**CLUSTERING- FINDING BEST DEAL WITH FLIGHTS**

## Group Name: A

## Group Members:

|  |  |  |
| --- | --- | --- |
| **First name** | **Last Name** | **Student number** |
| Ashish singh | Bisht | C0896264 |
| Ashok kumar | Vajravelu | C0894469 |
| Keerthana | Amalanathan | C0895321 |
| Priyanka | Kothapalli | C0894988 |
| Sai Kiran | Pilli | C0892418 |
| Thirumalan | Sathananthan | C0902020 |

**Submission date:** 12-12-2023

Table of Contents

[Abstract 3](#_bookmark0)

[Introduction 3](#_bookmark1)

[Methods 4](#_bookmark2)

[Discussion 14](#_bookmark3)

[Conclusions and Future Work 16](#_bookmark4)

[References 16](#_bookmark5)

[Acknowledgment 17](#_bookmark6)

**CLUSTERING- FINDING BEST DEAL WITH FLIGHTS**

# Abstract

Our main goal for this project is to use clustering techniques ("Clustering," n.d.) to examine and form clusters to give the best deals with flights, giving us essential insights into pricing trends and travel behavior. For this, we used the website ‘Kayak’ ("KAYAK," n.d.) for extracting data using scraping methods with Python library Beautiful Soup (GeeksforGeeks, 2022). After doing clustering techniques, the clusters were made by analyzing the flight travel duration, layover time, and price details. This resulted in clusters that the flight ticket even with less travel and less layover duration fell under the best deal and vice versa.

# Introduction

The project's primary goal is to do away with manual cognitive analysis in favor of an advanced technique that selects the optimal option from a range of possibilities. By using ML techniques, it is possible to suggest or advise the best airline carrier to travel from a source to a destination when applying this to a trip planner use case.

The goal is to analyze all the scenarios and aspects of the factors that would impact choosing an flight that carries passengers from one place to another and produce the best recommendation.

Working with real and relevant data is essential for this use case, so a web scraping approach has been implemented to fetch real-time data from online to address the real data constraints and complexities. The data set from the "Kayak" ("KAYAK," n.d.) website includes all flight data from a source to a destination over four months from January 2024 to April 2024. Those data are extracted by each person in the group and added in GitHub (P. K ,2023). This gives us details on several carrier types that provide diverse options for itineraries, costs, luggage allowances, etc.

Factors such as name of the carrier, layovers, number of bags, total journey time, cost, and other criteria are all taken into account to determine the optimal airline recommendation based on a thorough analysis of all the scenarios and elements that could influence the decision to select a carrier and are examined, prioritized, clustered, treated suitably for model training, and conclusions are drawn.

The general steps involved in data wrangling include cleaning, organizing, resolving issues with data distribution, and handling outliers. Afterwards, to guarantee the best model prediction, pre-modeling principles like feature co-relationships, converting all features to an appropriate informative format, and PCA analysis are followed.

Since the data are grouped into multiple groups according to different inter-feature associations or similarities, the clustering technique stands out as the most crucial component of the entire modeling process. This is beneficial since the separated format helps the model analyze and comprehend the data

more thoroughly. Consequently, the optimal option determined by the criteria is proposed, making the selection process evident, logical, and justifiable.

# Methods

The analysis is conducted within the context of examining flight data, focusing on understanding pricing trends and travel behavior. The setting involves utilizing web scraping methods to extract relevant data from the 'Kayak' ("KAYAK," n.d.) website.

### Dataset details:

The 'Kayak' ("KAYAK," n.d.) website is the project's primary data source, from which flight ticket information is taken. The dataset includes airline, flight, and baggage details, such as travel duration, layover times, and associated costs. Using **Beautiful Soap** (GeeksforGeeks, 2022, we scraped websites and collected data for four months, from **January to April 2024**. The dataset we obtained included **rows around 70k and 13 columns**, with **six distinct source-to-destination** countries. The following are the columns that we extracted from the website.

1. Date
2. Airline
3. Duration
4. Source
5. Destination
6. Departure Time
7. Arrival time
8. Layover time
9. Layover cities
10. Total stops
11. Carry bags count.
12. Checking bag count
13. Price

A bar graph with different colored bars

Description automatically generated

After importing the dataset, missing values are handled, and the features with less scope are removed. The categorical columns missing values are replaced by mode. We categorized departure time and arrival time as Midnight, Early Morning, Morning, Afternoon, Evening, and Night, because the customers always categorize like this instead of seeing the exact time.

