A green and grey logo

Description automatically generated with medium confidence

**Clustering Technique**

**Unsupervised machine learning algorithm**Investigate passenger statistics from San Francisco International airport

**Teacher: Ciro Martins**

**Students:** Claudia Maria Ciurea (117730)

Contents

[1. Introduction 3](#_Toc178858613)

[1.1 Aim of the project 3](#_Toc178858614)

[1.2 Description of Dataset 3](#_Toc178858615)

[1.3 Algorithm used 4](#_Toc178858616)

[2. Feature engineering 4](#_Toc178858617)

[2.1 K means 7](#_Toc178858618)

[2.2 Hierarchical agglomerative clustering 11](#_Toc178858619)

[3. Conclusions by comparing the 2 algorithms 12](#_Toc178858620)

[Bibliography 14](#_Toc178858621)

# Introduction

## 1.1 Aim of the project

The aim of the project is to analyze the flow of Passenger Traffic Statistics by Airline in San Francisco International Airport. To do so, an unsupervised machine learning algorithm was chosen to **group operating airlines into clusters with similar traits to see the pattern of traffic passenger statistics based on time of the year, how long is the flight,** price of the flight.

Clustering is a technique that groups similar data points such that the points in the same group are more like each other than the points in the other groups. The group of similar data points is called a Cluster**.** (Patlolla, 2024)

## 1.2 Description of Dataset

The dataset was taken from the website data.World: <https://data.world/data-society/air-traffic-passenger-data>

There are 15,007 rows, 16 columns. The data includes the number of passengers, the operating airline, the published airline, the geographic region, the activity type code, the price category code, the terminal, the boarding area, and the year and month of the flight. In Table 1 there is a description of the column of the dataset.

|  |  |
| --- | --- |
| **Column name** | **Description** |
| **Activity Period** | The date of the activity. (Date) |
| **Operating Airline** | The airline that operated the flight. (String) |
| **Operating Airline IATA Code** | The IATA code of the airline that operated the flight. (String) |
| **Published Airline** | The airline that published the fare for the flight. (String) |
| **Published Airline IATA Code** | The IATA code of the airline that published the fare for the flight. (String) |
| **GEO Summary** | A summary of the geographic region. (String) |
| **GEO Region** | The geographic region. (String) |
| **Activity Type Code** | The type of activity. (String) |
| **Price Category Code** | The price category of the fare. (String) |
| **Terminal** | The terminal of the flight. (String) |
| **Boarding Area** | The boarding area of the flight. (String) |
| **Passenger Count** | The number of passengers on the flight. (Integer) |
| **Adjusted Activity Type Code** | The type of activity, adjusted for missing data. (String) |
| **Adjusted Passenger Count** | The number of passengers on the flight, adjusted for missing data. (Integer) |
| **Year** | The year of the activity. (Integer) |
| **Month** | The month of the activity. (Integer) |

Table 1 Description of the features of the dataset

## Algorithm used

In order to find the best clustering algorithm for the dataset and propose the optimal clustering number, multiple algorithms were used: K means and Hierarchical clustering.

**K-means clustering algorithm.**

K-means clustering is the most used clustering algorithm. It's a centroid-based algorithm and the simplest unsupervised learning algorithm.

This algorithm tries to minimize the variance of data points within a cluster. It's also how most people are introduced to unsupervised machine learning.

K-means is best used on smaller data sets because it iterates over all the data points. That means it'll take more time to classify data points if there are a large amount of them in the data set. (McGregor, 2024)

**What is Hierarchical Clustering?**

[Hierarchical clustering](https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/) is a method of [cluster](https://www.geeksforgeeks.org/clustering-in-machine-learning/) analysis in data mining that creates a hierarchical representation of the clusters in a dataset. The method starts by treating each data point as a separate cluster and then iteratively combines the closest clusters until a stopping criterion is reached. The result of hierarchical clustering is a tree-like structure, called a dendrogram, which illustrates the hierarchical relationships among the clusters.

