically on households, who have been turned down for credit or feared being denied credit. By identifying distinct segments within this group, we aim to gain insights into their financial behaviors and characteristics.

Objectives

- To perform exploratory data analysis (EDA) to understand the distribution and relationships between key features in the dataset.
- To pre-process the data, handling missing values and standardizing numerical features.
- To select relevant features for clustering, utilizing dimensionality reduction 3
- To apply K-Means clustering to segment households based on their financial characteristics.
- To visualize the results and deploy an interactive Dash web application for further exploration of household segments.
- To provide insights that can assist in policy making, financial product development, and economic research.

Overview

In this project, we utilize the Survey of Consumer Finances (SCF) dataset, which includes various features related to household finances such as income, assets, debts, and demographic information. The key indicator of interest is the 'TURNFEAR' column, which identifies household who have been turned down for credit or feared being denied credit. Additionally, we focus on households whose net worth is under $2 million, as the dataset includes some outliers in the form of ultra-wealthy households that could skew the analysis.

We begin with a comprehensive exploratory data analysis to understand the data distribution, identify missing values, and visualize relationships between features. Following this, we pre-process the data by handling missing values and standardizing numerical features to ensure consistency.

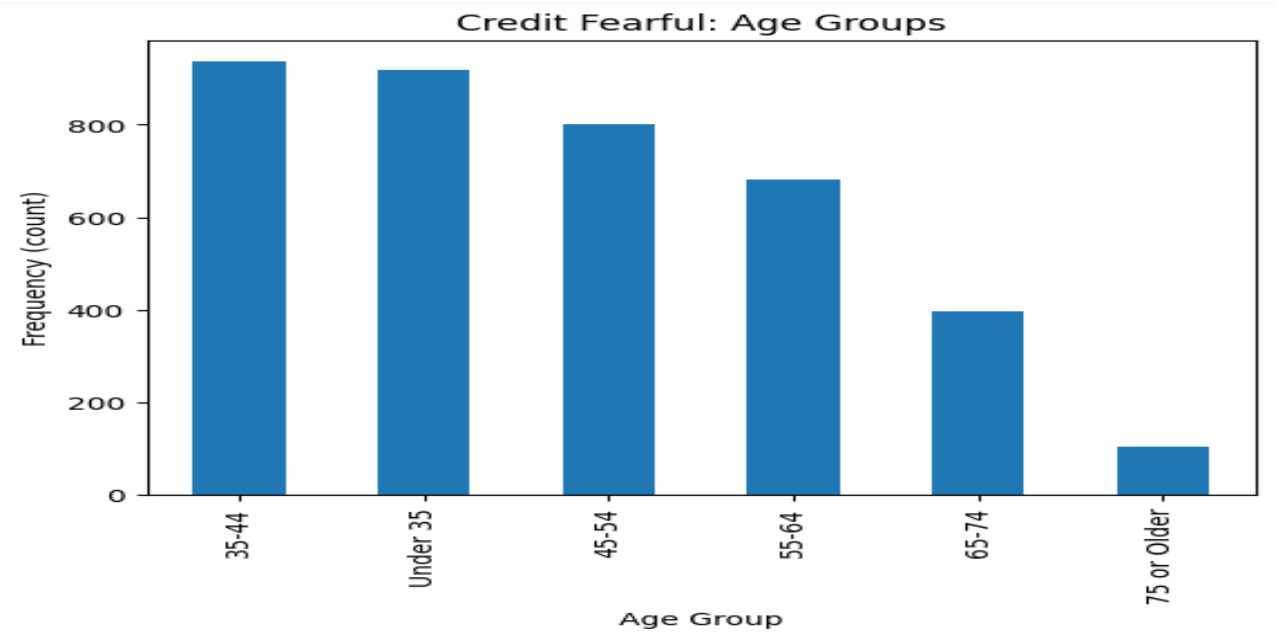
Feature selection is performed to identify the most relevant features for clustering. Principal Component Analysis (PCA) is employed for dimensionality reduction, helping to retain the most significant information while reducing the complexity of the data.

