Customer Segmentation-Retail

PRESENTED BY:

SIMIDOLA LAWANI

DATE:

19/07/2024

Objective

- The goal of this project is to strategically categorise our clientele so that the marketing division may customise product introductions and sales initiatives accordingly.
- By doing this, we can successfully target client segments based on their preferences and optimise resource allocation, saving time and money.
- Data was collected using loyalty cards used during checkouts.

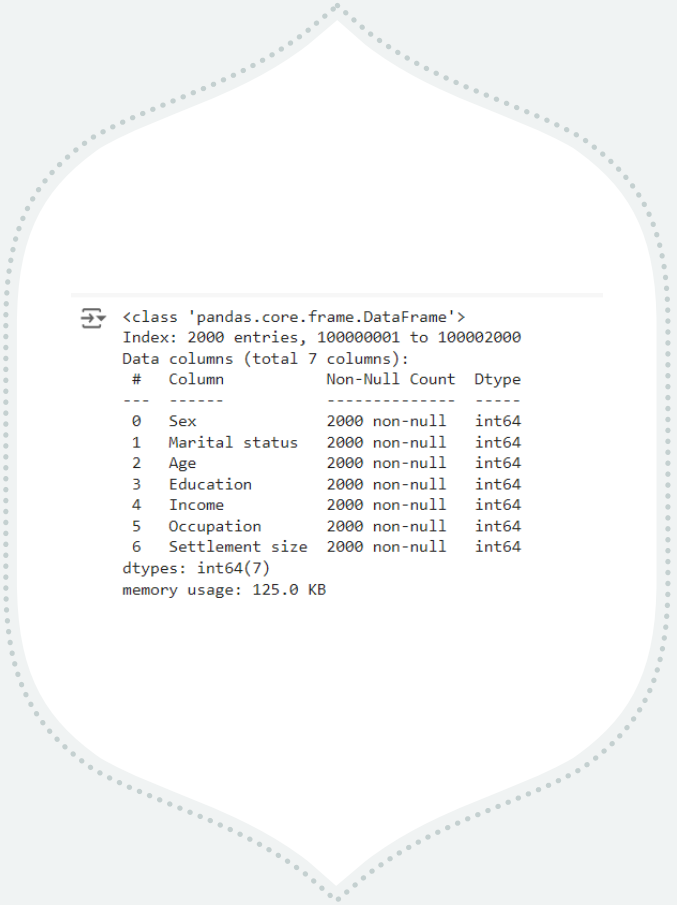
Steps followed

- Data Cleaning to removed null values, remove duplicates and investigate outliers.
- Performed Exploratory Data Analysis (EDA).
- Data Preprocessing to standardize dataset using Sk-learn's Standardscaler package in python.
- Applied K-Means and PCA algorithms to investigate how we can define new grouped customers.

The Dataset

Variables:

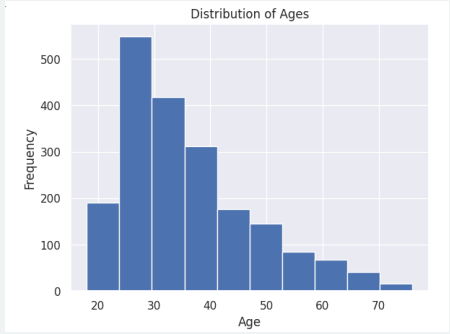
1. ID: Displays a customer's distinct identify.
2. Sex: A customer's biological sex, or gender. There are just two possibilities in this dataset. (0: male 1: a woman)
3. Status married: A customer's status married. (0: one 1: not single (married, widower, separated, or divorced)
4. Age: The customer's age in years, determined by subtracting their birth year from the current year at the time the dataset was created. (The lowest age found in the dataset is 18 Min. The maximum age recorded in the dataset, 76 Max)
5. Education: The customer's educational attainment.(0:unknown / other, 1: high school, 2: University, 3: Postsecondary education)
6. Income: The customer's self-reported annual income expressed in US dollars. (The dataset's lowest income, 35832 Min, was observed, 309364 Max value, which is the dataset's highest income recorded)
7. Occupation: The customer's category of employment. (0: Jobless or unskilled,1: knowledgeable worker or official, 2: officer, manager, independent contractor, and highly qualified worker)
8. Settlement size: The size of the customer's home city. (0: a tiny city, 1: a city of moderate size, 2: a large city)

