Bumpy Airways

Would you fly with us?

Mini Project 3 by I. Font

Dataset
Method
Aim
Results
Future improvements

Aim

Use unsupervised learning (clustering and PCA) to uncover patterns in airline passenger behaviour and satisfaction levels, and identify meaningful passenger segments.

Back

The Dataset

airline_passenger_satisfaction.csv from Maven Analytics

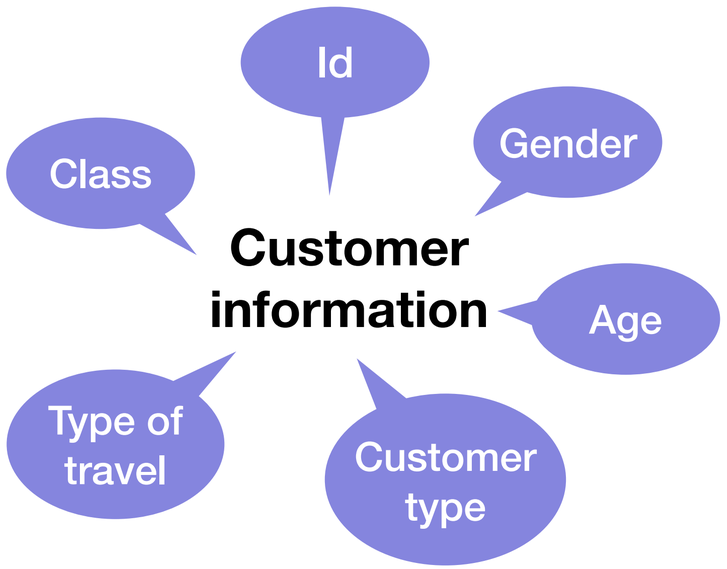
The Dataset

airline_passenger_satisfaction.csv from Maven Analytics

Customer information

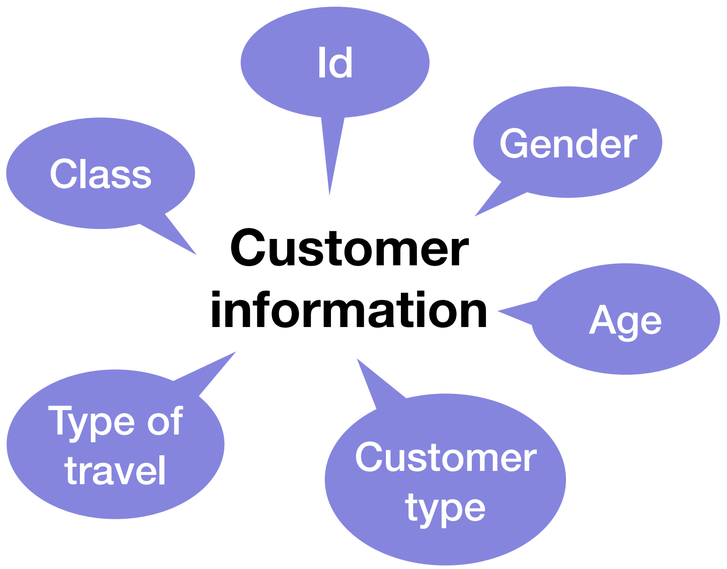
The Dataset

airline_passenger_satisfaction.csv from Maven Analytics



The Dataset

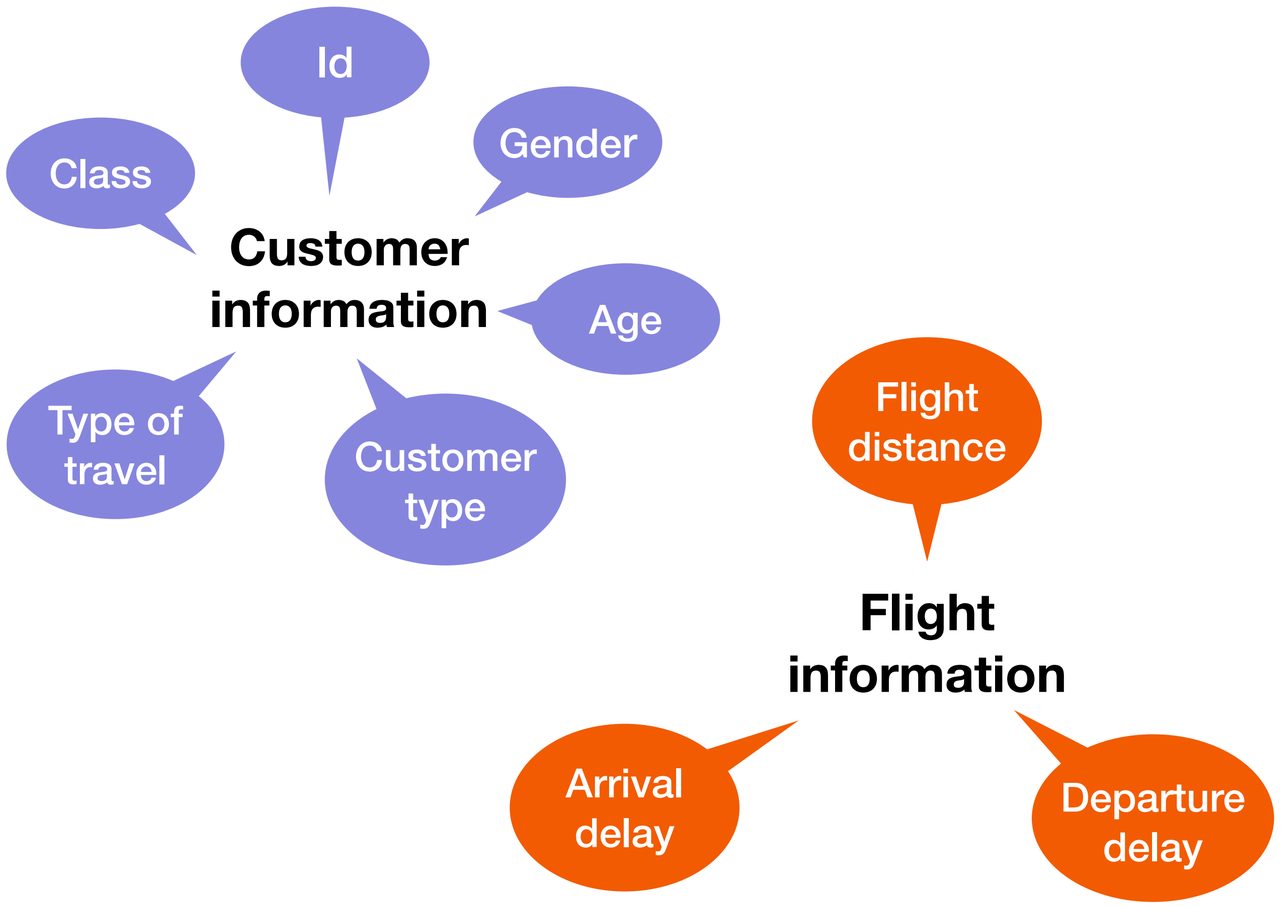
airline_passenger_satisfaction.csv from Maven Analytics



Flight information

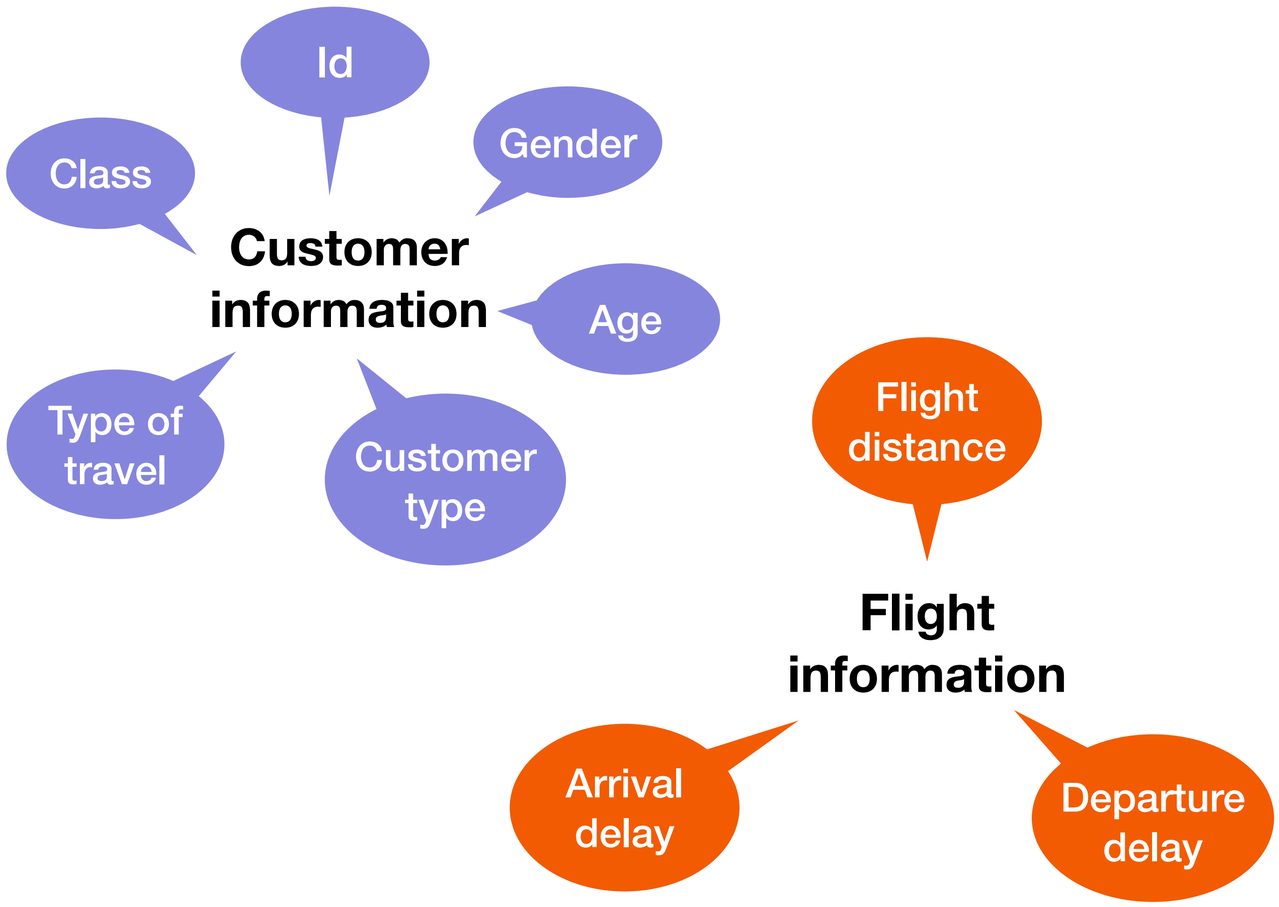
The Dataset

airline_passenger_satisfaction.csv from Maven Analytics



The Dataset

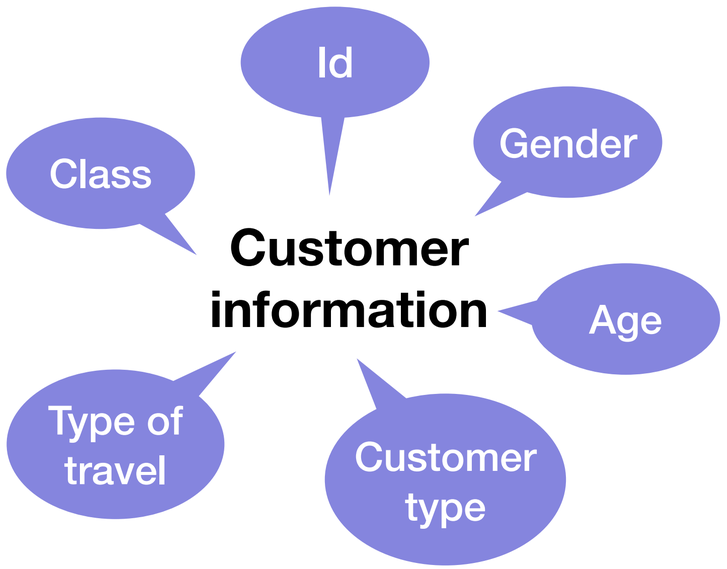
airline_passenger_satisfaction.csv from Maven Analytics