The K-Means clustering algorithm is then applied to segment the households into distinct clusters. We determine the optimal number of clusters using the Elbow Method and Silhouette Score, ensuring meaningful segmentation.

Finally, we visualize the clusters in 2D and 3D plots and deploy a Dash web application for interactive exploration of the household segments. The insights gained from this analysis provide valuable information for policymakers, financial institutions, and researchers, aiding in the development of targeted financial products and policies

Exploring The Data

Age group distribution of households who fear being denied or have been denied credit



The bar chart indicates that a significant proportion of individuals who fear being denied credit or who have been denied credit are younger. This trend suggests that younger people may perceive higher barriers to obtaining credit.

Education Qualifications for both credit fearful and non-fearful households

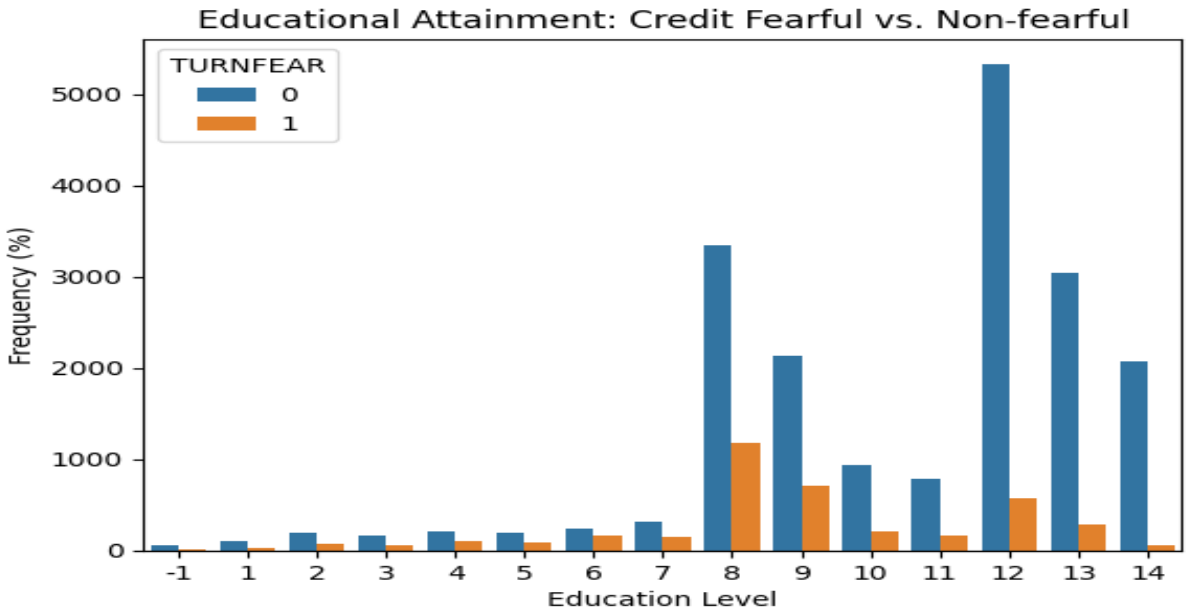
EDUC **Highest completed grade by reference person**

Percent	N	Value	Label
0.2	2,047,870	-1	LESS THAN 1ST GRADE
1.2	14,091,109	1	1ST, 2ND, 3RD, OR 4TH GRADE
1.7	20,550,083	2	5TH OR 6TH GRADE
3.3	40,655,677	3	7TH OR 8TH GRADE
2.4	28,723,036	4	9TH GRADE
3.3	40,204,164	5	10TH GRADE
3.8	46,941,671	6	11TH GRADE
1.4	16,652,539	7	12TH GRADE, NO DIPLOMA
28.1	343,109,104	8	HIGH SCHOOL GRADUATE - HIGH SCHOOL DIPLOMA OR EQUIVALENT
17.8	216,726,717	9	SOME COLLEGE BUT NO DEGREE
1.5	17,848,650	10	ASSOCIATE DEGREE IN COLLEGE - OCCUPATION/VOCATION PROGRAM
5.7	68,953,489	11	ASSOCIATE DEGREE IN COLLEGE - ACADEMIC PROGRAM
18.2	221,577,418	12	BACHELOR'S DEGREE (FOR EXAMPLE: BA, AB, BS)
8.1	98,335,500	13	MASTER'S DEGREE
3.6	43,526,402	14	DOCTORATE OR PROFESSIONAL SCHOOL DEGREE
	29,398	.	(No Data)
100.0	**1,219,972,825**		**Total**

