A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Francisca Andrea Argandona Alvarado |
| *Student Number* | 2024247 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA 3 Final Submission |
| *Assessment Due Date* | 18th May 2025 |
| *Date of Submission* |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

# Presentation Video Link

<https://youtu.be/JTXSTBi71W0>

Contents

[Introduction 1](#_Toc197273869)

[Motivation 1](#_Toc197273870)

[Research Problem 2](#_Toc197273871)

[Research Question 2](#_Toc197273872)

[Research Hypothesis 3](#_Toc197273873)

[Research Objectives 3](#_Toc197273874)

[Literature Review 4](#_Toc197273875)

[Clustering Techniques 4](#_Toc197273876)

[Methodology 5](#_Toc197273877)

[Project Aim 5](#_Toc197273878)

[Methodological Framework 5](#_Toc197273879)

[Business Understanding 5](#_Toc197273880)

[Data Understanding 5](#_Toc197273881)

[Data Preparation 6](#_Toc197273882)

[Modelling 8](#_Toc197273884)

[Evaluation 9](#_Toc197273885)

[Deployment 10](#_Toc197273886)

[Architectural Diagram 10](#_Toc197273887)

[Results and Discussion 11](#_Toc197273888)

[ANOVA Test 11](#_Toc197273889)

[Adjusted Rand Index 12](#_Toc197273890)

[Target Segment Identification 13](#_Toc197273891)

[Discussion 15](#_Toc197273892)

[Conclusion 15](#_Toc197273893)

[References 16](#_Toc197273894)

# Introduction

## Motivation

Over the last decade, Iceland has experienced a remarkable transformation in its tourism industry. Once considered a relatively isolated and marginal destination, Iceland has become a highly attractive location for internal visitors (Figure 1). As Johannesson (2010) discusses, tourism has played an increasingly important role in shaping Iceland’s economy, social structures, and development strategies. The country’s unique natural beauty, ranging from glaciers and volcanoes to geysers and waterfalls, has been a major draw for travellers seeking extraordinary landscapes and outdoor experiences. Moreover, Iceland’s reputation for safety, environmental consciousness, and cultural richness has further enhanced its global appeal. This rapid growth in tourism has brought significant economic benefits, contributing substantially to Iceland’s GPD (Figure 2) and creating employment opportunities across multiple sectors (Statista, 2025). However, it has also posed considerable challenges in terms of resource management, infrastructure development, and strategic planning (OECD, 2020). As the number of visitors continues to rise, it becomes increasingly important to develop a deeper understanding of tourist profiles to manage growth sustainably and maintain the quality of visitor experiences. One crucial step towards achieving this goal is the ability to segment tourists based on meaningful characteristics. Understanding who visits Iceland, for how long, and what economic background they represent allows tourists stakeholders to customise marketing strategies, optimise services, and implement public policies that better align with the need of different tourist groups. In this context, data analysis has become an indispensable tool for transforming raw information into actionable insights. Among the various analytical methods available, clustering techniques stand out for their ability to group observations based on patterns in the data without requiring predefined categories. Clustering offers a way to discover hidden structures in tourist data, providing new perspectives that are often not visible through traditional descriptive statistics (Dolnicar, 2002). Applying these techniques to study the behaviour of tourists visiting Iceland can yield valuable information for enhancing the effectiveness of tourism management strategies. “As a consequence, providers of tourism services are not wasting marketing dollars on market segments that are not interested in their offer and have the opportunity to develop a competitive advantage in the segments they target, Tourists who belong to these segments will be more satisfied and may therefore return to the tourism destinations or business and share their positive experiences with friends and family” (Dolnicar, 2013, pp 2).

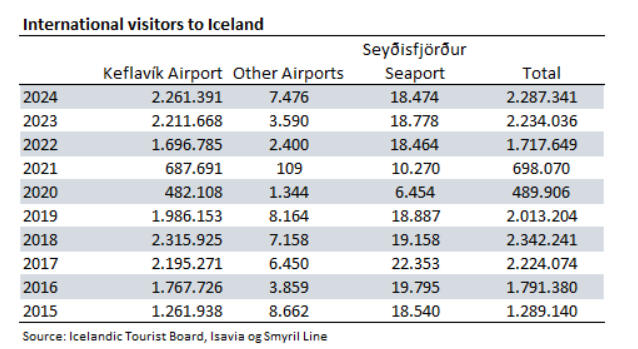


Figure International Visitors to Iceland from 2015 to 2024, Icelandic Tourist Board