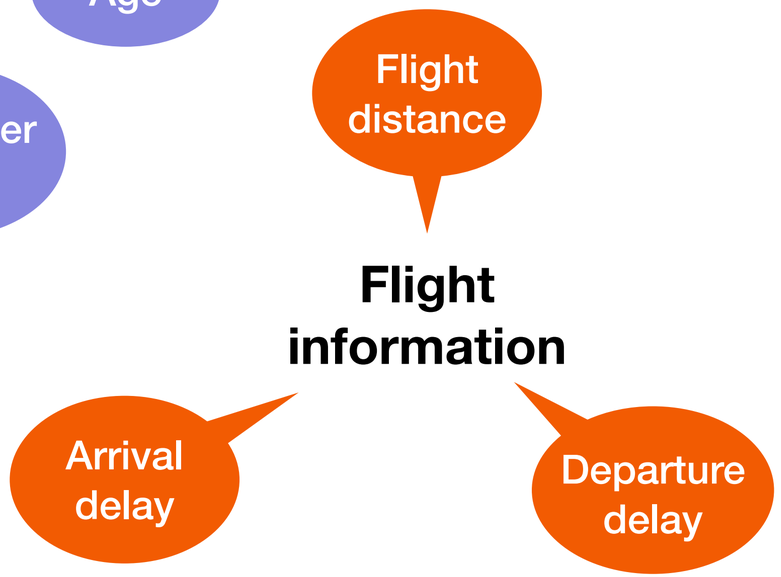


Satisfaction levels

The Dataset

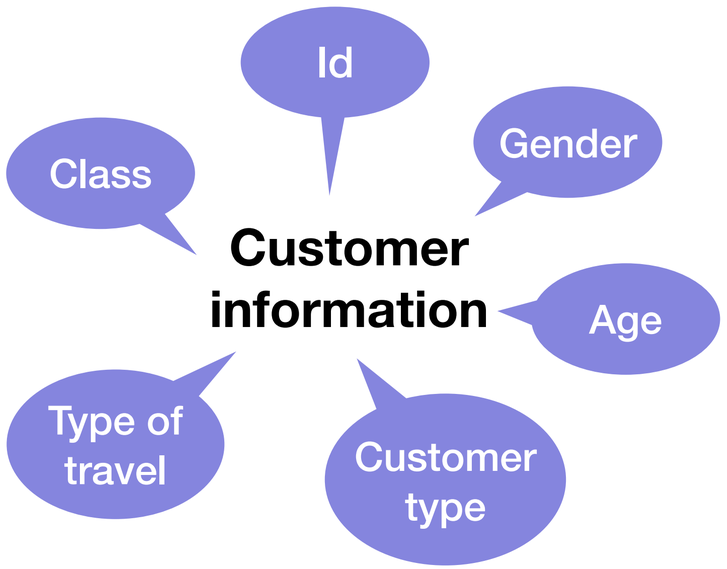
airline_passenger_satisfaction.csv from Maven Analytics

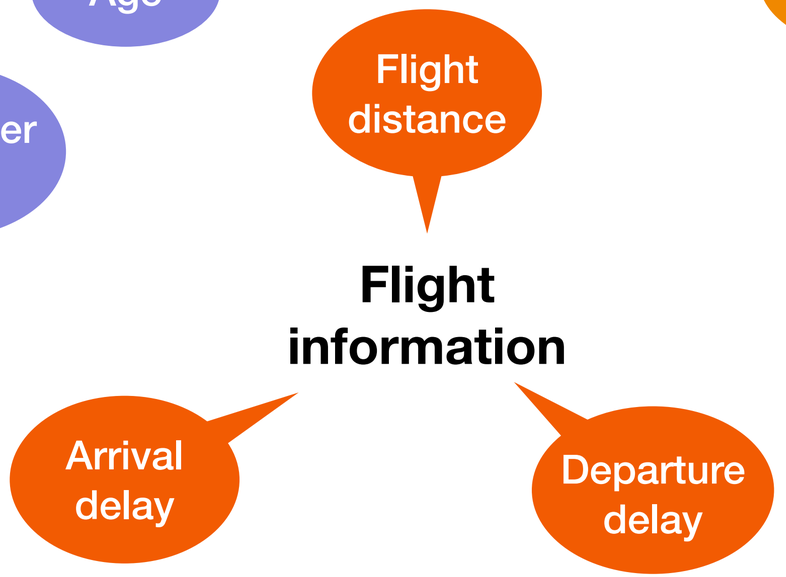




The Dataset

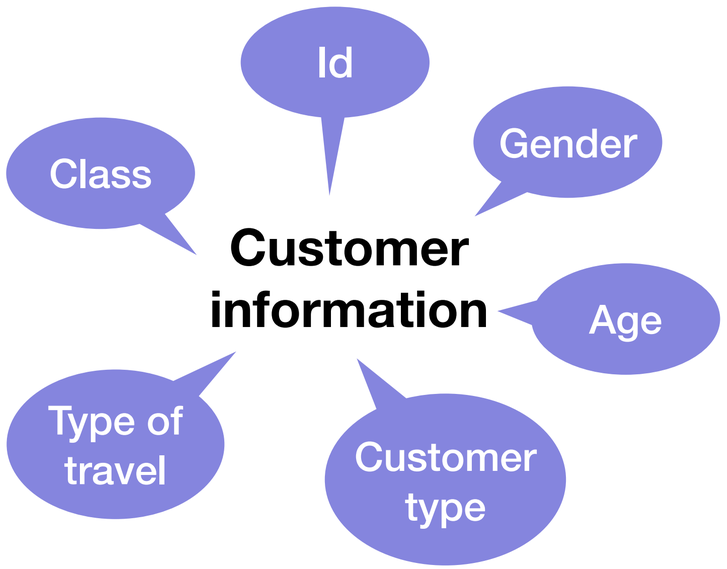
airline_passenger_satisfaction.csv from Maven Analytics

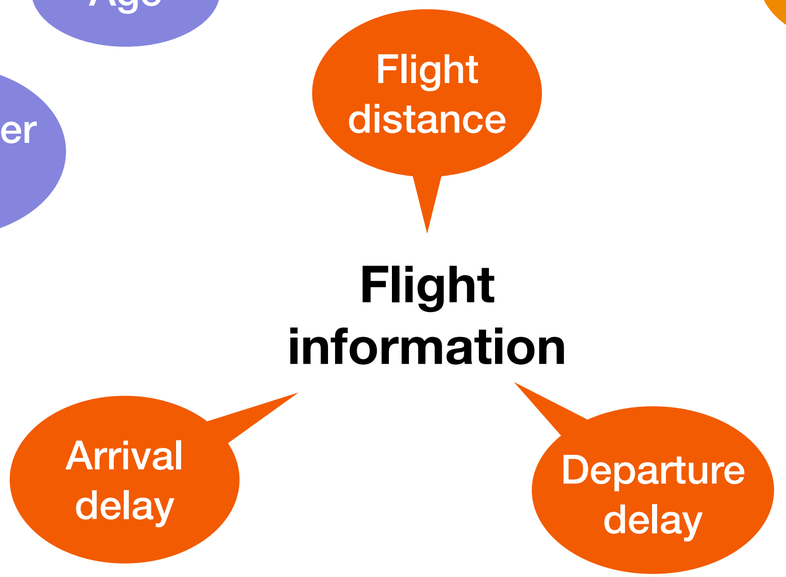




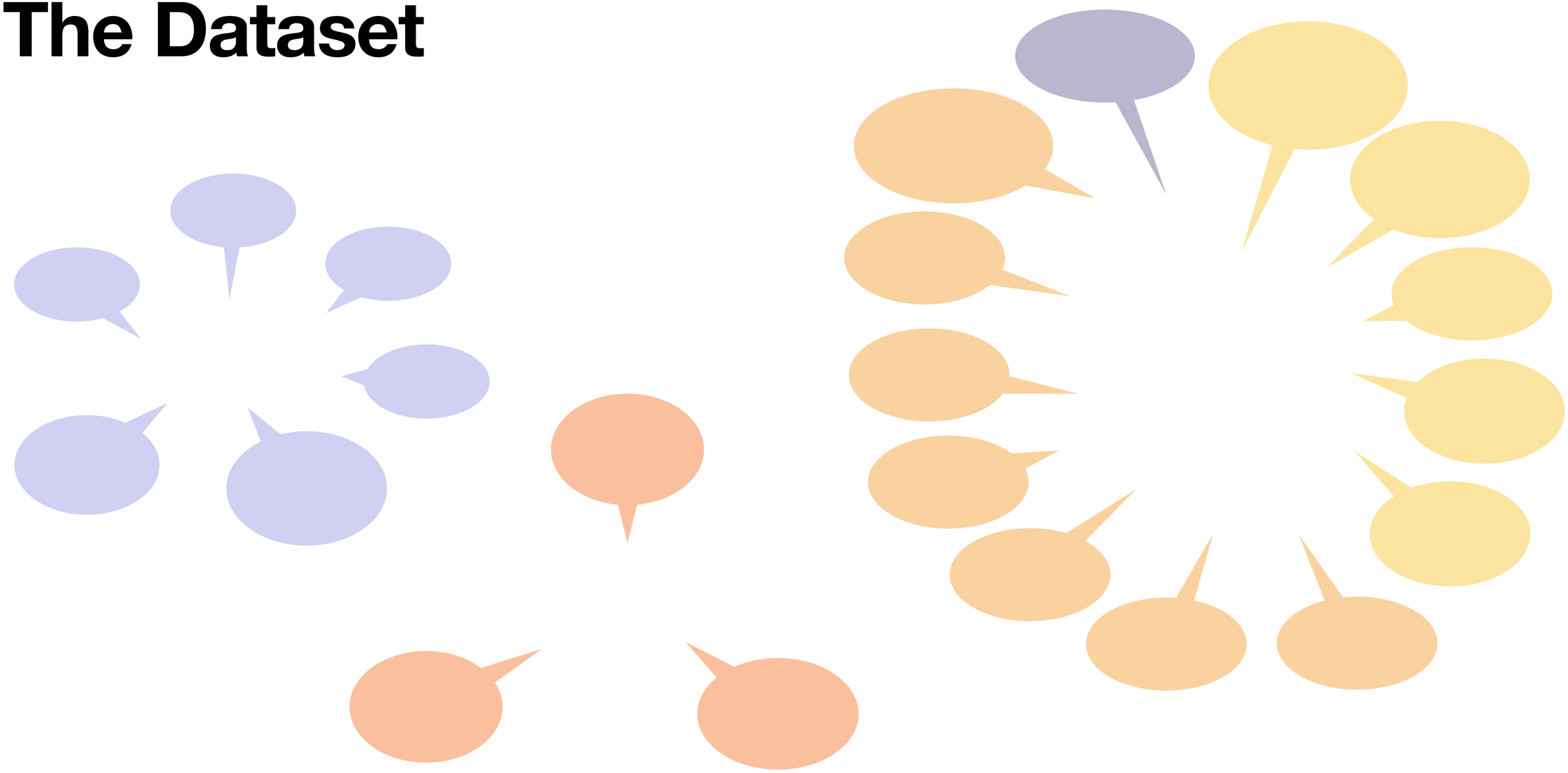
The Dataset

airline_passenger_satisfaction.csv from Maven Analytics





The Dataset



Back

The Dataset

- 129880 samples and 22 features

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped
- No gender bias

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped
- No gender bias
- 82% returning customers, 69% business travellers. Possible bias toward business travellers

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped
- No gender bias
- 82% returning customers, 69% business travellers. Possible bias toward business travellers
- Satisfaction values of 0 replaced with NaNs for EDA, then with mean

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped
- No gender bias
- 82% returning customers, 69% business travellers. Possible bias toward business travellers
- Satisfaction values of 0 replaced with NaNs for EDA, then with mean
- Added features for EDA: avg_service_rating, Satisfaction_binary

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped
- No gender bias
- 82% returning customers, 69% business travellers. Possible bias toward business travellers
- Satisfaction values of 0 replaced with NaNs for EDA, then with mean
- Added features for EDA: avg_service_rating, Satisfaction_binary
- Labels removed

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped
- No gender bias
- 82% returning customers, 69% business travellers. Possible bias toward business travellers
- Satisfaction values of 0 replaced with NaNs for EDA, then with mean
- Added features for EDA: avg_service_rating, Satisfaction_binary
- Labels removed
- Categorical variables -> one-hot encoding

