The Clustering Dataset

The dataset I’m using for this clustering task is: Airline Passenger Satisfaction (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction). The goal for this dataset is usually to predict if a customer was satisifed or not, but since this is the ground-truth, I’ve removed it from the dataset. The rest of the datasets contains scores Airline passengers have given different aspects of the flight experience.

The Final Dataset

Data cleaning

Normalizing

I chose to normalize the columns Departure Delay in Minutes, Arrival Delay in Minutes and Flight Distance

Encoding

I encoded four columns:

- Gender to be represented by either 0 or 1. Female = 0, Male = 1
- Customer type to be represented by either 0 or 1. Loyal Customer = 0, Disloyal Customer = 1
- Type of Travel to be represented by either Personal Traver or Business Travel. Personal Travel = 0, Business Travel = 1
- Class to be represented by 0, 1 or 2. Eco plus = 0, Business = 1, Eco = 2

Feature selection

I also chose to remove some columns. These are:

- Satisfaction
- Unnamed: 0

- id

EDA

There wasn’t too much Exploratory Data Analysis to do, but what I got from the data was:

- The dataset is pretty balanced
- Most columns has a right-leaning normal distribution
- On average most people are pretty satisfied