61070363 นายโสภณ วงศ์ใหญ่

K-Means, GMM Homework

Dataset

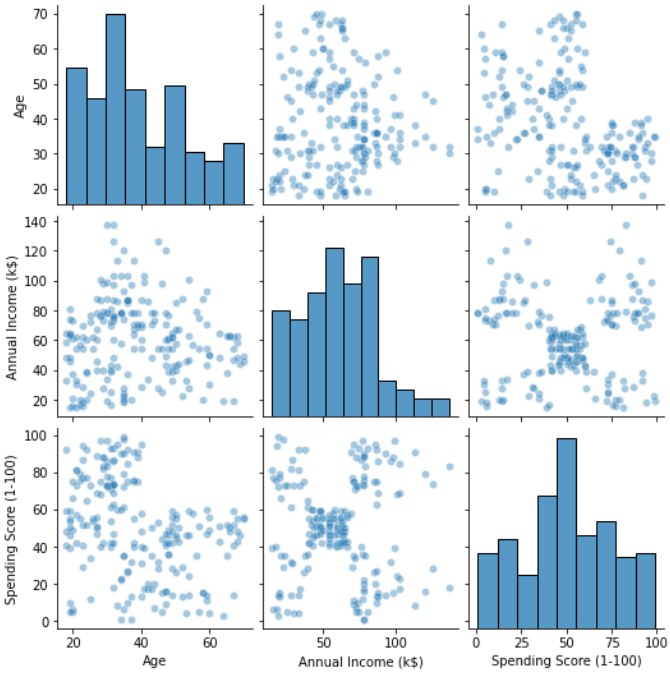
Dataset: Mall Customer Segmentation Data
URL: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python
Dataset Detail:

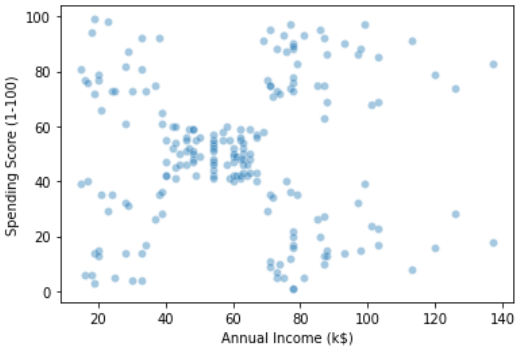
เป็นข้อมูลเกี่ยวกับลูกค้าของห้างสรรพสินค้าแห่งหนึ่งที่ต้องการจะจัดกลุ่มลูกค้า

- CustomerID: รหัสลูกค้า
- Gender: เพศ
- Age: อายุ
- Annual Income (k$): รายได้ต่อปี (มีหน่วยเป็น พันดอลลาร์สหรัฐ)
- Spending Score (1-100): คะแนนการใช้จ่าย หรือ คะแนนที่ห้างสรรพสินค้ากำหนดตามพฤติกรรมของลูกค้าและลักษณะการใช้จ่าย (มีหน่วยเป็น 1 - 100 คะแนน)

Feature Selection

เลือก feature มาเปรียบเทียบ โดยเปรียบเทียบ feature ที่น่าสนใจ เพื่อเลือก feature ที่เหมาะจะทำ clustering ได้แก่ Age, Annual Income (k$), Spending Score (1-100)

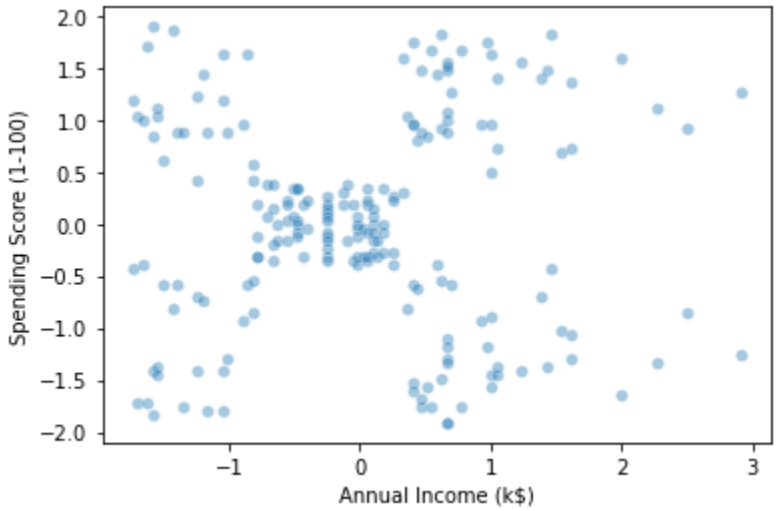




จากกราฟ pairplot (ด้านซ้าย) ทำให้ทราบว่า ควรเลือก Annual Income (k$) และ Spending Score (1-100) เนื่องจากมีลักษณะข้อมูลแบ่งเป็นกลุ่มค่อนข้างชัดเจนดังกราฟ (ด้านขวา)

