Clustering and Fitting

Student ID:23024049

Name: Ajay Santhosh Kavitha Veeramani

GitHub link:
https://github.com/ajaysanthoshkv/ADS-Clustering-and-Fitting.git

Abstract:

To perform simple mathematical calculations and acquire insights from the data, create graphs like line, box, and heatmaps. To analyze the number of clusters and perform linear regression using the social media user data set.

	age	gender	time_spent	platform	interests	location	demographics	profession	income	indebt	isHomeOwner	Ow
0	56	male	3	Instagram	Sports	United Kingdom	Urban	Software Engineer	19774	True	False	
1	46	female	2	Facebook	Travel	United Kingdom	Urban	Student	10564	True	True	
2	32	male	8	Instagram	Sports	Australia	Sub_Urban	Marketer Manager	13258	False	False	
3	60	non-binary	5	Instagram	Travel	United Kingdom	Urban	Student	12500	False	True	
4	25	male	1	Instagram	Lifestlye	Australia	Urban	Software Engineer	14566	False	True	
...	...	...	...	...	...	...	...	...	...	...	...	...
995	22	female	8	Instagram	Lifestlye	United Kingdom	Rural	Marketer Manager	18536	False	True	
996	40	non-binary	6	YouTube	Travel	United Kingdom	Rural	Software Engineer	12711	True	False	
997	27	non-binary	5	YouTube	Travel	United Kingdom	Rural	Student	17595	True	False	
998	61	female	4	YouTube	Sports	Australia	Sub_Urban	Marketer Manager	16273	True	True	
999	19	female	8	YouTube	Travel	Australia	Rural	Student	16284	False	True	

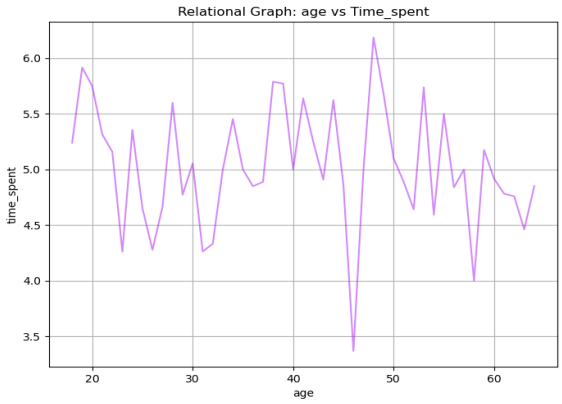
1000 rows × 12 columns

The data set comprises 1,000 entries, exhibiting an average age of 41 years, a mean time spent of 5.03, and an average income of around 15,015, according to descriptive statistics. With a standard deviation of 13.5 years, the age distribution is broad. The income has a standard deviation of 2,959, which indicates a higher spread. Age and time spent have little association (-0.034), whereas age and wealth have even less of a link (-0.087). Income varies little among age groups, with the highest average income occurring at age 18.

Graphical Datas:

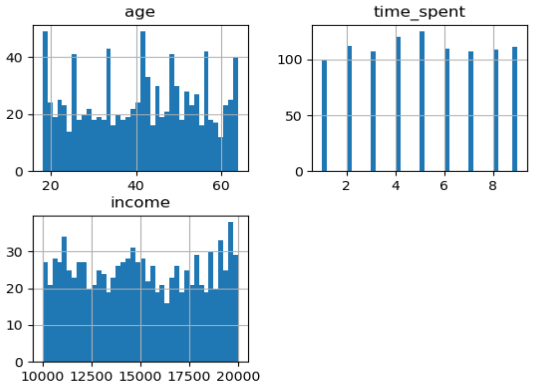
Line Graph:

The x-axis shows age, spanning from 20 to 60, while the y-axis shows time spent, with values ranging from 3.5 to 6.0. The line plot illustrates the differences in time spent at different ages. Those over 45 appear to utilize social media far less than other age groups.



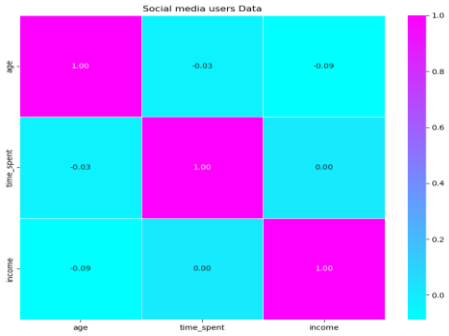
Histogram:

The first graphic, which ranges from about 0 to 70 years, shows the distribution of ages within a population or group. Time intervals ranging from 0 to 10 units may be used in the second chart to represent the amount of time spent on media. The income distribution among a group could also be represented by an income chart, with values ranging from around 10,000 to 20,000 units of money.



