
# Market Segmentation in Tourism

## An analysis of Australian data with model-based clustering methods

Bachelor's Thesis in Economics

Lucas Paul Unterweger

Spring 2022

Supervision by

Malsiner-Walli, Gertraud, Mag.Dr., M.Stat.
Knaus, Peter, MSc (WU)

# Abstract

The Covid19 pandemic has shown how dependent the tourism sector is on unrestricted mobility. Comprehensive travel restrictions, nationwide lockdowns, and general uncertainty about the future development of those measures paralyzed a crucial branch of every modern economy. This thesis aims to apply model-based clustering methods on a dataset containing survey data on tourist activities in Australia. Using the extracted clusters, insights from a meaningful market segmentation will be retrieved to find possible hidden structures in the tourism sector. It can be seen that there is indeed a market structure containing six market segments present in the data, each with its own key characteristics and features. These insights can be used to efficiently target consumer segments in the further course.

# Contents

# 1 Introduction

Hans Christian Andersen once wrote in *The Fairy Tale of My Life* that "to travel is to live", which beautifully captures the importance that travel has for life and an open mind. Since the beginning of civilization, humanity has longed to see every inch of this beautiful planet, and beyond, and consistently came up with new ideas on how to reach places that where hidden to us until that point. For centuries, cultures from all over the world managed to overcome huge distances to find new land, trade with foreign countries and to enlarge ones own sphere of influence. Nevertheless, travelling in a conventional, more modern sense as a form of pleasure and vacation has long been out of reach for the majority of the world's population. However, things changed in the 17th century, when wealthy individuals of the European social and economic nobility began to use long distance travel as a form of self development, commonly referred to as *"The Grand Tour"* (Towner, 1985). Travelling has become more than a proxy for expanding ones political and economic power. Cultural, educational and social reasons for overcoming huge distances emerged and shaped the course of the centuries to come. Be that as it may, long distance travel remained an experience that was mostly reserved for wealthy individuals, until the industrialisation accompanied by advancements in the transportation sector in the later 18th century opened the world's gates to a larger part of the world's population (Towner, 1985).

The *Atlas of the Historical Geography of the United States* by Paullin (1932) is a remarkable example of how innovation and technological advancements shaped the relative proximity of two places over time. In 1800, it took more than 6 weeks to get from New York to the west side of Lake Michigan, whereas it took only about three days in 1857 and just 2 hours and roughly 40 minutes today using an aircraft from New York to Chicago (NETSTEL Software, 2022). Similar cases can be found throughout history, which brings us to today. Technological progress and recent - in historical terms - economic growth created a sector which is now a crucial part of every modern economy: the tourism sector.

A vibrant and versatile tourism sector has established itself as a core part of many developed and developing countries. The *World Tourism Organization (UNWTO)* - a specialized agency of the United Nations - emphasizes this by pointing out that "modern tourism is closely linked to development and encompasses a growing number of new destinations. These dynamics have turned tourism into a key driver for socio-economic progress" (World Tourism Organization, 2022). Looking at data from the World Bank[1], one can see that the expenditures of international tourists accounted for 6% of total imports on average in 2019 with countries like Australia (14%), Argentina (14.8%) or the Russian Federation (11.5%) acquiring more than one in every ten dollars of their imports from international tourists (The World Bank Group, 2022a). In absolute terms, this accounted for approximately 1.439 Trillion USD globally in expenditures in 2019 (The World Bank Group, 2022b). The International Labour Organization estimates that, in 2016, the tourism sector roughly generated 108 million jobs, hence 3.6% of total global employment. That figure rises to 292 million jobs if one includes jobs that indirectly support the core elements of the tourism industry (International Labour Organization (ILO), 2017).

Having these key figures in mind, one can clearly see the significant macroeconomic influence a malfunctioning tourism branch could have on an economy, while the Covid-19 pandemic demonstrated how dependent the tourism sector is on unrestricted mobility. Comprehensive travel restrictions, nationwide lockdowns, and general uncertainty about the future development of those measures paralyzed a crucial branch of a functioning economic system. Concurrent to ongoing vaccination initiatives and the rollback of extensive state interventions, companies were and still are trying to restart their businesses to get the economy up and running once again.

With the precarious global situation still in mind, businesses located in

[1]Note: To avoid distortion from the temporary impact of the Covid-19 pandemic, I chose to have a look at data prior to Spring 2019.

popular travel destinations must come up with new concepts to efficiently return to the pre-pandemic status. A first step forward might be the thorough analysis of the market a company, public institution or government operates in and several quantitative and qualitative tools have emerged to perform such an analysis. A common practice to achieve said goal is *market segmentation*, which - according to Dolnicar et al. (2018) - "offers an opportunity to think and rethink" and in turn "leads to critical new insights and perspectives." A well-performed analysis of the structure of the tourism market and a profound knowledge of touristic habits can help re-engage consumers in a more efficient and profound way and also lead to a "competitive advantage in the selected target segement(s)" (Dolnicar et al., 2018). To demonstrate the advantages and possible insights of market segmentation, this thesis aims to apply a suitable method for market segmentation on a data set consisting of activities undertaken by tourists in the *Land down under*: Australia.

