Francisca Argandona, CCT College Dublin, 2025

Introduction

In recent years, Iceland has experienced a rapid increase in international tourism, transforming it into a key economic sector. This growth has brought both opportunities and challenges. To respond effectively, it is essential to understand who the visitors are and their needs. However, most current profiling focuses only on country of origin or total arrivals. This study aims to fill that gap by applying clustering algorithms to segment tourists based on age, length of stay, and income level. By identifying visitor profiles, this research contributes to more informed tourism strategies and service planning in Iceland.

The main objective is to segment international tourists who visited Iceland in 2023 using unsupervised machine learning techniques. The study aims to compare the performance of three clustering algorithms and evaluate their ability to group tourist profiles, the research seeks to contribute to more personalised services and better planning.

Research Question

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

Hypothesis

Null Hypothesis (H_0): There are no clear segmentation patterns among tourists based on the analysed variables.

Alternative

Hypothesis: (H_a): There are significant segmentation patterns among tourist, and these can be identified using unsupervised clustering algorithms.

Methodology

This project followed the CRISP-DM framework and used data from the Icelandic Tourist Board (2023), based on airport surveys. After cleaning and preparing the data, three clustering methods were applied: K-Means, Hierarchical, and DBSCAN. Cluster evaluation was conducted using the Silhouette Score, Davies-Bouldin Index, and visual inspection through heatmaps and dendrograms. After applying the clustering algorithms, Principal Component Analysis (PCA) was applied to reduce dimensionality and enable the cluster visualisation. Besides, interactive visualisations and ANOVA test were used to validate and understand the segmentation results.

Results

K-Means clustering with four groups was selected as the final model based in internal evaluation metrics and interpretability. The heatmap of cluster centres reveal clear differences, especially in age and length of stay variables. ANOVA tests confirmed those variables showed statistically significant differences across clusters, whereas income variables did not. Hierarchical clustering produced similar patterns with three clusters, and DBSCAN identified fewer dense groups with some noise. The visual and statistical results support the validity of the segmentation.

Conclusion