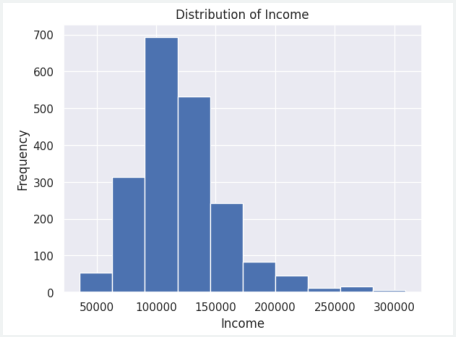


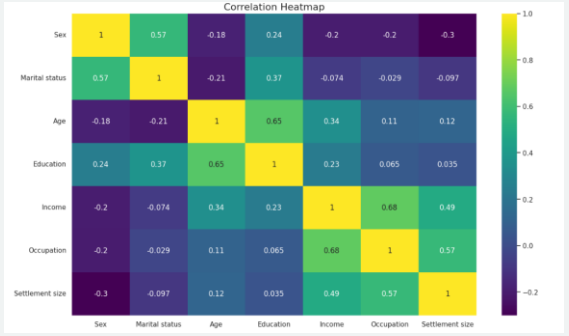
```
<class 'pandas.core.frame.DataFrame'>
Index: 2000 entries, 100000001 to 100002000
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Sex              2000 non-null   int64
 1   Marital status   2000 non-null   int64
 2   Age              2000 non-null   int64
 3   Education        2000 non-null   int64
 4   Income           2000 non-null   int64
 5   Occupation       2000 non-null   int64
 6   Settlement size  2000 non-null   int64
dtypes: int64(7)
memory usage: 125.0 KB
```

EDA

- Age and Education: Age and education have a positive correlation, indicating that older individuals tend to have higher education levels.
- Income and Occupation: Income and occupation are positively correlated. Meaning people with certain occupations tend to have higher incomes.
- Settlement Size and Marital Status: There is a weak negative correlation between settlement size and marital status. Larger cities may have more single individuals.
- Income and Age: Income and age have a positive correlation. Older individuals tend to have higher incomes.
- Education and Occupation: Education and occupation are positively correlated. Higher education often leads to more skilled or specialized jobs.



