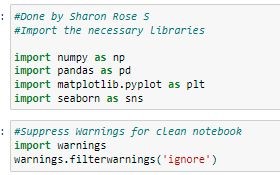
IBM Machine Learning Professional

Certificate

IBM\_Clustering - Assignment 4

**We'll be using the Mall Customer Dataset from Kaggle for this project.**

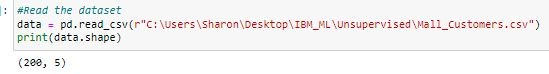
The dataset includes the following characteristics: Customer ID, Customer Gender, Customer Age, Customer Annual Income (in Thousand Dollars), and Customer Spending Score (based on customer behaviour and spending nature).

The following assignment problem statement :

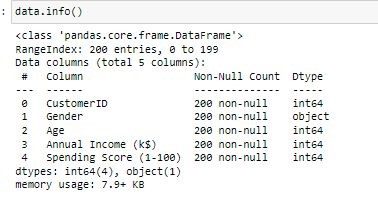
You own the shopping centre and want to learn more about the customers, such as who can easily converge into distinct clusters, so that the marketing team can correctly develop their approach.

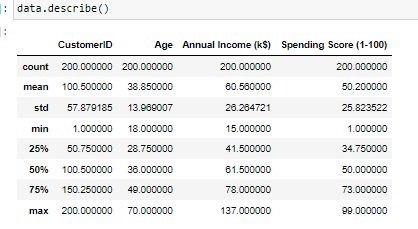
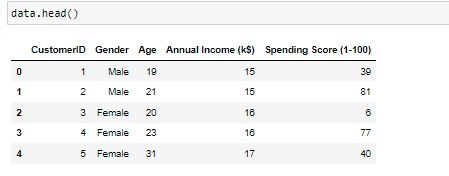
Let's import the necessary libraries:

Let's start by reading the Mall Customers.csv file into a pandas dataframe and looking at the shape of it.



Let's look at the dataset's details with the info(), head(), and description() functions.

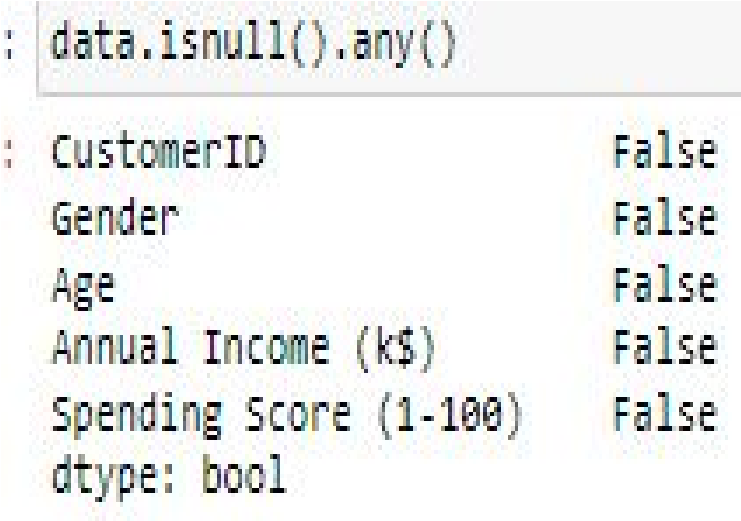




***Exploratory data analysis***

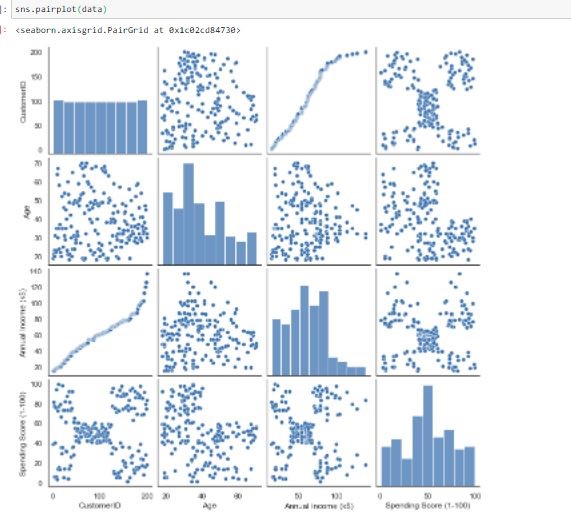
Let's get some data exploration going! We'll plan by looking for gaps in the data.

Missing Data: We'll use the built-in function isnull(), which is available within the dataframe, to check for missing data.