Similar to other advanced economies, Australia is home to a strong and dynamic tourism sector. According to the *Tourism Satellite Account* published yearly by the *Australian Bureau of Statistics*, the tourism industry contributed 60.8 billion USD to Australia's economy which roughly translates to 3.1% of national GDP in 2018-19. Additionally, the *Australian Bureau of Statistics* estimates that about 660 000 people were employed by a tourism-related enterprise in 2018-19, accounting for 5.2% of the total Australian workforce. (Tourism Research Australia, 2019) Furthermore, the UNWTO estimates that roughly 9.5 million people arrived in Australia in 2019 (excluding nationals residing abroad and crew members) and the recent trend has shown a steep increase on a year on year basis from roughly 4 million in 1995. (See figure 1)

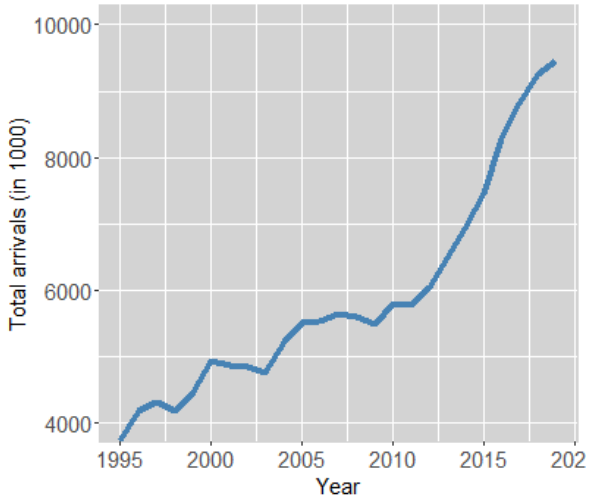


Figure 1: Total tourist arrivals (in 1000)

The aim of this thesis now is to apply model-based clustering methods on a dataset gathered by Katie Cliff (nèe Lazarevski) containing activities undertaken by Australian domestic tourists. An introductory market segmentation analysis using this dataset and distance-based methods has been conducted by Dolnicar et al. (2018). First, the available literature and recent publications will be reviewed in chapter 2 with a special focus on market segmentation and the current methods in model-based clustering. Afterwards, the dataset will be analysed on a descriptive level in chapter 3 to gain key insights that will be used to derive an empirical framework in chapter 4. Subsequently, the empirical analysis - i.e. the extraction of the cluster segments - will be conducted in chapter 5. The thesis will close by presenting the results and describing the segments in chapter 6. All in all, meaningful market segments should be extracted and profiled to gain valuable insights for agents in the Australian tourism sector. A conclusion will sum up the gained results and provide an outlook on the future.

# 2 Literature Review

It is evident that the importance of the tourism sector and henceforth the thorough analysis of it stretches far beyond pure academic curiosity. Both the private, public and academic sector gain from key insights on market structure, price developments and macroeconomic trends. The increase in data availability and data collection in recent decades has opened doors for a widespread analysis of different consumer and producer markets. Hence, it comes as no surprise that both academics and corporate researchers have published several articles, journal entries and books on market segmentation and methods for performing it.

A general intuition for the incentive behind market segmentation can be found in Goryushkina et al. (2019), where the authors state that "each service consumer has their own habits, tastes, financial opportunities and desires". Furthermore, it is apparent that the supply side of tourism (or more generally any) market has to "meet expectations of a great number of [often heterogeneous] consumers" and that that is the "key reason for a tourist company's necessity to identify a group of consumers with similar characteristics" (Goryushkina et al., 2019). Those characteristics - according to Goryushkina et al. (2019) - can roughly separated in five larger categories: demographical, geographical, socio-economic, psycho-graphic, behavioural or any combination of the above.

There are several entry points into the field of market segmentation. A comprehensive summary of a wide range of topics can be found in Dolnicar et al. (2018). This introductory textbook includes topics like a general introduction into the relevance of and the incentive behind market segmentation, followed by a ten step procedure on how to apply the theoretical tools in real world scenarios. (For a schematic display of this structure, see figure 2.) The concepts are accompanied by several coding examples using the *statistical computing language* R (R Core Team, 2018) and a complete case study at the end of the book. More generally, Dolnicar published several papers on

market segmentation with a special focus on tourism. A more recent one can be found in Volume 75(1) of the *Tourism Review* (Dolnicar, 2020) where she summarizes the history of the field in the past 75 years and gives an outlook on the next 75. She also published chapters in books like the *Handbook of e-Tourism*, where she covers the issue of carrying market segmentation over into the 21st century (Dolnicar, 2021). She argues that several studies "discuss the aspects relating to e-tourism but are based on information provided by tourists in surveys" and "that there has not been a substantial uptake of e-data in the tourism market segmentation" (Dolnicar, 2021).

Clear and well-defined markets are a rarity in economics, although of high interest for not just industrial and competition economists, which it might not be as trivial as it seems to clearly define the tourism sector. An extensive introduction into the tourism sector as a whole is given by the book *Travel Marketing, Tourism Economics and the Airline Product* by Camilleri (2018). There he - in reference United Nations Conference on International Travel and Tourism - defines tourists as "temporary visitors staying at least 24h in a destination", however this excludes domestic tourism (tourism within a country), which is why he proceeds by instead using the definition proposed by the *Tourism society*. They describe "tourism [as] the temporary short-term movement of people to destinations outside the places where they normally live and work." Camilleri (2018)

On a methodological level, cluster analysis has proven itself to be an efficient method for the market segmentation of cross-sectional data (Dolnicar et al., 2018). Cluster analysis itself was first introduced by Harold E. Driver and Alfred L. Kroeber in 1932 in their book "Quantitative Expression of Cultural Relationships" and hence - more or less - originated from the field of anthropology. Yet cluster analysis has been applied to several scientific fields so far, ranging from image detection in medicine to network analysis in computer science. As we will see later in chapter 4, which focuses on the empirical framework of this thesis, there no single "go to" clustering method that fits to every dataset. A good starting point to find the optimal methodology is by narrowing the choices down to suitable subgroups of methods.