## Research Problem

Despite the importance of understanding the tourist population in Iceland, current profiling practices remain limited. Most available data tend to focus on basic demographic indicators such as country of origin, gender, activities or general expenditure levels (Icelandic Tourist Board). However, these broad statistics fail to capture more elaborated aspects of visitor behaviour, such as how long different groups stay, how age influences patterns, or how income affects tourism activities. The absence of more detailed segmentation prevents tourism agencies and business from designing targeted marketing campaigns and developing personalised tourist experiences. Without deeper insights into visitor characteristics, public policy strategies also risk being less effective or missing opportunities to promote sustainable tourism practices. Therefore, there is a clear need to move beyond simple country or origin analysis towards more sophisticated methods of tourist segmentation. This project addresses this gap by applying unsupervised clustering algorithms to classify tourists based on three key variables: age, length of stay, and income.

A graph of blue bars

AI-generated content may be incorrect.

Figure Total contribution of travel and tourism to gross domestic product in Iceland from 2019 to 2023, Statista 2025

## Research Question

Given the above problem, the research question which will guide this research is: How can clustering techniques be applied to identify relevant tourist segments visiting Iceland, based on age, length of stay and average income? This question highlights the dual focus of the projects: both the methodological exploration of clustering techniques and the practical goal of generating meaningful tourist segments to support tourism development strategies.

## Research Hypothesis

Null Hypothesis (H0): There are no clear segmentation patterns among tourists based on the analysed variables.

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

Testing these hypotheses allows us to assess whether clustering is a valid and effective approach for analysing Iceland tourism data. The statistical analysis performed throughout the project will provide evidence to accept or reject the null hypothesis.

## Research Objectives

The main objective of this project is to explore the effectiveness of unsupervised machine learning techniques in segmenting international tourists visiting Iceland according to age, length of stay, and income. To achieve this overall goal, the following specific objectives are defined:

* Analyse and prepare a structured tourism dataset sourced from credible secondary data (Visitor Surveys, 2023) to ensure it is suitable for clustering analysis.
* Apply three different clustering algorithms, K-Means, Hierarchical clustering, and DBSCAN, to the data, comparing their outputs.
* Evaluate the performance and quality of each clustering method by using appropriate metrics and statistical analysis.
* Assess the stability of clustering outcomes through Adjusted Rand Index.
* Visualise and interpret the clusters generated, identifying the main characteristics that define each tourist segment.
* Discuss the implications of the findings for Iceland’s tourism strategies.

# Literature Review

The use of machine learning techniques in tourism research has grown significantly over the past years, offering new ways to analyse tourist behaviour, segment markets, and support strategic decision-making. Unsupervised learning techniques have gained attention due to their ability to uncover patterns without the need for labelled data. Tourism is a complex industry influenced by diverse demographic, behavioural and psychographic factors (Dolnicar, 2007). Traditional statistical methods have often fallen short in capturing this complexity. This has encouraged researchers to adopt machine learning models to improve decision-making in areas such as tourist flow prediction, experience personalisation, and visitor segmentation. According to The United Nations World Tourism Organization (World Tourism Organization and Saxion University of Applied Sciences, 2025, pp.19) “AI algorithms analyse vast amounts of data, including user preferences, past behaviours and trends, to provide personalized recommendations for destinations, accommodations, activities and attractions. This enhances the user experience and increase customer satisfaction”. According to Dolnicar (2002), traditional segmentation approaches in tourism often oversimplify tourism behaviours, highlighting the need for more data-driven and sophisticated techniques. Regarding the tourism in Iceland, some of the tourism- specific policy actions of Iceland are the “Improvement of data collection arrangements via the Icelandic Tourist Board, including the recently established Tourism Data Dashboard” and the “development of a digital toolbox to make it easier for tourism businesses to analyse their technology needs and find solutions” (OECD, 2020). Besides, Johannesson concluded that “the rapid growth in tourist arrivals poses serious challenges that past directions in terms of policy and practice have not dealt with in an adequate manner.” (Jóhannesson, G.T., Huijbens, E.H. and Sharpley, R, 2010)

## Clustering Techniques

Clustering is a type of unsupervised machine learning that groups data points into clusters based on similarity. It is particularly well-suited to tourism studies, where traveller behaviour can be segmented based on attributes such as demographics, spending habits, and travel motivations. Several studies have demonstrated the value of clustering for segmenting tourist markets. For instance, Dolnicar reviewed numerous data-driven segmentation studies and concluded that clustering techniques offer more meaningful tourist profiles than traditional demographic approaches (Dolnicar, 2002). Moreover, research by the OECD points out that advanced data analytics, including clustering, can improve tourism competitiveness by allowing to target and manage visitor flows more effectively (OECD, 2020).

