Big Data

LAB-03

Timings: 11:30-2:30

Lab Protocols:

1. Carefully read and follow all instructions.
2. You need to do all 3 tasks (attendance + Evaluation + Submission)
3. You can search the basics of python, concepts, and syntax online.
4. TA isn't ment to resolve your PC or internet issues. TA is only here to guide you through lab.
5. No evaluation would be done after Lab's timing. So, keep the track of time.
6. Do keep in mind that sharing the code, discussing it during lab or looking for online solution is highly unethical, and all actions would be considered as plagiarism.
7. Plagiarism will result in serious penalty

Task1 – Do some preprocessing steps on data. Visualize your results (10 Marks)

Part (A): (5 Marks)

Perform The following necessary tasks for data cleaning.

1. Load the csv file
2. Do you analysis
3. Without loosing any data, give most suitable replacement for null values.
   Hint: each gender has its own pattern
4. Remove Duplicates
5. Remove any unnecessary feature from the data
6. Prepare it for visualization.

Expected results:

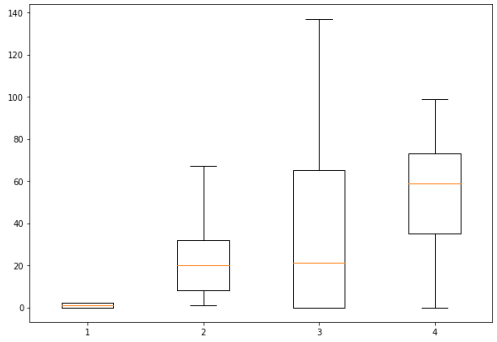
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 243 entries, 3 to 213
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Gender                  243 non-null    int64
 1   Age                     243 non-null    float64
 2   Annual Income (k$)      243 non-null    float64