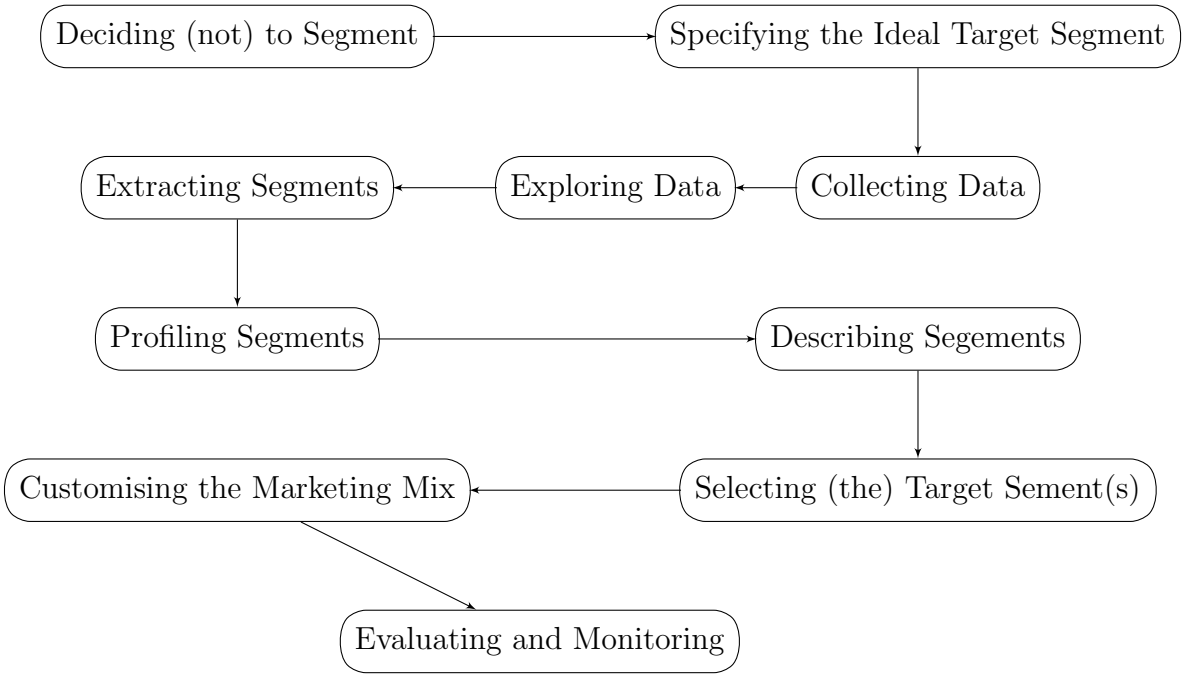


Figure 2: A schematic illustration of the proposed 10 step procedure by Dolnicar et al. (2018)

A possible way to distinguish between subgroups of clustering algorithms is by differentiating between distance-based and model-based clustering methods. Kamthania et al. (2018), for example, apply a distance-based approach - specifically k-mode clustering - and propose a framework for business intelligence (BI) applications. A comprehensive textbook on model-based approaches in clustering and classification is given with Bouveyron et al. (2019). The authors provide an extensive overview on finite mixture models in combination with extensions for e.g. discrete data and different non-Gaussian distributions. (It is a beautiful book with a lot of coding examples and visualizations.) Another interesting report was published by Preud'homme et al. (2021), in which the authors provide a "head-to-head comparison" of clustering methods. All in all, they compare four model-based approaches, including Latent-class models, and five distance-based approaches like k-means and hierarchical clustering. The authors conclude that "model-based tools (p.e.

the Kamila and LCM packages implemented in R) usually perform better than distance-based tools (except K-prototypes packages implemented in R) in the setting of heterogeneous data" (Preud'homme et al., 2021).

Having gathered an overview on available literature, it is time to return to the purpose of this thesis: applying model-based clustering methods on the aforementioned dataset to extract and profile meaningful clusters/market segments. As we already have a dataset at hand and in accordance to the 10-step procedure proposed by Dolnicar et al. (2018), it is necessary to explore the data before actually extracting segments. Hence, a comprehensive descriptive data analysis will now be performed.

# 3 Descriptive Analysis

Dolnicar et al. (2018) repeatedly point out in their book that "data-driven market segmentation is exploratory by nature." There exists no such thing as the ideal procedure to identify consumer segments of any kind, as the choice of the used statistical tools is dependent on the structure of the data itself, how it has been gathered, what it tries describe and several additional factors. That is is why a descriptive analysis of the data at hand might be in order before setting up a theoretical framework in chapter 4. The dataset which will be used for this thesis was collected by Katie Cliff (nèe Lazarevski) in 2007 for the purpose of her PhD thesis and was funded by the Australian Research Council under the grant *DP0557769* (Dolnicar et al., 2018). It contains 1003 observations of 45 variables containing data on activities of adult Australian residents. All 45 variables, let us denote them as $A_i$ with $i \in \{1, ..., 45\}$ and $A_i$ being the $i$-th column of the data set, have been coded as binary variables, meaning that

$$A_i = \begin{cases} 1 & \text{if the person has undertaken the activity} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A quick overview on the dataset can be found in table 1 below.

| **Category** | **Description** |
|---|---|
| Description: | Vacation activities |
| Number of observations: | 1003 |
| Number of variables: | 45 |
| Type of variables: | Binary, every oberservation $\in 0, 1$ |
| Collected by: | Katie Cliff (nèe Lazarevski) |
| Collected in: | 2007 |

Table 1: Summary of the dataset

| **Undertaken activities** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Bushwalk | Beach | Farm | Whale | Gardens | Camping | Swimming | |
| 0 | 0 | 1 | 0 | 0 | 1 | 0 | ... |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 0 | 1 | 0 | 1 | 1 | 0 | 1 | ... |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | ... |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | ... |
| 0 | 1 | 0 | 0 | 1 | 0 | 1 | ... |
| 1 | 0 | 1 | 0 | 1 | 0 | 1 | ... |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | ... |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |

Table 2: The first ten rows and seven columns of the dataset

Before thinking about suitable tools for a descriptive analysis, it might be useful having a look at a sample of the dataset, which is provided above in table 2. (Note: Only seven variables and ten observations will be shown to improve readability. The complete data set can be found in Dolnicar et al. (2018).) The first key characteristic to notice is that it is a purely binary dataset, which has a crucial implication on the choice of model in the further course, yet this will be covered in Chapters 4 and 5. Be that as it may, it also reduces the explanatory power of general descriptive analysis tools like boxplots and histograms.