We had 25% of outliers in our dataset. So, we used three types of outlier handling (Quantile-based Flooring and Capping, Trimming, and Log Transformation), compared the results by analyzing the visualization, and

decided to use the Quantile-based Flooring and Capping. We added a new feature for the price per minute of the duration (price per minute) to add more insights.

We encoded categorical columns by doing one-hot-encoding. Month, year, and date separate the date column separated by moth, year, and date for better analysis. We used MultiLabelBinarizer to encode airlines.

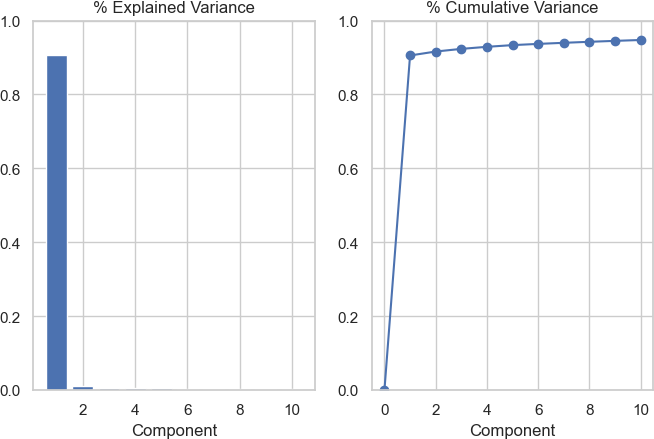
The analysis employs a clustering approach to analyze and categorize flight data. The clustering techniques are applied to explore patterns based on flight duration, layover time, and price details. Clustering techniques (K- Means Clustering, Agglomerative clustering) are applied to explore patterns based on flight duration, layover time, and price details.

The algorithms used focus on grouping flights with similar characteristics into clusters. The analysis considers the relationships between flight duration, layover time, and price details to form meaningful clusters. Both successful and unsuccessful clustering approaches are documented to provide a comprehensive understanding of the outcomes.

### DIMENSIONALITY REDUCTION

We end up with 177 features columns (encoded categorical columns included), which has increased the dimensionality of the data, and this will make it difficult for the model to converge optimally. Since many of these features are correlated, most of these are redundant. We can use a dimensionality reduction technique called Principal Component Analysis (**PCA**).

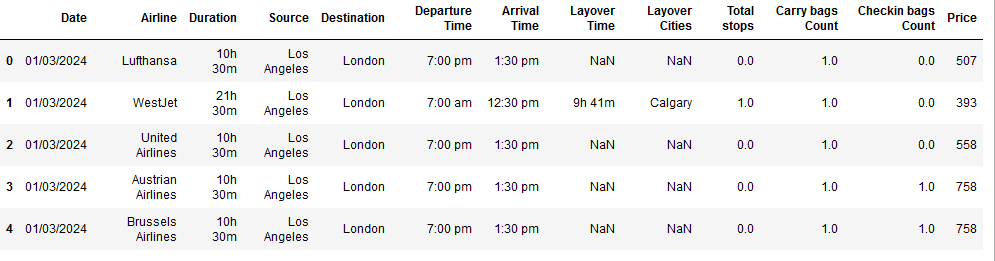
Principal components, a new collection of uncorrelated variables, are created from the original features by PCA. Most of the the data's variability is captured by these components. PCA aids in lowering the dimensionality of the data by keeping only a portion of these elements. This can improve computational efficiency and lower the chance of overfitting, making it especially helpful for datasets with a lot of features.



In our dataset using number of components as 10 we can explain ~95% of variance in the data and reduce the data dimensionality.

Results

After the data is scrapped from ‘Kayak’ ("KAYAK," n.d.) the data took the shape of 70,349 rows and 13 columns. Breaking down further, 4 categorical columns and 9 numerical columns including the date column.



Stepping into data wrangling stage, date column is converted to appropriate format. Consequently, the missing values of the data set is analyzed and found that around 11.77% of data is missing in 3 columns namely ‘Lay over time’, ‘Lay over cities’, ‘Carry bag counts’ and ‘Check-in bag counts.