Properties

Data type: numeric

Record/columns: 1/48-49



This side-by-side plot reveals a notable correlation between educational attainment and credit fearfulness. It shows that a higher proportion of respondents who express fear of credit have only a high school diploma. In contrast, those with university

degrees are more commonly found among respondents who do not fear credit. This pattern suggests a potential relationship where higher educational attainment is associated with greater confidence in securing credit. The data implies that individuals with advanced education may possess more financial literacy or resources, contributing to their enhanced confidence in credit transactions.

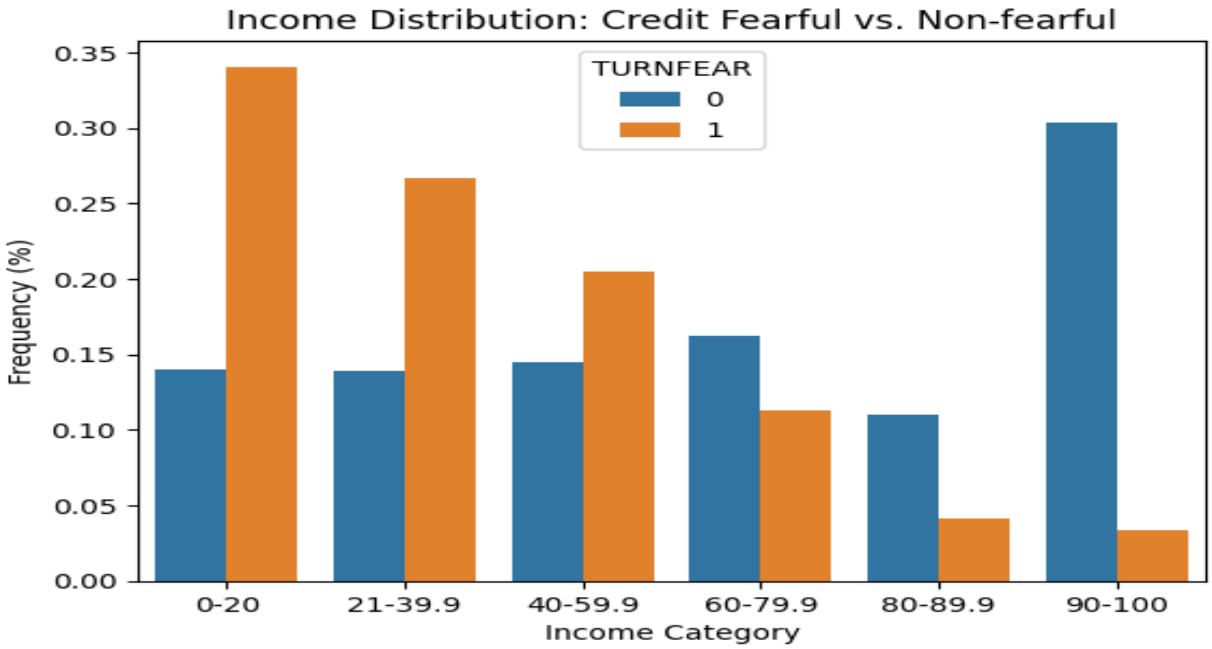
Income Distribution for both Credit Fearful and Non-fearful households