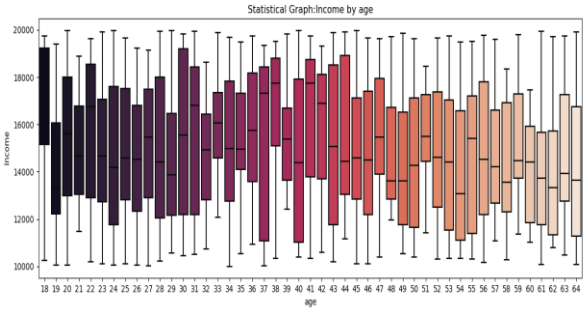
Heatmap:

The picture is a heatmap that shows the correlation data between users of social media. Age, time spent, and income are the factors that are being compared. As might be predicted, the diagonal values of 1.00 show a perfect association with one another. The correlation between the various variables is displayed by the off-diagonal values, all of which are near to 0, indicating little to no linear link between these components.



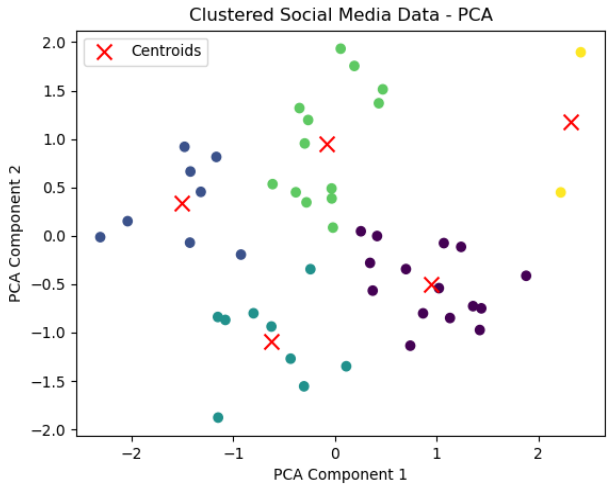
Box Plot:

A box plot shows the median, quartiles, and outliers of the income distribution for a specific age group. The age-related differences in whisker lengths and box sizes suggest that the income variation varies with age.



Clustering:

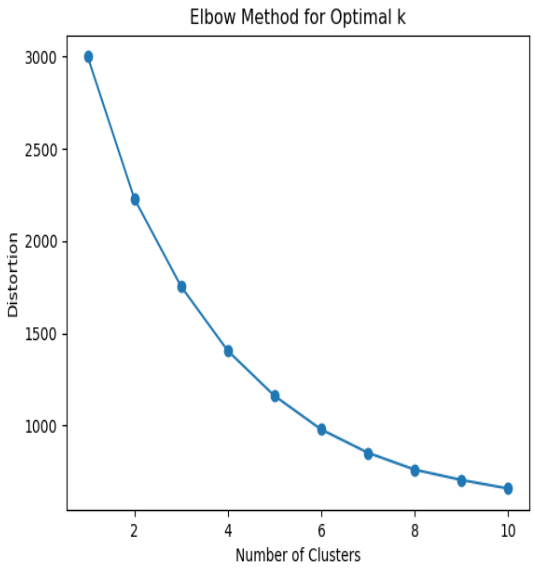
The labels "PCA Component 1" and "PCA Component 2" on the horizontal and vertical axes, respectively, imply that the high-dimensional original data has been divided into two principal components for the purpose of display. To find patterns or groupings in the social media data, cluster analysis can be performed to visualize how the data points are organized around the five input centroids.



Elbow method:

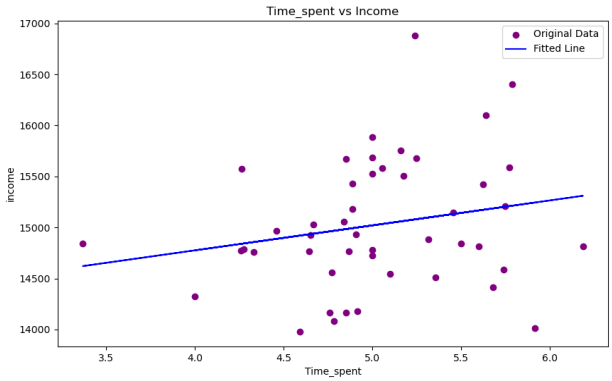
"Number of Clusters," with 2 to 10 possibilities. The best number of clusters (k) may lie at the location where the rate

of reduction abruptly changes, like an "elbow," according to the plot's falling curve, which starts high and then plateaus. The ideal equilibrium between the number of clusters and the within-cluster variance is reached at this stage, where adding more clusters does not result in a significant decrease in distortion.



Fitting:

- The fitted line, which displays the trend or relationship between time spent and income based on the data points, is represented by a blue line on the graph. In general, the pattern indicates that revenue tends to increase along with time spent. Original Data is represented by purple dots, Fitted Line by a blue line.



Silhouette:

When the number of clusters increases from 2 to 4, the silhouette score often improves on the graph, indicating improved cluster cohesion and separation. The silhouette score peaks at 4 clusters, and after that it slightly drops and swings as the number of clusters increases. This suggests that, depending on the type of data being examined, 4 clusters may be the ideal amount.