The present study compares three clustering methods (K-Means, Hierarchical, and DBSCAN) to a tourism dataset from Iceland in the year 2023. While prior work has largely focused on larger-scaled datasets, this research seeks to explore how even small, structured dataset can yield valuable insights into tourist behaviour based on age, length of stay, and income. By testing different clustering methods and looking at the groups they create, this study adds to the discussion on how data can help improve tourism planning. The results can contribute to the design of more personalised services, marketing strategies, and sustainable managements practices.

To sum up, many studies have shown that clustering methods can help understand and organise tourist data better than traditional ways. Researchers now use more detailed data, like behaviour and preferences, instead of just age or income. This project uses three clustering methods to study a real dataset. Even though the dataset is small, it can still give useful information about tourist groups. The literature shows that machine learning, especially unsupervised methods, can support better planning and decision-making in the tourism industry.

# Methodology

## Project Aim

The aim of this project is to explore tourist behaviour in Iceland during 2023 by applying different clustering algorithms to a structured dataset. By identifying distinct groups based on age, length of stay, and income, the project seeks to provide insights that can support more targeted tourism strategies and service designs.

## Methodological Framework

This project was developed following the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. The process involved six stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment, although the last stage will not be included in this project. Each stage contributed crucially to the project’s overall aim of exploring tourist behaviour in Iceland during 2023, by identifying meaningful segments based on demographic and behavioural patterns.

### Business Understanding

The initial step focused on defining the project’s objectives and understanding the underlying business problem. Tourism authorities and stakeholders in Iceland increasingly need to customise their services and marketing efforts to diverse tourist profiles. Understanding how tourist differ in age, stay duration, and income can guide better-targeted strategies, enhance visitor satisfaction, and promote sustainable tourism development. The project’s primary goal was thus to uncover distinct tourist segments through unsupervised learning techniques.

### Data Understanding

After clarifying the business goals, the next step involved familiarising with the available data. The dataset used in this project was obtained from the Icelandic Tourist Board’s official website, <https://www.ferdamalastofa.is/en/recearch-and-statistics/visitor-surveys>, specifically from the Visitor Surveys section. This site offers structured data gathered through departure surveys conducted at Keflavik International Airport. These surveys are performed year-round among foreign tourists departing from Iceland and aim to monitor changes and developments in the tourism market over time. For this project, the Excel file from 2023 was used, as it represents the most recent available year at the time of analysis. The original Excel file contains a rich collection of data divided across several sheets, each focused on different aspects of tourist’s behaviour and demographics. The main sections include age distribution of visitors across various nationalities, income brackets reported by tourists, length of stay in Iceland and type of accommodation, group size, activities undertaken during the visit, travel planning details, when and how the trip was booked, means of transport used, visitor satisfaction and attitudes regarding services and experiences. The nationalities included in the dataset represent the countries with the highest number of visitors to Iceland in 2023, according to the survey results compiled by the Icelandic Tourist Board.

### Data Preparation

Once the data was gathered from the Icelandic Tourist Board’s visitor surveys of 2023, an exhaustive data preparation phase was conducted to ensure that the dataset was suitable for clustering analysis. The original Excel file contained multiple sheets, each focused on a specific aspect of tourist behaviour such as age distribution, nationality, income levels, length of stay, type of accommodation, activities, and travel group size. For this project, only the sheets corresponding to Age, Income, and Length of stay were selected. These were chosen because they are directly relevant to understanding travel behaviour and segmentation tourists for strategic marketing purposes, another important reason behind the selection was the small size of the dataset, with only 21 observations (one for each nationality). In clustering analysis, having many variables compared to a small number of data points can lead to problems such as overfitting, curse of dimensionality, and reduced interpretability. To avoid creating an overly large and sparse data frame that would compromise the robustness and reliability of the clustering results, the number of features was intentionally kept concise. By focusing on a limited set of highly relevant variables, the analysis maintained a balance between complexity and interpretability, ensuring that the clusters generated were statistically meaningful. In the original Excel sheets, the data was presented as percentages summing to approximately 100% for each nationality, allowing for direct comparison between different tourist groups. The preparation process followed several key steps:

* **Merging relevant data sheets:** the sheets corresponding to age, income, and length of stay were manually extracted and merged into a single data frame. Care was taken to ensure that the nationality information was consistently aligned across the different sheets, as slight differences in labelling sometimes occurred. Nationality was then set as the index of the resulting data frame to facilitate easier reference and manipulation.
* **Cleaning the data:** some columns present in the original sheets, such as ‘All’ or ‘Total’, reflected aggregate percentages for all groups combined. Since these did not offer distinctive information for clustering purposes and could bias the segmentation, they were removed from the dataset.
* **Normalizing:** the variables extracted from the original sheets were already expressed in percentage format. A verification step confirmed that, for each nationality, the percentages across each variables set approximately summed to 100% thus confirming the internal consistency of the dataset. No further scaling or normalization was necessary at this stage because all features were already on a comparable scale (0 to1). Besides, preprocessing data should be avoided when applying clustering (Dolnicar, 2002).
* **Handling missing values:** during the data inspection phase, it was confirmed that there were no missing values in the selected sheets for the nationalities chosen. This eliminates the need for imputation techniques and allowed for direct modelling without introducing artificial estimations.
* **Final data structure:** after preparation, the final dataset consisted of 21 observations (each representing a nationality) and 17 features, divided as follows:
  + 5 age-related variables (< 24 years, 25-34 years, 35-44 years, 45-54 years, >55 years)
  + 7 length-of-stay variables (Did not stay overnight, 1 night, 2-3 nights, 4-5 nights, 6-8 nights, 9-12 nights, >13 nights).
  + 5 income variables (low, low average, average, high average, high).

These preparation steps ensured that the dataset was clean, consistent, and ready for clustering analysis. Following the data preparation process, an exploratory data analysis (EDA) was conducted to better understand the relationship between the selected variables. A correlation heatmap was generated to visually inspect the variables (Figure 3). This analysis helped confirm the selection of relevant features for clustering. Additionally, interactive bar plots were developed to visualise the distribution of each variable across different nationalities (Figure 4).

A diagram of heatmap

AI-generated content may be incorrect.

Figure Correlation Heatmap of Analysed Variables

### A screenshot of a graph AI-generated content may be incorrect.

Figure Top 5 Nationalities with the highest Proportion of Visitors Age over 55 to Iceland in 2023

### Modelling

Three clustering algorithms were applied to uncover patterns in the data:

* **Hierarchical Clustering**: this method built a hierarchy of clusters using a bottom-up approach (agglomerative clustering). A dendrogram was created to visually identify possible numbers of clusters.
* **K-Means Clustering**: a popular and robust method for partitioning data into k non-overlapping clusters. K-Means assumes spherical clusters and minimizes intra-cluster variance.
* **DBSCAN** *(Density-Based Spatial Clustering of Applications with Noise*): unlike K-Means and Hierarchical clustering, DBSCAN does not require pre-specifying the number of clusters and is able to detect noise points.

For K-Means, it was essential to determine an optimal number of clusters. The number of clusters (k) was explored between 2 and 5. The rationales behind this range was based on the relatively small size of the dataset (21 observations), a high number of clusters would overfit the data and result in meaningless or extremely small groups; practical considerations, tourism segmentation strategies typically prefer a limited number of targetable groups for operational practicability; and the exploratory dendrogram from hierarchical clustering suggested 3-4 natural divisions (Figure 5). To evaluate the optimal k value, two interval validation indices were used: silhouette score, which measures how similar an object is to its own cluster compared to other clusters, values close to 1 indicate well-clustered data; and Davies-Bouldin index, which measures the average similarity between clusters, where lower values represent better partitioning. By systematically applying both metrics, it was found that four clusters offered a strong balance between high internal consistency and interpretability. The Davies-Bouldin index minimized at k=4, and the silhouette score was second highest for four clusters, supporting the selection. Finally, DBSCAN was applied without assuming a predefined number of groups, it was executed by manually tuning the eps (radius) and min\_samples (minimum points to form a cluster) parameters. The selection of the optimal parameters was based on visual inspection.

A graph with different colored lines

AI-generated content may be incorrect.

Figure Dendrogram using Ward Method

### Evaluation

The clustering outcomes were evaluated using a combination of internal validation metrics, heatmap visualisations, and statistical testing. Heatmaps allowed a visual interpretation of how different clusters varied across age, length of stay, and income. To statistically assess whether the clustering solutions reflected meaningful group differences, ANOVA tests were conducted for each variable, with the F-statistic and p-values interpreted at a 95% confidence level. Additionally, cluster stability was assessed by computing the Adjusted Rand Index across multiple runs. This multi-faceted evaluation ensured that both the interpretability and robustness of the clustering results were rigorously tested.

### Deployment

Due to the project does not include a real-world implementation, this stage will not be part of the project.