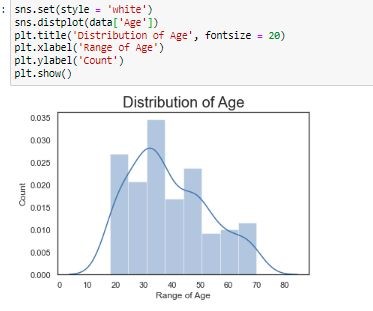


Therefore, there are no null values present in the dataset.

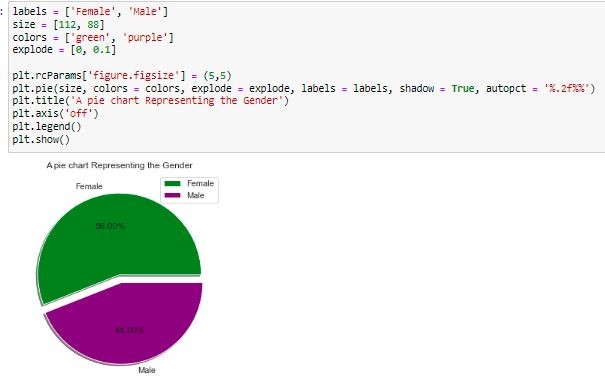
Let's use pairplot to visualise how the data is distributed.



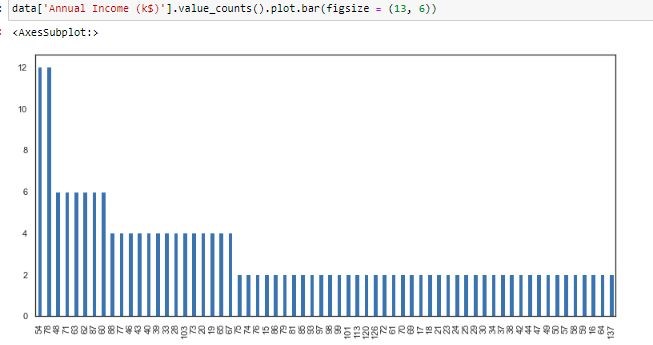
We can now use the distplot to see if the age follows a normal distribution or not, which will help us figure out what the average age of mall visitors is.



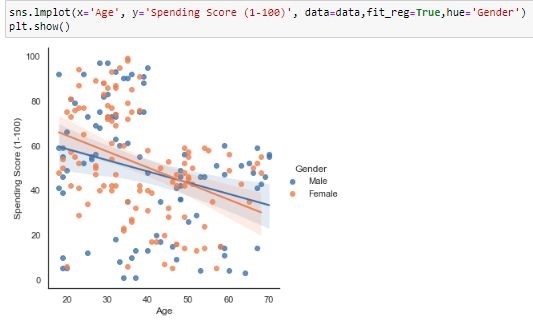
Looking into the gender distribution.