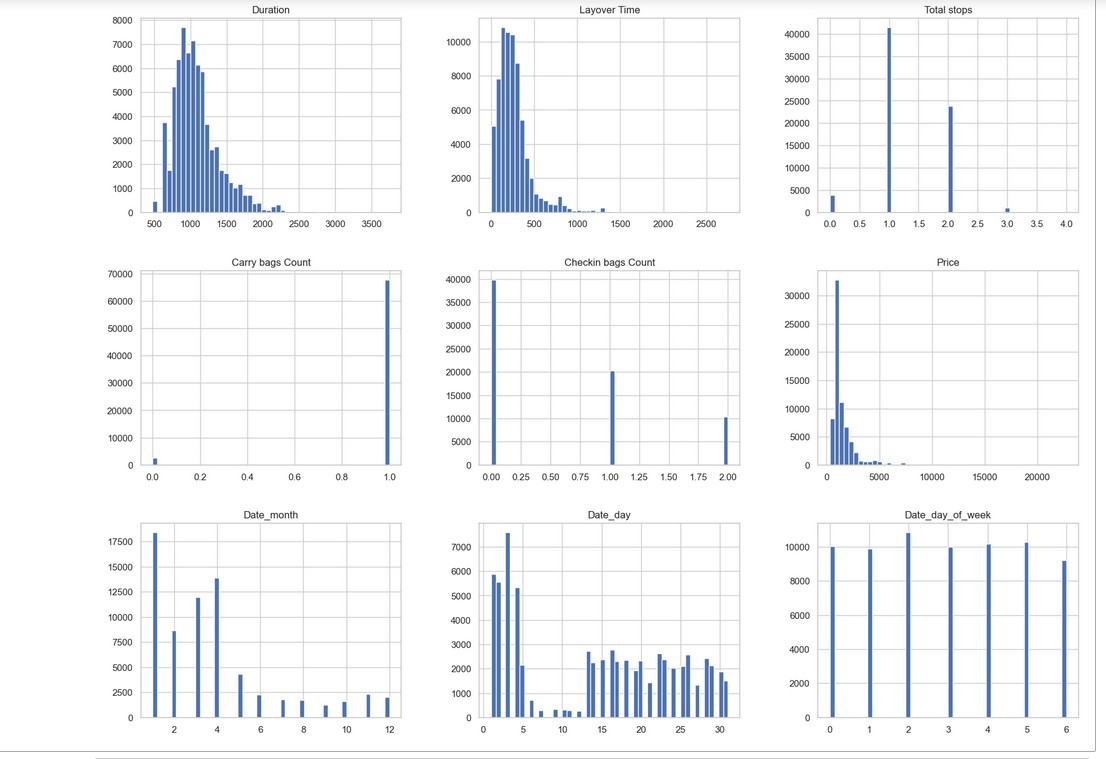
Additionally, it is noted that Total stops equals 0 whenever Layover cities and Layover Time values are both "NaN," indicating that if there are no stops, there won't be any Layover cities. We chose to remove this column because, like "Data\_year," it has little bearing on finding the best flight deals. Since the data we have is from 2024, the feature has a unique value. Layover Time ‘NaN’ values are made zero. Hence, proceeded after dropping ‘Lay over cities’ column.

To address the missing value in the above stated columns we considered replacing the missing values with

mode of the group since each airline will follow a certain custom for ‘Check-in bags’ and ‘Carry bags’.

In terms of feature engineering, the time columns such as ‘Duration’ and ‘Lay over time’ are converted to minutes format for easy processing. Furthermore, ‘Date’ column is broken into hours and minutes to identify parts of the day and thereby, constructing to new columns ‘Arrival Time New’ and ‘Departure Time New’ filling values such as ‘Morning’, ‘Evening’ being a little more insightful as they are also ordinal value columns of the data set.

The distribution thus achieved is,



Moving on to further assessment of the distribution, outliers were detected in columns ‘Lay over time,’ ‘Duration’ and ‘Price.’

Hence, best practices for handling outliers were all visualized before choosing the optimal solution.

Here we have added some visualizations to get the relationship between features.

A graph with blue lines

Description automatically generatedA graph of different colored squares

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated with medium confidence

In both of arrival and departure time price is high in the morning as most of the people like to be reached to the destination and started from the source.