In conclusion, this methodological framework enabled a comprehensive and systematic exploration of tourist segmentation in Iceland. The project demonstrated how structured data analysis, coupled with robust machine learning techniques, can reveal actionable insights to support destination management and planning. Future work could extend the methodology by incorporating larger datasets, monthly or more recent data, and by applying alternative clustering validation techniques to further strengthen the findings.

## Architectural Diagram

The overall analytical process followed in this study is illustrated in the diagram below, which adhered to the CRISP-DM framework (Figure 6)

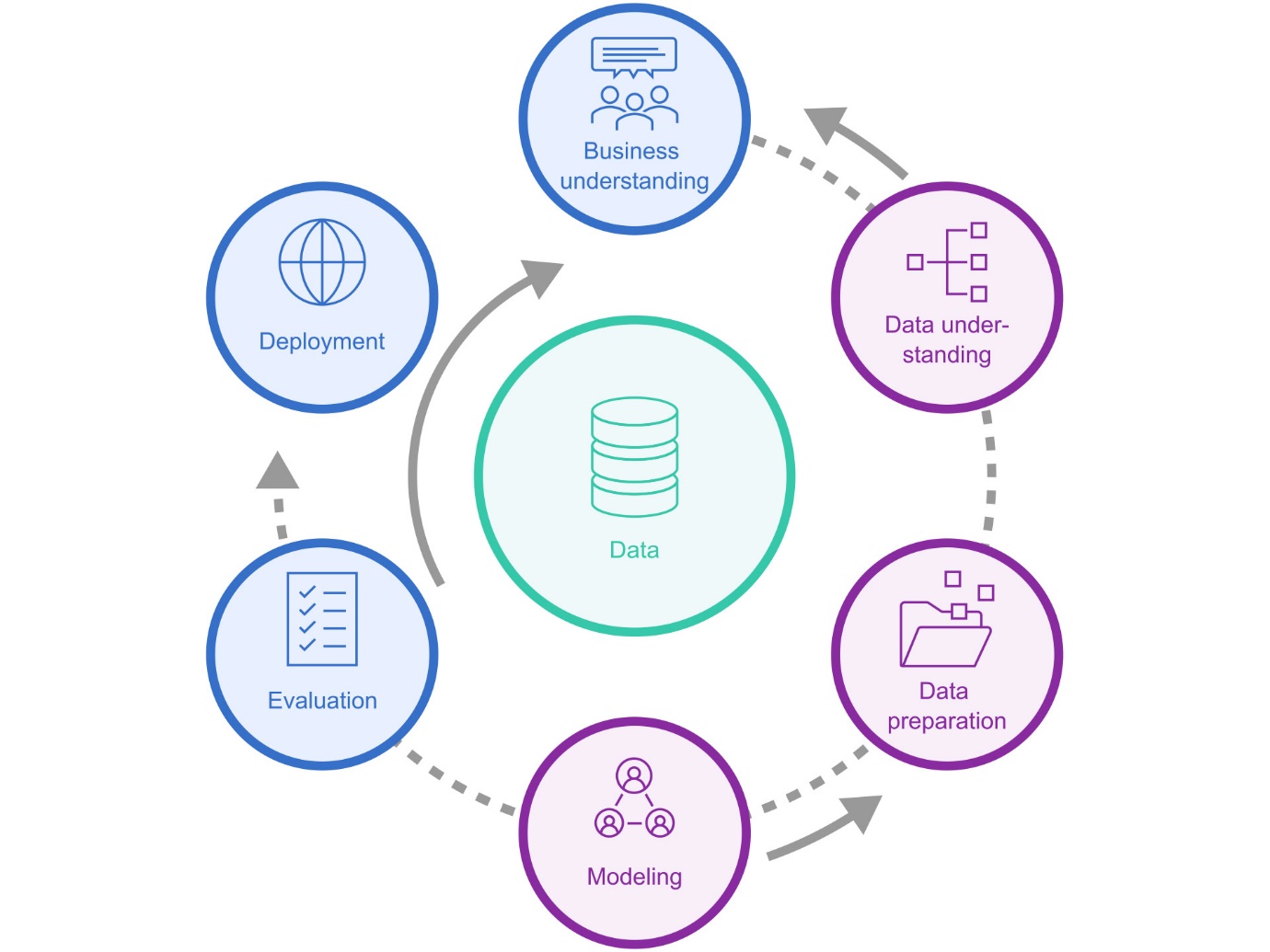


Figure Architectural Diagram

# Results and Discussion

Following the comparison of various clustering methods, K-Means clustering with four clusters was identified as the most suitable solution for segmenting tourists visiting Iceland in 2023. Heatmaps were generated to visually examine the distribution of demographic, behavioural, and economic variables across the clusters (Figure 7). The visual inspection indicated clear distinctions between clusters, particularly with respect to age groups and length of stay. Besides K-Means, hierarchical clustering and DBSCAN were also explored to determine the most suitable segmentation method. Hierarchical clustering was primarily employed to visualise the potential number of clusters through a dendrogram, which suggested a division into three groups. Initially, the heatmap revealed that hierarchical clustering did not segment the age categories effectively, while the income appeared nearly identical to the K-Means. Only the length of stay showed slightly more variation across clusters. DBSCAN was also tested as a density-based alternative, but it did not yield satisfactory results, the algorithm classified a large portion of the data as noise and only formed one cluster, indicating that the dataset’s characteristics were not ideal for density-based clustering. Even though, based on these visual impressions, K-Means appeared to segment the data more clearly and meaningfully, Hierarchical was also included in the analysis for further interpretation.

A screenshot of a graph

AI-generated content may be incorrect.

Figure Heatmap of Cluster Centres- K-Means

## ANOVA Test

To statistically verify the significance of differences across clusters in K-Means and Hierarchical, an ANOVA (Analysis of Variance) test was conducted. The results were interpreted at a 95% confidence level (α= 0.05), following common practice in the social sciences and applied data analysis.A comparison of the ANOVA results between the two methods (Table 1) revealed that Hierarchical clustering identified a greater number of variables with