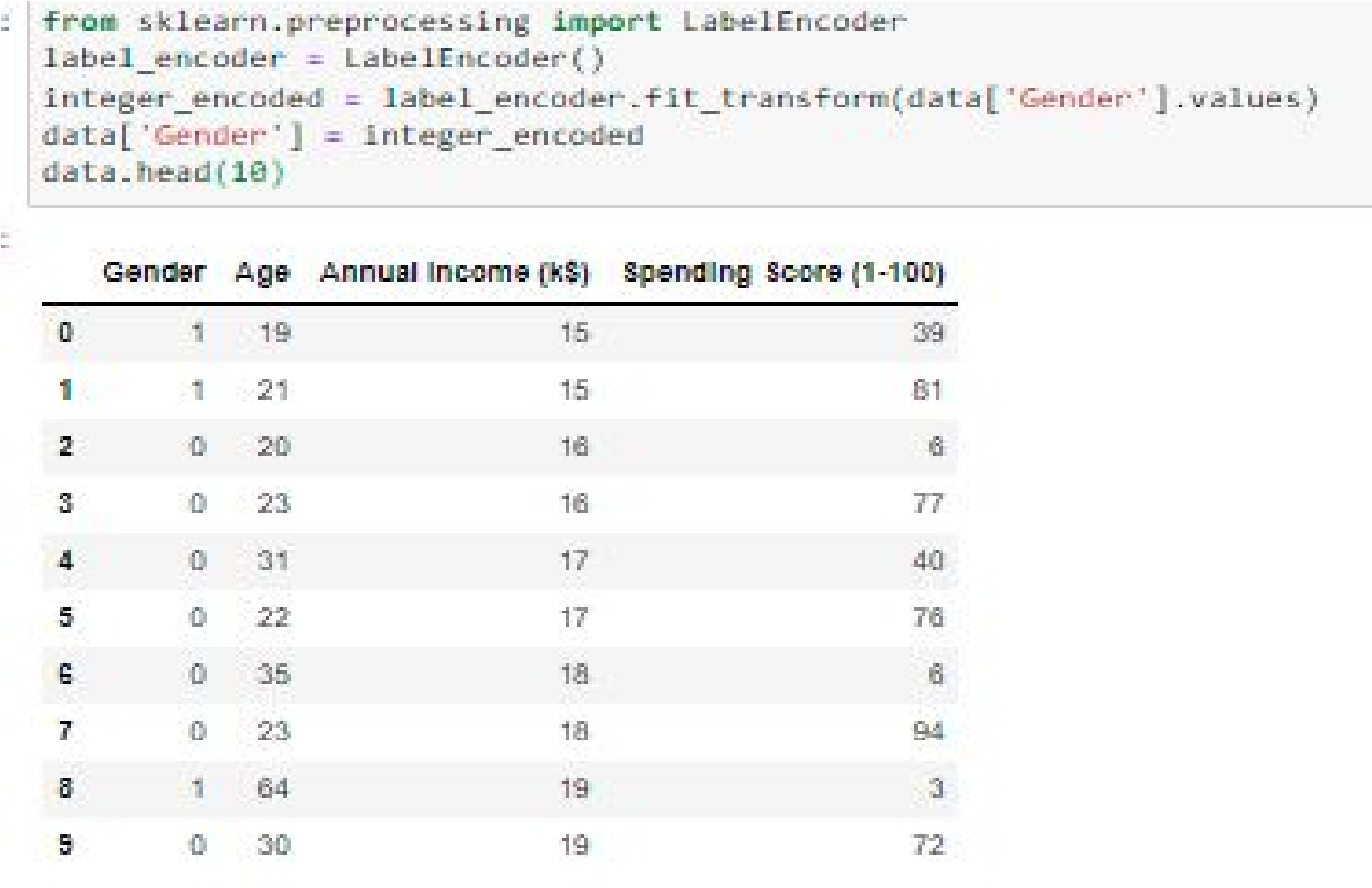
Let's utilise a bar plot to see the customers' annual income.



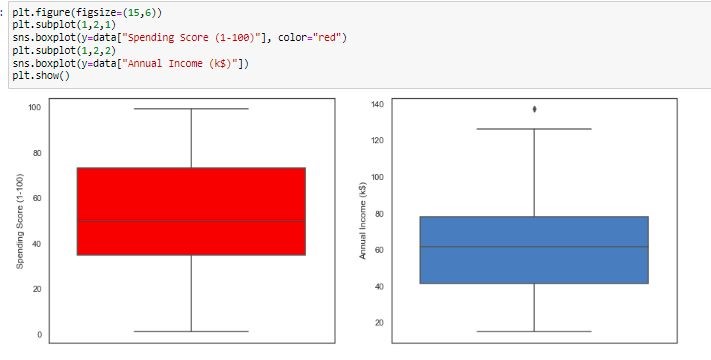
Let's look at the amount spent by male and female consumers in relation to their age using implot.



Encoding the categorical feature Gender is the next step.



Let's use boxplot to see whether there are any outliers in terms of income and spending score features:

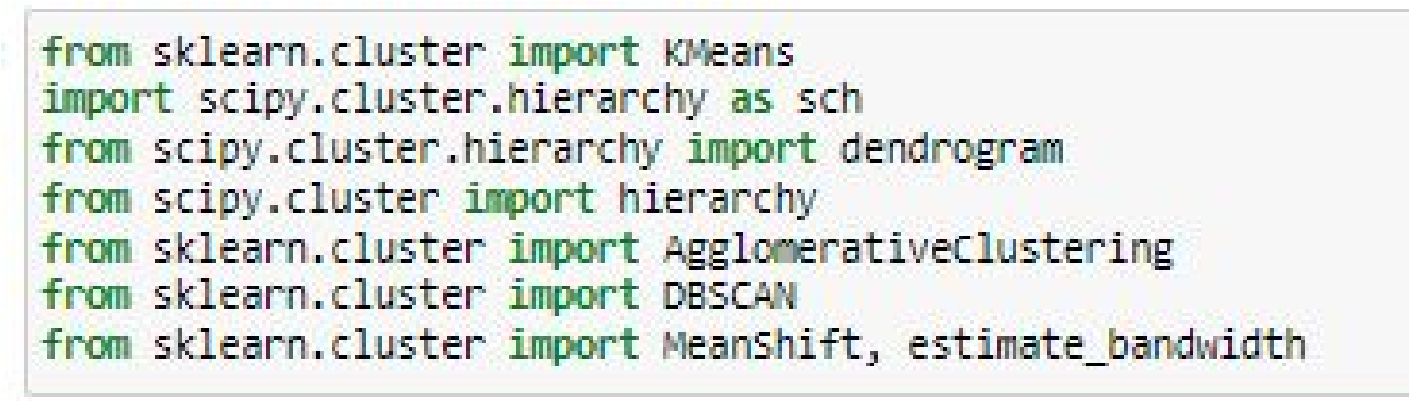


We can see which age groups of customers frequented the mall the most using a bar plot.