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped
- No gender bias
- 82% returning customers, 69% business travellers. Possible bias toward business travellers
- Satisfaction values of 0 replaced with NaNs for EDA, then with mean
- Added features for EDA: avg_service_rating, Satisfaction_binary
- Labels removed
- Categorical variables -> one-hot encoding
- Features standardised with RobustScaler

Back

The Dataset

- 129880 samples and 22 features
- 393 missing values in Arrival Delay -> dropped
- No gender bias
- 82% returning customers, 69% business travellers. Possible bias toward business travellers
- Satisfaction values of 0 replaced with NaNs for EDA, then with mean
- Added features for EDA: avg_service_rating, Satisfaction_binary
- Labels removed
- Categorical variables -> one-hot encoding
- Features standardised with RobustScaler

Back

KMeans cluster analysis

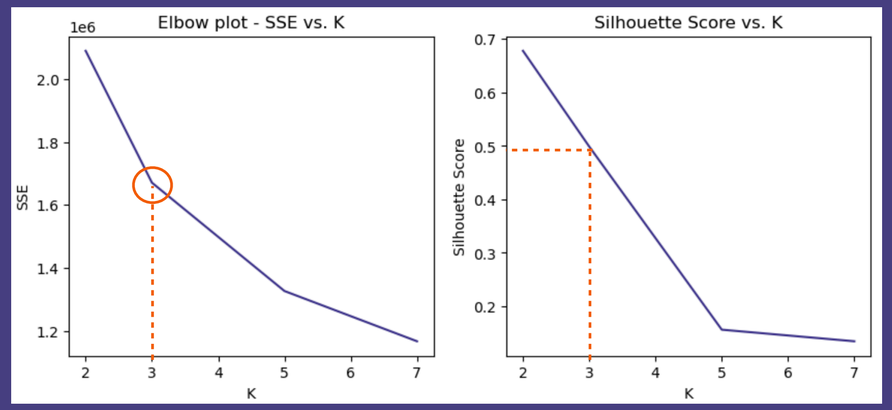
PCA dimensionality reduction analysis

HDBSCAN cluster analysis

Back

KMeans cluster analysis

- Elbow and silhouette scores -> 3 clusters fitted



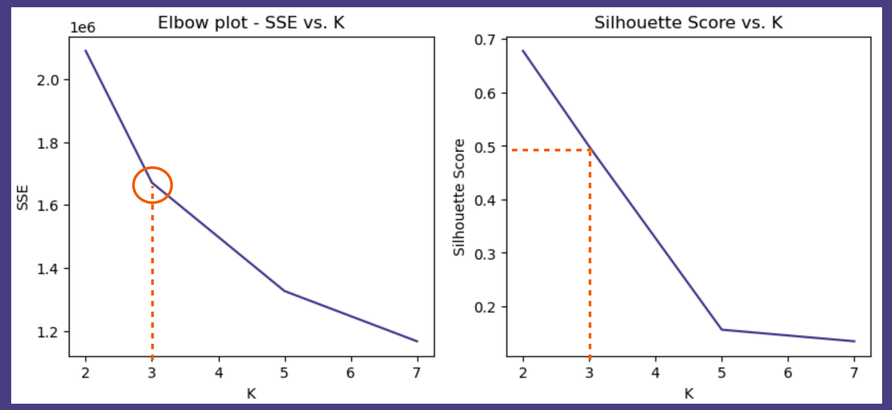
HDBSCAN cluster analysis

PCA dimensionality reduction analysis

Back

KMeans cluster analysis

- Elbow and silhouette scores -> 3 clusters fitted



HDBSCAN cluster analysis

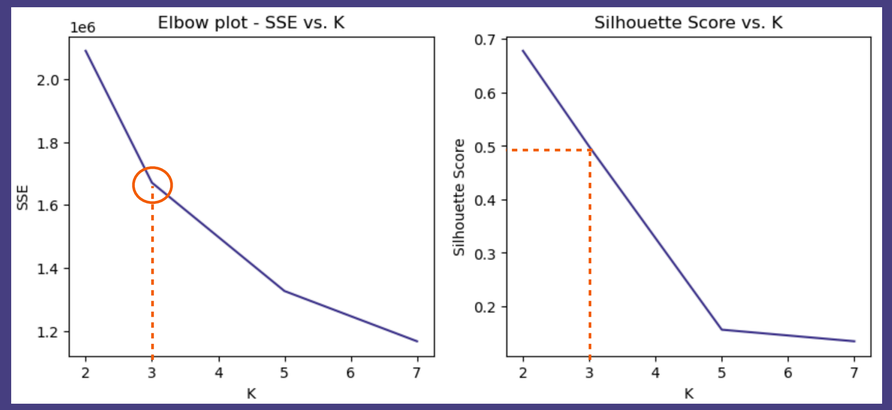
- Scalable version of DBSCAN for large datasets

PCA dimensionality reduction analysis

Back

KMeans cluster analysis

- Elbow and silhouette scores -> 3 clusters fitted



HDBSCAN cluster analysis

- Scalable version of DBSCAN for large datasets

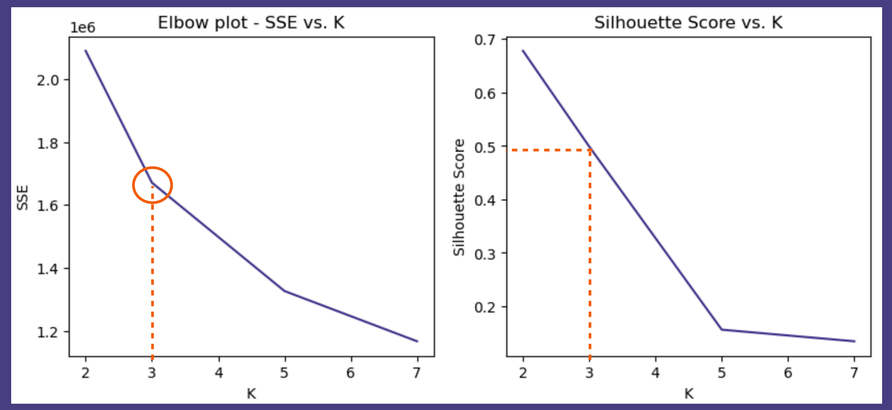
- min_cluster_size = 5, min_samples = 15 tuned on a random sample of 12000 using noise level and Silhouette score

PCA dimensionality reduction analysis

Back

KMeans cluster analysis

- Elbow and silhouette scores -> 3 clusters fitted

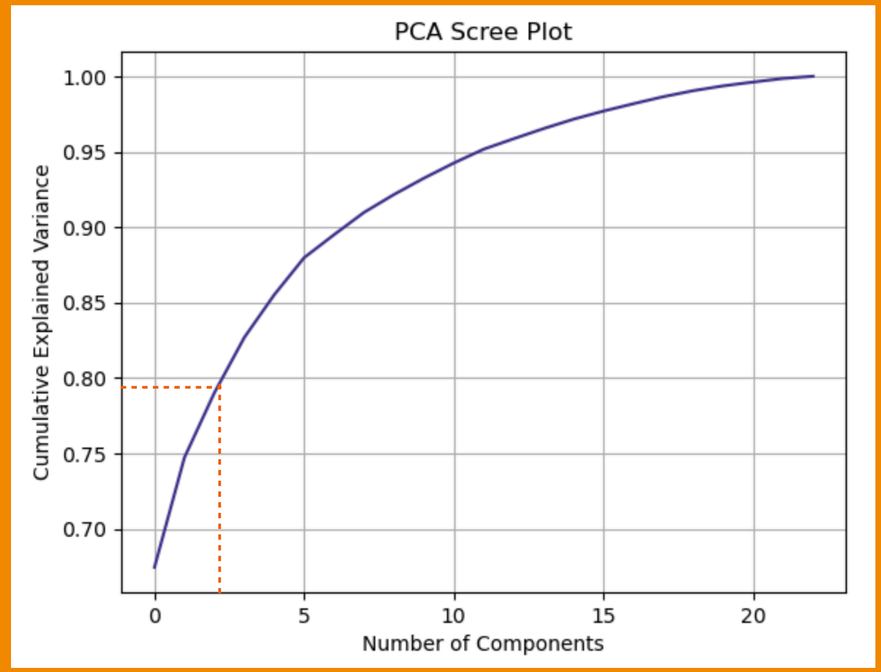


HDBSCAN cluster analysis

- Scalable version of DBSCAN for large datasets
- min_cluster_size = 5, min_samples = 15 tuned on a random sample of 12000 using noise level and Silhouette score

PCA dimensionality reduction analysis

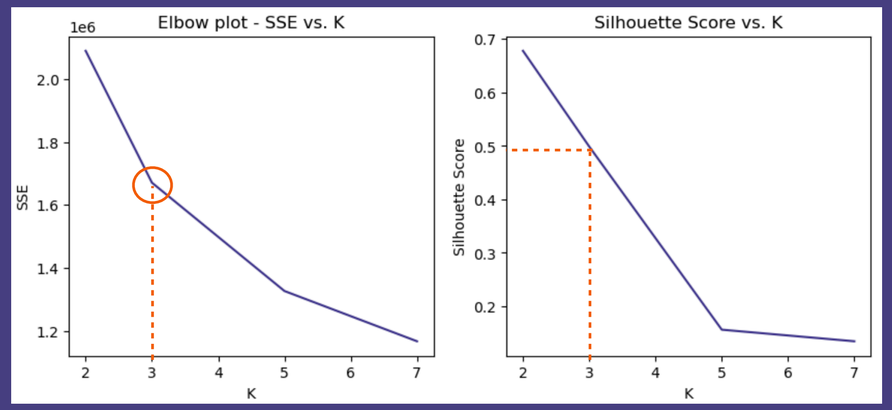
- Optimum number of components from Scree plot



Back

KMeans cluster analysis

- Elbow and silhouette scores -> 3 clusters fitted

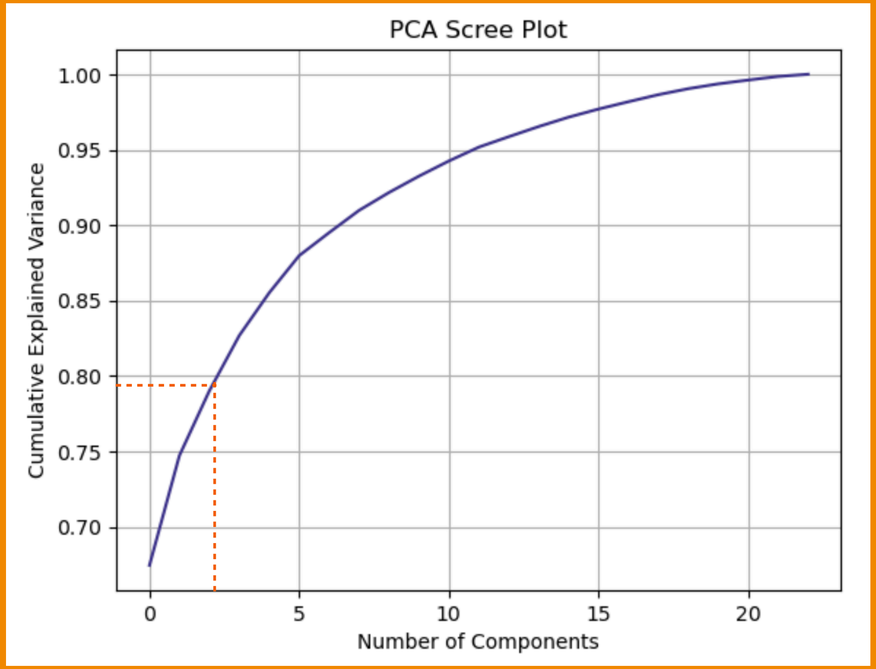


HDBSCAN cluster analysis

- Scalable version of DBSCAN for large datasets
- min_cluster_size = 5, min_samples = 15 tuned on a random sample of 12000 using noise level and Silhouette score