As previously mentioned, the columns of the dataset refer to specific activities, whereas the individual rows belong to the given answers of a person. The set of activities questioned in the survey includes a wide range of ventures from, outdoor sports like swimming to indoor activities like visiting a museum. (A complete list of the included activities can be found Dolnicar et al. (2018, page 304).) To start, one could have a look at the average number of activities per person as well as its distribution. Having a look at

figure 3, we can see that the average number of activities is 13.85 per person with 50% of all observations being between 9 and 17 activities. It can easily be seen that (1) the distribution is slightly right-skewed and (2) that some observation represent the extreme cases of either 0 total activities (the person has not undertaken any activity) or 45 activities (the person has undertaken any possible category). These results might mislead to a hasty conclusion that a person which did not participate in any activity cannot contribute to the following segmentation analysis. However, it is crucial to keep in mind that a cluster of "non-active" people might still be a cluster.

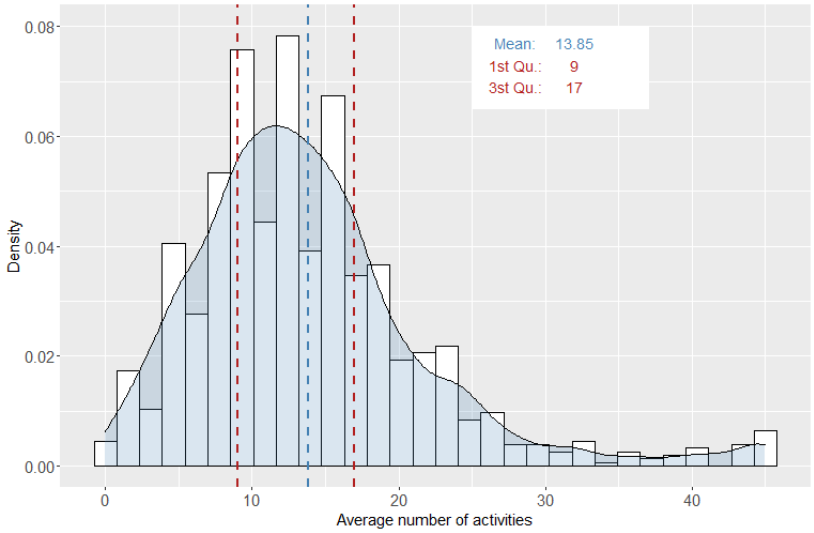


Figure 3: Distribution of average number of activities per person

Having analysed the dataset on an observational basis, it might be useful to have a look at the dataset from an activity-based point of view, hence looking at each individual activity and their occurrences throughout all observations of the sample. To do that, I computed the column sums of the dataset and sorted them in a descending order. In figure 4 one can the see the most common activities throughout the sample. "Relaxing" (which also includes simply doing nothing), "Eating", "Sightseeing" and "Shopping" top the list with 810, 806, 790 and 763 occurrences respectively, which aligns with intuition, as they capture the core of most journeys. Furthermore, activities like going to various kinds of markets ("Markets"), visiting friends and family

("Friends") or going to pubs ("Pubs") complete the top 10 (Dolnicar et al., 2018). It is immediately apparent that no activity was performed by every single person, which directly follows from the fact that some observations have an activity count of zero (See figure 3 above). Would such an activity exist, one could argue that said variable should be excluded as it would fall short of playing a role in distinguishing between different consumer segments and hence the following segmentation analysis.

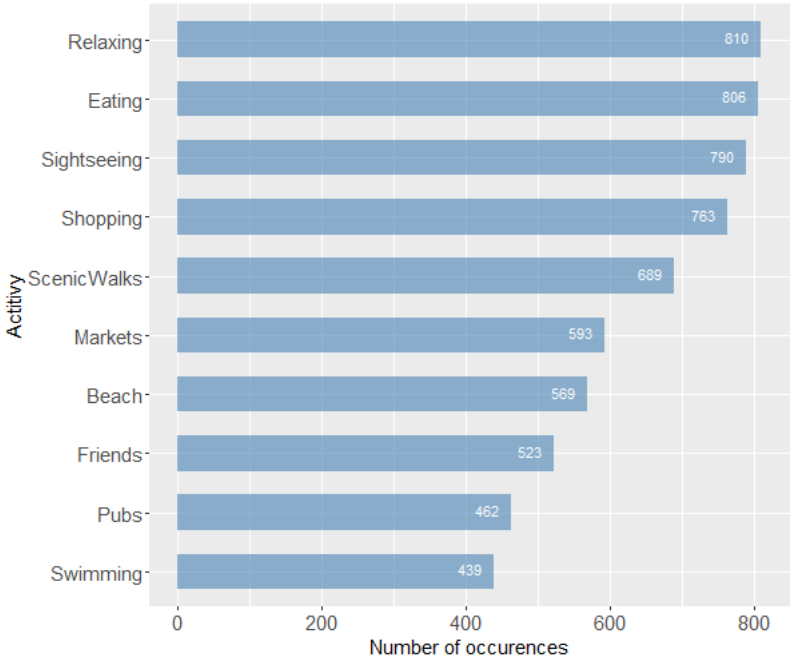


Figure 4: Most common activities

Nevertheless, it might also be helpful to have a look at the flip side of the table, i.e. the least common activities. In figure 5 we can see the top 10 least common activities in the dataset. "Skiing" tops this ranking - much to the regret of my skiing-loving Austrian mind - with just 42 people having undertaken this activity, which also make sense due to Australia's geographical situation and climate. The following categories are "Adventure" (this includes adventurous activities like bungee jumping, rafting, etc.), "Riding" and "Surfing". "Cycling", "Scuba Diving" and "Tennis" are pretty self ex-

planatory. The categories "Exercising" (which refers to exercise/gym/swimming at a local pool or river and functions as a "miscellaneous" category for exercising) as well as "Golf" and "Water Sports" complete the list of the 10 least undertaken activities. Analogous to the example before, an activity that has not been undertaken by any person might lack explanatory power and hence could have been excluded from the analysis. As this is apparently not the case, we can proceed without hesitation.

