

Iceland Tourism Segmentation 2023

Francisca Argandona Alvarado
CCT College Dublin

Introduction

Tourism in Iceland has grown rapidly in recent years, becoming a major contributor to the economy. However, most existing profiles only considers country of origin or overall arrival numbers. To address the lack of detailed profiling, this study applied clustering techniques to segment international visitors to Iceland by age, stay duration, and income.

The goal is to uncover meaningful visitor segments that can guide more effective tourism strategies.

International visitors to Iceland				
	Keflavik Airport	Other Airports	Seyðisfjörður Seaport	Total
2024	2.261.391	7.476	18.474	2.287.341
2023	2.211.668	3.590	18.778	2.234.036
2022	1.696.785	2.400	18.464	1.717.649
2021	687.691	109	10.270	698.070
2020	482.108	1.344	6.454	489.906
2019	1.986.153	8.164	18.887	2.013.204
2018	2.315.925	7.158	19.158	2.342.241
2017	2.195.271	6.450	22.353	2.224.074
2016	1.767.726	3.859	19.795	1.791.380
2015	1.261.938	8.662	18.540	1.289.140

Source: Icelandic Tourist Board, Isavia og Smyril Line

Research Question

How can clustering techniques be applied to identify relevant tourist segments visiting Iceland, based on age, length of stay and average income?

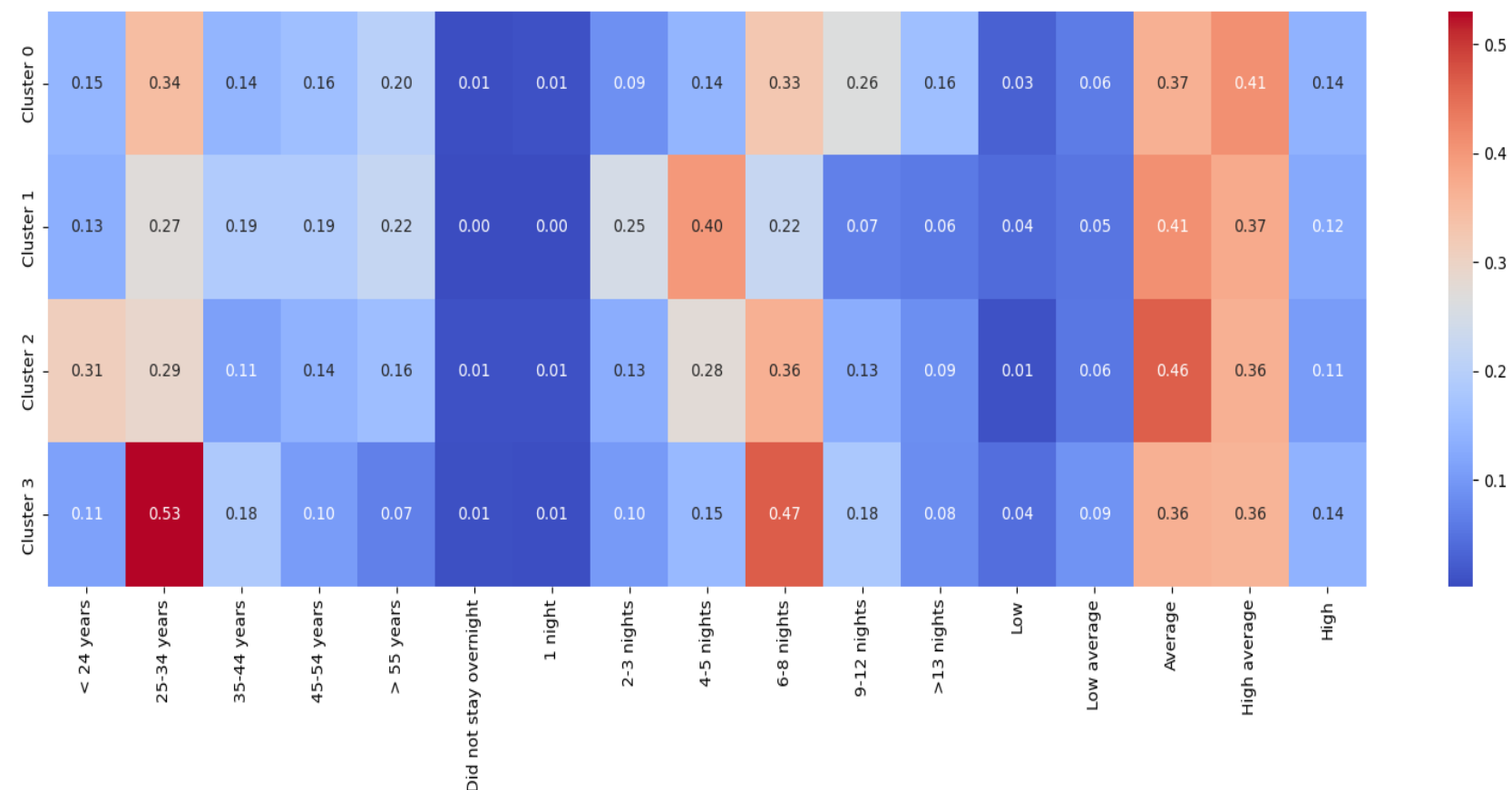
Objectives

Segment tourists using clustering techniques (K-Means, Hierarchical, DBSCAN); evaluate clustering performance; visualise and interpret tourist based on the analysed variables; provide practical insights to support data-driven tourism strategies.

Methodology

The study followed the CRISP-DM framework. The data was obtained from 2023 visitor surveys conducted at Keflavik International Airport. The selected variables were age, stay duration, and income.