PCA dimensionality reduction analysis

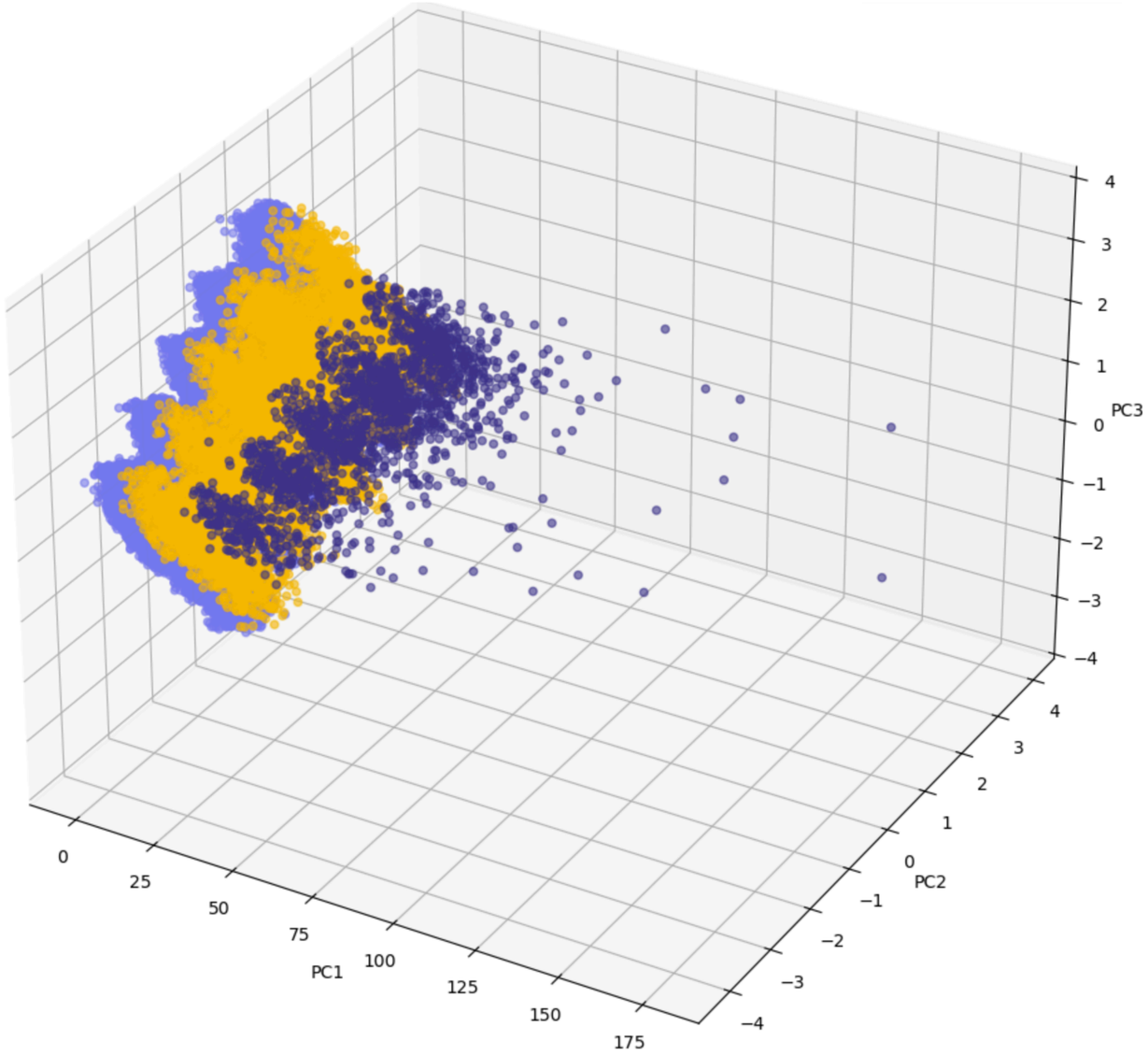
- Optimum number of components from Scree plot



- 3 components explain about 80% of variance

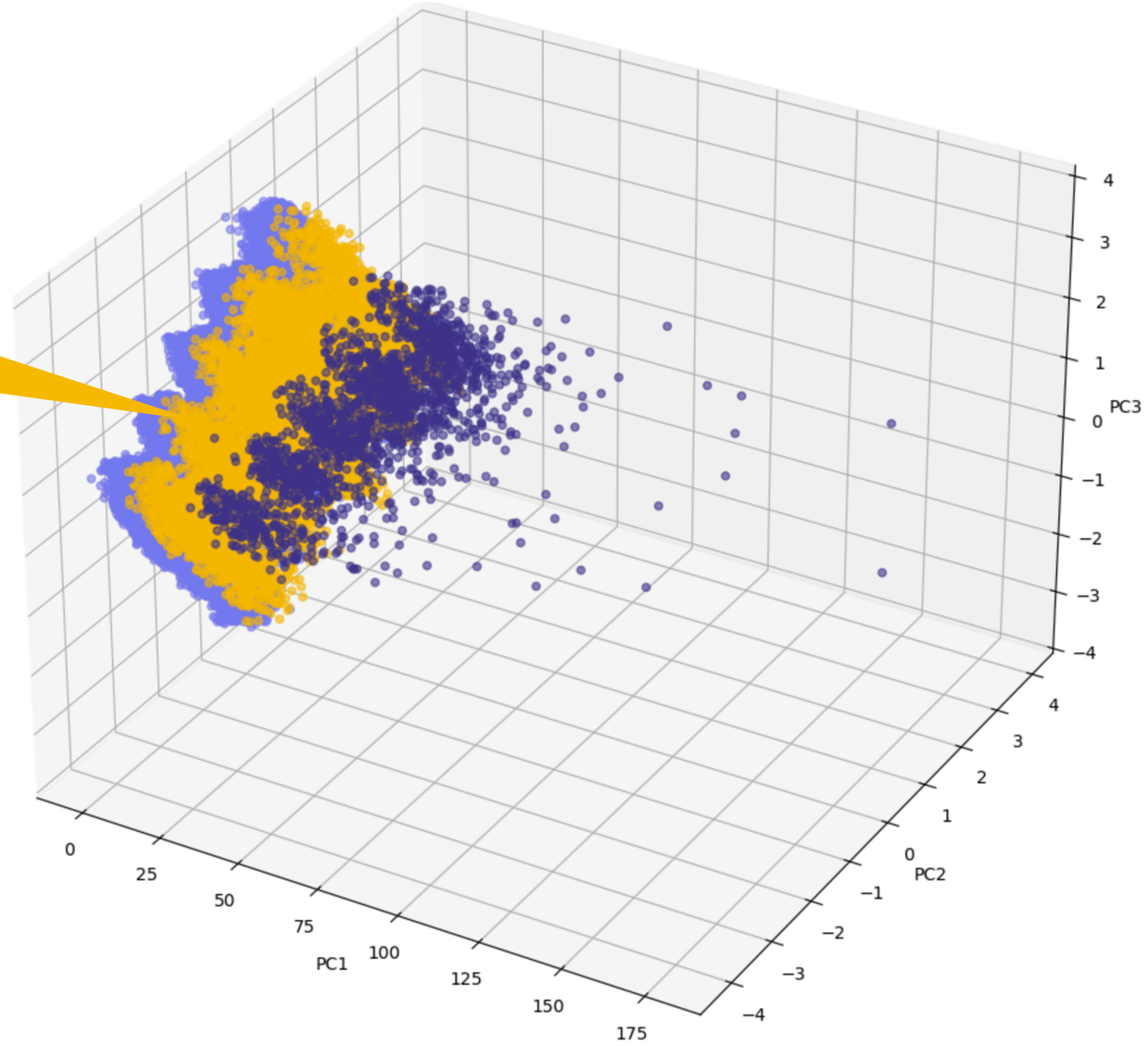
Back

KMeans clusters



KMeans clusters

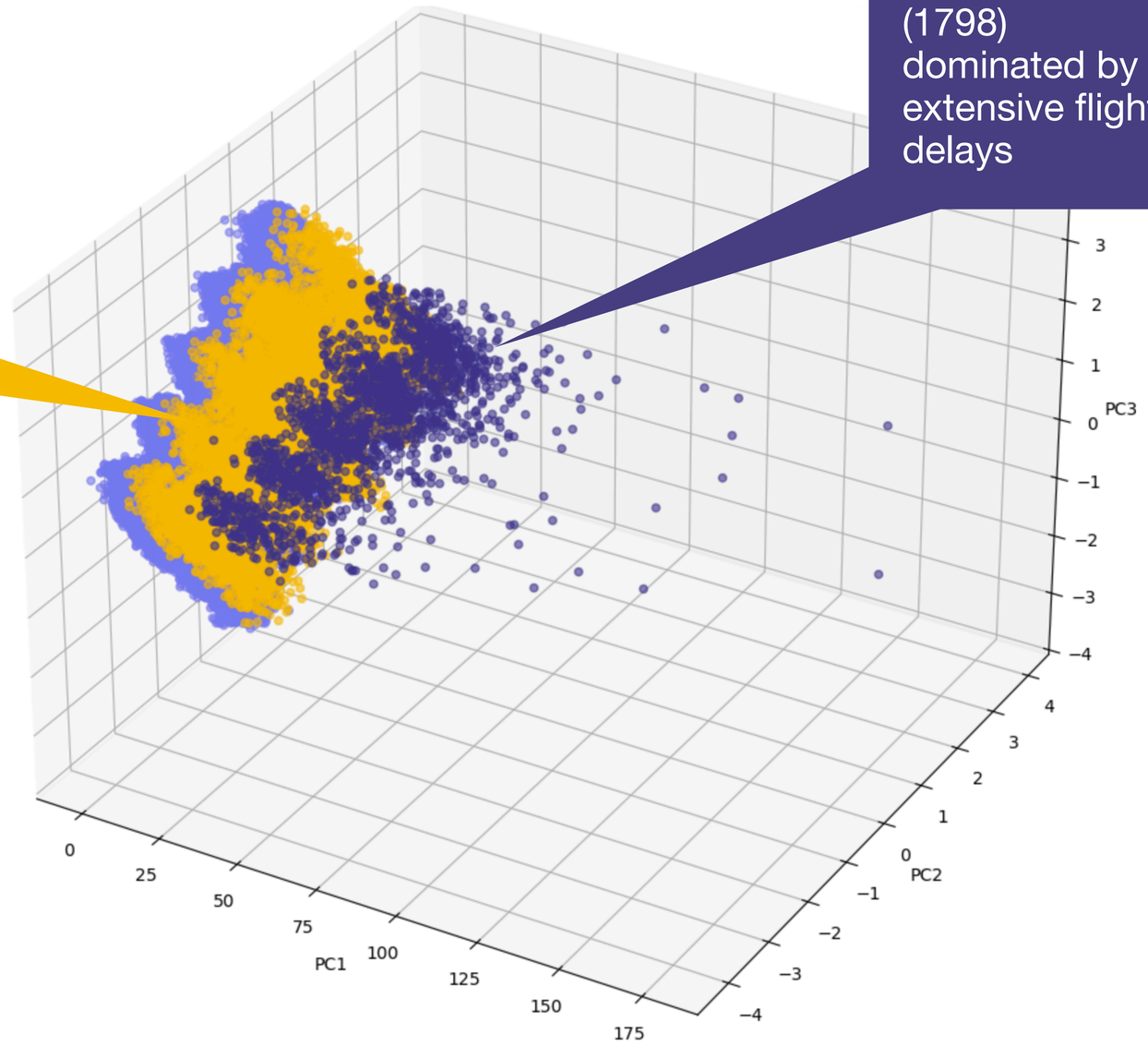
Big cluster (115711) dominated by business return customers