A line graph with numbers and a red line

Description automatically generated

Mostly people fly accross different countries will prefer two or more checkin bags, So the price and count if checking bags are highly correlated

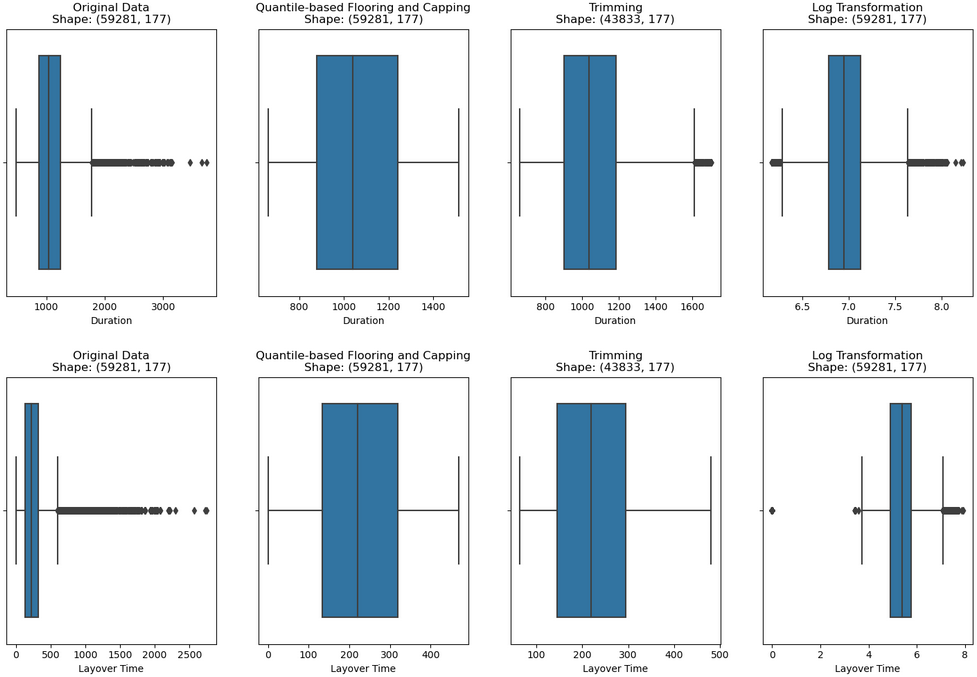
A green and red line graph

Description automatically generated

The regression lines shows, if the layover time increses, price also increses. SO they are positively correlated with each otherA screen shot of a graph

Description automatically generated

From the above visualization we can identify that British Airways has the highest no of flights in the dataset which is 8699



#### from above plots we can conclude the following points

#### By using the IQR method with q1 = 0.5 and q3 = 0.90, we can handle outliers.

#### By using trimming method, we can be able to handle outliers, but we are losing some data which will bias the model

#### By using Log transformation there are so many outliers which can't get handled. We choose to go with IQR method which seems to be appropriate for our dataset

The cleaned data set looked in great shape except the nominal columns such as ‘Source’, ‘Destination’,

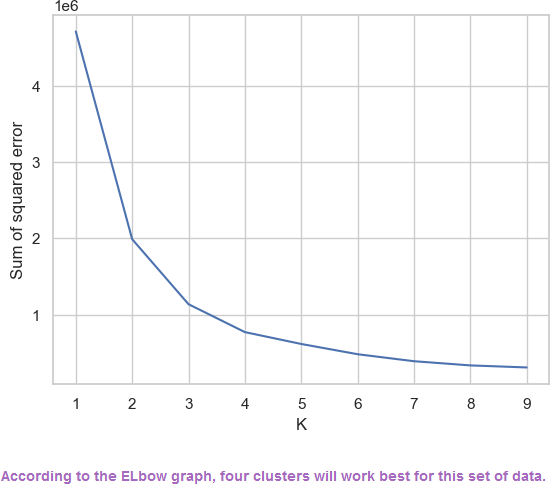
‘Date\_day’, ‘Date\_month’ and ‘Date\_day\_of\_week’ which is converted into beneficial columns using one-hot encoder, resulting increase in the total number of columns. Total columns have increased to 177 in total.

Since, one-hot encoder was used maximum of the columns would have values within the range of [0-1] hence it is wise to use Min-Max normalization method to scale the rest of the columns.

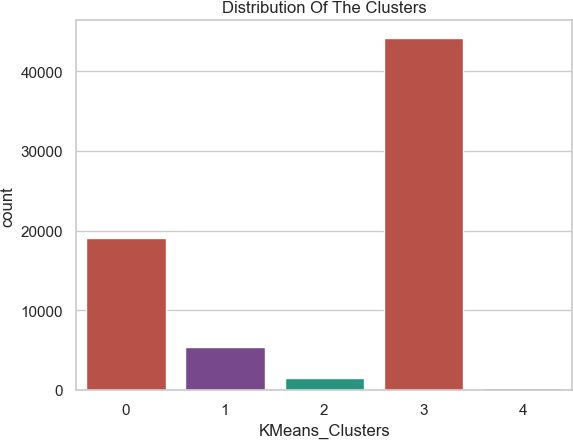
Additionally, since the size of the data set is now drastically increased, we used PCA to analyze just the principal components and narrowed down to just 2 components explaining 95% of the entire variance.