 3   Spending Score (1-100)  243 non-null    float64
dtypes: float64(3), int64(1)
memory usage: 9.5 KB
```

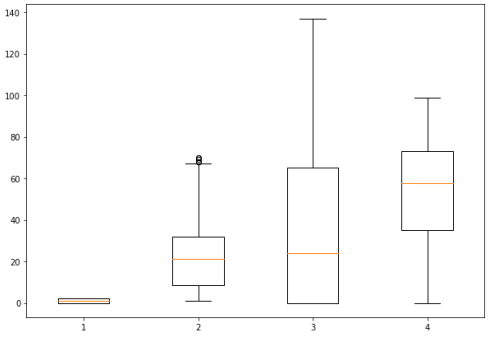
Part (B): (5 Marks)

Perform below activity

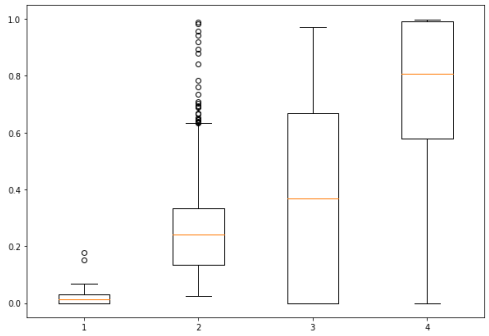
1. Visualize the data using box plot.
2. Remove outliers.
3. View data in scatter plot.

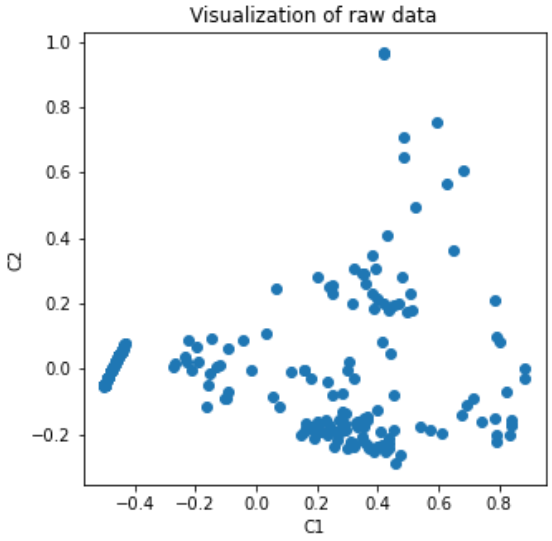
Expected results





Normalized





Task 2 – Apply Built in K mean Method (15 Marks)

Part (A): (2 Marks)

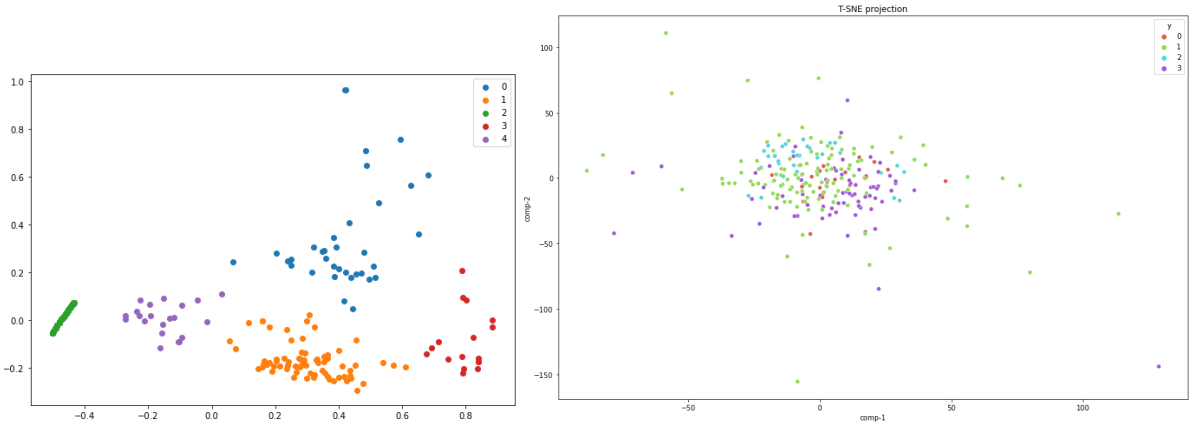
Do k-mean clustering on the data and pick k centers with hit and trial method.

- Visualize using scatter plot
- Visualize using TSNE

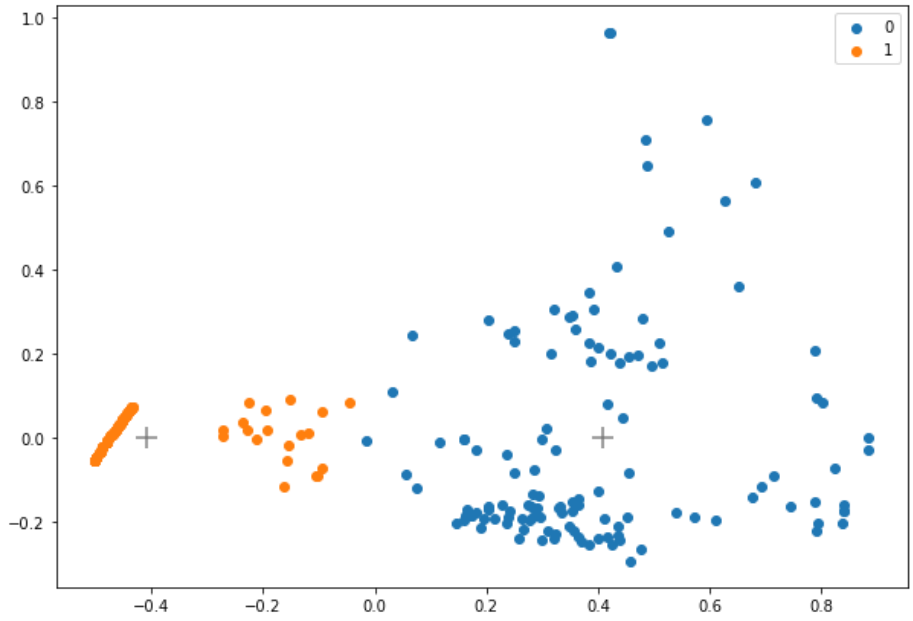
Do it in the end (10 Marks)

Visualize the results by placing X at cluster centers. (Use the above Matplotlib functionality)

Expected results:



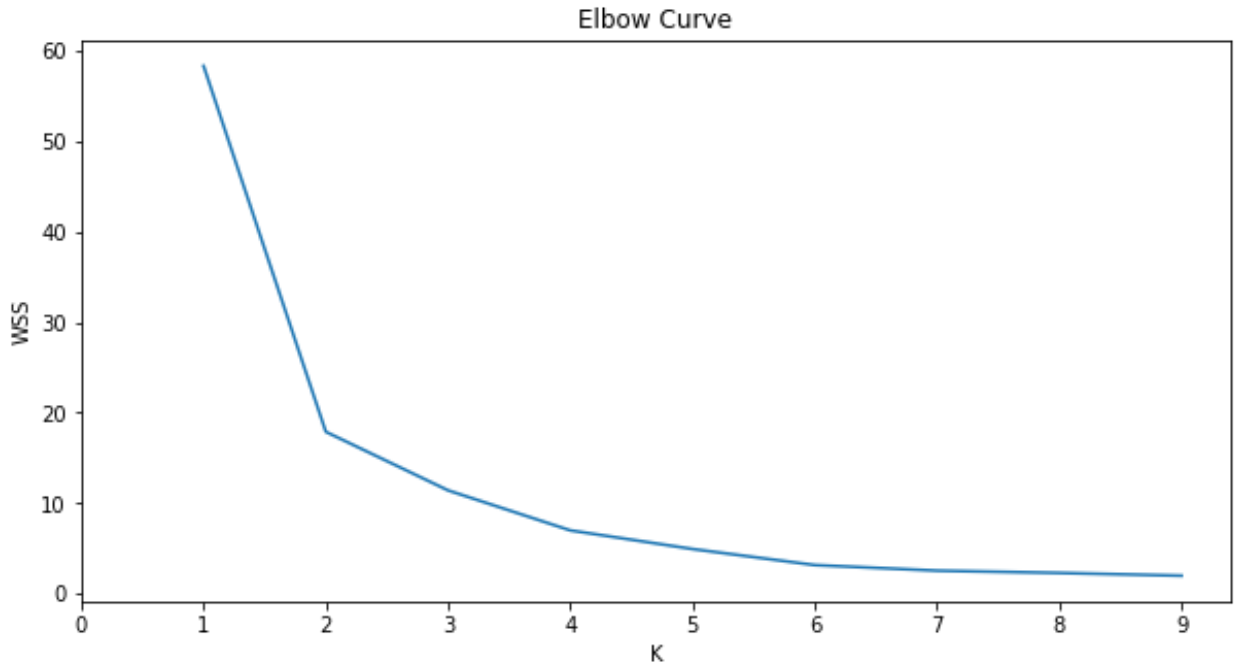
Do It in the end:



Part (B): (3 Marks)

Do k-mean clustering on the data and pick k centers with elbow Method. Custom implement, within sum of square functionality
Formula= (point- center) **2 for every point
Visualize the curve using matplotlib plotting.

Expected Results



Task 3 – Agglomerative Clustering Method (5 Marks)

Part (A): (3 Marks)

Use Dendro gram to represent appropriate cluster

Part (B): (2 Marks)

Apply Agglomerative clustering and display your results using scatter plot

Expected results