Assignment 5 Report

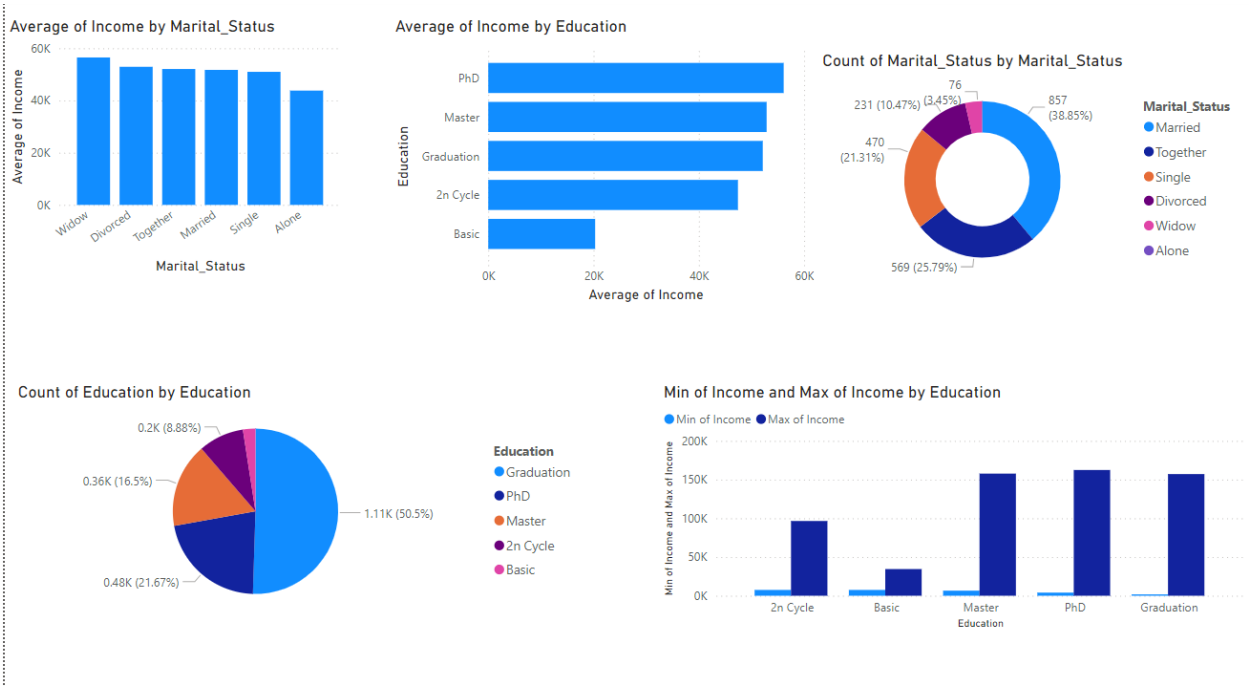
Team 51

Muskan Raina (2021101066)

Arghya Roy (2021115008)