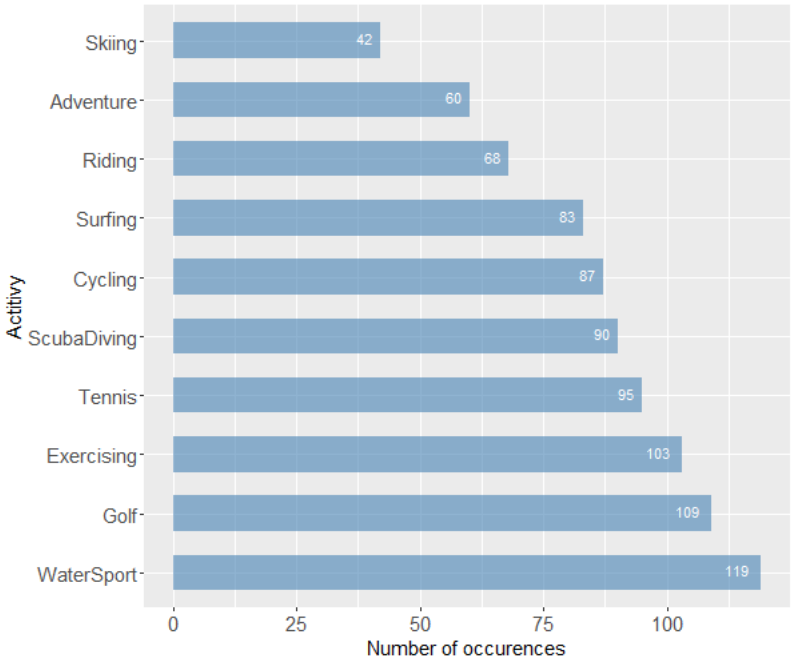


Figure 5: Least common activities

As mentioned above, the fact that it is a purely binary dataset restricts the descriptive analysis to very few statistical tools, however one last information might be useful for the further analysis. Two variables that are exactly alike could be reduced to one single variable, as adding the other variables would not add any valuable information to the analysis. In other words, highly correlated variables are interesting candidates for exclusion or aggregation. The fact that "using correlated segmentation variables makes it difficult for the segmentation algorithm to find the best solution" is also stressed by Dolnicar (2020). Hence, it is advisable to "explore [ones] data

before extracting market segments to ensure you have the smallest possible set of nonredundant segmentation variables" (Dolnicar, 2020). Several measures of correlation exist, such as the *Pearson correlation coefficient* or *Spearman's rank correlation coefficient*. However, these metrics are mainly constructed for continuous data, which is not present in the given dataset. A correlation metric for binary data is the *polychoric* (or in the case of a binary variable *tetrachoric*) correlation, which models the correlation between two categorical variables that are assumed to be normally distributed continuous latent variables in reality (Drasgow, 2014).

**Hamming Distances**

| Variable 1 | Variable 2 | Hamming Distance |
|---|---|---|
| Skiing | Riding | 50 |
| Skiing | Adventure | 50 |
| Riding | Cycling | 69 |
| Riding | Adventure | 70 |
| Skiing | Surfing | 71 |
| Skiing | Cycling | 73 |
| Riding | Surfing | 75 |
| Skiing | Tennis | 77 |
| Skiing | ScubaDiving | 78 |
| ScubaDiving | Adventure | 80 |

Table 3: The ten variable pairs with the lowest *Hamming distance*

Nonetheless, as we are talking about activities that the survey respondents either have done or not, this seems similarly unsuitable. A valid measure or likeness could be the so-called *Hamming distance* (a detailed definition can be found in chapter 4). To quickly summarize its main idea, two vectors

of length $n$ have a Hamming distance of $n$ if they have different values at every position of the vector, and a Hamming distance of 0 if they are identical. A list of the ten variable combinations with the lowest Hamming distances can be found in table 3. Keeping in mind that the dataset consists of 1003 variables, a value of 50 (the Hamming distance of *Skiing* and either *Riding* or *Adventure*) might be an indicator that those variables could be aggregated to one variable, however as we have seen in figure 5, these Hamming distance are present due to the fact that these variables are very rare and hence have a lot of zeros. Based on this, One could argue that one of those variables might be a strong indicator for a e.g. *Riding* market segment, which is why I will refrain from aggregating some variables.

# 4 Empirical Framework

Having seen some key characteristics of the dataset at hand, it is now time to finally take a step forward and take a look a it from an empirical perspective. As mentioned in chapter 2, cluster analysis has been applied to various different problems in scientific fields like psychology, biology and economic issues like market segmentation. Diverse as its areas of application is the tool set one has at hand when aiming to perform cluster analysis of any kind. Throughout the last century, the existing statistical tools have improved rapidly and new, never before seen methods have emerged.

Hence, it comes as no surprise that that an a priori categorization of available clustering methods turns out to be a non-trivial endeavour. One reason for this diversity might be the fact that the term "cluster" is not precisely defined. According to Estivill-Castro (2002), the definition often depends on the "philosophical points on the matter" and he refers to Aldenderfer and Blashfield (1984) who state that the "top-down view regards clustering as the segmentation of a heterogeneous population into a number of more homogeneous subgroups", as well as Duda et al. (1973) who clarify that the "bottom-up view defines clustering as finding groups in a data set by some natural criterion of similarity". He summarizes his paper with 11 recommendations, the probably most interesting one being: "Do not forget that clusters are, in a large part, on the eye of the beholder" (Estivill-Castro, 2002). Nevertheless, there happen to be various publications, books and introductory papers out there which have proven themselves to be invaluable companions in trying to get a pragmatic overview on landscape of clustering methods. The first noteworthy book - and this is no surprise, as it has been mentioned multiple times so far - is *Market Segmentation Analysis* by Dolnicar et al. (2018). The authors broadly categorized some of the most common clustering methods in two main groups: *distance-based* and *model-based methods* (Dolnicar et al., 2018).