Agglomerative Clustering: Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom-up method). At first, every dataset is considered an individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed. (Hierarchical Clustering in Data Mining, 2024)

# Feature engineering

Feature engineering is the process of selecting, manipulating and transforming raw data into features that can be used in [supervised learning](https://builtin.com/machine-learning/supervised-learning). It’s also necessary to design and train new [machine learning](https://builtin.com/machine-learning/machine-learning-basics) features so it can tackle new tasks. It consists of five processes: feature creation, transformations, feature extraction, exploratory data analysis and benchmarking. (Patel, 2024)

**Data analysis**

In order to investigate the dataset, I have been through every feature to better understand how many distinct values a feature (column) has Figure 1.

**Data cleaning**

In order to check for missing values, the command *print(df.isna().sum(*)) in Figure2 was applied. Only 54 rows out of 15,007 rows have 2 columns where the IATA Code is missing. For easiness, I decided to remove these rows as it will not affect the overall dataset.

In Figure 3 are the dimensions of the datasets.

A computer screen shot of a code

Description automatically generated

Figure 1 Dataset analysis

A screenshot of a computer screen

Description automatically generated

Figure 2 Missing values

A close up of a number

Description automatically generated

Figure 3 Dataset dimensions

**Data type check up**

It is important to check the data type of each column to ensure the correctness of the dataset. In Figure4 we can see that there are 12 columns of the type object and only 4 columns of the type integer.

A screenshot of a computer

Description automatically generated

Figure 4 Data type check

**Converts categorical variables**

This is an important step in the data preparation process. In order to apply any kind of clustering algorithm the dataset needs to have numeric variables. Bellow I will explain the features that I decided to go with for the clustering process. Some of them are categoric variables, therefore the function get\_dummies was used to encode them. In figure 3 we can see that the dataset that we will have 12 columns. In figure 5 we can see the columns of the dataframe.

## 2.1 K means

In order to perform clustering, I have created a new dataframe. I want to cluster the company Airlines based on the region they were coming from or going to, if the plane landed or took of, the price category, id it was expensive or not and the number of passengers. Therefore, I can find patterns in the dataset regarding the most popular destinations based on number of passengers, region, and price. It is important for an international airport like the San Francisco one to know and understand the passenger statistics and planes whining their airport.

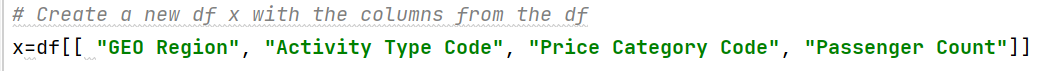


Figure 5 New Dataframe

Following steps:

1. Save the values from x to X\_train



1. Calculate Inertia for different numbers of clusters to find best K (cluster number)

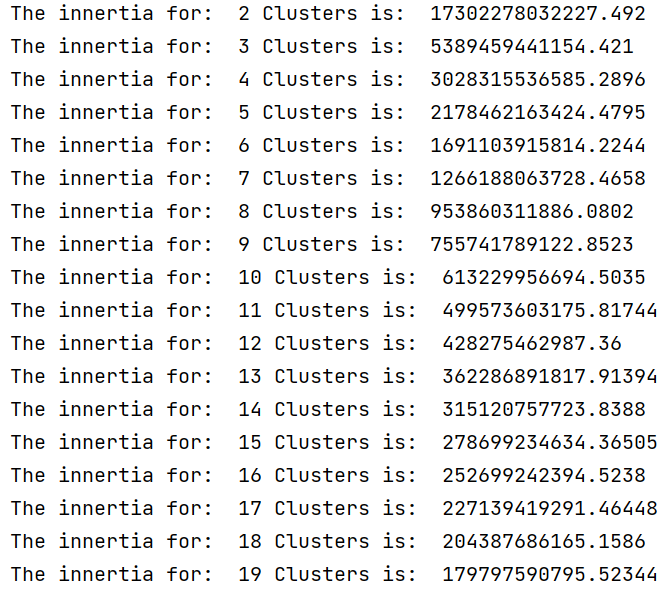


Figure Inertia for more clusters

1. Create a plot with the inertia values to understand where the curve isn’t too steep

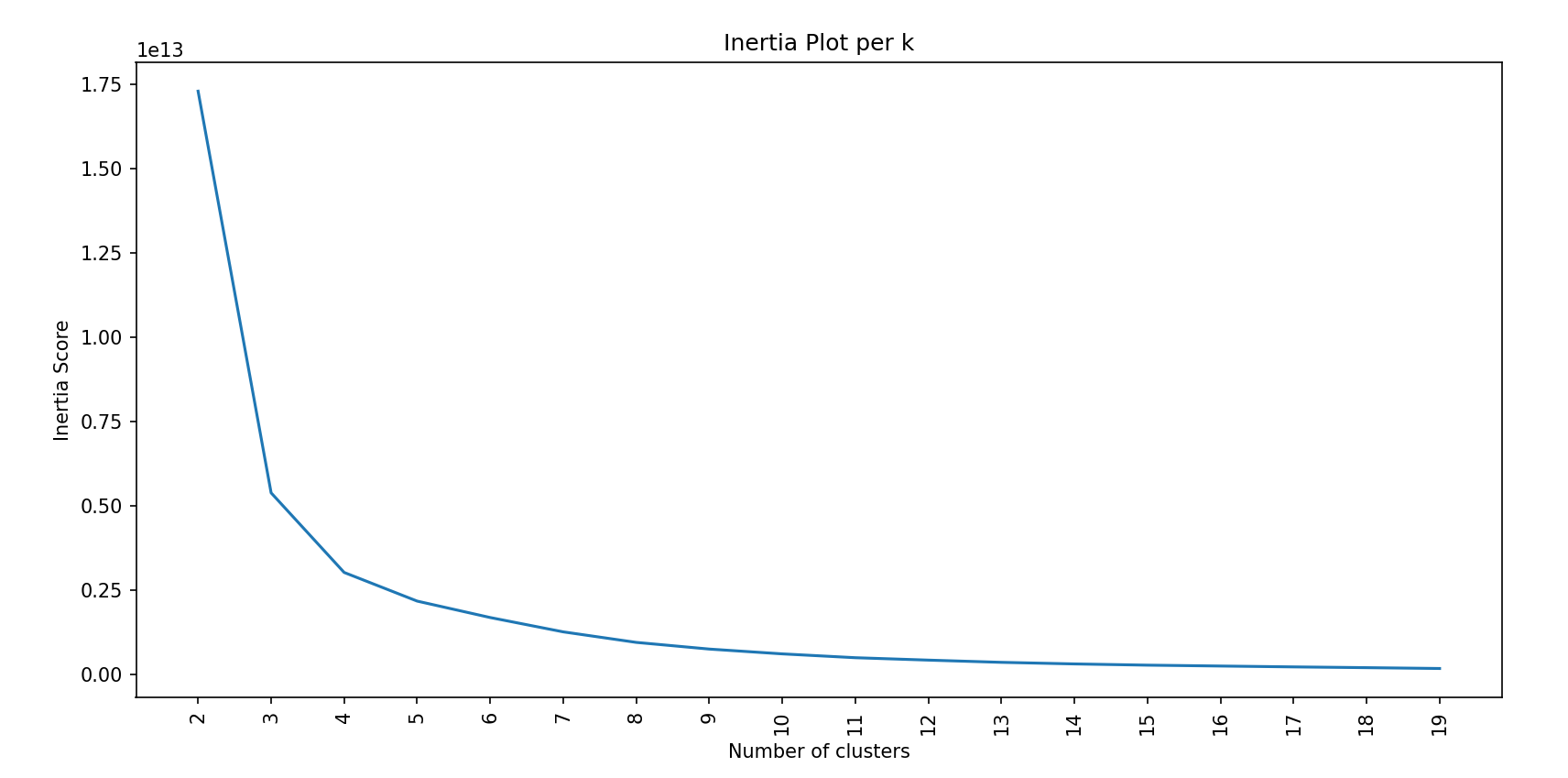


Figure Inertia graph

Inertia measures how well a dataset was clustered by K-Means. It is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster.

A good model is one with low inertia AND a low number of clusters (K). However, this is a tradeoff because as K increases, inertia decreases.

To find the optimal K for a dataset, use the Elbow method; find the point where the decrease in inertia begins to slow. (Clustering: K-Means, 2024)

* Based on the Inertia plot, optimal number of clusters is 5

Next steps:

1. Apply kMeans algorithm on 5 clusters

A close up of a text

Description automatically generated

Figure Kmeans algorithm

1. Save cluster labels and predictions for new dataset

cluster\_labels = kmeans.labels\_

predictions = kmeans.predict(X\_train)

1. Calculate how many objects are in each cluster

A black text on a white background

Description automatically generated

Figure

* Apply PCA to visualize the data and see if we can obtain better inertia scores
* Normalization before PCA: can help in achieving better PCA results, especially if the features are on different scales. This scales each feature to have a mean of 0 and a variance of 1.

