A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Francisca Andrea Argandona Alvarado |
| *Student Number* | 2024247 |
| *Module Title* |  |
| *Assessment Title* |  |
| *Assessment Due Date* |  |
| *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.

Contents

[Introduction 1](#_Toc158384946)

[Chapter 1 1](#_Toc158384947)

[Chapter 1.1 1](#_Toc158384948)

[Chapter 1.1.1. 1](#_Toc158384949)

[References 2](#_Toc158384950)

# Introduction

## Motivation

“Over the last decade, Iceland has experienced an unprecedented tourism boom, establishing itself as a highly attractive destination due to its unique natural beauty, culture, and safety”, as noted by Johannesson et al. (2010, p.278), “tourism in Iceland has evolved from a marginal activity to a central component of the national economy”. This growth in tourism has brought significant economic opportunities, but also important challenges related to strategic planning. Understanding the profiles of visitors travelling to Iceland non only allows for the improvement of tourism services but also supports the implementation of more effective public policies. In this context, data analysis has become an important tool for transforming information into applicable knowledge. Techniques such as clustering enable the segmentation without the need for pre-labelled data, helping to uncover hidden patterns. Applying these methods to the study of tourists visiting Iceland can offer new insights into their behaviours, thus contributing to more informed decision-making by industry stakeholders.

## Research Problem

Currently, there is limited profiling of tourist visiting Iceland. Available information often focuses on general statistics by country of origin, without deeper segmentation based on variables such as age, length of stay, and income. This lack of segmentation limits the ability of tourism institutions to develop targeted marketing campaigns and promote a more personalised experience.

## 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?

## 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.

## Research Objectives

The main objective of this project is to explore the effectiveness of unsupervised algorithms in segmenting tourists visiting Iceland according to 3 variables: age, length of stay, and income. To achieve this general aim, there are some other objectives such as, analyse and prepare a tourism dataset, apply and compare 3 clustering models, evaluate their performance, visualise and interpret the segments generated by each method to identify relevant tourist profiles, and discuss the implications of the findings.

# 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 economic factors. Traditional statistical methods have often fallen short in capturing this complexity. According to the UNWTO (2023), “big data and advanced analytics offer the opportunity to shift from reactive to proactive destination management”. 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.

## 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 (2022) reviewed numerous data-driven segmentation studies and concluded that clustering techniques offer more meaningful tourist profiles than traditional demographic approaches. Her work highlights the importance of including behavioural and psychographic data, such as travel motivations and preferences, in addition to basic variables like age or income.

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 help with making better decisions and creating more suitable services for visitors in Iceland.

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, which provides a robust and systematic framework for carrying out data mining tasks. The process involved six stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. 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, 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, a thorough 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. The specific variables selected for this project were focused on age, length of stay, and income. 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 both statistically meaningful and practically actionable. 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. Similarly, any non-numeric columns or headers were carried over from the Excel formatting were cleaned out to ensure the dataset was purely numerical.
* 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).
* 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. In addition, focusing on these key demographic and behavioural features positioned the project to generate actionable insights relevant for tourism strategy and segmentation.

### Modelling

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

* 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.
* 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.
* 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. 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.

### Evaluation

The clustering outcomes were assessed not just validation metrics, but also through heatmaps, where cluster centres were visualised to observe how different nationalities grouped across age, stay duration, and income variables; and ANOVA testing, to statistically verify if the formed clusters significantly differed across the variables, this involved examining the F-statistic and p-values for each variable between clusters. It was found that significant differences existed primarily in age-related variables and length of stay, while income showed little to no significant variance among clusters. Hierarchical clustering, explored primarily via dendrogram analysis, reinforced the selection of 3-4 cluster divisions. However, when conducting ANOVA on clusters derived from hierarchical clustering, the results showed fewer significant differences compared to K-Means clustering. Thus, K-Means was chosen for the final analysis. DBSCAN was tuned using manual adjustments of eps (radius) and min\_samples (minimum points to form a cluster). However, a portion of the dataset was consistently labelled as noise (indicated by cluster label -1). Given the compact nature of the dataset (small number of nationalities), DBSCAN’s performance was suboptimal compared to K-Means. It was concluded that DBSCAN was the least suitable for this context

### Deployment

Once the final clusters were established, an in-depth interpretation was conducted, clusters were profiles according to their demographic and behavioural features and strategic recommendations were derived, suggesting how Icelandic tourism authorities might customise services and marketing efforts based on these profiles.

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, unstructured data sources (like social media posts), and by applying alternative clustering validation techniques to further strengthen the findings.

## Architectural Diagram

# Results and Discussion

This section presents the outcomes of the clustering analysis and subsequent ANOVA tests conducted to validate the hypothesis. 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. The selection was grounded in the evaluation of internal validation indices: Davies-Bouldin index highlighted four clusters as the optimal choice, while silhouette score corroborated this by ranking it as the second-best option. Heatmaps were generated to visually examine the distribution of demographic, behavioural, and economic variables across the clusters. 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. However, this method was not pursued further for final clustering due to less distinct separation between groups when interpreted the heatmap. 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 a few small clusters, indicating that the dataset’s characteristics were not ideal for density-based clustering. Based on these findings, K-Means was selected as the final method for segmentation, offering better interval validation metrics, clearer and more interpretable clusters structures, and a more practical basis for informing targeted marketing and destination management strategies.