Feature Scaling

ใช้ StandardScaler() เพื่อช่วยให้ data มี scale เล็กลง ได้ผลลัพธ์ดังนี้

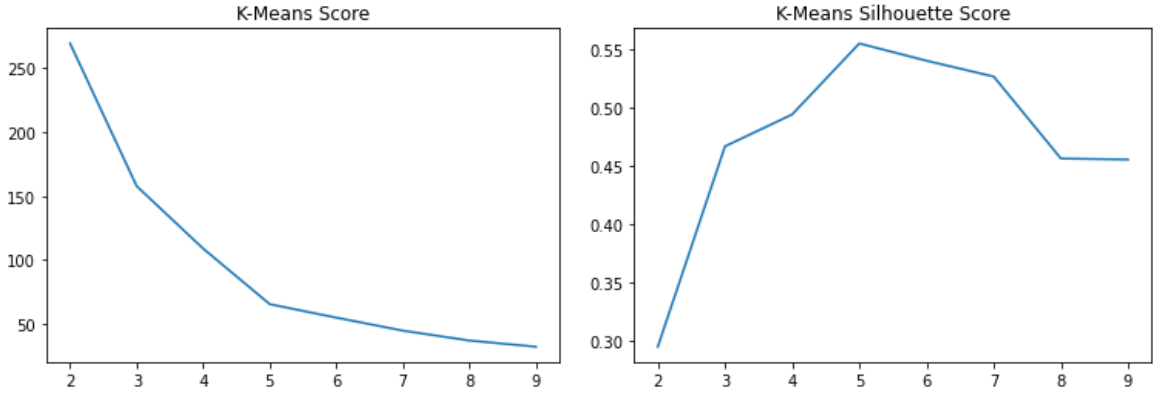


Modeling

ได้ทำการกำหนด random_state เพื่อให้ได้ผลลัพธ์คงเดิม โดยที่นี้ให้กำหนดเท่ากับ 0

K-Means Clustering

หา k ที่เหมาะสมสำหรับ K-Means

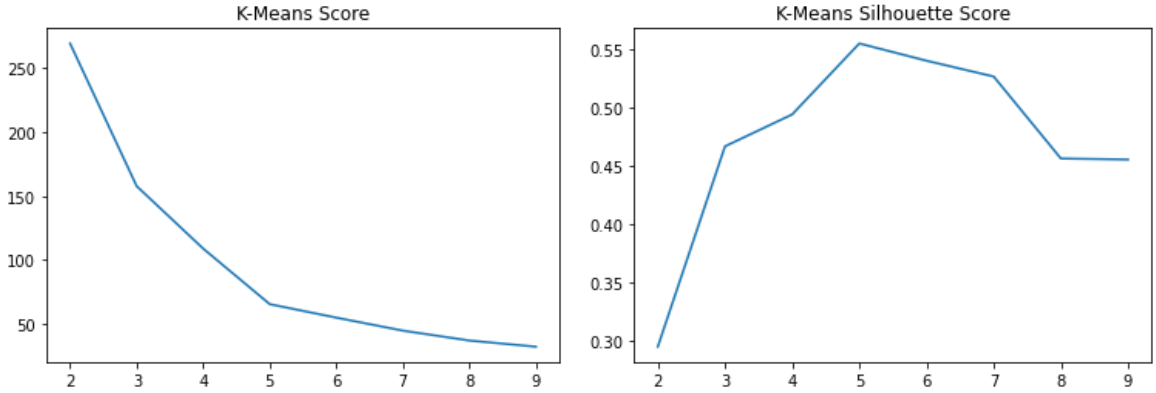


- Elbow Method (กราฟซ้าย): 5 โดยดูจากจุดที่หักที่สุด
- Silhouette Score (กราฟขวา): 5 โดยดูจากค่าสูงสุดที่ได้

ดังนั้น k ที่เหมาะสมสำหรับ K-Means คือ 5

Gaussian Mixture Model (GMM)

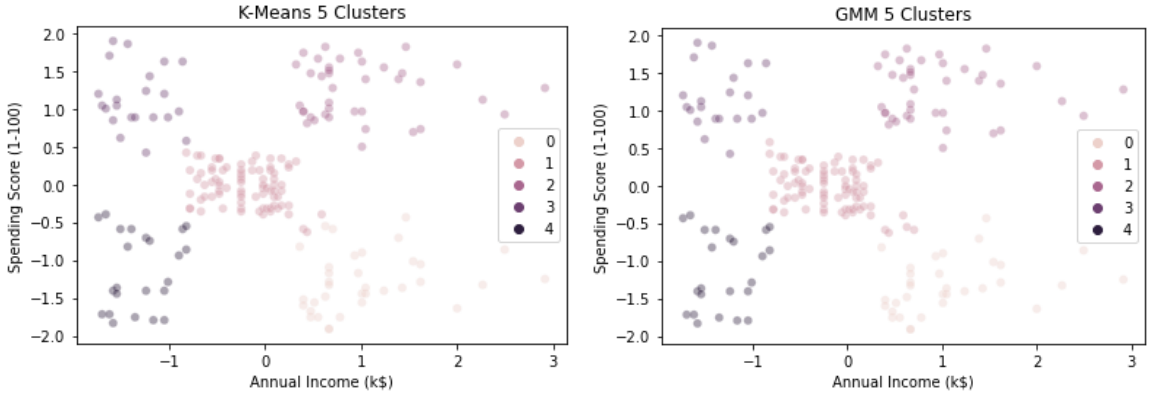
หา k ที่เหมาะสมสำหรับ GMM



- Elbow Method (กราฟซ้าย): 5 โดยดูจากจุดที่หักที่สุด
- Silhouette Score (กราฟขวา): 5 โดยดูจากค่าสูงสุดที่ได้

ดังนั้น k ที่เหมาะสมสำหรับ GMM คือ 5

k = 5



Conclusion

สรุปจากการทดลอง พบว่า ค่า k ที่ได้จากการหาค่า k ที่เหมาะสมของ K-Means และ GMM โดยใช้ Elbow Method และ Silhouette Score ได้ค่าเท่ากัน คือ k = 5 และผลลัพธ์จากการทำ clustering ของ model ทั้ง 2 ที่ k = 5 ให้ผลลัพธ์ใกล้เคียงกัน