To eventually implement clustering technique, used K-Means elbow method to determine the no of clusters suitable for the chosen data set.

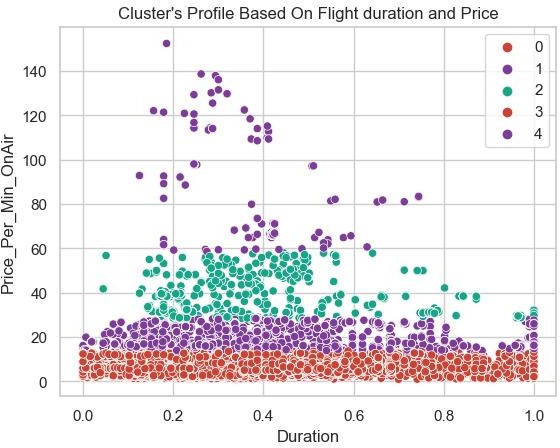
According to the graph 4 clusters are needed,



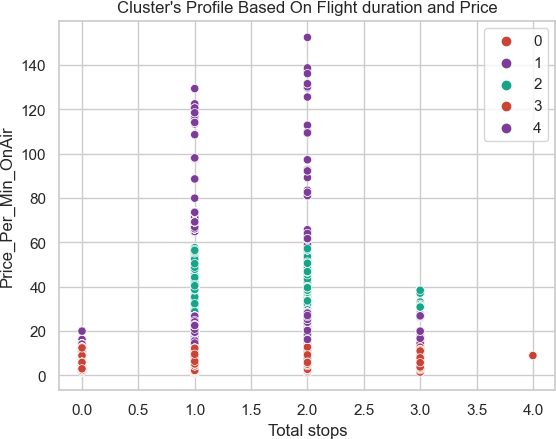
Distribution of each cluster were,



To get more detailed insights creating profile against features,

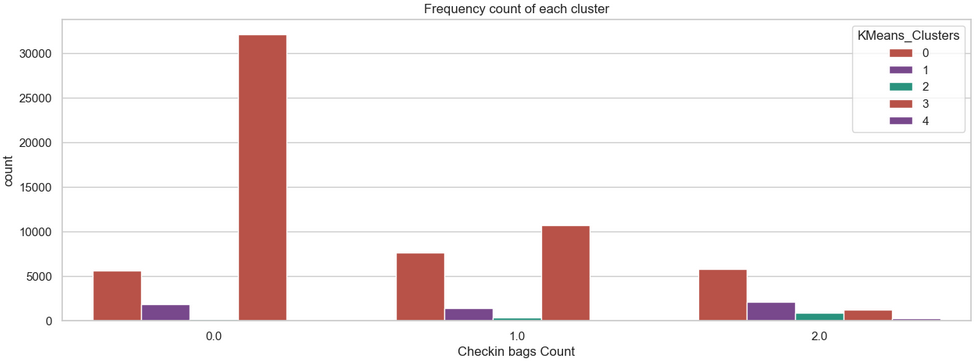


Cluster: Duration VS Price\_Per\_Min\_OnAir

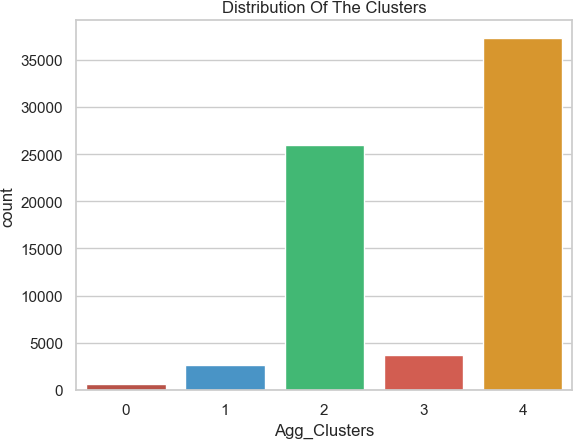


Cluster: Total Stops VS Price\_Per\_Min\_OnAir

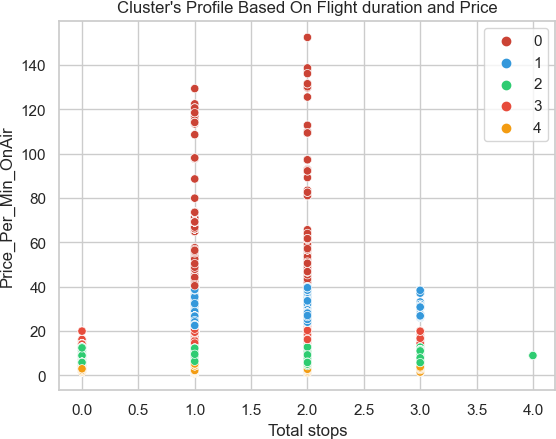
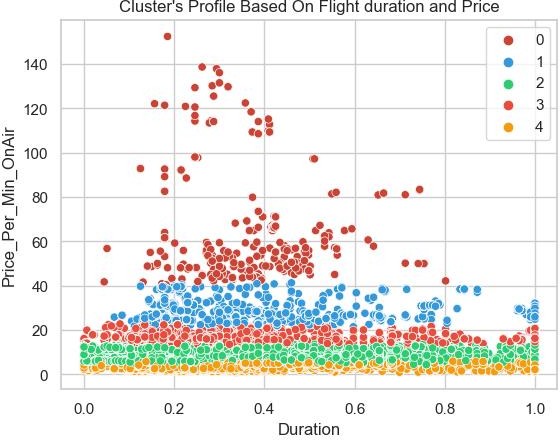
Frequency count of each cluster,



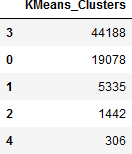
Implementing agglomerative clustering to find if this provides better cluster groups,



Comparing the results of cluster profiling of agglomerative clustering to K- Means clustering,



Comparing distribution of clusters,

Both the types of cluster methods require and form the same exact number of clusters even leading to almost close distribution of data within the clusters

### Concluding best clustering method:

To decide the best clustering method, we have compared all the three metrics (Zuccarelli, 2021) between the clustering technique outputs.