To statistically verify the significance of differences across clusters in K-Means, an ANOVA (Analysis of Variance) test was conducted. The results were interpreted at a 95% confidence level (α= 0.05).A comparison of the ANOVA results between K-Means and Hierarchical clustering further supported the selection of K-Means for this study. Although both methods revealed significant differences in key variables such as the 25-34 years and >55 years groups, as well as lengths of stay from 2 to 12 nights, K-Means provided stronger and more consistent statistical evidence. K-Means captured significant differences among the youngest visitors (< 24 years) and those staying for longer durations (> 13 nights), which hierarchical clustering did not. Furthermore, while hierarchical clustering detected some variation in income categories, the differences were weaker and less aligned with the study’s initial focus on age and behavioural patterns. In addition, the p-values obtained with K-Means were generally lower, indicating clearer separation between clusters.

The outcomes in K-Means are summarised as follows:

* Age group
  + < 24 years: significant difference among clusters (p= 0.0102).
  + 25-34 years: highly significant difference (p= 0.0002).
  + 35-44 years and 45-54 years: no significant differences observed (p> 0.05).
  + > 55 years: significant difference (p= 0.0190).
* Length of stay
  + 2-3 nights: significant difference (p= 0.0011).
  + 4-5 nights: very strong significance (p< 0.0001)
  + 6-8 nights and 9-12 nights: significant differences (p= 0.0016 and p= 0.0002).
  + >13 nights: significant difference (p= 0.0368)
* Income
  + Low, low average, average, high average and high: no significant differences were observed (p> 0.05).

The findings of this study offer valuable insights into the segmentation of tourists visiting Iceland. By applying K-Means clustering, four distinct tourist profiles were identified, differing notably in demographic characteristics and patterns of stay. The significant differences observed among the younger age group (<24 years) and the 25-34 age category underline age as key differentiator among tourist segments. This aligns with prior studies (Dolnicar, 2022) that underscore the critical role of demographic variables in market segmentation. Younger tourists typically display different travel motivations, preferences, and expenditure behaviours compared to the older people. The significant variation observed in the > 55 years age category further emphasis the need for age-specific tourism strategies. Older visitors may prioritise comfort, accessibility, and cultural experiences, while younger tourists might favour adventure, affordability, and social opportunities. In terms of the length of stay, clear distinctions among clusters were apparent. Significant differences in stay 2-3 nights, 4-5 nights, 6-8 nights, and 9-12 nights suggest the presence of short-term visitors and those opting for medium to extend stays. Destination managers can leverage this knowledge to customise services, accommodation options, and marketing strategies to the needs of different segments. The lack of significant differences for ‘Did not stay overnight’ and 1-night stays implies a more homogeneous distribution for very short-term visitors, possibly due to stopover travellers. Unexpectedly, the analysis revealed no significant differences in income across clusters. Given Iceland’s high living costs and the limited range of budget options, it is plausible that tourists, regardless of demographics background, tend to exhibit similar spending behaviours. Future studies could integrate more granular economic variables, such as types of expenditures (e.g. accommodation, dining, activities), to assess whether differences emerge at a more detailed level.

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 and confirmed by the ANOVA test results.

Cluster 0: this group is mainly composed of tourists aged between 25 and 34 years old, with a relatively even presence across other age groups as well. Regarding the length of stay, these visitors tend to prefer medium length stays, especially between 4 to 5 nights and 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 stays in Iceland.

Cluster 1: tourists in this cluster are also largely from the 25-34 years age group, although there is a slightly broader age distribution compared to cluster 0. They show a stronger preference for 4-5 nights stays and moderate representation in 2-3 nights as well. 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 6-8 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 longer period compared to other clusters.

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 longer stays, especially 6-8 nights and 9-12 nights. Their income distribution is like other clusters, with a majority in the average to high average range. This group represents younger, relatively well-off tourists who prefer to spend more days in Iceland, likely seeking a deeper travel experience.

Overall, the clusters confirm that Iceland’s tourism in 2023 was dominates by younger travellers (mainly 25-34 years old) with the medium to high average income levels, and that medium-length stays (4-8 nights) were the most popular. These findings align with the general patterns observed during the analysis and support the development of tourism strategies targeting young adults with moderate to strong spending power, particularly for trips lasting around one week.

## Hypothesis Testing Conclusion

Based on the ANOVA results, the null hypothesis can be rejected for several variables, particularly age groups like 25-34 years and length of stay categories such as 4-5 nights and 6-8 nights, as their p-values were below the 0.05 significance threshold. This indicates that there are significant differences between at least two clusters for these variables. However, for other variables such as < 24 years, did not stay overnight, and the income categories, the p-values were greater than 0.05, meaning that for these variables, the null hypothesis could not be rejected. Overall, the clustering identifies meaningful differences among tourist groups, especially regarding age and length of stay, while income was less influential in distinguishing between clustering.

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 and more actionable for practical tourism management purposes.

Chapter 5: Conclusion (brief about research problem, aim, solution, research, contribution to the body of knowledge, and future work).

# References