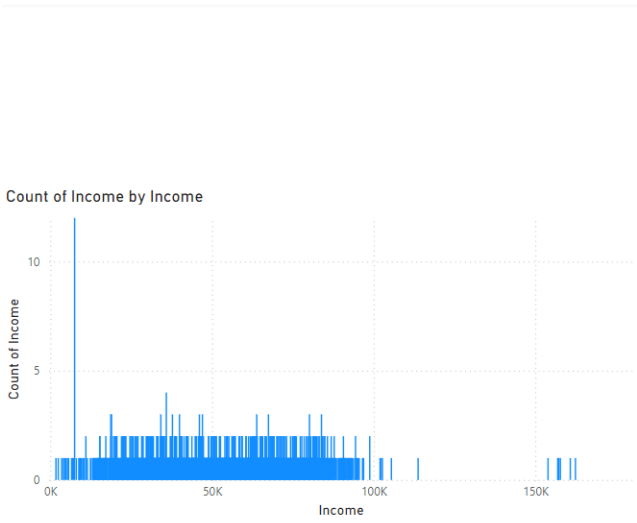
Part 1: Power BI

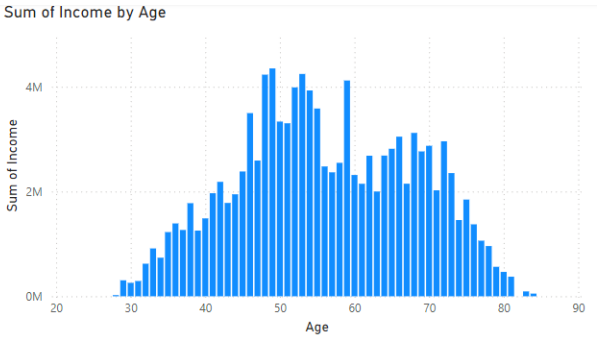
Visualizations

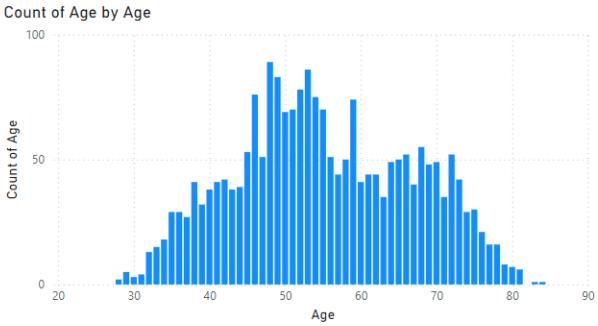


Insights based on the visualizations

- Income generally increases with higher education levels.
- Marital status affects income, with widowed and divorced individuals earning the most on average.
- Most people have a graduation-level education, which correlates with moderate to high income levels.

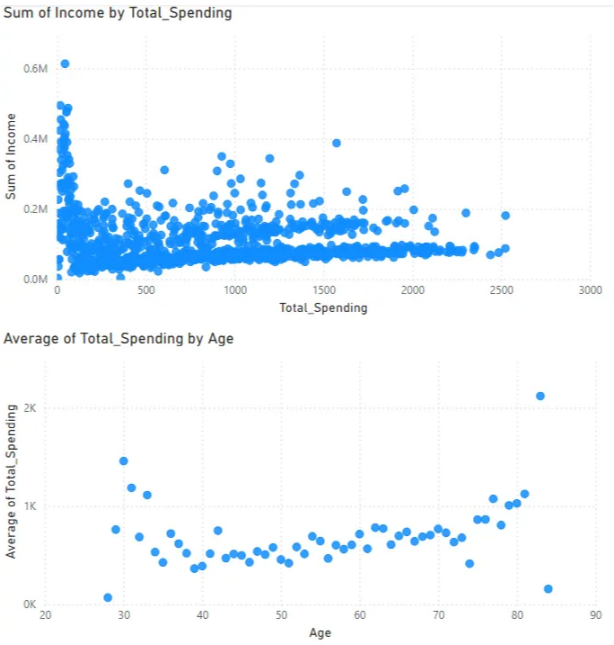






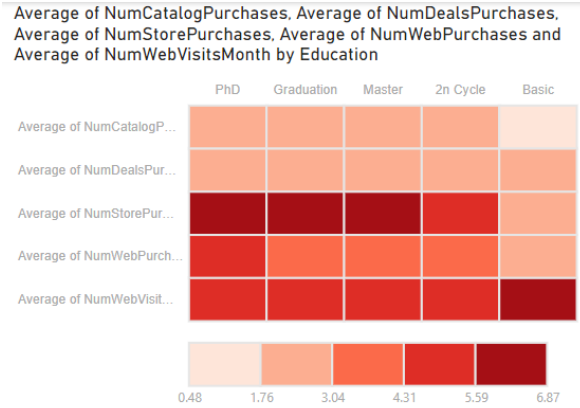
Insights based on the visualizations

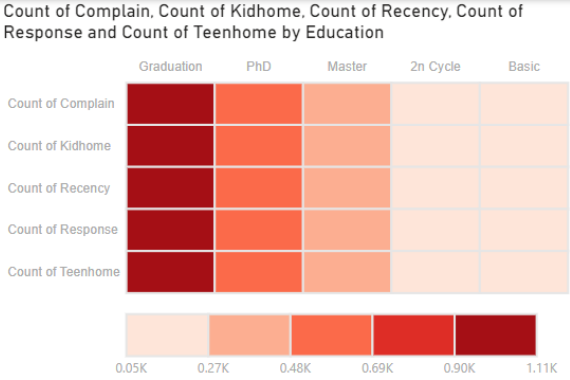
- Most individuals earn less than $50K, with very few in the high-income bracket.
- People in their 40s tend to have the highest earning potential, both in terms of total contribution and frequency.
- The workforce is predominantly middle-aged, with a balanced distribution between younger and older individuals.

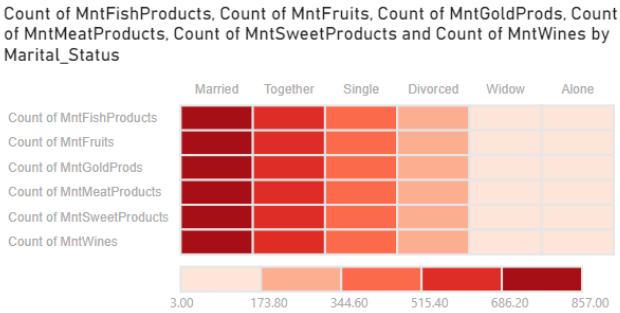


Insights based on the visualizations

- Most people maintain moderate spending levels regardless of their income
- There's no strong evidence that people automatically spend more just because they earn more
- The vast majority of spending activity clusters in the lower ranges, suggesting conservative spending habits are common
- Spending habits seem to follow life stages rather than showing a simple linear progression
- There are natural peaks and troughs that align with typical life events:
    - Higher spending in early adulthood (settling down phase)
    - More controlled spending in middle age (family/saving phase)
    - Increased spending in later years (retirement/healthcare phase)