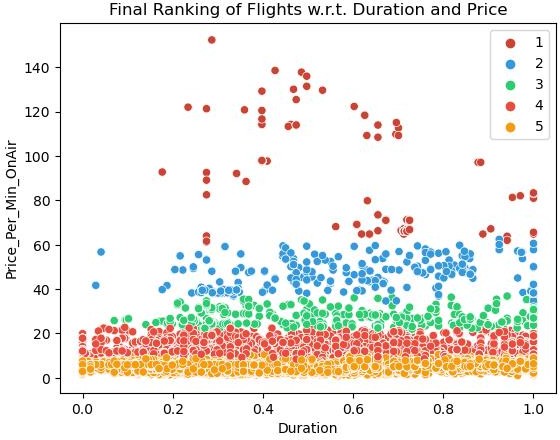
Clusters formed with agglomerative clustering model have better Silhouette Score than K means clusters. It indicates that clusters formed with Algometric model are better at matching a given object to its own cluster than its neighboring ones.

1. Clusters formed with Algometric clustering model have better Silhouette Score than K means clusters. It indicates that clusters formed with Algometric model are better at matching a given object to its own cluster than its neighboring ones.
2. According to the Davies-Bouldin Index got for both techniques, K means clusters got value of 0.7 which is lesser than 0.8 got with Algometric clustering. This indicates that K means clustering is performing better according to this index.
3. The higher index of Calinski-Harabasz with k means clustering(50k) indicates that clusters formed with this technique are better than algometric clustering.

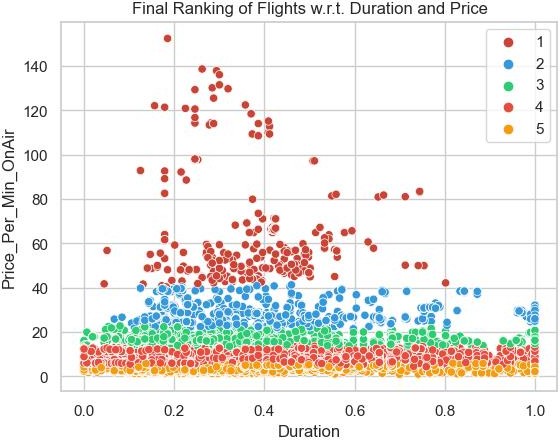
Comparatively metrics we got from both methods do not have much difference, but we can say that K means has performed little better than Algometric clustering for our dataset. So, we decided to stick with Algometric clustering.