INCCAT **Income percentile groups**

Text of this Question or Item

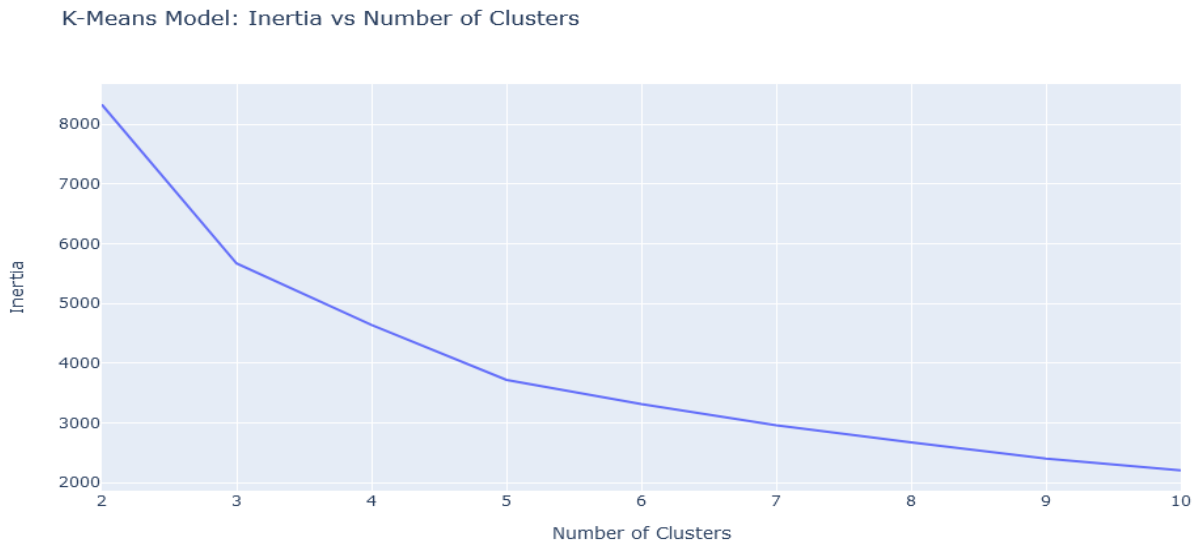
```
Income percentile groups
1 = 0-20
2 = 20-39.9
3 = 40-59.9
4 = 60-79.9
5 = 80-89.9
6 = 90-100
```

	Percent	N	Value	Label
	20.0	243,967,557	**1**	0-20
	20.0	243,991,241	**2**	20-39.9
	20.0	243,980,482	**3**	40-59.9
	20.0	243,993,848	**4**	60-79.9
	10.0	122,005,192	**5**	80-89.9
	10.0	122,034,506	**6**	90-100
	100.0	**1,219,972,825**		**Total**



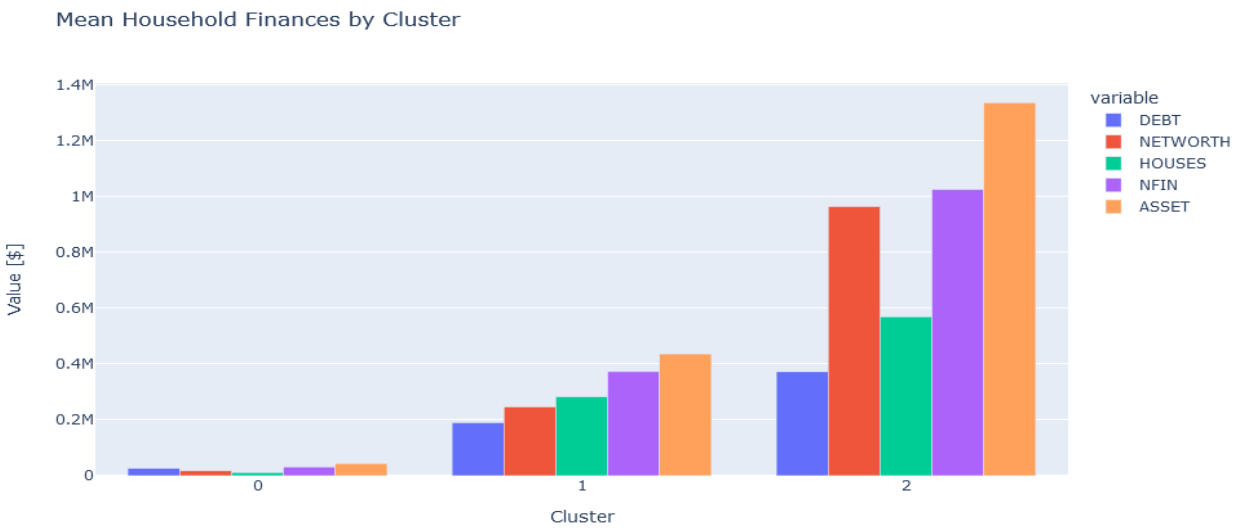
Comparing the income categories across the fearful and non-fearful groups, we can see that credit-fearful individuals are much more common in the lower income categories. In other words, individuals that fear credit denial tend to have lower incomes.