A close up of words

Description automatically generated

Figure

* Applied PCA on 2 components

A black and purple text

Description automatically generated

Figure

The output indicates that the first two principal components explain about 26.59% of the variance, which suggests that my data's variance is spread across more components

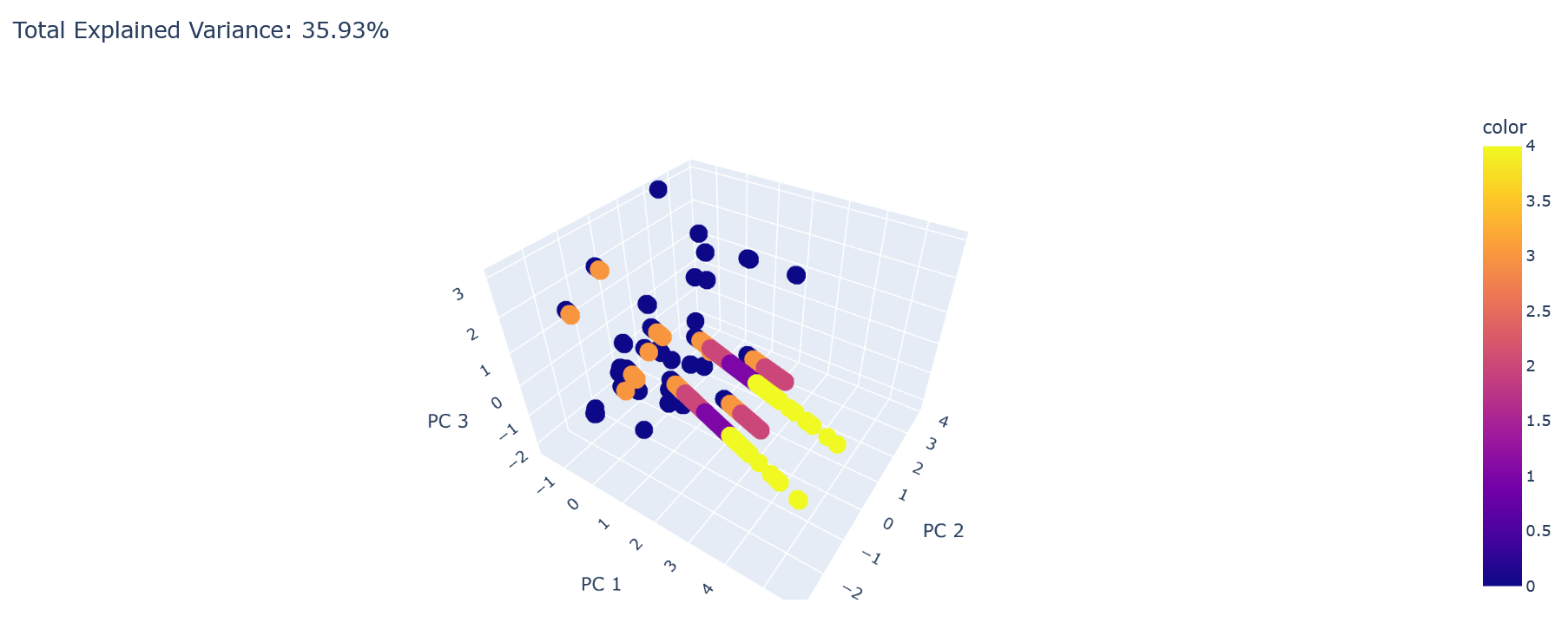
* Applied PCA on 3 components

A close up of a word

Description automatically generated

Figure

Now for 3 components we can see that the total Explained Variance: 35.93%.



Figure

* Run PCA on more components

A close-up of a computer code

Description automatically generated

Figure

The goal is to hit 95% of the explained variance by the components.

A screenshot of a computer code

Description automatically generated

Figure

The total variance is 12 and we want to have 11.4 therefore, we can see that at component 10 approximately we can obtain 95% of the explained variance.

