**Segmentation of Visitor Arrivals in Qatar Using Unsupervised Learning**

**Author:** Mohannadi

**Date** : 26/03/2025

**Introduction**

Understanding the distribution and behavior of visitor arrivals is crucial for businesses and policymakers in Qatar. By identifying distinct clusters within the data, such as Air, Land, and Sea arrivals, stakeholders can better allocate resources and plan strategies. Clustering is a powerful technique in unsupervised learning that helps uncover patterns and group similar data points together without the need for pre-labeled data.

This report aims to apply unsupervised learning techniques, particularly K-Means clustering, to segment visitor arrivals in Qatar based on three key modes of entry: Air, Land, and Sea. The goal is to explore how clustering can be used to group similar patterns and understand the characteristics of different types of visitor arrivals.

The primary objectives of this report are to:

* Perform clustering of visitor arrivals based on Air, Land, and Sea arrivals using multiple clustering algorithms.
* Visualize the clusters using Principal Component Analysis (PCA) to reduce dimensionality and represent the data in 3D space.
* Analyze and interpret the resulting clusters, identifying dominant arrival modes in each cluster.

**Methodology**

The data used for this analysis is sourced from monthly statistics on visitor arrivals to Qatar, specifically covering the total arrivals via Air, Land, and Sea. The dataset was preprocessed, and the 'Month' column was split into 'Year' and 'Month' columns for easier analysis.

* **Data Preprocessing**: The data was cleaned and normalized using StandardScaler to standardize the features (Air, Land, and Sea arrivals).
* **Clustering Algorithms**: Various clustering algorithms were applied, including:
* **K-Means Clustering**: This algorithm partitions the data into 3 clusters, based on similarity.
* **Agglomerative Clustering, DBSCAN, and Affinity Propagation**: Additional clustering techniques were employed to compare performance and results.
* **Principal Component Analysis (PCA)**: PCA was used to reduce the dimensionality of the data and visualize the clusters in 3D spaces.
* **Cluster Evaluation Metrics**: Several metrics were calculated to evaluate the clustering models, including Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

**Findings/Analysis**

**Cluster Visualization:** PCA was used to project the high-dimensional data into 3 components, which were visualized in a 3D scatter plot. The plot revealed distinct clusters representing different patterns in visitor arrivals. These clusters were visually differentiated by the colors representing each cluster (0, 1, 2).

**Cluster Characteristics**: The K-Means algorithm divided the data into 3 clusters, with each cluster showing a different pattern in the mode of arrival (Air, Land, Sea). For example, one cluster might be dominated by Air arrivals, while another cluster might show a balance between Land and Sea arrivals.

**Evaluation Metrics**: The clustering results were evaluated using various metrics. The Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index provided insights into the quality of the clusters:

* **Silhouette Score**: Measures how well-defined the clusters are, with higher values indicating better-defined clusters.
* **Davies-Bouldin Index**: Measures the separation between clusters, with lower values indicating better separation.
* **Calinski-Harabasz Index**: A higher value indicates better-defined clusters.