### Rating each cluster on Agglomerative cluster:

#### To rate the clusters, we grouped the whole data according to the AgglomerativeClustering and then found the average Price\_Per\_Min\_OnAir of each cluster and then the cluster with least average is given Rating 5 and the highest is given Rating 1.

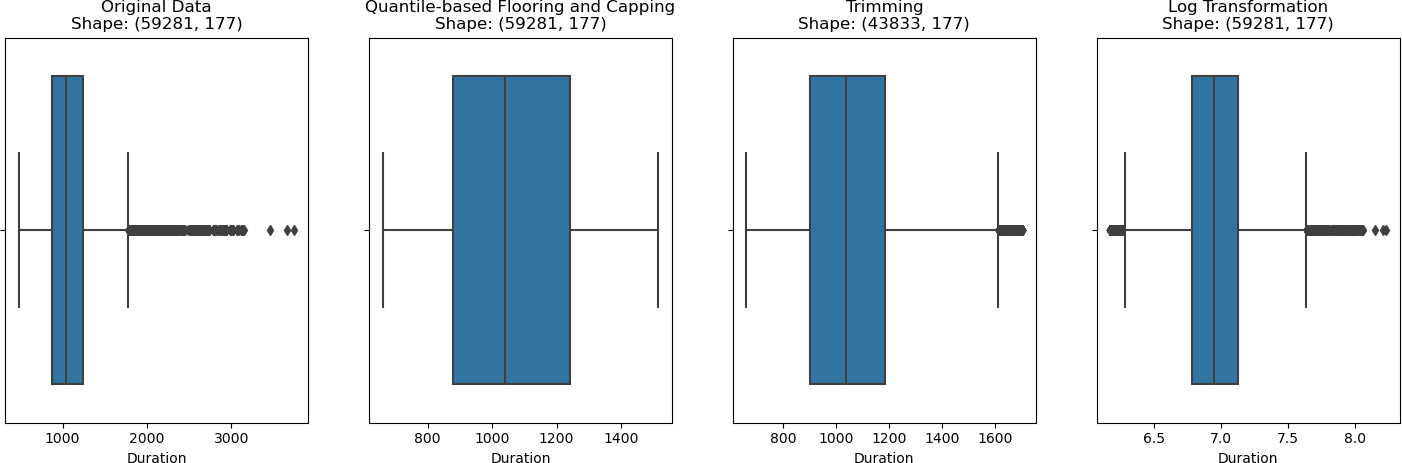


# Discussion

* **Objective Fulfillment:** By analyzing the clustering distribution, the best deal for the customers fell under the condition either less or more duration with affordable and lower price shown below .
* **Data Source and Realism:** Working with real-time data from Kayak ("KAYAK," n.d.) helped us to get more insights. However, we faced some problems while extracting data using beautiful soup. So,

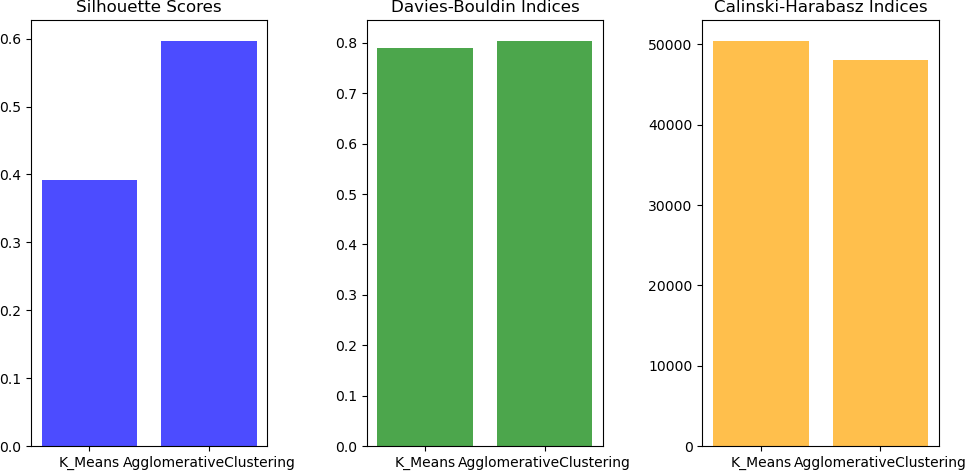
instead of sticking with one method, we tried executing JavaScript codes to get data from the web instead of beautiful soup. Three to four hours were spent on the data extraction process, which included six different source-destination pairs and four months. The length of time was due to the careful extraction from several sources. Each team member committed time to extracting data from a particular source-destination combination, which helped create the comprehensive dataset.