Insights based on the visualizations

- Higher education generally correlates with more active purchasing and customer engagement
- More educated customers tend to interact more with the company (both purchases and customer service)
- Being in a relationship (married/together) strongly correlates with higher purchasing across all product categories
- Purchase activity gradually decreases from married couples to single individuals to those living alone
- Both education and relationship status appear to be strong predictors of consumer behavior
- The most active customers are typically highly educated and in relationships
- This suggests these demographic factors could be valuable for market segmentation and targeting

Data Preprocessing

1. Made the `Age` column in the following way by subtracting `Year_Birth` from 2024.
2. Made the `Total_Spending` column by adding the columns `MntWines`, `MntFruits`, `MntMeatProducts`, `MntFishProducts`, `MntSweetProducts` and `MntGoldProds`.
3. Handling missing values

   Only a very minute percentage of datapoints had missing values, mostly for the `Income` attribute, so they were removed.
4. Removing outliers and noisy data

   a. `Age`

      There were just three datapoints with strikingly high age, so they were removed.

   b. `Income`

      There was one datapoint which was way higher than anything else, so it was filtered out.

   c. `Marital_Status`

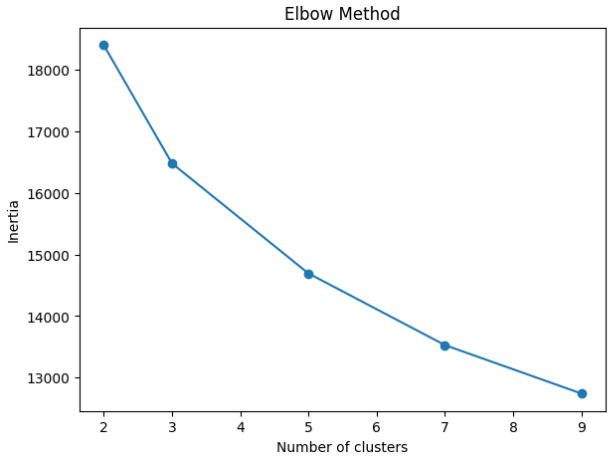
      There were a small number of datapoints with the values *Absurd* and *YOLO* so they were removed.
5. Removing redundant attributes

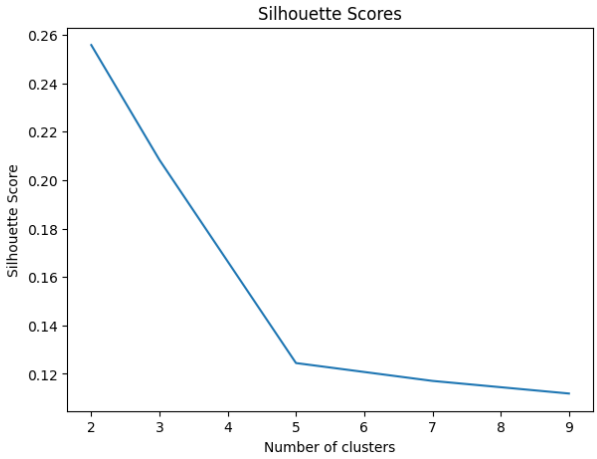
   The attributes `Z_CostContact` and `Z_Revenue` both have a cardinality of 1 so the columns were removed.

Part 2: K-Means Clustering Implementation

We implemented the K-Means clustering algorithm from scratch and applied it to the preprocessed data we got from the previous part.

Values of `k` tried: $2, 3, 5, 7, 9$





Silhouette Score Analysis

- The silhouette score measures how well points fit within their cluster compared to others, with higher scores indicating better-defined clusters.
- The score steadily decreases from $k = 2$, staying decently high around $k = 3$, and dropping sharply from $k = 5$.