statistically significant differences across clusters. These included more age categories, and two income levels. However, K-Means also demonstrated strong significance in key behavioural indicators that were directly aligned with the study’s segmentation goals. It detected clearer differences in the <24 years age group and in visitors staying for more than 13 nights, distinctions that Hierarchical did not capture. Additionally, K-Means produced lower p-values in many of the shared significant variables, suggesting a more distinct separation between clusters. Table 1 shows the ANOVA p-values for both algorithms. Shaded cells highlight variables where significant differences were found between clusters (p< 0.05), which means they reject the null hypothesis.

Table ANOVA Results -Comparison of p-values- K-Means/Hierarchical

|  |  |  |
| --- | --- | --- |
| **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 |

## Adjusted Rand Index

“The stability of market segmentation solutions across repeated calculations is a key quality indicator if a segmentation solution” (Hajibaba, Grun, Dolnicar, 2019), for that reason the stability of the algorithms was evaluated using Adjusted Rand Index (ARI), which measures the consistency of cluster assignments across multiple runs. For both methods, a total of five runs were performed, as the data contains only 21 observations. For K-Means, the algorithm was executed using different random states, while for Hierarchical clustering, different data samples were used. The resulting ARI score were averaged to obtain a stability metric for each method. K-Means achieved a mean ARI of 0.61, indicating a moderate to high level of clustering consistency. In contrast, the Hierarchical model produced a much lower mean ARI of 0.011, suggesting that its clusters were highly sensitive to changes in the input data. These results further support the selection of K-Means as the more robust and reliable method for segmenting in this study. Additionally, the ARI was also computed for different numbers of clusters, ranging from 2 to 5. The results, presented in Table 2, show that K-Means achieves its highest stability with 4 clusters, while in Hierarchical clustering exhibited consistently low ARI values, indicating poor stability regardless of the number of clusters.

Table Comparison of average ARI results

|  |  |  |
| --- | --- | --- |
| **Algorithm/ Number of clusters** | **K-Means** | **Hierarchical** |
| 2 | 0.582 | -0.016 |
| 3 | 0.308 | 0.011 |
| 4 | 0.611 | 0.021 |
| 5 | 0.602 | 0.019 |

## Target Segment Identification

Each of the four clusters identifies through K-Means clustering represents a distinct tourist profile based on age and length of stay, as suggested by the heatmap of cluster centres (Figure 7) and confirmed by the ANOVA test results.