* **Factors Considered:** We consider the quantity of checked and carried bags because customers may need to bring additional items, particularly when relocating internationally. Despite the cost, most customers will think less of a layover. Thus, we examined these factors in depth.
* **Outlier Handling and Feature Engineering:** Even though the Trimming method for outliers worked well as shown in the below figure for Duration column, we selected quantile-based flooring and capping, as trimming resulted in almost 50% of data loss. Adding a new feature price\_permin\_OnAir which will give the price for each flight per minute during the travel means removing layover time and multiplying it with the ratio of baggage and number of stops. This is to normalize the price factor for all flights despite their duration and can be used as a deciding factor at the end to give ratings to the flight deal.



* **Categorical Column Encoding:** The utilization of label encoding for airlines, departure time, and arrival time significantly enhanced clustering, yielding valuable insights and revealing intricate patterns within the data. This encoding method contributed to a more effective analysis, enabling the extraction of meaningful patterns and relationships from the dataset.
* **Clustering Techniques and Outcomes:** We have done K-Means clustering and Agglomerative clustering. Based on the metrics shown in below figure we can conclude the points below.
  1. Clusters formed with Agglometric clustering model have better Silhouette Score than K means clusters. It indicates that clusters formed with Agglometric model are better at matching a given object to its own cluster than its neighboring ones.
  2. According to the Davies-Bouldin Index got for both techniques, K means clusters got value of

0.7 which is lesser than 0.8 got with Agglometric clustering. This indicates that K means clustering is performing better according to this index.

* 1. The higher index of Calinski-Harabasz with k means clustering(50k) indicates that clusters formed with this technique are better than agglometric clustering.Comparatively metrics we got from both methods do not have much difference, but we can say that K means has performed little better than Agglometric clustering for our dataset.

# Conclusions and Future Work

The clustering technique ("Clustering," n.d.) implemented can segment the data into proper use groups from which certain characteristics can be used to arrive at the required solution in this case of achieving the best deal on airline from point A to point B. Factors such as fare, luggage count and duration are considered in this case to determine a cluster that meets all the minimal requirements of the above.

Future improvisation of the above analysis could be that a suggestive mechanism could be implemented given the requirements to advise the best carrier.

# References

*Search for Cheap Flights & Airline Tickets |* KAYAK. <https://www.ca.kayak.com/flights>

P. K. (2023). *GitHub - priyankaa-k/Kayak\_flight\_data\_clustering*. GitHub. [https://github.com/priyankaa-](https://github.com/priyankaa-k/Kayak_flight_data_clustering) [k/Kayak\_flight\_data\_clustering](https://github.com/priyankaa-k/Kayak_flight_data_clustering)

*Implementing Web Scraping in Python with BeautifulSoup*. (2022, June 28). GeeksforGeeks. <https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/>

*How to convert time object into total minutes python*. (n.d.). Stack Overflow. <https://stackoverflow.com/questions/27475583/how-to-convert-time-object-into-total-minutes-python>

*2.3. Clustering*. (n.d.). Scikit-learn. [https://scikit-](https://scikit-learn.org/stable/modules/clustering.html#%3A~%3Atext%3D2.3.1.%20Overview%20of%20clustering%20methods%20%C2%B6%20%20%2Cnearest-neighbor%20gr%20) [learn.org/stable/modules/clustering.html#:~:text=2.3.1.%20Overview%20of%20clustering%20methods%20%C](https://scikit-learn.org/stable/modules/clustering.html#%3A~%3Atext%3D2.3.1.%20Overview%20of%20clustering%20methods%20%C2%B6%20%20%2Cnearest-neighbor%20gr%20) [2%B6%20%20,nearest-neighbor%20gr%20.](https://scikit-learn.org/stable/modules/clustering.html#%3A~%3Atext%3D2.3.1.%20Overview%20of%20clustering%20methods%20%C2%B6%20%20%2Cnearest-neighbor%20gr%20) . .%20%208%20more%20rows%20

Zuccarelli, E. (2021, December 29). *Performance Metrics in ML - Part 3: Clustering | Towards Data Science*. Medium. [https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-](https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6) [d69550662dc6](https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6)