Three clustering models were applied: Hierarchical, K-Means, and DBSCAN. Clustering quality was assessed using the Silhouette Score, Davies-Bouldin Index, dendrogram visualisation and ANOVA. Dimensionality reduction with PCA helped to visualise cluster separation, besides the heatmaps on cluster centres. Adjusted Rand Index was also used to test stability across multiple runs.



Data Preparation

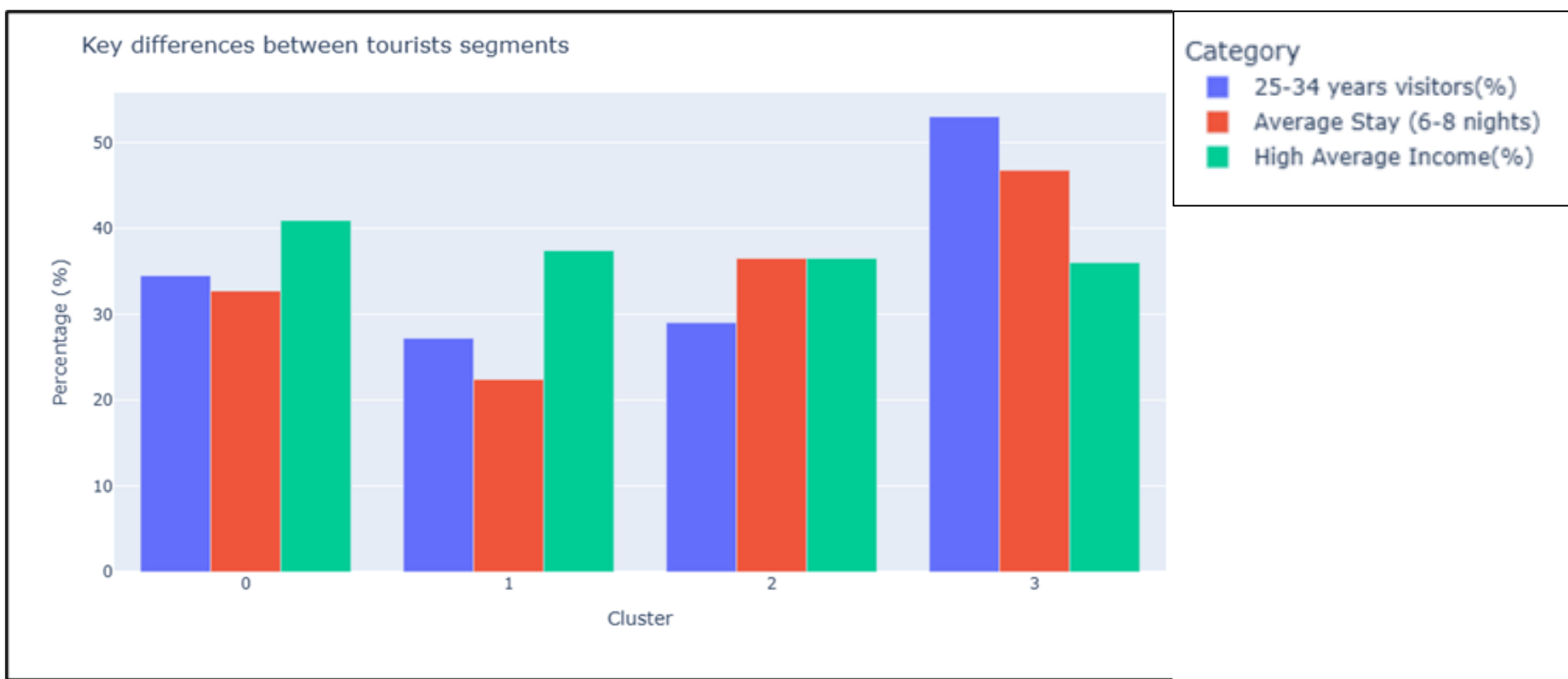
Data was extracted from multiple sheets of the original Excel and merged into one dataset. The final dataset included 21 nationalities and 17 features, all expressed as proportions. The variables were not scaled and no missing values were found. A heatmap and interactive bar plot helped to understand relationships between variables.

Results

K-Means with four clusters outperformed the other algorithms based on both visual and quantitative evaluation. Hierarchical with three clusters showed some separation, but the clusters were less distinct and less stable. DBSCAN failed to produce meaningful clusters and labelled most data as noise. The ANOVA test confirmed significant differences in age and length of stay, while income showed no clear distinction. ARI values indicated K-Means has stronger cluster stability (mean ARI= 0.61) compared to Hierarchical (mean ARI= 0.011).

Clustering Algorithm/ p-values	K-Means	Hierarchical
< 24 years	0.0102	0.4176
25-34 years	0.0002	0.0000
35-44 years	0.0997	0.0222
45-54years	0.0780	0.0360
>55 years	0.0190	0.0032
Did not stay overnight	0.7384	0.5282
1 night	0.7597	0.4639
2-3 nights	0.0011	0.0001
4-5 nights	0.0000	0.0005
6-8 nights	0.0016	0.0000
9-12 nights	0.0002	0.0493
>13 nights	0.0368	0.2511
Low	0.4456	0.0282
Low Average	0.1508	0.0725
Average	0.1124	0.3044
High Average	0.2431	0.0100
High	0.4197	0.9944

Two clusters (2 and 3) stood out for targeting young travellers with longer stays and average to high income levels, representing valuable segments for tourism strategies.



Conclusion

K-Means clustering was chosen for its clearer, more stable, and actionable segments. The results reveal distinct profiles, particularly among young tourists with moderate to long stays. This information can help stakeholders design tailored services and marketing. However, since the dataset is small and represents only one year, findings should be seen as a snapshot due to market segments change over time. Future work should include larger samples and additional behavioural variables.