## 4.1 Distance-based methods

The beauty of distance-based methods lies in their intuitive and straightforward nature. Essentially, as Dolnicar et al. (2018) points out, it all comes down to some sort of "notion of similarity or dissimilarity [between two observations], mathematically speaking: a distance measure." This measure of (dis-)similarity between two points $A = (a_1, a_2, ..., a_n) \in \mathbb{R}^n$ and $B = (B_1, b_2, ..., b_n) \in \mathbb{R}^n$ can be denoted as $d(A, B)$. According to Dolnicar et al. (2018), some properties must hold when handling such metrics, which are *symmetry*

$$d(A, B) = d(B, A),$$

and the *identity of indiscernibles:*

$$d(A, A) = 0.$$

It does not take much time to realize that there cannot be a universal distance measure, as the applicability of such a metric is heavily dependent on the structure of the observed data. This is why several distance measures and related concepts have been introduced so far. A widespread metric is the so called *Minkowski distance* - named after the German mathematician Hermann Minkowski (Deza & Deza, 2009). Using the previously, arbitrarily defined points $A$ and $B$ and some $p \in \mathbb{Z}$, the *Minkovski distance of order p* is defined as

$$d(A, B) = \left( \sum_{i=1}^{n} |a_i - b_i|^p \right)^{1/p}. \tag{2}$$

However, this metric is more commonly known under the names *Euclidean distance* or *Manhattan distance* which correspond to the cases $p = 2$ and $p = 1$ respectively. To better understand the intuition behind this approach, it might be useful to have a look at an example of those two distance metrics in the case of $n = 2$. Let us assume we have two data points of the $\mathbb{R}^2$, or more specifically $X = (x_1, x_2)$ and $Y = (y_1, y_2)$. Each data point represents a point in $\mathbb{R}^2$. Fortunately, it is easy to display these points visually; see

figure 6.

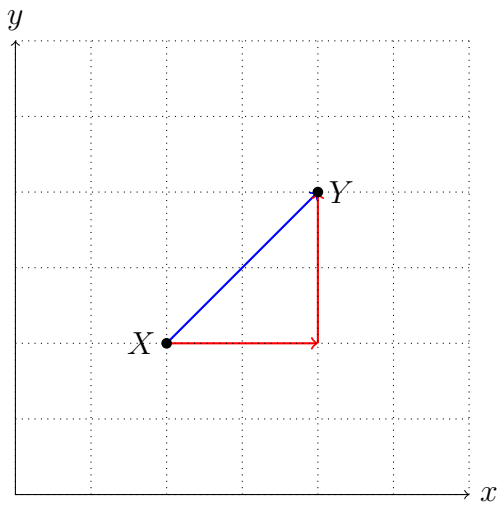


Figure 6: Distance measure example

Looking at the Euclidean distance ($p = 2$, displayed as the blue line) and the Manhattan distance ($p = 1$, displayed as the red line), the Minkovski distance reduces to

$$d_{euclid}(A, B) = \left( \sum_{i=1}^{n} |a_i - b_i|^2 \right)^{1/2}, \tag{3}$$

$$d_{manhattan}(A, B) = \sum_{i=1}^{n} |a_i - b_i|. \tag{4}$$

Returning to our dataset, we remember that our all activities (*features*) are denoted as binary variables, meaning that any point $P$ (*observation*) is of the following structure: $P = (p_1, p_2, ..., p_{45})$ with $p_i \in \{0, 1\}, \forall i \in \{1, 2, ..., 45\}$. In such a case, the Manhattan distance is called *Hamming distance* and equation (4) can be reduced to the number of times, where $a_i \neq b_i$. Before going on, one last distance measure might be important to mention, which is specifically useful for pure binary data and it is called the *Jaccard similarity coefficient* $J$ and was first introduced by Swiss botanist Paul Jaccard (Jaccard, 1912). For two arbitrarily chosen points $A$, $B$ with $n$ binary attributes,

it is commonly defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

These measures of distance enable us to compare different observations based on a chosen metric. An algorithm could now use these measures of (dis-)similarity to cluster a population of points into a specific number of sub populations. Returning to Dolnicar et al. (2018), they differentiate between two main approaches: *hierarchical* and *partitioning methods.*

### 4.1.1 Hierarchical methods

Hierarchical methods follow a very natural and intuitive pattern of clustering data, mainly because they "mimic how a human would appriach the task of dividing a set of $n$ observations (consumers) into $k$ groups (segments)" (Dolnicar et al., 2018). A general differentiation can be made between *divisive* and *agglomerative* hierarchical clustering. The first one uses a top-down approach by splitting the entire data set into the two most distinctive market segments and those two segments again into two respective segments and so on. The agglomerative approach starts with $n$ distinct clusters - one for each observation - and then iteratively merges segments to create a more efficient solution. In other words, it applies a bottom-up approach. Either approach uses a predefined distance metric (see the previous chapter) and a linkage method. Roughly speaking, the "linkage method generalises how, given a distance between pairs of observations, distances between groups of observations are obtained" (Dolnicar et al., 2018). *Single linkage*, for example, chooses the "nearest neighbour". For more on hierarchical clustering see e.g Nielsen (2016).