KMeans clusters

Big cluster (115711) dominated by business return customers

Small cluster (1798) dominated by extensive flight delays

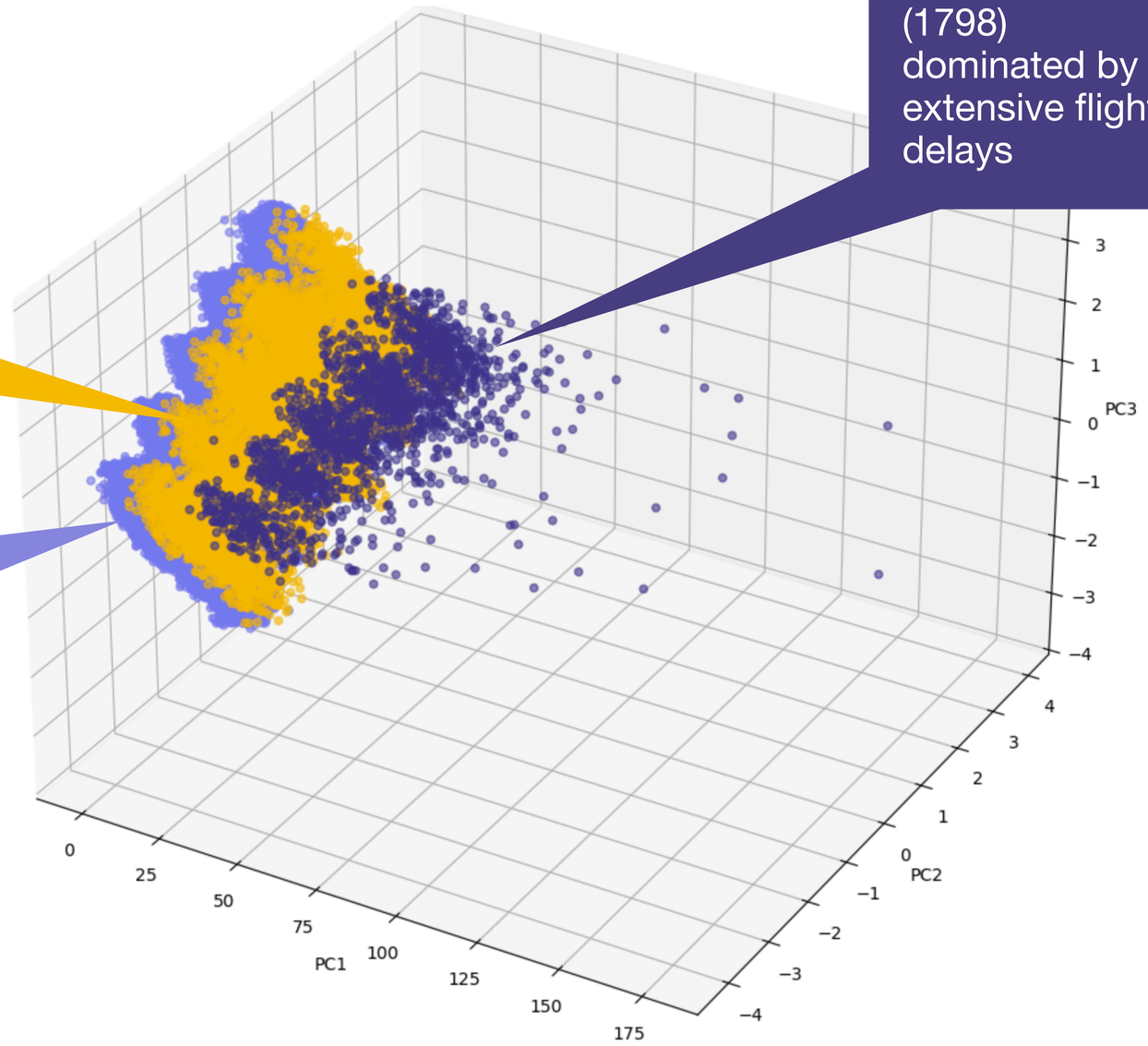


KMeans clusters

Big cluster (115711) dominated by business return customers

Small cluster (1798) dominated by extensive flight delays

Medium cluster (11978) very similar to big cluster

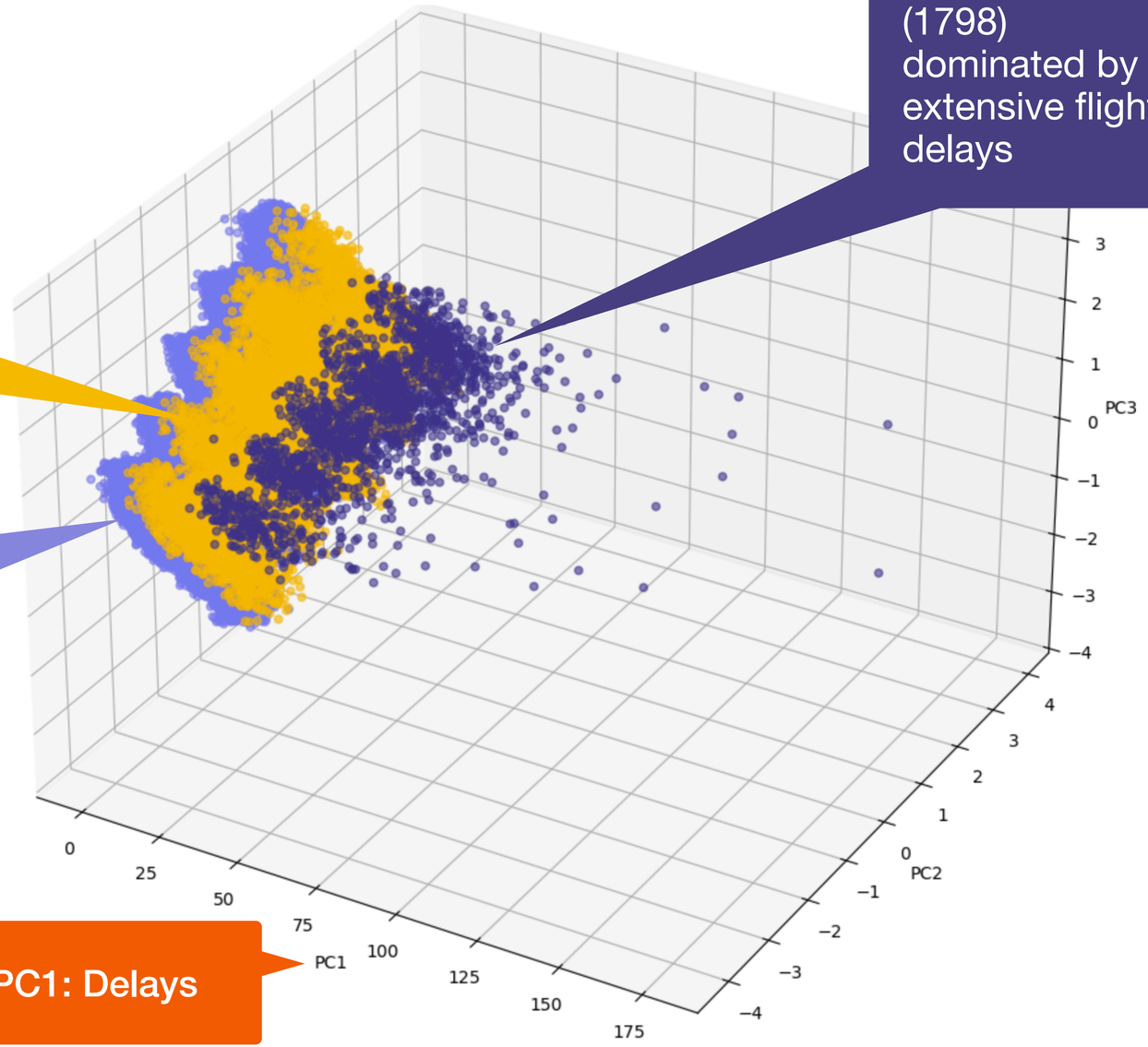


KMeans clusters

Big cluster (115711) dominated by business return customers

Small cluster (1798) dominated by extensive flight delays

Medium cluster (11978) very similar to big cluster

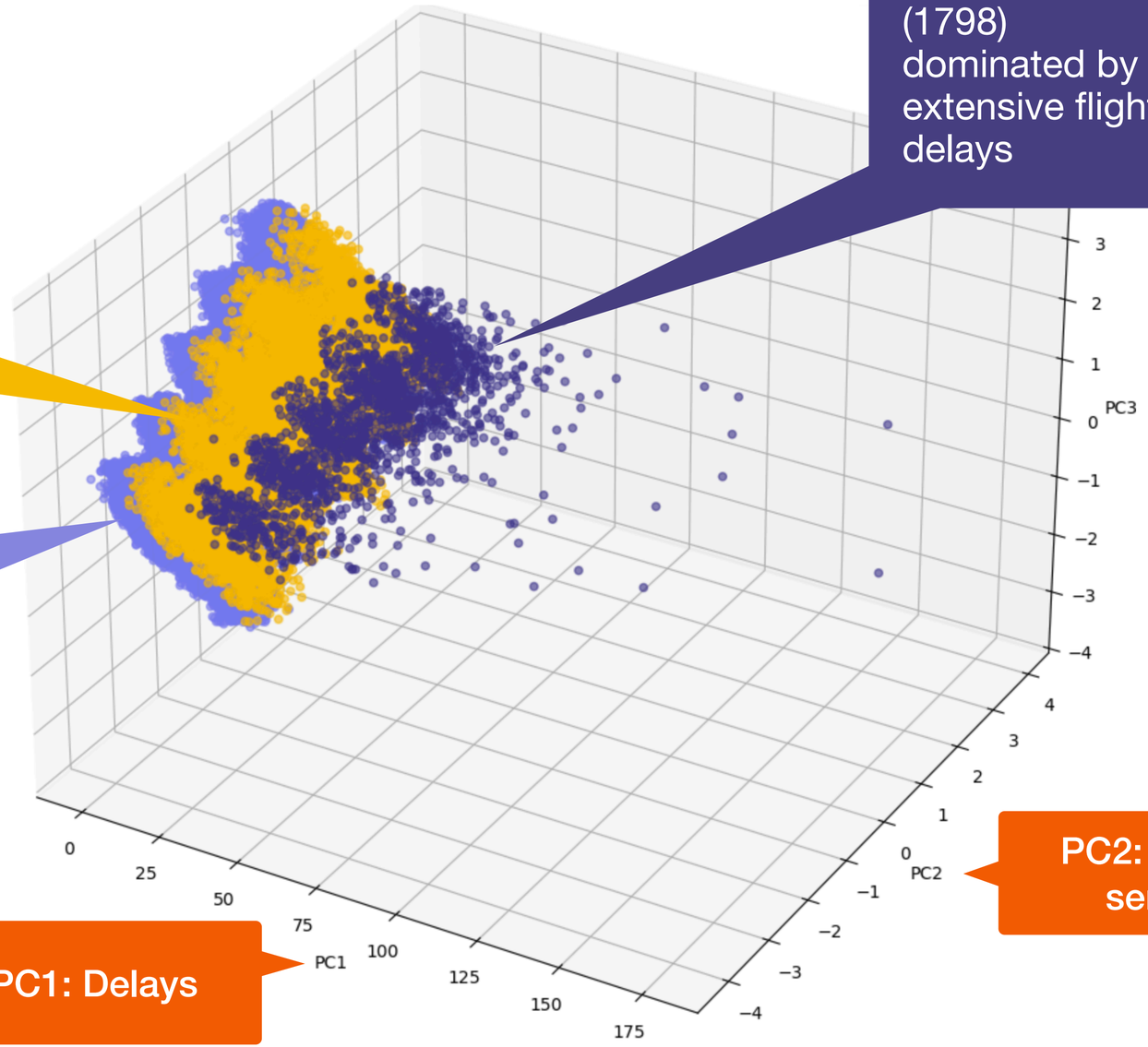


PC1: Delays

KMeans clusters

Big cluster (115711) dominated by business return customers

Small cluster (1798) dominated by extensive flight delays