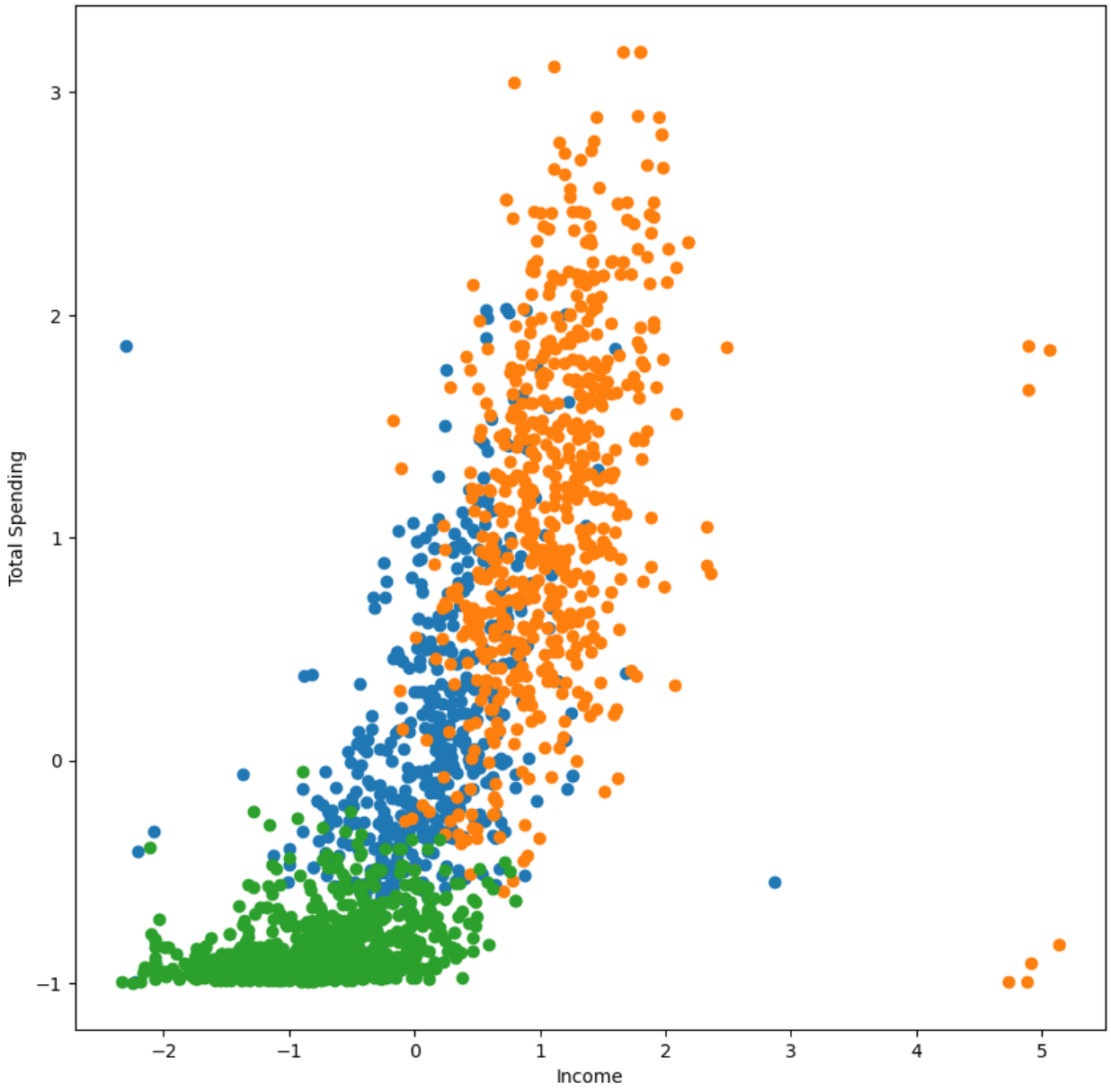
Elbow Method Analysis

- The elbow method identifies the point where the improvement in distortion slows down, indicating fewer benefits from adding more clusters.
- The elbow graph shows a sharp decrease between $k = 2$ and $k = 3$, with the slope steadily decreasing after $k = 3$. This suggests $k = 3$ is an appropriate choice for the "elbow" point.

Optimal Number of Clusters

$k = 3$ stands out as the optimal choice based on the elbow method, and it also achieves a strong silhouette score, indicating well-formed clusters.

Analyzing the Clusters

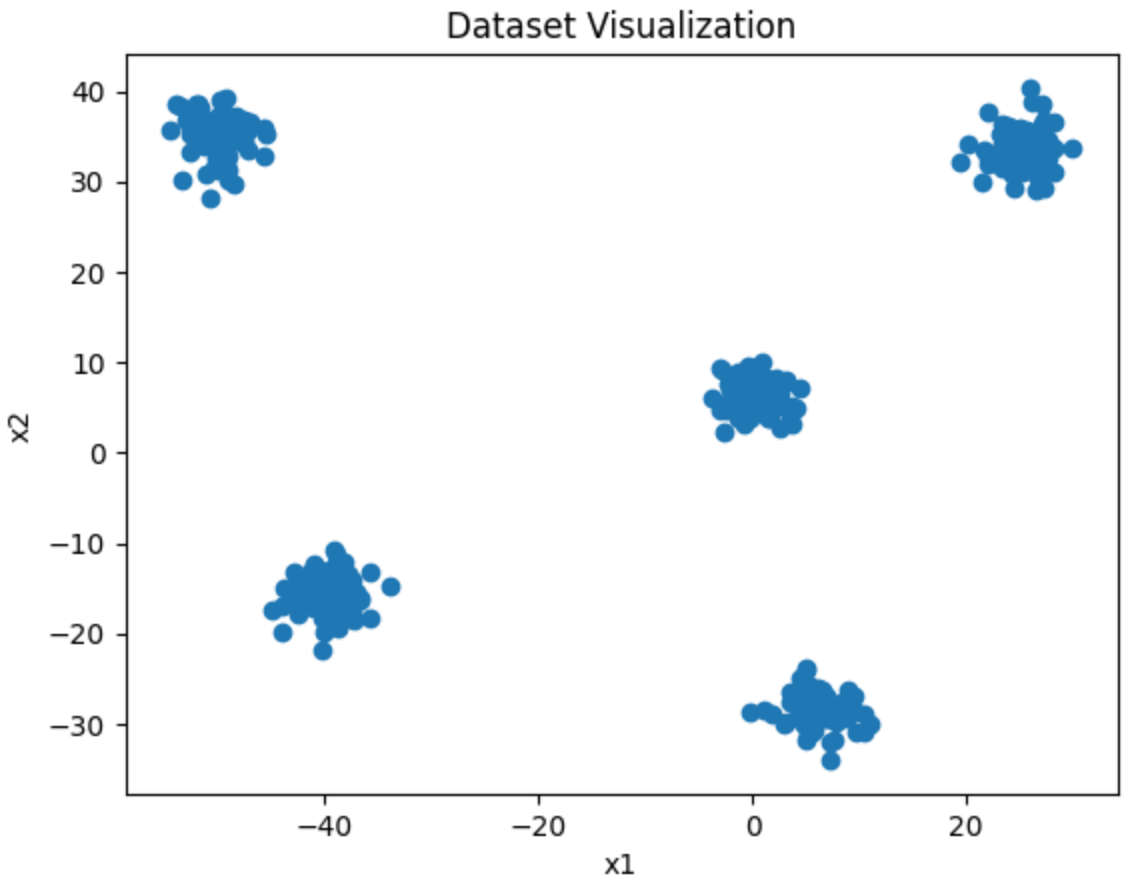


It is difficult to fully analyze given the high dimensionality of the data but here, if we just look at the clusters from the point of view of two of the features, `Income` and `Total Spending` of the customer, we can clearly see how the clusters are well defined as those with similar `Income` and `Total Spending` are grouped together.

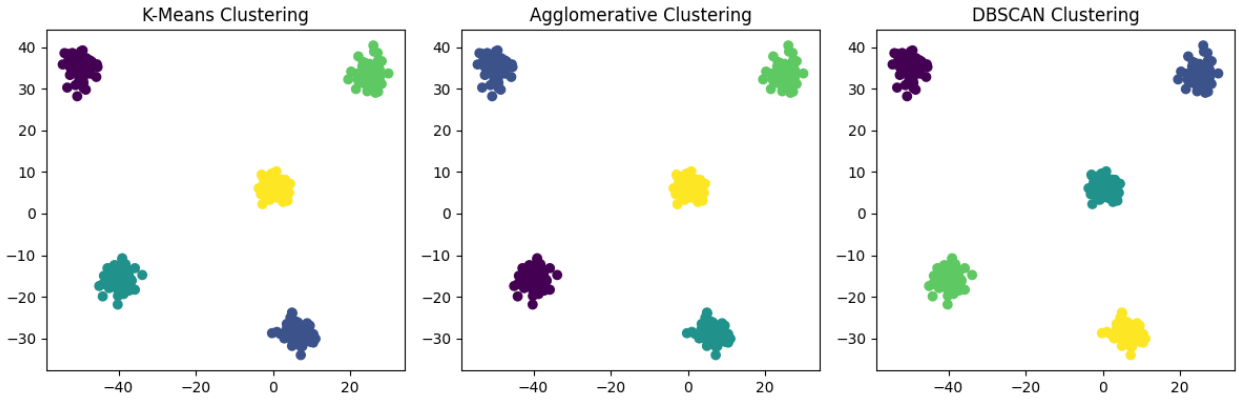
Part 3: Clustering with Different Algorithms

`Compact.csv`

Dataset



Applying all 3 clustering algorithms



Metrics

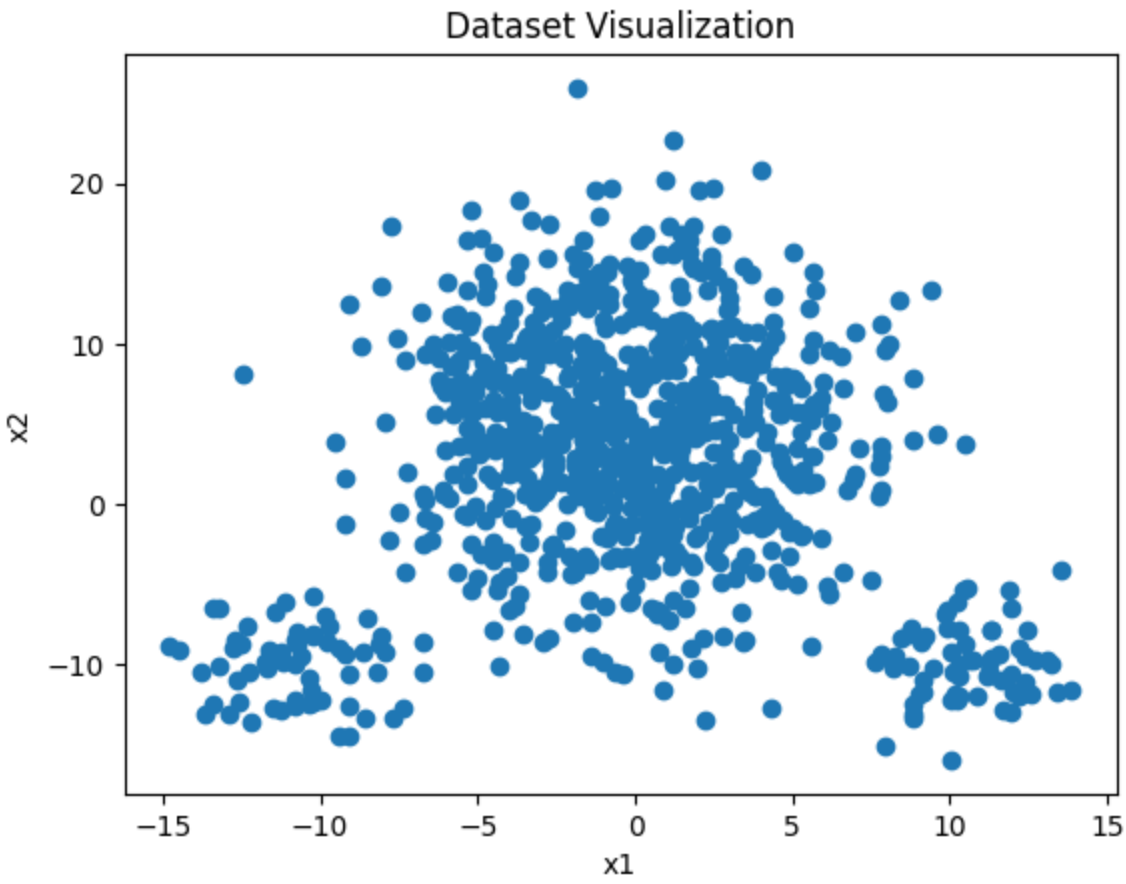
	Silhouette Score	Intra-Cluster Distance	Noise Points
K-Means	0.91178	3163.693808	0
Agglomerative Clustering	0.91178	1987.446642	0
DBSCAN	0.91178	3163.693808	0

Analysis

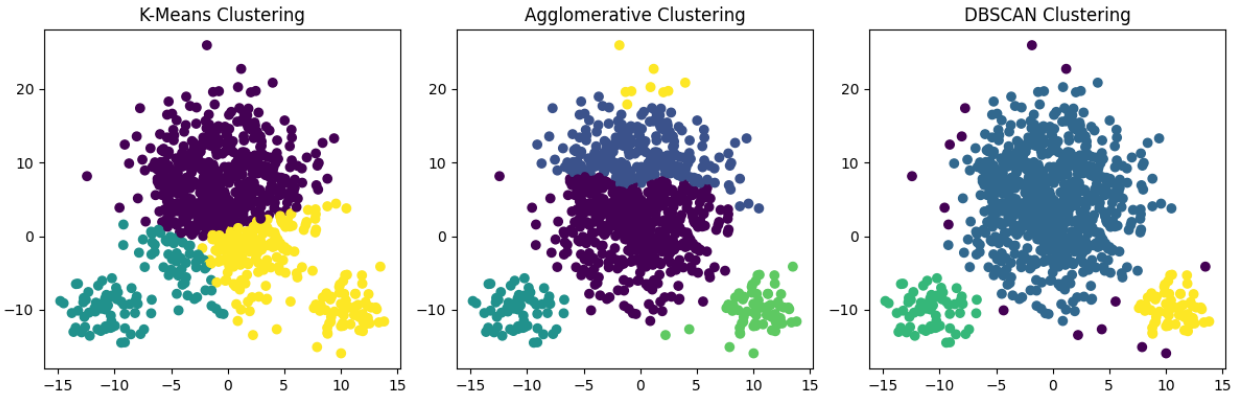
It appears that Agglomerative clustering is the best method for this dataset. While the silhouette score is same for all the methods and even visually, all the methods seem to have done a good job, if we look at inter-cluster distance, we can see that Agglomerative clustering has the least inter-cluster distance. This means that the clusters are more compact and well separated. This is the reason why Agglomerative clustering is the best method for this dataset.

`Skewed.csv`

Dataset



Applying all 3 clustering algorithms



Metrics

	Silhouette Score	Intra-Cluster Distance	Noise Points
K-Means	0.429143	31014.871887	0

	Silhouette Score	Intra-Cluster Distance	Noise Points
Agglomerative Clustering	0.348568	21640.576044	0
DBSCAN	0.498486	38525.676717	15

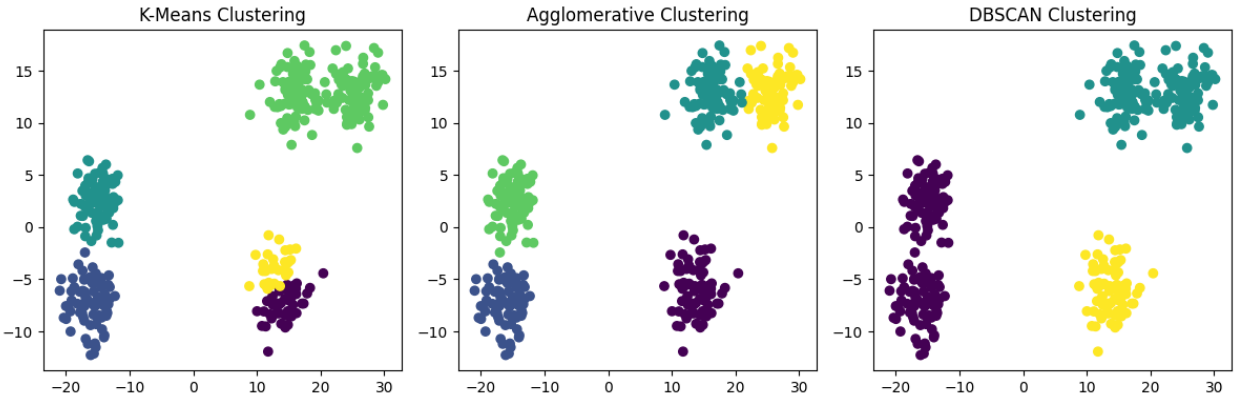
Analysis

It appears that DBSCAN is the best method for this dataset. While it has the highest inter-cluster distance, it also has the highest silhouette score, and visually, it is the only method that has been able to separate the clusters well. This is the reason why DBSCAN is the best method for this dataset.

`Subclusters.csv`

Dataset

Applying all 3 clustering algorithms



Metrics

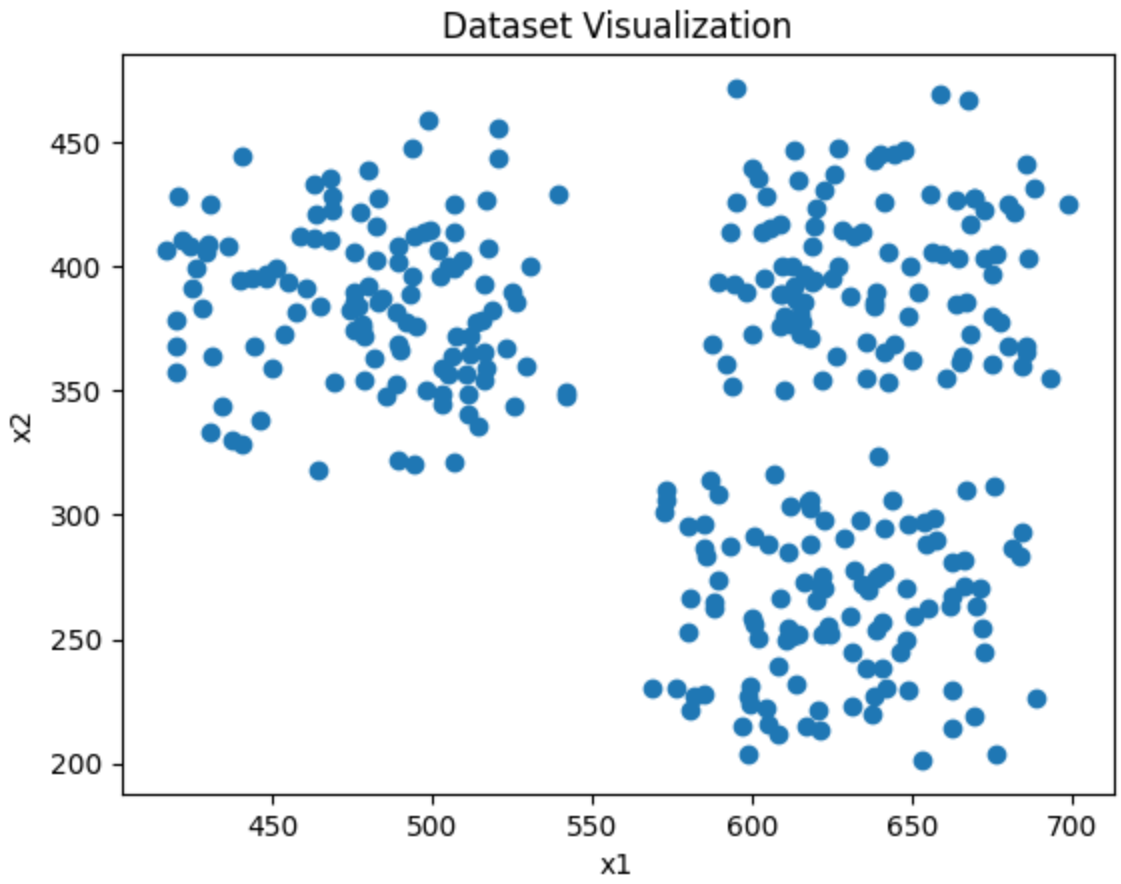
	Silhouette Score	Intra-Cluster Distance	Noise Points
K-Means	0.586788	6626.821232	0
Agglomerative Clustering	0.668312	2058.387038	0
DBSCAN	0.738949	10795.396187	0

Analysis

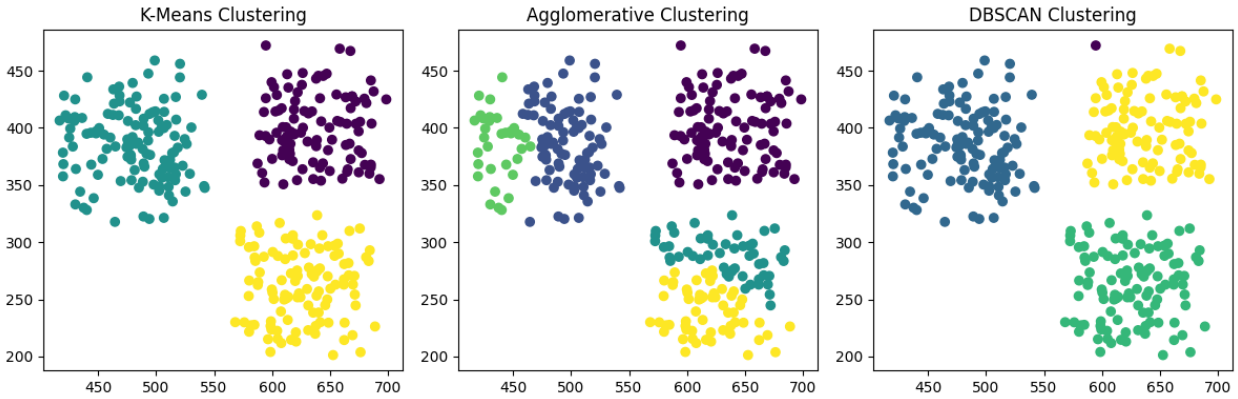
In this case, it is hard to say which method is the best. K-means has the lowest silhouette score, an average intra-cluster distance, and visually, it does a good job except for one of the clusters. DBSCAN has the highest silhouette score, but also the highest intra-cluster distance, and visually, while it does separate the clusters well, it seems like it can't capture the subclusters well. Agglomerative clustering has an average silhouette score, but it has the lowest intra-cluster distance and visually, it has separated the clusters well, including the subclusters. This is the reason why Agglomerative clustering is the best method for this dataset.

`wellSeparated.csv`

Dataset



Applying all 3 clustering algorithms



Metrics

	Silhouette Score	Intra-Cluster Distance	Noise Points
K-Means	0.613946	629926.614046	0

	Silhouette Score	Intra-Cluster Distance	Noise Points
Agglomerative Clustering	0.396587	376185.033656	0
DBSCAN	0.614976	622758.765930	1

Analysis

Both K-means and DBSCAN do a really good job in this case while Agglomerative clustering is not able to separate the clusters that well and has the lowest silhouette score, though it has the lowest intra-cluster distance, but that is possibly because it has over-clustered the data. Both K-means and DBSCAN have almost equally high silhouette scores and also almost equal intra-cluster distances. However, DBSCAN does validly identify one noise point, which K-means doesn't. This is the reason why DBSCAN is the best method for this dataset.

Silhouette Scores across all Datasets

	Compact	Skewed	Subclusters	wellSeparated
K-Means	0.91178	0.408248	0.668823	0.613946
Agglomerative Clustering	0.91178	0.348568	0.668312	0.396587
DBSCAN	0.91178	0.498486	0.738949	0.614976