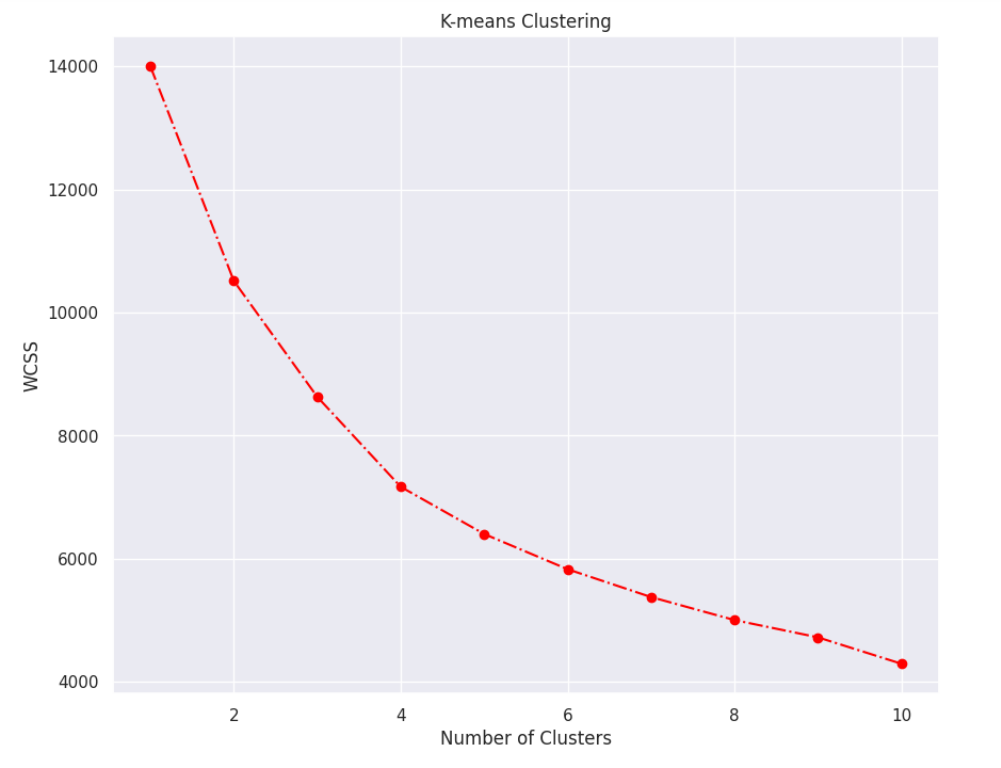




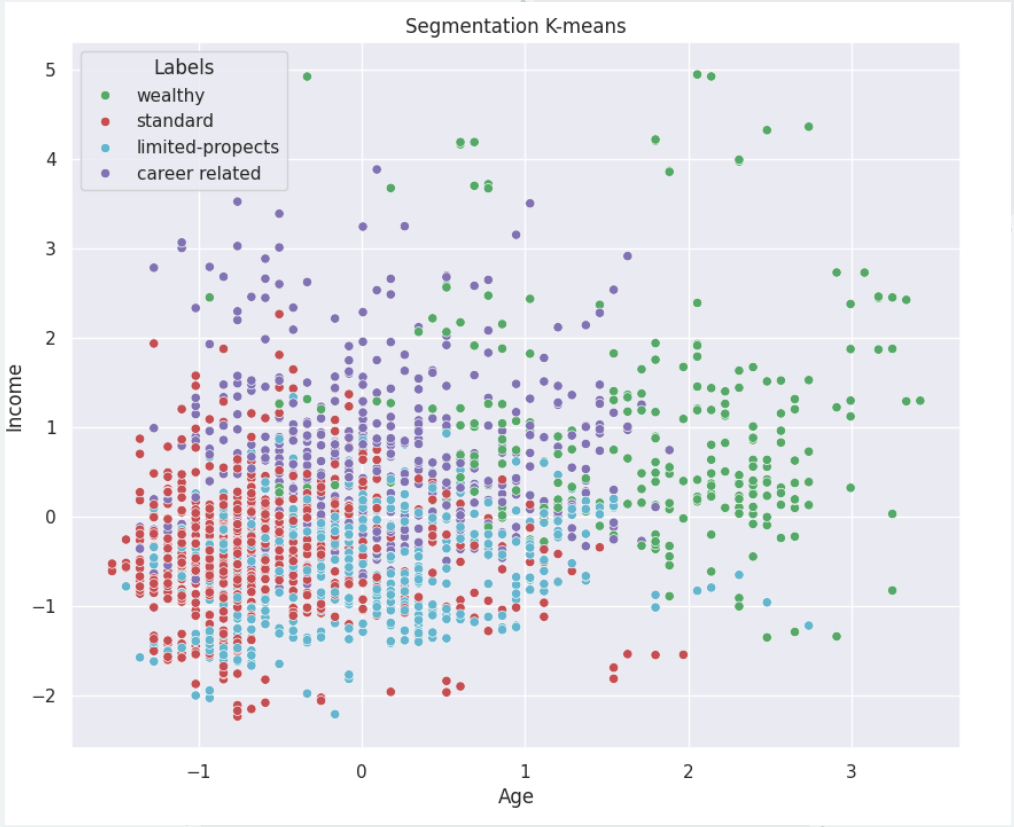
Elbow Method For Clustering

Determined number of clusters using elbow method after using Kmeans for clustering as can be observed in the graph on the right. Picked 4 clusters based on the results obtained from the elbow method chart on the right.

K-means clustering reduces the distance between each data point and the closest cluster centre (centroid), resulting in the grouping of data points into a predetermined number (k) of clusters. The first positions of these centroids in the data space are arbitrary. Afterwards, the algorithm repeats the process until convergence (minimum centroid movement) by repeatedly assigning each data point to the closest centroid, recalculating the centroid location based on its assigned points, and so on.



Kmeans Clustering Data Visualisation

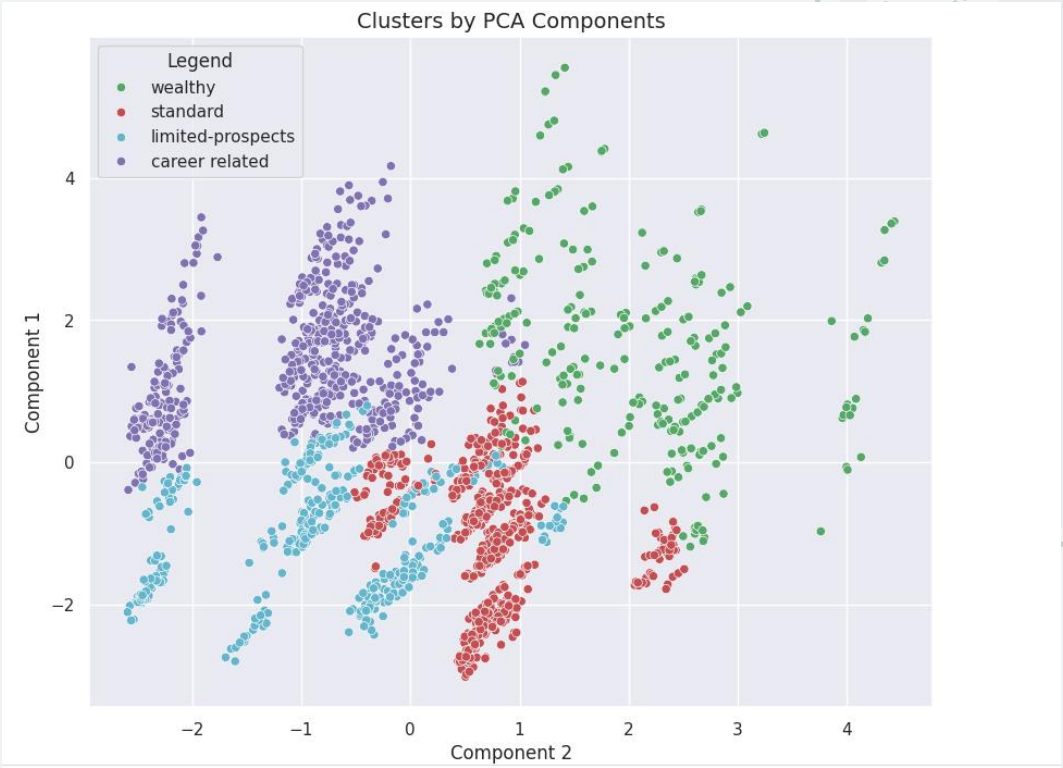


The average features of the individuals within each segment are represented by numerical values (centroids) found in the row of that segment. The clustered were labeled based on the data findings.
The "wealthy" group earns more money and has more education.
The majority of the factors in the "limited-prospects" section have below-average ratings.
In most respects, the "standard" portion is close to average.
The emphasis in the "career-related" section is on occupation and education.

Segment K-means	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
wealthy	0.090136	0.391040	1.689452	1.819919	0.981226	0.499317	0.457039
limited-prospects	-0.209147	-0.954062	-0.028257	-0.485711	-0.606168	-0.754190	-0.856438
standard	0.796753	1.001351	-0.592830	0.050173	-0.398834	-0.276394	-0.389380
career related	-0.857528	-0.645647	-0.023378	-0.508091	0.531869	0.722760	0.964888

Kmean Clustering with PCA

Only the green segment could be distinguished when we plotted the K means clustering solution without PCA, but the component-based division is considerably more noticeable.



Conclusion

- Our segmentation did better after combining Kmeans clustering unsepervisered learning technique PCA. (Principal Component Analysis).
- The information derived from this customer segmentation technique make it easier to target specific groups of customers with tailored products, services, and marketing strategies.