KMeans clustering



In clustering analysis, particularly in k-means clustering, selecting the appropriate number of clusters is crucial for accurately capturing the underlying patterns in the data. The elbow method is a commonly used technique for determining the optimal number of clusters. In this plot, the "elbow" appears at the point where the drop in inertia becomes less dramatic. This point indicates the optimal number of clusters. For this analysis, the elbow plot indicated that 3 clusters were optimal. Selecting the number of clusters at this point ensures that the clusters are well-defined without overfitting the data

Distribution of mean household finance across different clusters



This bar chart visually depicts the mean household finances across three clusters, labeled 0, 1, and 2. The financial metrics displayed include:

- **Debt:** Indicates the average debt level of households within each cluster.
- **Net Worth:** Represents the total value of assets minus liabilities for each cluster.
- **House:** Refers to the average number of houses owned by households in each cluster.
- **NFIN:** Represents the total value of non-financial assets held by households, such as property, vehicles, and tangible possessions.
- **Asset:** The total value of all assets, including both financial and non-financial assets.

Key Observations:

- Cluster 2 Dominates: Cluster 2 exhibits the highest values across all financial metrics, indicating that it comprises households with significantly greater wealth and asset ownership compared to the other clusters.
- Cluster 0 Lags Behind: Cluster 0 consistently displays the lowest values across all financial metrics, indicating a lower financial standing relative to

the other clusters. This cluster shows reduced levels of both financial and non-financial assets.

- Cluster 1 Occupies a Middle Ground: Cluster 1 falls between the affluent Cluster 2 and the less fortunate Cluster 0.
    - Intermediate Wealth: The financial metrics for Cluster 1 (debt, net worth, houses, NFIN, and assets) are positioned between those of the other two clusters, suggesting a moderate level of financial stability.
    - Potential for Growth: Although not as wealthy as Cluster 2, Cluster 1 might have greater potential for financial growth due to its lower starting point.