Medium cluster (11978) very similar to big cluster

PC1: Delays

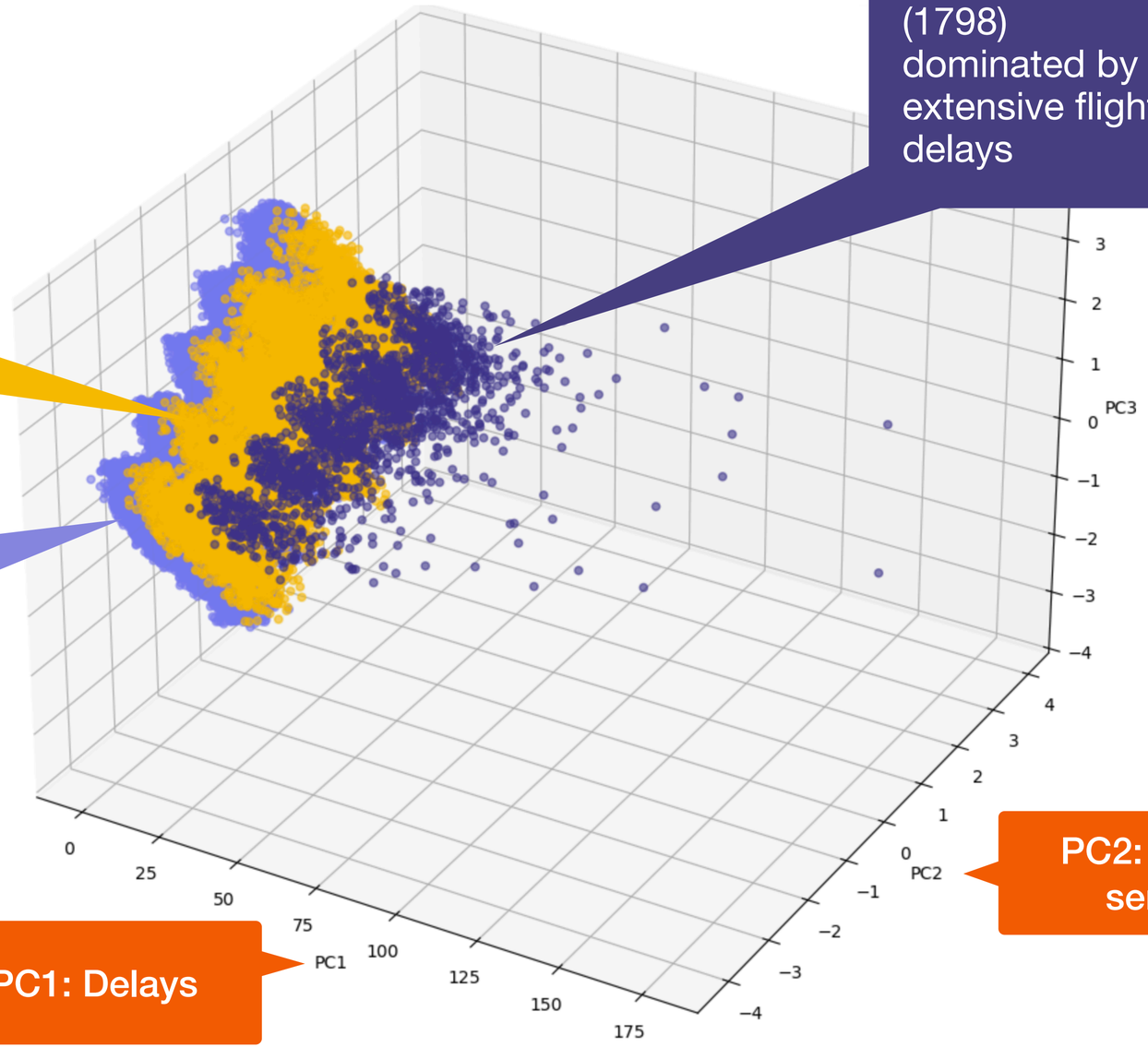
PC2: Overall service

KMeans clusters

Big cluster (115711) dominated by business return customers

Small cluster (1798) dominated by extensive flight delays

Medium cluster (11978) very similar to big cluster

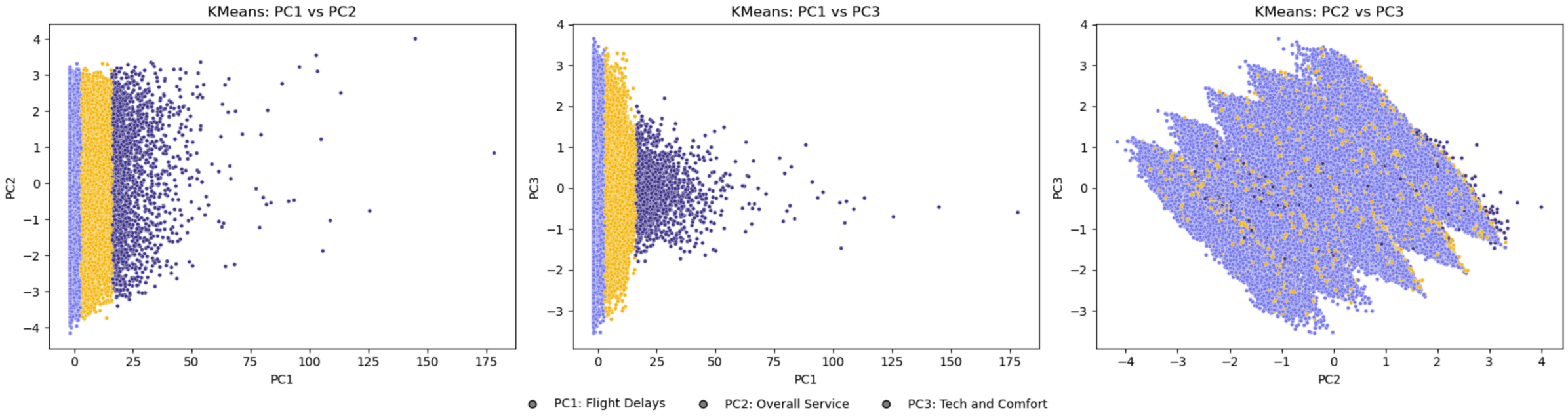


PC3: Tech & comfort

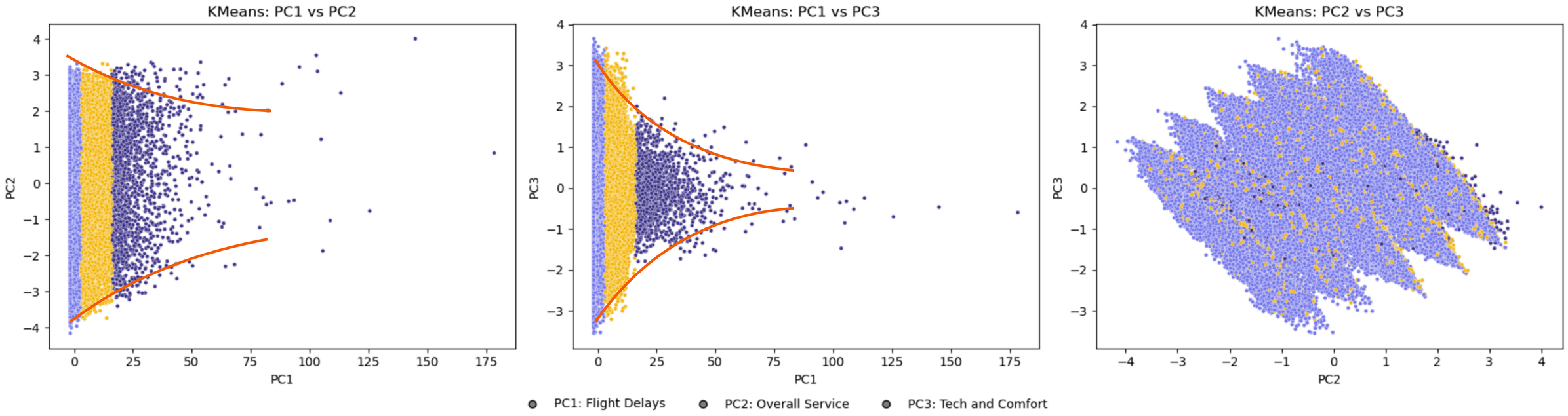
PC2: Overall service

PC1: Delays

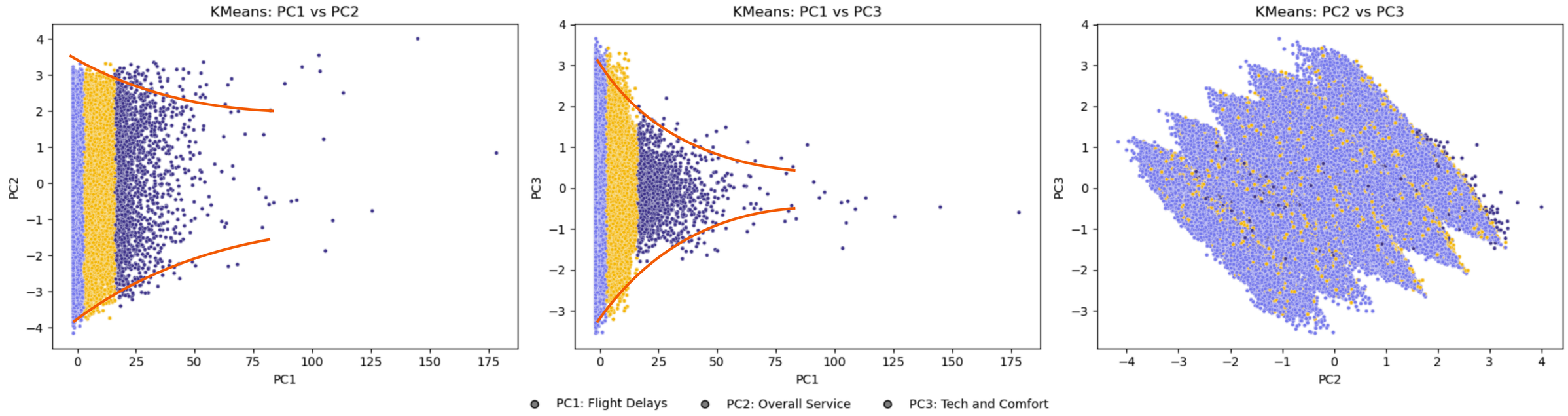
KMeans clusters



KMeans clusters



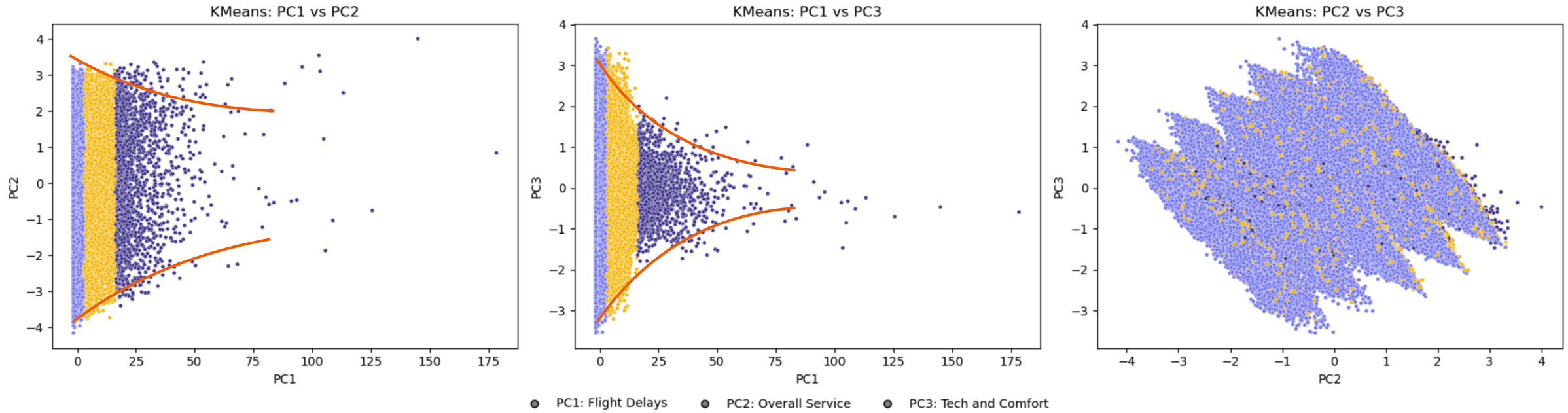
KMeans clusters



For short flight delays some passengers are:

- very satisfied → good flight + good service = delighted.
- very dissatisfied → maybe expectations were high, or service was bad despite punctuality.