A screenshot of a computer code

Description automatically generated

Figure

A red line graph with white background

Description automatically generated

Figure

* Run PCA on 10 components to minimize inertia
* Run K Means again

We can see that the inertia is smaller that at the beginning, so we manage to obtain better results by minimizing inertia

A screenshot of a computer

Description automatically generated

Figure

A graph with a line

Description automatically generated

Figure

## 2.2 Hierarchical agglomerative clustering

Using the same dataset X, I have created a dendrogram to see the clusters however the dataset is too big, therefore I have selected randomly 10% of the data to create a dendrogram. We can see that the optimal number of clusters is 3 if we draw a horizontal line at distance 3.

A diagram with green lines

Description automatically generated

Figure

A screenshot of a graph

Description automatically generated

Figure

* I applied the algorithm on 3 clusters



# Conclusions by comparing the 2 algorithms

As science says, there is no perfect algorithm that fit one dataset. There is no wrong output, but there is always room for improving the outputs, either by applying PCA to reduce the number of components, or by trying to normalize the data or trying different other algorithms that creates clusters in different ways.

In order to compare the 2 algorithms I calculated 3 metrics for each algorithm.

A screenshot of a computer

Description automatically generated

Silhouette Score

KMeans: 0.7812305849505805

Hierarchical: 0.8107300438200656

Interpretation:

The Silhouette Score measures how similar an object is to its own cluster compared to other clusters. Higher values indicate better-defined clusters. Here, Hierarchical clustering has a higher Silhouette Score, suggesting it forms better-defined clusters compared to KMeans.

Davies-Bouldin Index

KMeans: 0.4779926551667796

Hierarchical: 0.4338461376222593

Interpretation:

The Davies-Bouldin Index is a ratio of within-cluster distances to between-cluster distances. Lower values indicate better clustering. Hierarchical clustering has a lower Davies-Bouldin Index, indicating better cluster separation and less within-cluster variance compared to KMeans.

Calinski-Harabasz Index

KMeans: 83736.19337595851

Hierarchical: 58761.957825114136

Interpretation:

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values indicate better-defined clusters. KMeans has a significantly higher Calinski-Harabasz Index, suggesting it forms more distinct clusters than Hierarchical clustering.

Summary

**Silhouette Score: Hierarchical (0.8107) is better than KMeans (0.7812).**

**Davies-Bouldin Index: Hierarchical (0.4338) is better than KMeans (0.4780).**

**Calinski-Harabasz Index: KMeans (83736.19) is better than Hierarchical (58761.96).**

Considering two out of three metrics (Silhouette Score and Davies-Bouldin Index), Hierarchical clustering performs better than KMeans in terms of forming well-defined and well-separated clusters. However, KMeans performs significantly better in the Calinski-Harabasz Index, indicating more distinct clusters in terms of variance.

Conclusion:

Hierarchical clustering generally performs better based on the Silhouette Score and Davies-Bouldin Index, suggesting better overall cluster quality. However, KMeans demonstrates superior performance in the Calinski-Harabasz Index, indicating stronger distinctness between clusters. The choice between the two algorithms may depend on the specific importance of these metrics for your application.

# Bibliography

*Clustering: K-Means*. (2024, 05). Retrieved from Code Academy: https://www.codecademy.com/learn/dspath-unsupervised/modules/dspath-clustering/cheatsheet

*Hierarchical Clustering in Data Mining*. (2024, 05). Retrieved from Geeks for geeks: https://www.geeksforgeeks.org/hierarchical-clustering-in-data-mining/

McGregor, M. (2024, 05). *8 Clustering Algorithms in Machine Learning that All Data Scientists Should Know*. Retrieved from Free Code Camp: https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/

Patel, H. (2024, 05). *Feature Engineering Explained*. Retrieved from Built in: https://builtin.com/articles/feature-engineering#:~:text=Apr%2029%2C%202024-,Feature%20engineering%20is%20the%20process%20of%20selecting%2C%20manipulating%20and%20transforming,used%20in%20a%20predictive%20model.

Patlolla, C. R. (2024, 05). *Understanding the concept of Hierarchical clustering Technique*. Retrieved from Medium: https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec