Data Science deviated people into groups by clustering:

- Created a tool that to deviated people into groups by clustering to help company

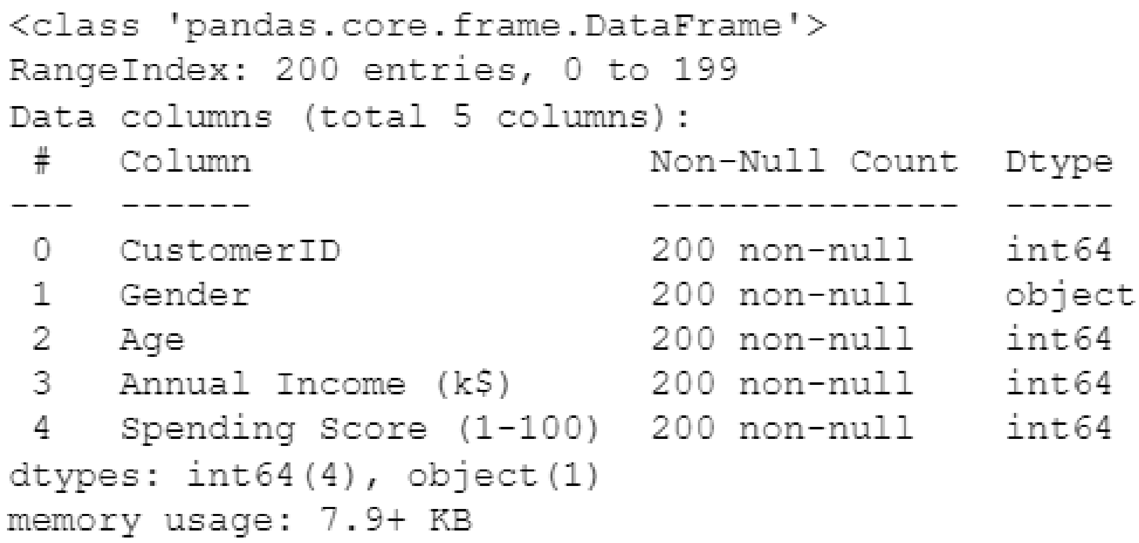
Packages: pandas, numpy, sklearn, matplotlib, seaborn

The solving mechanism

- build machine learning model using python

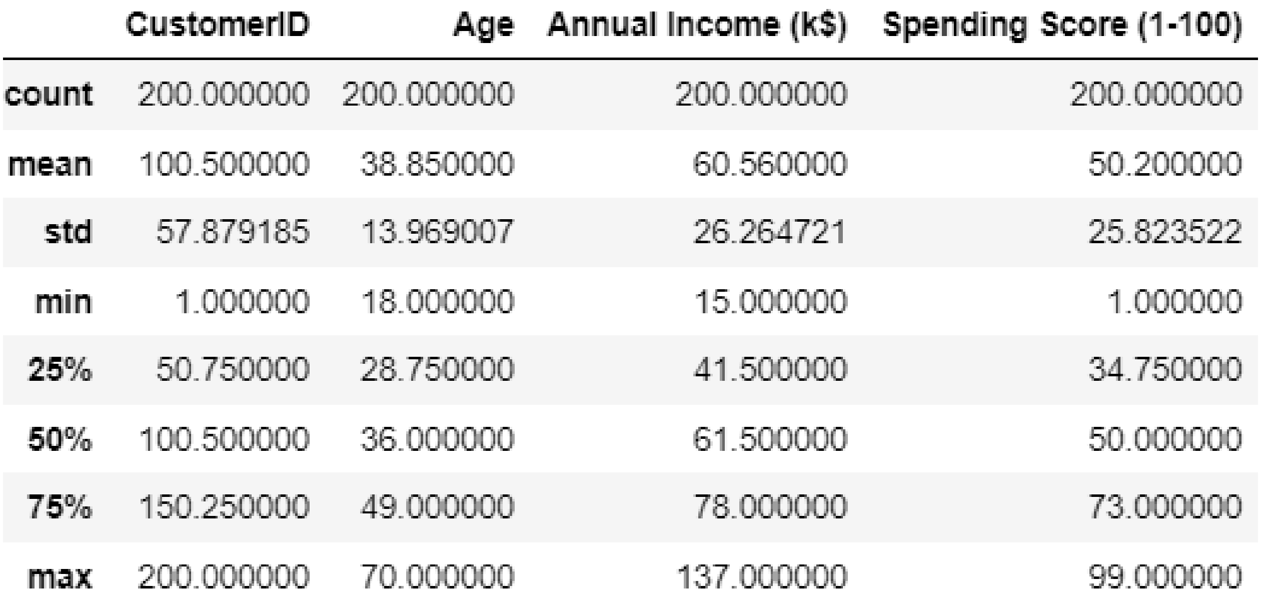
Describe the dataset

- Data source:
  - Mall Customer Segmentation Data | Kaggle
- Data description
- I use pandas library to description dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

  - form output I know number of rows (13320) and num of columns (9)
  - name of columns and data type for each column
  - number of null values in columns (Ex: bath has 53 sell null)



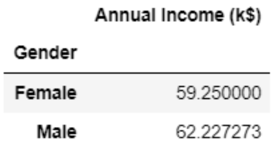
	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

- I conclude from this table count , mean , min , median , max , standard deviation
- From this information I know count of value in each column
- Std mean standard deviation it help us to know the spread of values
- Max , Min , mean , Median of each column

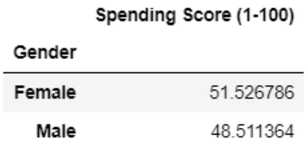
descriptive statistics and data distribution charts

I looked at the distributions of the data and the value counts for the various categorical variables. Below are a few highlights from the pivot tables.

Table 1



	Annual Income (k$)
Gender	
Female	59.250000
Male	62.227273



	Spending Score (1-100)
Gender	
Female	51.526786
Male	48.511364

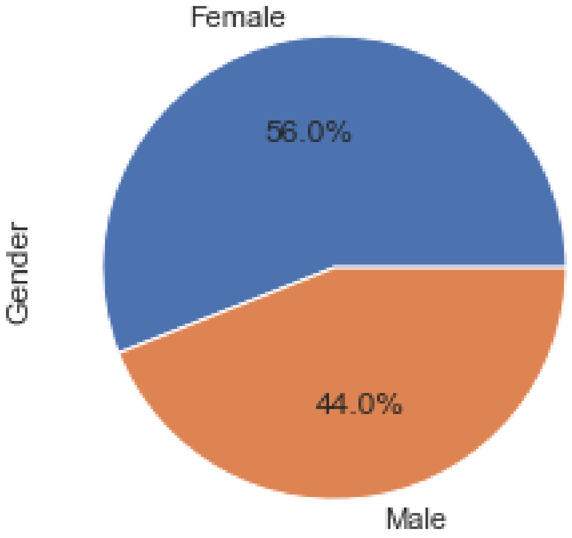


Figure 1

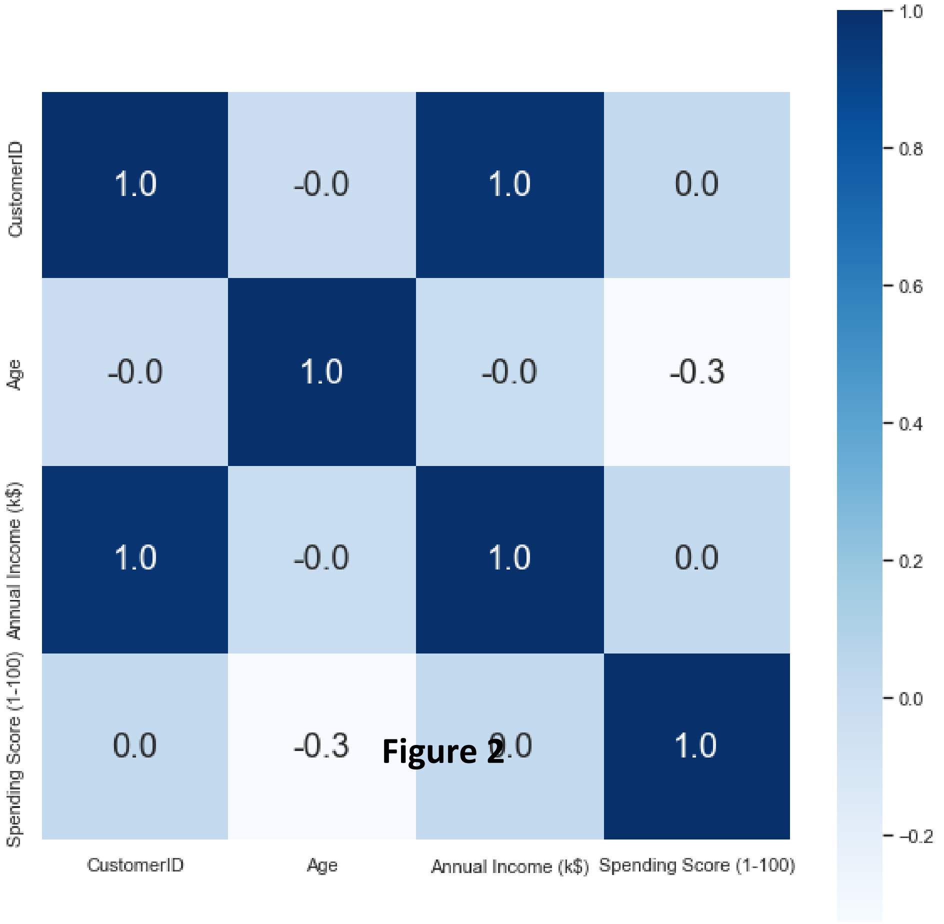


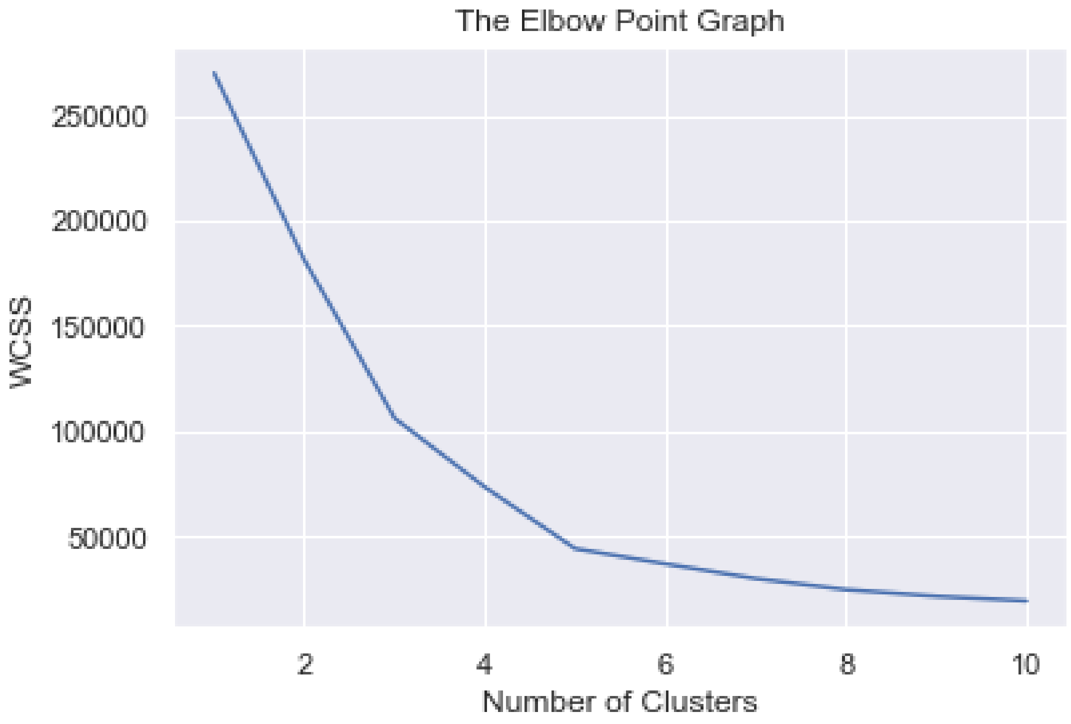
Figure 2

التعليق

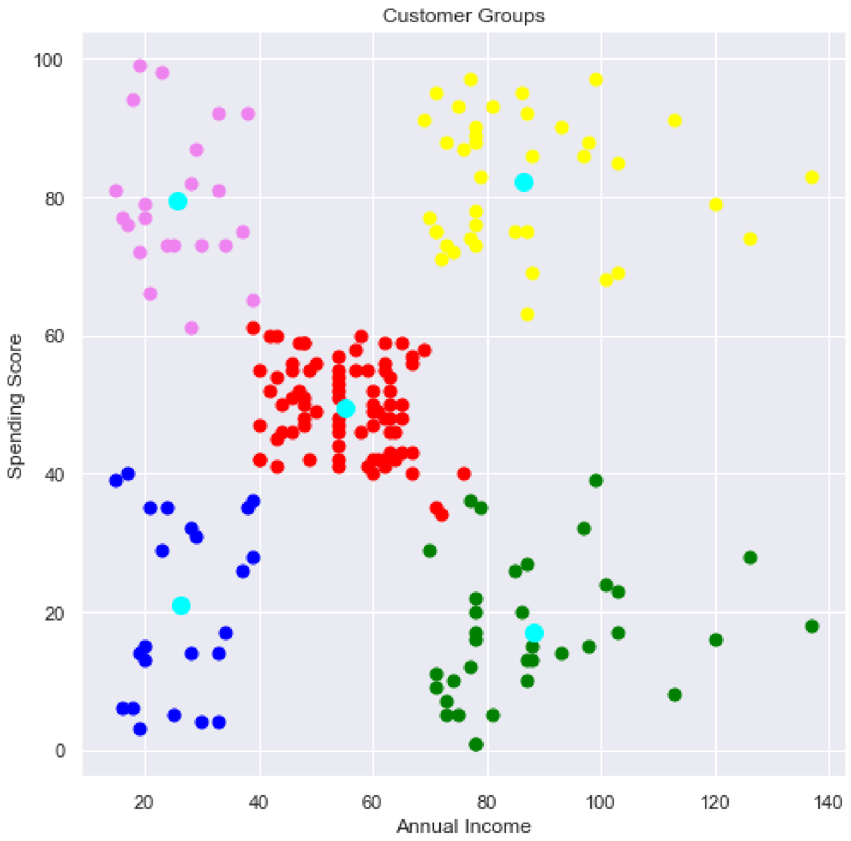
- (Table 1) pivot table that continent tow columns fistr column name location oher column price home in this location
- (figure 1) barplot in X name location Y count house in location
- (figure 2) this chart dis correlation between [total_sqft , bath , price , bhk]

Model Building

- First, Choosing the Annual Income Column for X variable.
- Then , Spending Score column , and Finding wcss value for different number of clusters



- Optimum Number of Clusters = 5
- Training the k-Means Clustering Model
- return a label for each data point based on their cluster for Y variable.
- 5 Clusters - 0,1,2,3,4
- Visualiing all the clusterrs


Customer Groups
Spending Score
Annual Income