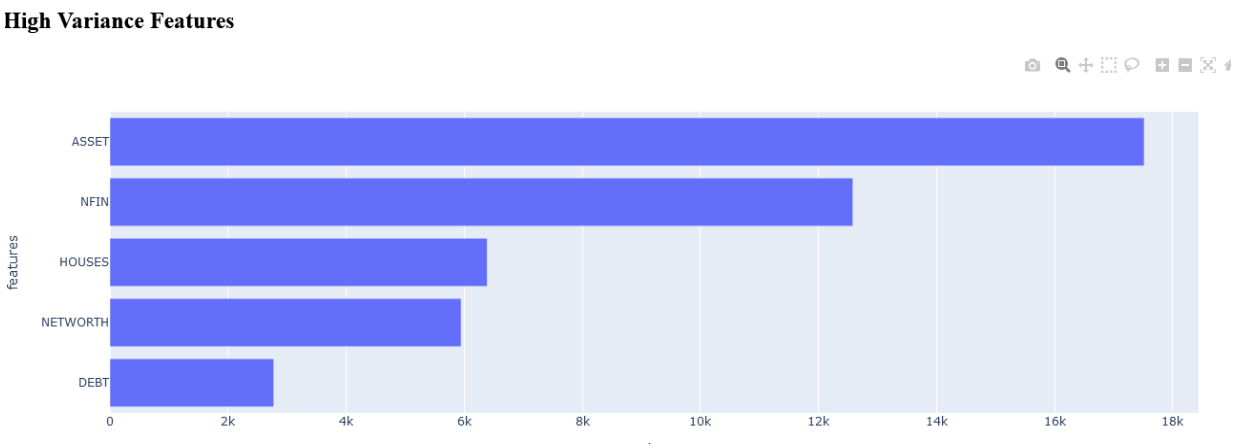
PCA Representation of clusters

This chart visualizes the results of a Principal Component Analysis (PCA) applied to a dataset. PCA is a dimensionality reduction technique that simplifies data by transforming it from a high-dimensional space into a lower-dimensional one while retaining as much variance as possible. The PC1 and PC2 axes represent the main directions of variation in this reduced space. The data points are clearly clustered into three distinct groups, each represented by a different color (blue, red, and green). This clustering indicates the presence of three distinct categories or classes within the dataset, suggesting meaningful groupings in the data's structure.

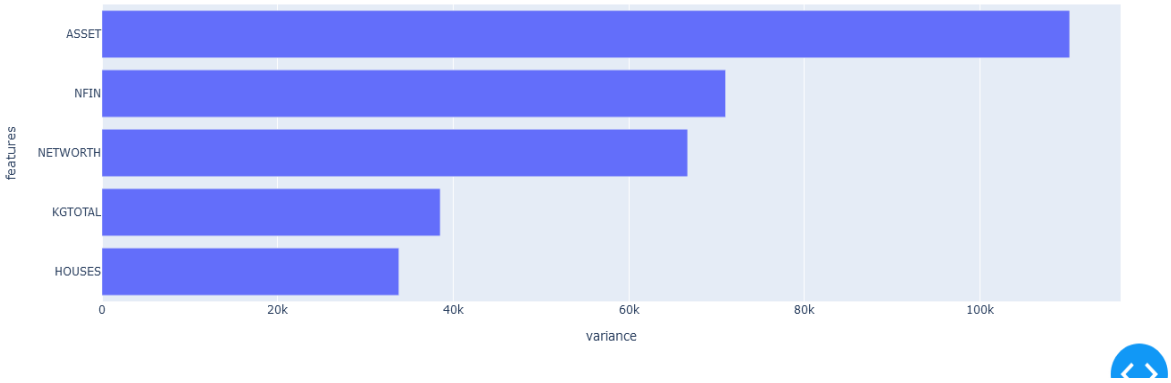
Interactive Dashboard on the Survey of Consumer Finances (SCF)

To provide a deeper and more engaging analysis of the Survey of Consumer Finances (SCF), I created an interactive dashboard. This tool allows users to explore various financial metrics and demographic factors within the dataset. This dashboard also allows users to choose their own features, build a model, and evaluate its performance through a graphical user interface. It also empowers users to construct and assess predictive models without the need for coding knowledge, making advanced analytics accessible to a broader audience.

Key features of the dashboard include:



- **High Variance Features Plot:** Visualizes features with the highest variance, highlighting the most significant variables in the dataset.



- **Trimmed Variance Features Plot:** Displays features with trimmed variance, helping to focus on relevant variables while reducing noise.

- **Cluster Control Slider:** Allows users to select the number of clusters, dynamically adjusting the clustering analysis based on user input.

- **Inertia and Silhouette Score:** Provides metrics for evaluating the quality of the clustering, helping users understand the effectiveness of their chosen number of clusters.

- **PCA Representation of Clusters:** Visualizes the clusters in a reduced-dimensional space, making it easier to see the distinct groupings within the data.

This dashboard serves as a powerful tool for visualizing the complexities of consumer finance and empowers users to perform sophisticated analyses, making data science more accessible and actionable