* **Cluster 0**: this group is mainly composed of tourists aged between 25 and 34 years old. Regarding the length of stay, these visitors tend to prefer medium-long length stays, especially between 6 to 8 nights. In terms of income, the majority belong to the average to high-average categories. This cluster can be described as young travellers with moderately strong purchasing power, opting for medium-long stays in Iceland.
* **Cluster 1**: tourists in this cluster are also largely from the 25-34 years age group. They show a stronger preference for 4-5 nights stays. Their income levels are mainly concentrated in the average and high- average categories. Therefore, this group represents mid-aged travellers with solid economic resources, often choosing medium length visits.
* **Cluster 2**: this group has a more balanced composition, with notable representation from both the <24 years and 25-34 years segments. Their stays are slightly longer, with a noticeable preference for 6 to 8 nights. In terms of income, most travellers belong to the average to high average groups. These are younger tourists, possibly students or early-career professionals, who tend to stay long period.
* **Cluster 3**: this cluster stans out clearly from the others, most of its members are aged 25-34 years, which is the highest concentration among all clusters. These tourists also show a significant preference for medium-long stays, especially 6-8 nights. Their income distribution is like other clusters, with a majority in the average to high average range. This group represents mid-aged tourists, relatively well-off tourists who prefer to spend more days in Iceland, likely seeking a deeper travel experience.

Building on the heatmap interpretation and ANOVA results, clusters 2 and 3 emerged as particularly relevant for deeper analysis. These segments represent young, economically active travellers with a tendency to stay longer than average, suggesting they may be especially valuable to Iceland’s tourism industry. According to Dolnicar (2007), the third foundation of data-driven segmentation highlights that even a single valid and managerially relevant segment can justify a segmentation solution. In this context, clusters 2 and 3 offer meaningful differentiation not only in demographic characteristics but also in behavioural aspects. To better visualise and communicate the distinguishing characteristics of cluster 2 and 3, bar charts were created highlighting their variables (Figure 7 and 8). These visualisations offer clearer, more intuitive representation of each segment’s profile, making the results more accessible to non-technical stakeholders and facilitating strategic decision-making. Although no cluster was predominantly composed of tourist over 55, the ANOVA results indicated significant differences in this age group across clusters. This suggest that older visitors, while not forming the majority in any specific segment, represent a distinct subgroup with unique characteristics that should not be disregarded. A similar pattern was observed for certain lengths of stay, which did not define any cluster but still showed statistically significant variation. These findings highlight that even variables which may appear less influential in visual representations can carry meaningful differences.

A graph of different colored bars

AI-generated content may be incorrect.

Figure Bar Charts of K-Means clustering results (Variables: 25-34 years old, 6-8 nights, High Average Income)

A graph of different colored bars

AI-generated content may be incorrect.

Figure Bar Charts of K-Means clustering results (Variables: < 24 years old, 6-8 nights, Average Income)

Overall, the clusters confirm that Iceland’s tourism in 2023 was dominates by younger travellers with the medium to high average income levels, and that medium-long length stays (4-8 nights) were the most popular. The selection of K-Means clustering over hierarchical clustering was well-supported by internal validation indices and the interpretability of the results. Although hierarchical clustering yielded reasonably distinct grouping, the segments derived from K-Means were clearer, more stable, and more actionable for practical tourism management purposes. This clarity enhances the ability of stakeholders to identify and target specific tourist profiles with tailored strategies.

## Discussion

The clustering analysis revealed meaningful tourist segments based on age and length of stay, highlighting that visitors to Iceland are not a homogeneous group. These patterns, confirmed by ANOVA testing, suggest opportunities for more targeted strategies by tourism authorities. However, the study’s small sample size limits generalisability. Future research with larger datasets and additional variables, such as spending behaviour or travel motivations, could enhance understanding. Moreover, as Dolnicar notes “the market changes all the time. Therefore, market segments change over time and have to be continuously monitored” (Dolnicar, 2013, pp.6). This study reflects only a snapshot of tourist’s profiles, based on the most recent publicly available data. Although just two years have passed since data collection, visitor preferences may have changed, underlining the importance of periodic re-segmentation to maintain relevance.

# Conclusion

Understanding the profiles of tourists visiting Iceland has become increasingly important due to the country’s rapid growth as a global destination. However, despite this importance, previous research and available data have often been limited to basic descriptive indicators such as country of origin, gender, or broad expenditure levels. This lack of deeper segmentation impedes the ability of tourism agencies, policymakers, and businesses to design more targeted marketing strategies, offer personalised services, and implement sustainable tourism practices. Addressing this gap, the main aim of this project was to apply unsupervised clustering techniques to develop a more detailed segmentation of visitors based on age, length of stay, and income. The solution proposed involved analysing a dataset obtained from surveys conducted at the main airport in Iceland in 2023, as collected by the official tourism website. Although the dataset had limitations, most notably a relatively small number of observations and a focus on only three variables, it provided a valuable opportunity to explore the potential of clustering methods in generating more insightful visitor profiles. K-Means clustering was selected as the final algorithm due to its strong performance in forming interpretable and statistically supported. The results confirmed the existence of meaningful segmentation patterns among tourists. Through the heatmap visualisation and ANOVA test, it was observed that there were statistically significant differences across clusters for various age groups and length of stay. However, no significant differences were found based on income levels. This partially supports the alternative hypothesis, as clear patterns were found in age and length of stay but not in income. Consequently, the null hypothesis was rejected for two out of three dimensions studies. This study contributes to the body of knowledge by demonstrating that even small-scale datasets can provide actionable insights when combined with appropriate unsupervised learning techniques. It shows that moving beyond simple demographic breakdowns allows for the identification of groups with different behaviours, which is critical for creating more tailored tourism strategies. While previous analyses often stopped at country-of-origin statistics, this project highlights the added value of clustering based on behavioural and socio-economic factors. Nevertheless, this research also acknowledges its limitations. The small sample size and the limited number of variables limit the generalisability of the results. Future research should aim to work with larger datasets, include a broader range of behavioural and psychographic variables, such as motivations for travel, activity preferences, satisfaction levels, and possibly compare different clustering techniques to validate and enhance the robustness of the findings. Moreover, other studies could explore how tourist profiles change over time, particularly as Iceland’s tourism evolves in response to global trends.

In conclusion, this project represents a step forward in refining the profiling of tourists in Iceland. It illustrates the potential of data driven methods to move past traditional analysis, offering deeper insights into visitor characteristics. By applying clustering techniques, tourism stakeholders can better target their offerings and policies, contributing not only to economic growth but also to a more sustainable and personalised visitor experience in Iceland’s unique tourism landscape.

# References

Dolnicar, S. (2002). A Review of Data-Driven Market Segmentation in Tourism. *Journal of Travel & Tourism Marketing*, 12(1), pp.1–22. doi: <https://doi.org/10.1300/j073v12n01_01>.

Dolnicar, S. (2008). *Market Segmentation in Tourism*. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/30387969_Market_Segmentation_in_Tourism>.

Dolnicar, S. (2013). Tourism Market Segmentation - A Step by Step Guide. pp.1–14.

Ferðamálastofa Icelandic Tourist Board. (n.d.). *Numbers of foreign visitors*. [online] Available at: <https://www.ferdamalastofa.is/en/recearch-and-statistics/numbers-of-foreign-visitors#overnight-visitors-all-entry-points>.

Field, A. (2009). *Discovering Statistics Using IBM SPSS Statistics*. 3rd ed. Los Angeles: Sage Publications.

Hajibaba, H., Grün, B. and Dolnicar, S. (2019). Improving the stability of market segmentation analysis. *International Journal of Contemporary Hospitality Management*, 32(4), pp.1393–1411. doi: <https://doi.org/10.1108/ijchm-02-2019-0137>.

Jóhannesson, G.T. and Huijbens, E.H. (2010). Tourism in times of crisis: exploring the discourse of tourism development in Iceland. *Current Issues in Tourism*, 13(5), pp.419–434. doi: <https://doi.org/10.1080/13683500.2010.491897>.

Jóhannesson, G.T., Huijbens, E.H. and Sharpley, R. (2010). Icelandic Tourism: Past Directions—Future Challenges. *Tourism Geographies*, 12(2), pp.278–301. doi: <https://doi.org/10.1080/14616680903493670>.

Lopez, A.M. (2025). *Iceland: travel & tourism’s GDP contribution 2012-2028*. [online] Statista. Available at: <https://www.statista.com/statistics/786578/travel-and-tourism-s-total-contribution-to-gdp-in-iceland/>.

O’Neill, A. (2024). *Iceland - employment 2015-2025| Statista*. [online] Statista. Available at: <https://www.statista.com/statistics/795247/employment-in-iceland/>.

OECD (2020). *OECD Tourism Trends and Policies 2020*. [online] Doi.org. Available at: <https://doi.org/10.1787/6b47b985-en>.

scikit-learn (2010). *2.3. Clustering — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/clustering.html>.

Szabo, B. (2020). *How to Create a Seaborn Correlation Heatmap in Python?* [online] Medium. Available at: <https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e>.

UNWTO (2025). *UN Tourism Startup Competitions*. [online] Unwto.org. Available at: https://www.unwto.org/innovation-digital-tranformation/ai-for-good-in-tourism [Accessed 26 Apr. 2025].

Visit Iceland (2023). *Visit Iceland*. [online] www.visiticeland.com. Available at: <https://www.visiticeland.com/>.

World Tourism Organization and Saxion University of Applied Sciences (2025). Artificial Intelligence Adoption in Tourism – Key Considerations for Sector Stakeholders. doi: <https://doi.org/10.18111/9789284426065>.