### 4.1.2 Partitioning methods

As easy as these sort of approaches might seem, it is apparent that they can easily raise issues when dealing with large data sets. A second approach

according to Dolnicar et al. (2018) are *partitioning methods*. These type of methods do not iteratively partition the data set to get the desired results, but rather partition the data set just once. For a hierarchical algorithm to work, one would have to compute the chosen metrics of (dis-)similarity for every pair of observations in the data set, resulting in $(n(n-1))/2$ distances. (Remember the symmetry property: $d(A, B) = d(B, A)$.) A partitioning algorithm calculates the distance to the respective center segments and, more or less, chooses the nearest one (depending on the chosen algorithm, see Dolnicar et al. (2018)).

Two very common partitioning algorithms are the *k-means* and the *k-modes* algorithm. The first starts by choosing $k$ random observations from the data set with $k$ being the desired number of segments. These $k$ observations are used as pre-determined cluster centres and the remaining observations are assigned accordingly to the nearest center. After that, new cluster centres are computed to minimize the in-cluster distances. After that, the cluster memberships are revoked and the algorithm starts anew with the recalculated cluster centres. The algorithm ends with a specified number of iterations, or as soon as no further improvements can be made by rerunning the algorithm (Géron, 2020).

It is important to note that, despite its efficiency, the k-means algorithm runs into issues when dealing with categorical data in general and binary data specifically. Huang (1998) provides some extensions to the existing algorithm, which they eventually call the $k$-modes algorithm. All in all, they (1) switch from the euclidean distance measure often used in k-means to a measure suited for categorical data, (2) use a frequency based approach when updating the cluster centres and (3) use modes instead of cluster means (Huang, 1998).

### 4.1.3 Hybrid Methods

A final contender for clustering that is worth mentioning is called *bagged clustering* (Leisch, 1999). Essentially, it combines the approaches of *hierarchical* and *partitioning* clustering with the process of bootstrapping. The algorithm starts by using a partitioning method, e.g k-means, on a randomly drawn subset of the data. The gathered centroids are then used in hierarchical clustering to gather a more efficient solution. The advantage of this method is that (1) it is not endangered by large data sets due to partitioning and (2) the gathered centroids in combination with hierarchical clustering are able to identify niche segments in data. (The interested reader is referred to Dolnicar et al. (2018) or Leisch (1999).)

## 4.2 Model-based methods

Distance-based measures - as power- and useful they are - can lead to complications when dealing with datasets which consist of different kinds of features (categorical, continuous, binary, etc.). However, there is a second broad category of clustering methods that take an entirely different approach: *model-based clustering methods.* A beautiful review of said models can be found in Bouveyron et al. (2019). Right at the beginning, the authors neatly summarize the incentive behind coming up with a different approach to cluster analysis.

> "At the same time, they[2] left several practical questions unresolved, such as which of the many available clustering methods to use? How many clusters should we use? How should we treat objects that do not fall into any group, or outliers? How sure are we of a clustering partition, and how should we assess uncertainty about it?" (Bouveyron et al., 2019, page 3)

To summarize, model-based clustering tries to introduce modern methods of statistical inference into the field of cluster analysis. The basic idea is to

[2]Referring to algorithms that were based on a matrix of (dis-)similarity, similar to the previously mentioned distance-based methods.

set up a probability model which consists of a finite mixture of multivariate distributions - generally called *finite mixture models* - and find a set of distributions that are most likely to have generated the given data set (Bouveyron et al., 2019). The scientific field of statistical mixture analysis can with certainty be described as a rabbit hole one could easily get lost in. A good companion for such an undertaking is the *Handbook of Mixture Analysis* by Frühwirth-Schnatter et al. (2019). Nevertheless, it might be beneficial to recap the basic model setup of such a finite mixture model.

Let us assume we have several multivariate data points $y_1, ...y_n$, each with $k$ different features, hence $y_i = (y_{i,1}, y_{i,2}, ..., y_{i,k})$. Then, according to Bouveyron et al. (2019), "a finite mixture model represents the probability distribution or density function of one multivariate observation, $[y_i]$, as a finite mixture of weighted average of $G$ probability functions". These probability functions are called mixture components:

$$p(y_i) = \sum_{g=1}^{G} \tau_g f_g(y_i|\theta_g). \tag{5}$$

Here, $\tau_g$ refers to the probability that observation $y_i$ has been generated by the mixture component $g$, with $\sum_{g=1}^{G} \tau_g = 1$ and $\forall g : \tau_g \geq 0$. The function $f_g(.|\theta_g)$ denotes the density of the $g$-th component and its parameters. (e.g. $\mu$ and $\sigma$ in case of a normal distribution). Again, returning to our dataset at hand, we realize that every feature $k$ is a binary variable, hence we would need a distribution which returns either a value 1 or a value 0.

A random variable of such kind is called *Bernoulli* variable, named after Swiss mathematician *Jakob Bernoulli*. The probability of such a variable $X$ being 0 can be written as $P(X = 0)$, or $P(X = 1)$ in the case of $X$ being 1. To simplify, this probability can be written as $\pi$:

$$\begin{aligned} \pi := P(X = 1), \quad \pi \in (0, 1), \\ 1 - \pi = P(X = 0). \end{aligned}$$

The probability mass function then can be written as

$$f(\pi) = \pi \cdot (1 - \pi).$$

More or less, this distribution with its parameter $\pi$ can be used to describe how likely an event is going to happen, or in case of this empirical analysis, how likely it is that a certain activity has been undertaken. This can be used in a finite mixture setup to cluster binary data, commonly known as *Bernoulli Mixture Model* (BMM).