KMeans clusters



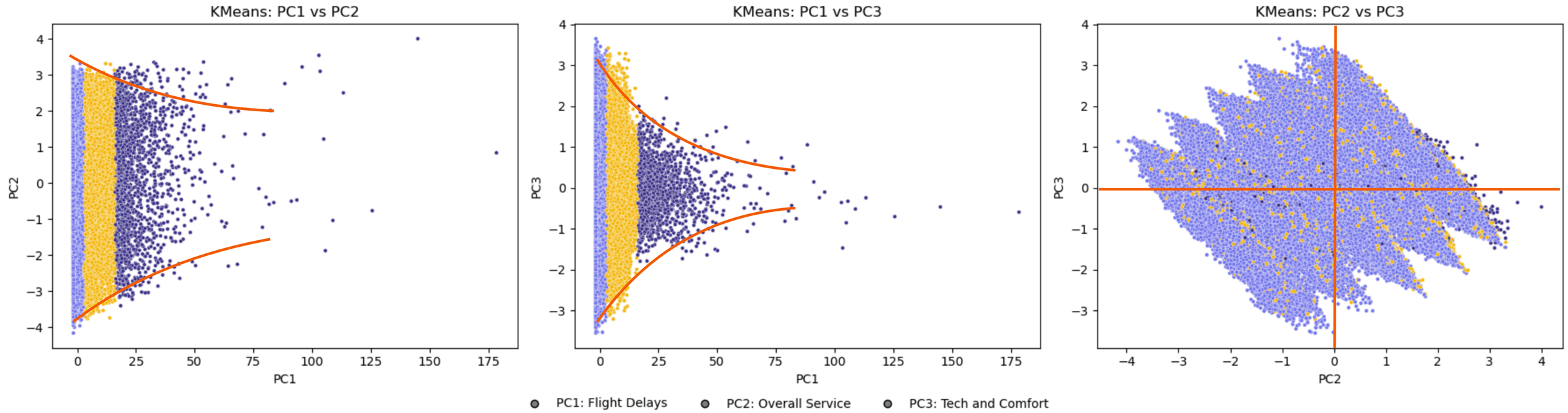
For short flight delays some passengers are:

- very satisfied → good flight + good service = delighted.
- very dissatisfied → maybe expectations were high, or service was bad despite punctuality.

As flight delays increase, satisfaction levels tend to be more neutral:

- tech and comfort does not improve the passengers' experience.

KMeans clusters



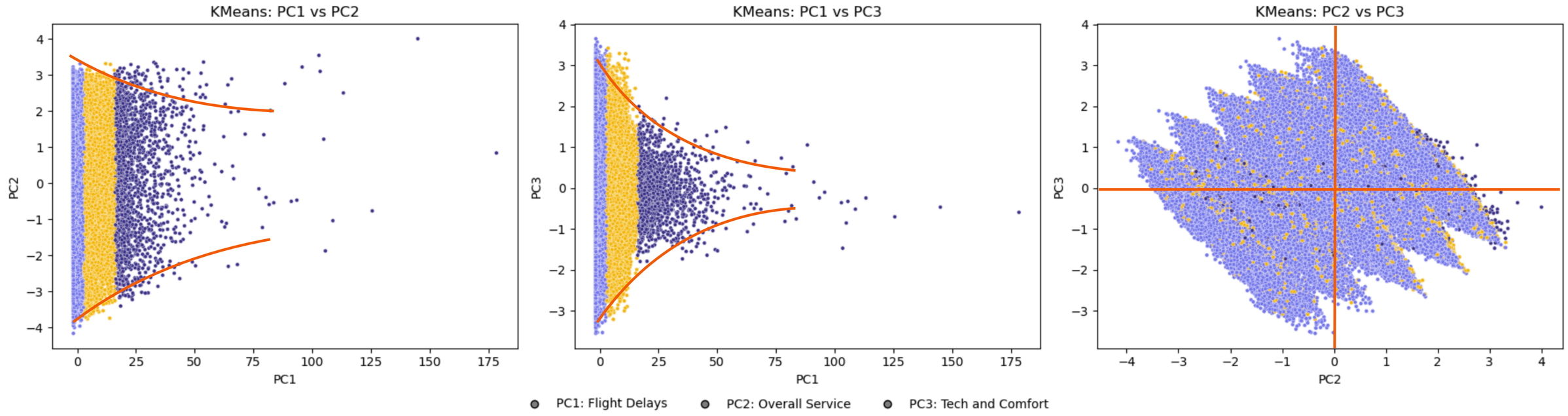
For short flight delays some passengers are:

- very satisfied → good flight + good service = delighted.
- very dissatisfied → maybe expectations were high, or service was bad despite punctuality.

As flight delays increase, satisfaction levels tend to be more neutral:

- tech and comfort does not improve the passengers' experience.

KMeans clusters



For short flight delays some passengers are:

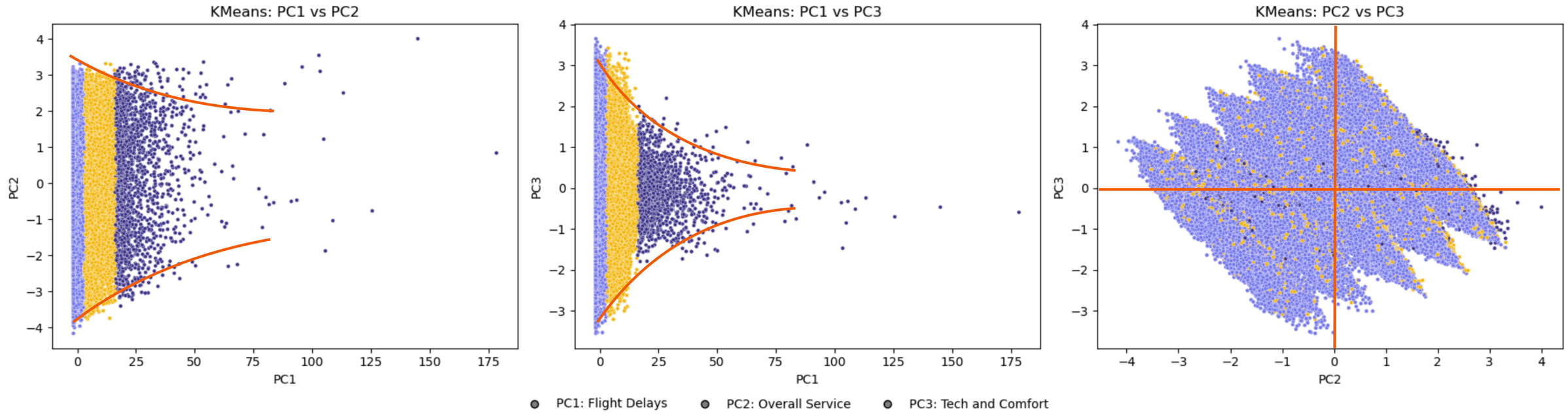
- very satisfied → good flight + good service = delighted.
- very dissatisfied → maybe expectations were high, or service was bad despite punctuality.

As flight delays increase, satisfaction levels tend to be more neutral:

- tech and comfort does not improve the passengers' experience.

There seems to be a tradeoff between overall service and tech/comfort satisfaction. Biggest groups:

KMeans clusters



For short flight delays some passengers are:

- very satisfied → good flight + good service = delighted.
- very dissatisfied → maybe expectations were high, or service was bad despite punctuality.

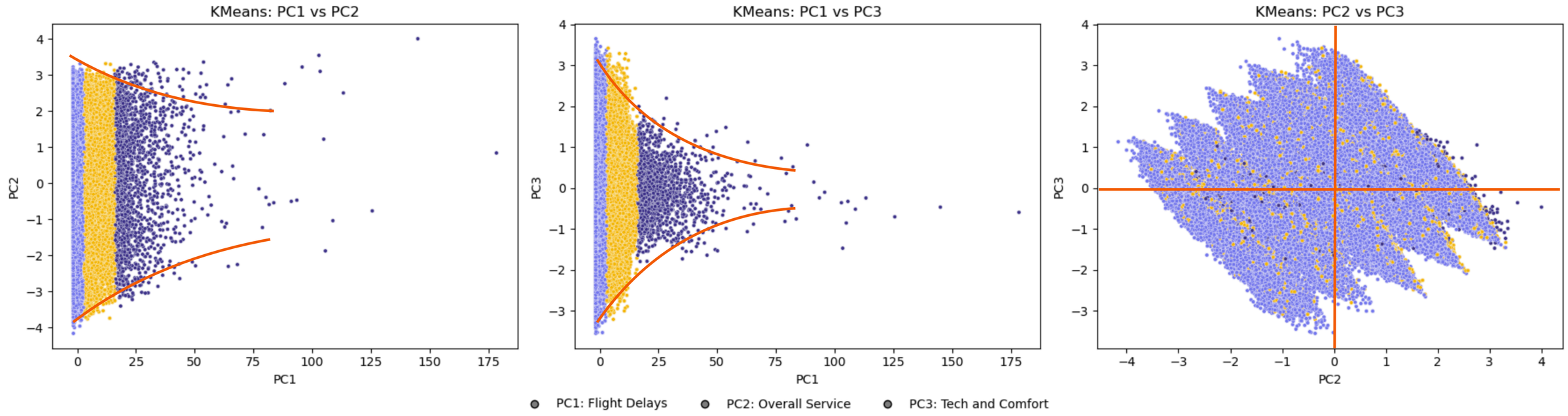
As flight delays increase, satisfaction levels tend to be more neutral:

- tech and comfort does not improve the passengers' experience.

There seems to be a tradeoff between overall service and tech/comfort satisfaction. Biggest groups:

- negative overall service with positive tech/comfort satisfaction

KMeans clusters



For short flight delays some passengers are:

- very satisfied → good flight + good service = delighted.
- very dissatisfied → maybe expectations were high, or service was bad despite punctuality.

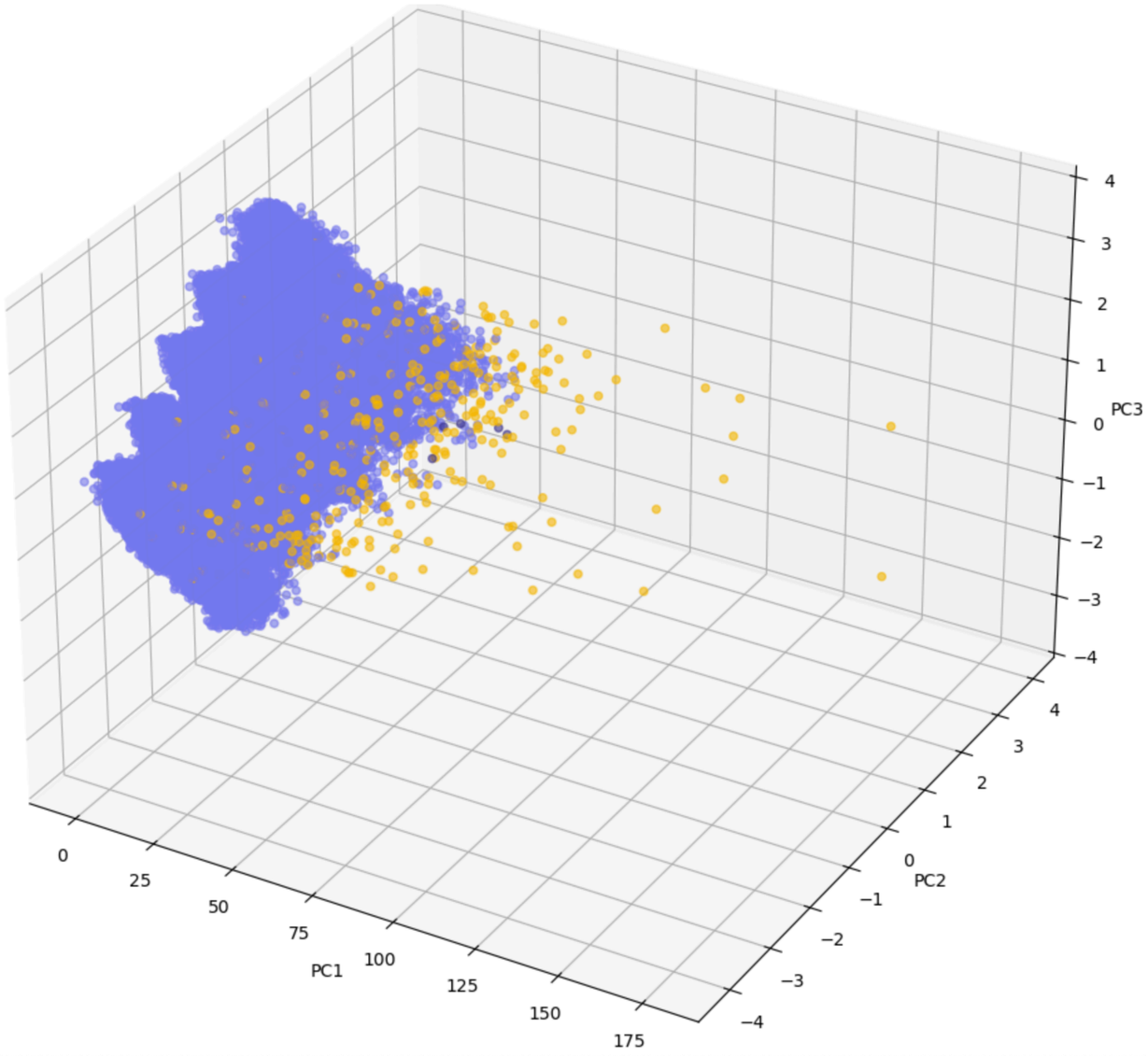
As flight delays increase, satisfaction levels tend to be more neutral:

- tech and comfort does not improve the passengers' experience.