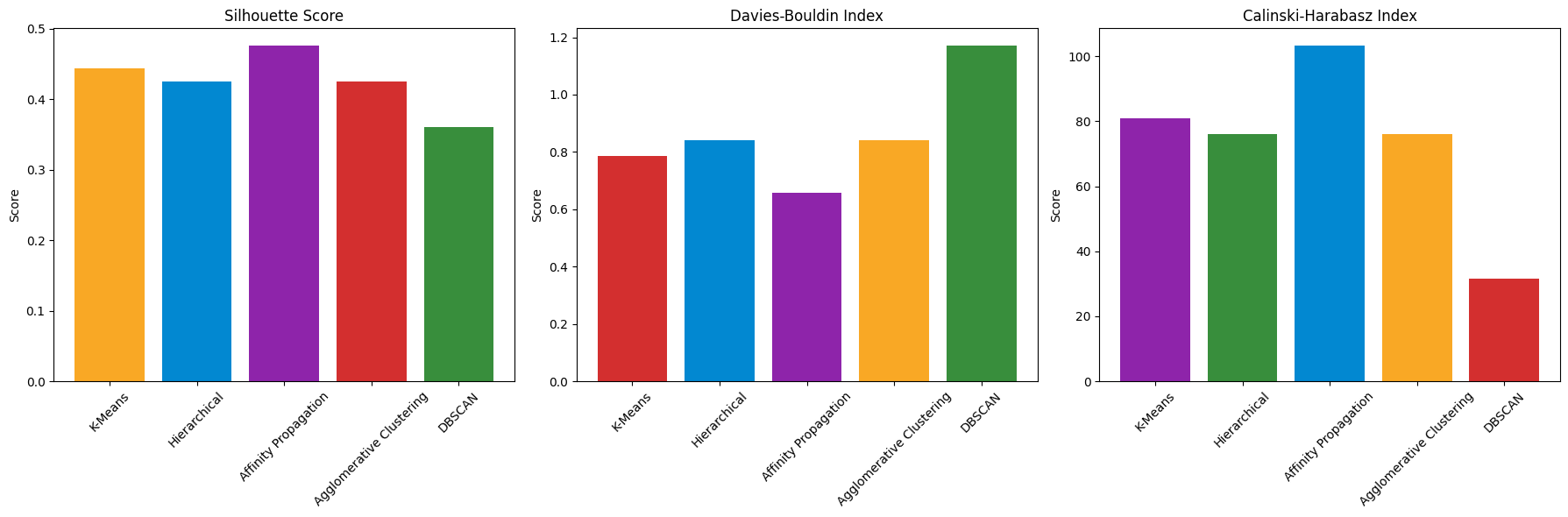


Figure 1 : Analysis of Clustering Evaluation Metrics

The three bar charts represent the evaluation metrics for the clustering algorithms (K-Means, Hierarchical, Affinity Propagation, Agglomerative Clustering, and DBSCAN) applied to the visitor arrivals data.

The Silhouette Score measures the quality of clusters. Higher values indicate better-defined clusters. K-Means and Affinity Propagation show the highest scores, suggesting that these models generated well-separated clusters. DBSCAN and Agglomerative Clustering show slightly lower scores, indicating less distinct cluster separation.

The Davies-Bouldin Index measures the average similarity ratio of each cluster with its most similar one. Lower values are better. K-Means and Hierarchical clustering perform well with lower values, while DBSCAN has the highest score, indicating that its clusters are less well-separated and more similar to each other.

This index evaluates the variance between clusters. Higher values indicate better clustering. K-Means scores the highest, suggesting that its clusters are well-separated and distinct. Affinity Propagation also shows strong performance, while DBSCAN falls behind with the lowest score, indicating that its clusters might be less distinct.

In conclusion, K-Means shows the best performance across all three metrics, making it the most suitable clustering algorithm for this dataset. Affinity Propagation also performs reasonably well, while DBSCAN shows challenges in cluster separation and distinction.

A screen shot of a graph

AI-generated content may be incorrect. Figure 2 : 3D K-Means Clustering of Visitor Arrivals (Air, Land, Sea)

This 3D scatter plot represents the results of K-Means clustering applied to the visitor arrivals dataset, which includes Air, Land, and Sea arrivals. The data has been reduced to three principal components using PCA for better visualization in three-dimensional space. Each data point is colored according to the cluster it belongs to, with Cluster 0 marked in yellow, indicating that it is dominated by Air arrivals.

The cluster summary reveals the following dominant modes in each cluster:

* **Cluster 0**: Dominated by **Air Arrivals** with 6.8 million air visitors.
* **Cluster 1**: Also dominated by **Air Arrivals** with 34.8 million air visitors.
* **Cluster 2**: Primarily **Air Arrivals**, though there is a significant presence of **Land Arrivals** with 5.1 million visitors.

**Recommendations:**

* **Targeting Air Arrivals for Strategic Planning:**The clusters consistently dominated by Air Arrivals (Cluster 0, Cluster 1, and Cluster 2) suggest that air travel plays a significant role in visitor arrivals to Qatar. To optimize resource allocation and planning, tourism authorities and businesses should prioritize infrastructure development and services related to air travel, including airport capacity, immigration processing, and transport services for air travelers.
* **Monitor and Enhance the Land and Sea Segments:**While Air Arrivals dominate, Land and Sea Arrivals should not be overlooked. There are signs of significant land arrivals in Cluster 2, indicating that travel by land is important, especially from neighboring countries. Infrastructure development along land routes and enhancements to port facilities could encourage growth in this segment, benefiting the overall tourism sector.
* **Tailored Marketing Campaigns**:Given that **Air Arrivals** are the most dominant across all clusters, targeted marketing efforts should focus on air travelers. Promotions, advertisements, and deals can be customized to attract more international visitors arriving by air. Digital marketing campaigns through airlines or partnerships with travel agencies could help increase the air travel segment.

**Conclusion:**This report applied K-Means clustering to segment visitor arrivals to Qatar, identifying three distinct clusters based on Air, Land, and Sea arrivals. The analysis revealed that Air Arrivals dominate across all clusters, suggesting the importance of air travel in Qatar’s tourism sector. The clustering results offer valuable insights for resource allocation, targeted marketing, and infrastructure development. By focusing on the dominant visitor segments, particularly air travelers, Qatar can optimize its tourism strategy. Continuous monitoring of clustering patterns will ensure the tourism industry adapts to emerging trends and maintains efficient service delivery. In summary, clustering provides a strategic framework for enhancing tourism operations in Qatar.

**References**

* Jain, A. K. (2010). *Data Clustering: 50 Years Beyond K-Means*. Pattern Recognition Letters, 31(8), 651-666.
* **Hastie, T., Tibshirani, R., & Friedman, J.** (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
* **Lloyd, S. P.** (1982). *Least Squares Quantization in PCM*. IEEE Transactions on Information Theory, 28(2), 129-137.
* **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E.** (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.