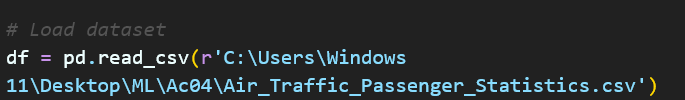
Air Traffic Clustering

## 1. การนำเข้าข้อมูลและการตรวจสอบข้อมูลเบื้องต้น

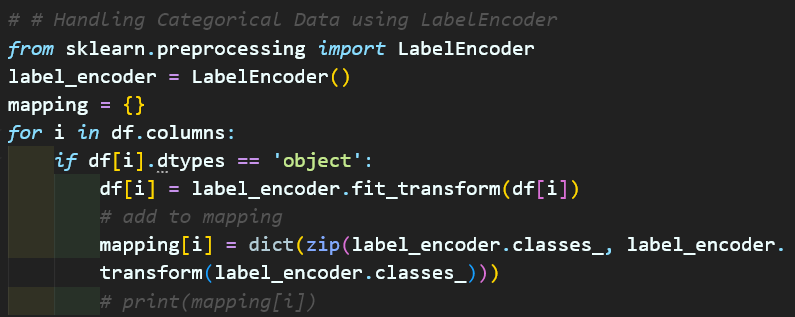
ในส่วนนี้จะเริ่มจากการนำเข้าข้อมูลจากไฟล์ Air\_Traffic\_Passenger\_Statistics.csv และแสดงข้อมูลตัวอย่างเบื้องต้น รวมถึงทำการตรวจสอบสถิติของข้อมูล เช่น ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน และอื่นๆ เพื่อให้เราเข้าใจโครงสร้างข้อมูลมากขึ้น

2. การเตรียมข้อมูล (Preprocessing)  
ในขั้นตอนนี้จะเริ่มจากการลบคอลัมน์ที่ไม่จำเป็น เช่น 'Activity Period' ตรวจสอบข้อมูลที่ขาดหายและลบแถวที่มีข้อมูลขาดหายออกเพื่อให้ข้อมูลที่ใช้ไม่มีช่องว่าง  
A screen shot of a computer program

Description automatically generated

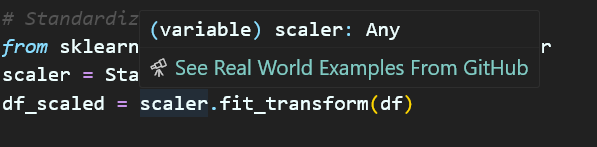
## 3. การจัดการข้อมูลเชิงหมวดหมู่ (Categorical Data)

ใช้ LabelEncoder เพื่อแปลงข้อมูลเชิงหมวดหมู่ (เช่น ชื่อสายการบิน) ให้เป็นตัวเลข ซึ่งเป็นขั้นตอนสำคัญในการเตรียมข้อมูลสำหรับการทำ Clustering เนื่องจากอัลกอริธึมที่ใช้ต้องการข้อมูลในรูปแบบตัวเลขเท่านั้น



## 4. การปรับมาตรฐานข้อมูล (Standardization)

ใช้ StandardScaler เพื่อปรับมาตรฐานข้อมูล ทำให้ค่าต่างๆ ในแต่ละคอลัมน์มีค่าเฉลี่ยเป็นศูนย์และค่าเบี่ยงเบนมาตรฐานเป็นหนึ่ง ซึ่งจะช่วยให้อัลกอริธึมสามารถแบ่งกลุ่มได้อย่างมีประสิทธิภาพมากขึ้น



## 5. การแบ่งกลุ่มด้วย K-means Clustering

ในขั้นตอนนี้เราจะใช้โมเดล K-means โดยกำหนดจำนวนกลุ่มเป็น 3 (n\_clusters=3) จากนั้นทำการฝึกโมเดลกับข้อมูลที่ผ่านการปรับมาตรฐานแล้ว และเพิ่มคอลัมน์ 'cluster' ใน DataFrame เพื่อเก็บผลการแบ่งกลุ่ม สุดท้ายจะประเมินคุณภาพการแบ่งกลุ่มด้วยค่า Silhouette Score

A screen shot of a computer program

Description automatically generated

## 6. การแบ่งกลุ่มด้วย Hierarchical Clustering

ใช้ AgglomerativeClustering ซึ่งเป็นวิธีการแบ่งกลุ่มแบบ Hierarchical โดยกำหนดจำนวนกลุ่มเป็น 3 เช่นกัน จากนั้นจะประเมินผลการแบ่งกลุ่มด้วย Silhouette Score เหมือนกับ K-means

A screen shot of a computer program

Description automatically generated

## 7. การแบ่งกลุ่มด้วย DBSCAN

ใช้ DBSCAN ซึ่งเป็นวิธีการแบ่งกลุ่มที่ไม่ต้องกำหนดจำนวนกลุ่มล่วงหน้า โดยใช้ค่า eps และ min\_samples ในการควบคุมผลลัพธ์ จากนั้นจะประเมินการแบ่งกลุ่มด้วย Silhouette Score แต่เนื่องจาก DBSCAN จะแบ่งกลุ่มตามความหนาแน่นของข้อมูล ผลลัพธ์การประเมินด้วย Silhouette Score อาจไม่เหมาะสมหาก DBSCAN สร้างกลุ่มที่มีแค่กลุ่มเดียว

A screen shot of a computer program

Description automatically generated

## 8. การลดมิติและการแสดงผลลัพธ์การแบ่งกลุ่ม

ในขั้นตอนนี้เราจะใช้ PCA เพื่อลดมิติข้อมูลจากหลายมิติให้เหลือเพียง 2 มิติ จากนั้นจะแสดงผลลัพธ์การแบ่งกลุ่มในรูปแบบกราฟกระจาย (Scatter Plot) โดยใช้สีเพื่อแสดงกลุ่มที่ได้จาก K-means Clustering และ Hierarchical Clustering  
A black background with white text

Description automatically generated

## 9. ภาพผลลัพธ์

A screenshot of a computer

Description automatically generated

A black screen with white text

Description automatically generated

## 10. อธิบายผลลัพธ์

จากผลลัพธ์ที่ได้จากการประเมินคุณภาพของการแบ่งกลุ่ม (Clustering) ด้วยค่า Silhouette Score:

* **K-means Clustering**: Silhouette Score = 0.1968
* **Hierarchical Clustering**: Silhouette Score = 0.1500
* **DBSCAN Clustering**: Silhouette Score = 0.1818

โมเดลที่ดีที่สุดตามค่า Silhouette Score คือ **K-means Clustering** เนื่องจากมีค่า Silhouette Score สูงที่สุด แต่ขึ้นอยู่กับลักษณะของข้อมูล ถ้าข้อมูลมี outliers หรือการกระจายตัวไม่สม่ำเสมอ DBSCAN อาจเป็นตัวเลือกที่ดีกว่าเพราะสามารถจัดการกับ outliers ได้ดี

**Fact:**

K-means Clustering:

* K-means เป็นอัลกอริธึมที่ทำงานได้ดีในกรณีที่ข้อมูลมีรูปแบบที่แบ่งเป็นกลุ่มชัดเจน โดยที่จำนวนกลุ่มถูกกำหนดล่วงหน้า