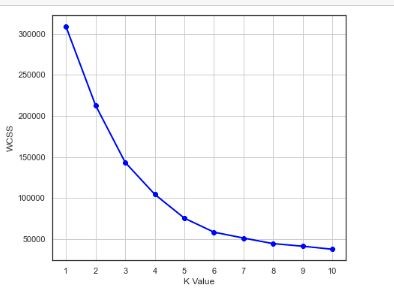
Customers aged 26 to 35 tend to be the most regular visitors to the mall, according to the graph above.

Let's get the libraries we'll need.



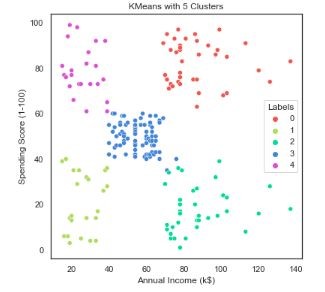
Train the following models based on the data's Annual Income and Spending Score: K Means, DBSCAN, Agglomerative Hierarchical Clustering andMeanshift

1. K Means 

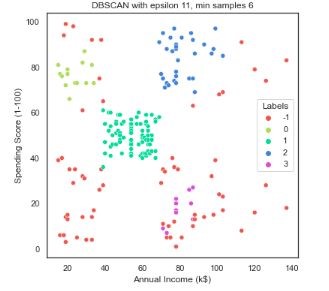
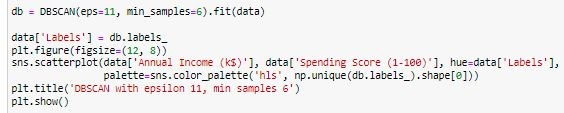


We can see from the elbow approach that the optimal number of clusters is 5.

Let's see how K Means clustering works in practise.

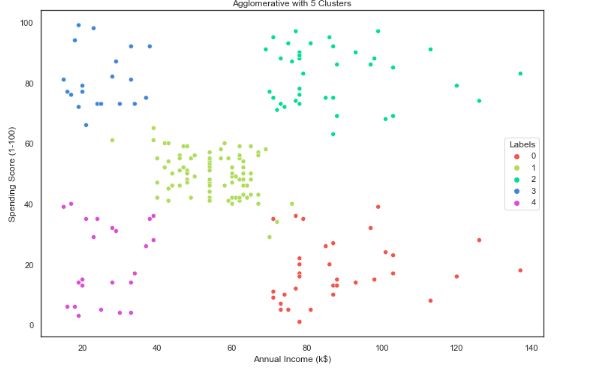


1. DBSCAN Clustering



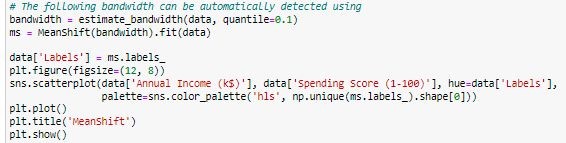
As we can see, DBSCAN doesn't perform well because the density of our data isn't extremely high. The term -1 implies outliers, thus it will be displayed as such. We could have done better if we had more data.

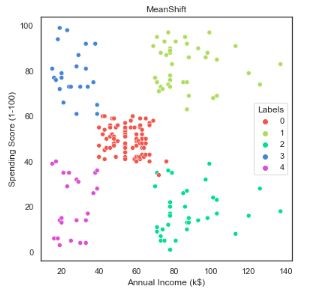
1. Agglomerative hierarchy Clustering



1. Mean Shift Clustering

Meanshift: The approach establishes the number of clusters automatically rather than relying on a parameter bandwidth to decide the size of the zone to search over. This value can be set manually, but it can also be computed using the estimate bandwidth function, which is used when the bandwidth isn't supplied.





Conclusion and suggestion:

As a result, as we showed in the previous section, the clustering results for the same dataset are nearly identical. If we have a really large dataset, the shape of clusters is likely to change slightly. As a result, we can apply any clustering algorithm to our problem statement other than DBSCAN clustering, which does not perform well due to the low density of our data. Outliers are implied by the phrase -1, thus it will be displayed as such. If we had additional data, we could have done a better job.