There seems to be a tradeoff between overall service and tech/comfort satisfaction. Biggest groups:

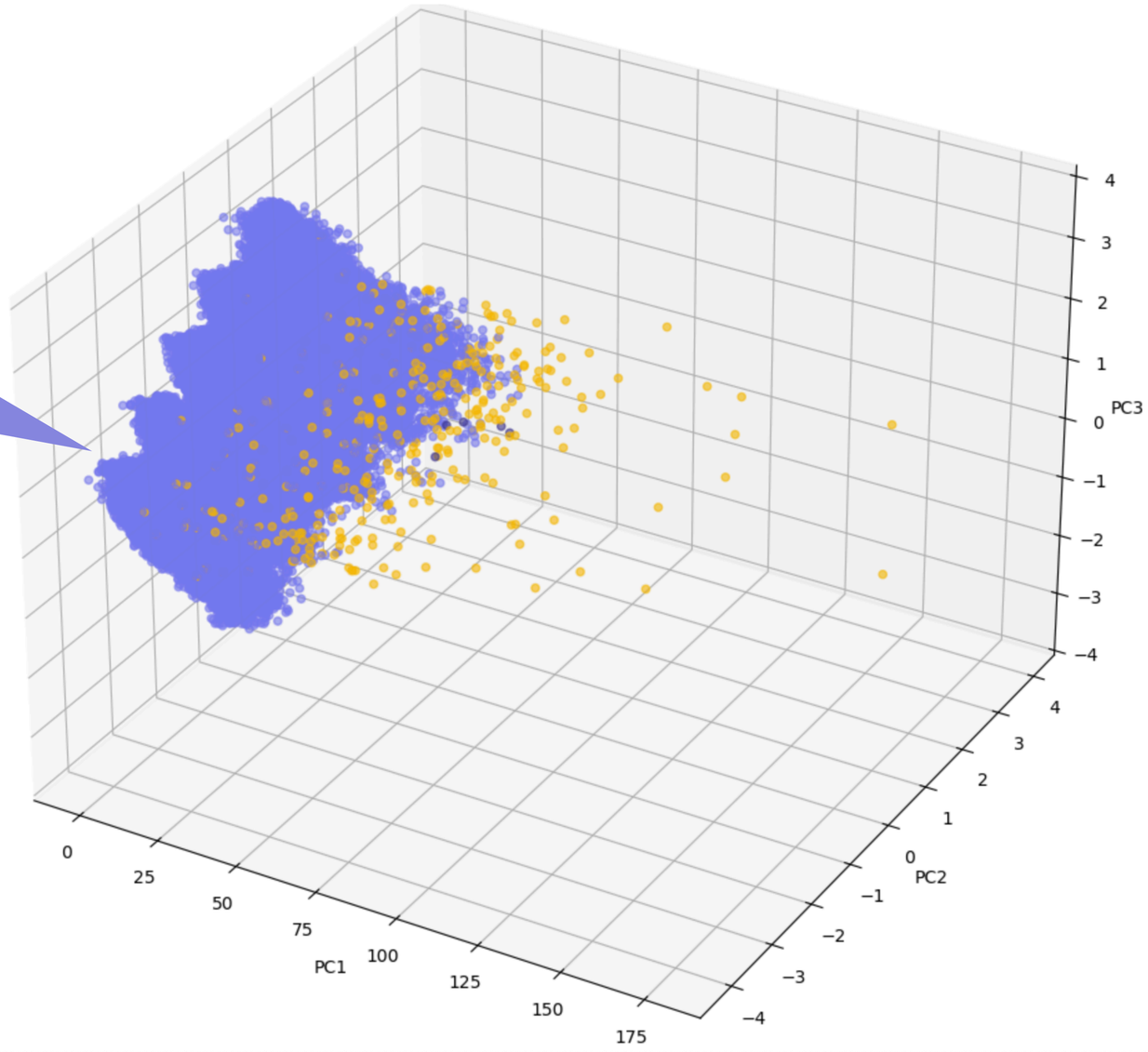
- negative overall service with positive tech/comfort satisfaction
- negative overall service with negative tech/comfort satisfaction

HDBSCAN clusters



HDBSCAN clusters

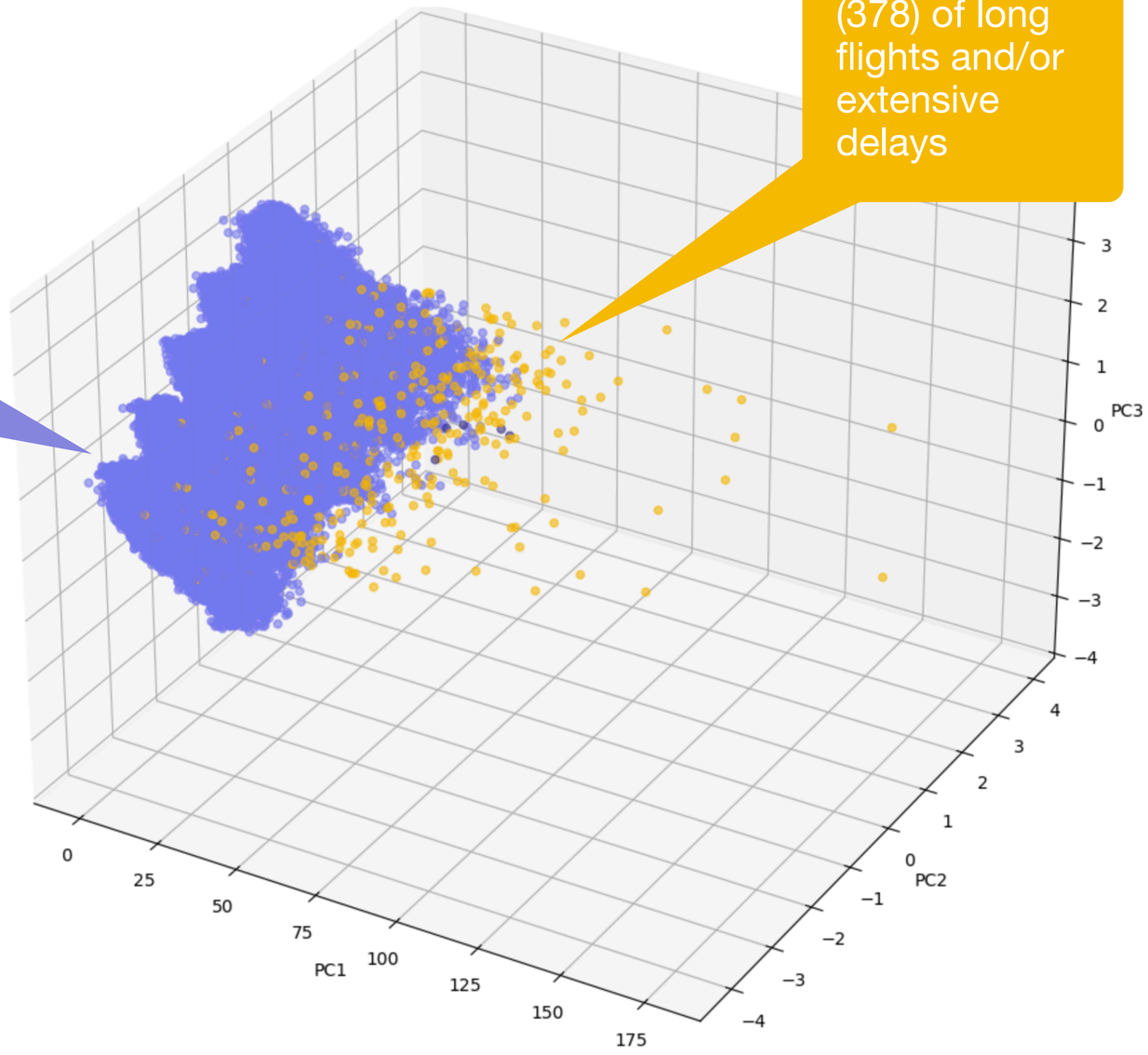
One big cluster
Not interpretable



HDBSCAN clusters

One big cluster
Not interpretable

Noise/outliers (378) of long flights and/or extensive delays



Recommendations to Bumpy Airways

Back

Recommendations to Bumpy Airways

Invest in punctuality — delays are deeply tied to negative overall experience and cluster heavily.

Back

Recommendations to Bumpy Airways

Invest in punctuality — delays are deeply tied to negative overall experience and cluster heavily.

Enhance overall service, specially the check in service and inflight entertainment and particularly for economy passengers

Back

Recommendations to Bumpy Airways

Invest in punctuality — delays are deeply tied to negative overall experience and cluster heavily.

Enhance overall service, specially the check in service and inflight entertainment and particularly for economy passengers

Prioritize digital convenience and connectivity across the customer journey — from booking to onboard, especially for business travellers.

Back

Recommendations to Bumpy Airways

Invest in punctuality — delays are deeply tied to negative overall experience and cluster heavily.

Enhance overall service, specially the check in service and inflight entertainment and particularly for economy passengers

Target customer segments with tailored strategies (i.e. ticket discounts for first time travellers, comfort upgrades for returning customers, free wifi for business passengers)

Prioritize digital convenience and connectivity across the customer journey — from booking to onboard, especially for business travellers.

Back

Future improvements

Future improvements

- Further investigation of the KMeans clusters:

Future improvements

- Further investigation of the KMeans clusters:
    - Examine correlation heat-maps of service-related features within each cluster.

Future improvements

- Further investigation of the KMeans clusters:
  - Examine correlation heat-maps of service-related features within each cluster.
  - Identify which customer types (i.e. business vs. personal, returning vs. new) are most dominant in each cluster.

Future improvements

- Further investigation of the KMeans clusters:
  - Examine correlation heat-maps of service-related features within each cluster.
  - Identify which customer types (i.e. business vs. personal, returning vs. new) are most dominant in each cluster.
  - Analyse the cluster centroids to uncover deeper insights about the typical passenger in each group.

Future improvements

- Further investigation of the KMeans clusters:
  - Examine correlation heat-maps of service-related features within each cluster.
  - Identify which customer types (i.e. business vs. personal, returning vs. new) are most dominant in each cluster.
  - Analyse the cluster centroids to uncover deeper insights about the typical passenger in each group.

- Repeat the analysis after removing the delay features to see if other underlying patterns become more evident.

Future improvements

- Further investigation of the KMeans clusters:
  - Examine correlation heat-maps of service-related features within each cluster.
  - Identify which customer types (i.e. business vs. personal, returning vs. new) are most dominant in each cluster.
  - Analyse the cluster centroids to uncover deeper insights about the typical passenger in each group.

- Repeat the analysis after removing the delay features to see if other underlying patterns become more evident.
- Explore soft clustering techniques:

Future improvements

- Further investigation of the KMeans clusters:
  - Examine correlation heat-maps of service-related features within each cluster.
  - Identify which customer types (i.e. business vs. personal, returning vs. new) are most dominant in each cluster.
  - Analyse the cluster centroids to uncover deeper insights about the typical passenger in each group.

- Repeat the analysis after removing the delay features to see if other underlying patterns become more evident.
- Explore soft clustering techniques:
  - Gaussian Mixture Models (GMM): assign each passenger a probability of belonging to each cluster, rather than forcing a hard assignment.

Future improvements

- Further investigation of the KMeans clusters:
  - Examine correlation heat-maps of service-related features within each cluster.
  - Identify which customer types (i.e. business vs. personal, returning vs. new) are most dominant in each cluster.
  - Analyse the cluster centroids to uncover deeper insights about the typical passenger in each group.

- Repeat the analysis after removing the delay features to see if other underlying patterns become more evident.
- Explore soft clustering techniques:
  - Gaussian Mixture Models (GMM): assign each passenger a probability of belonging to each cluster, rather than forcing a hard assignment.
  - Soft clustering is particularly useful when clusters overlap or boundaries are unclear.

Questions ?