Such a setup might look like as follows:
We are looking at 45 different binary variables $v_i, i = 1, ..., 45$ and their respective Bernoulli distributions:

$$f(v|\pi_g) = \prod_{i=1}^{45} \pi_{ig}^{v_i}(1 - \pi_i)^{1-v_i}. \tag{6}$$

This can be used in a Finite Mixture Model as follows:

$$p(v|\pi, \tau) = \sum_{g=1}^{G} \tau_g f_g(v|\pi_g). \tag{7}$$

The remaining question is about how to estimate such a model and its parameters. According to Bouveyron et al. (2019), this is usually done by performing Maximum Likelihood estimation, more specifically by using the *Expectation-Maximization (EM)* algorithm. The algorithm starts with an $E$-step, in which "the conditional expectation of the complete data log-likelihood given the observed data and the current parameters estimates is computed." (Bouveyron et al., 2019) The second step is the $M$-step, in which "parameters that maximize the expected log-likelihood from the E-step are determined" (Bouveyron et al., 2019). The EM-algorithm is often implemented in software packages containing mixture models. A more comprehensive explanation of that can be found in Bouveyron et al. (2019).

# 5 Empirical Analysis

We have seen the key characteristics of the dataset, set up an empirical framework and can now finally start with the empirical analysis. There are many ways to perform statistical computations - e.g. Python, SPSS, SAS and R - just to name a few. R (See here for more: www.r-project.org) has many advantages, some of which are (1) being free and easy to use and (2) the aim of pursuing an open-source policy, which in turn leads to the existence of countless packages and implementations of statistical methods. These features are the reasons why the following computations will be computed with R. The complete code can be found on a public GitHub webpage under the following in link: https://github.com/therealLucasPaul/EconThesis2022

We will start by using a model-based approach to clustering and market segmentation. A package which has implemented a variety of mixture models into R is *flexmix* - short for Flexible Mixture Modelling - authored by *Bettina Grün* and *Friedrich Leisch*. Some vignettes and further information can be found in e.g. Grün, Leisch, et al. (2007) or Leisch (2003). I have already stated that "data-driven market segmentation is exploratory by nature" (Dolnicar et al., 2018) and hence it is (1) recommended to compute several different models upfront and (2) use a metric of comparison to find the most appropriate one. To satisfy the first, I computed several *Finite Mixtures of Bernoulli* with one to fifteen clusters. A model with one cluster would use one set of distributions to generate every one of the 1003 data points in the dataset. A model with 1003 clusters would fit one individual set of distributions for every data point. Each case of extremes lacks explanatory power for obvious reasons, which is why it is advisable to find a sweet spot in-between. The second property can be satisfied by making use of the log-likelihood. Two popular metrics in econometric analysis in general are the *AIC* (Akaike's Information Criterion) and the *BIC* (Bayes Information

Criterion)

$$AIC = -2 \cdot ln(\mathfrak{L}(\hat{\theta}|\text{data})) + 2 \cdot K,$$
$$BIC = -2 \cdot ln(\mathfrak{L}) + K \cdot ln(n).$$

Here, $\mathfrak{L}$ refers to the likelihood of the estimated model with the maximum likelihood estimator $\hat{\theta}$ and $K$ to the number of estimable parameters (Burnham & Anderson, 2004). Both criteria use the log-likelihood and a certain correction term for $K$. A third criterion, which is specifically useful for model-based clustering methods, is the *Integrated Completed Likelihood* (ICL) and can be defined as "the ordinary BIC penalized by the substraction of the estimated mean entropy" (Biernacki et al., 2000). All in all, the lower the metric, the better the model might be. Additionally, these measures penalize models for "being too complex", or in other terms, they penalize models for the estimation of additional parameters. Hence, a model with seven clusters faces a higher penalty than a model with just five clusters.

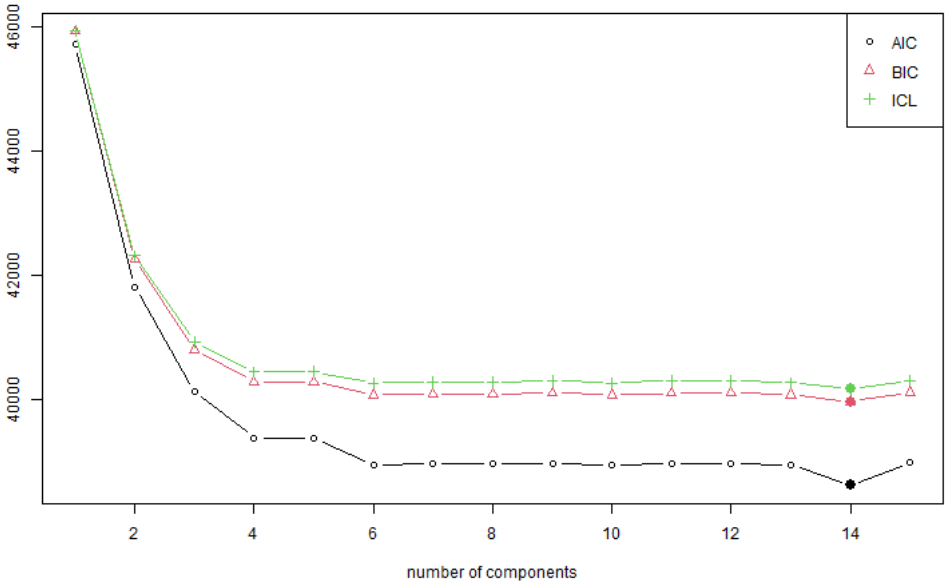


Figure 7: Finite Mixture of Bernoulli - Model Comparison for 1 to 15 clusters

The three metrics of comparison for the 15 different models can be found in figure 7. The first thing to notice is that the metrics follow an overall - although only weakly - decreasing trend with an increasing number of clusters. Disregarding the aforementioned penalty, this somewhat makes sense, as a model with 1003 clusters would perfectly describe the given dataset. A common method for choosing the right model is by using the *Elbow Method* (Géron, 2020). It can be described as choosing the number of clusters that "breaks the curve" and does not add significant improvement of the chosen metric to the model. In this case, it can be argued that either four or six clusters satisfy this requirement. Four clusters, because it is the first number of clusters that on this imagined horizontal line of BIC and ICL and six clusters, because it can be seen that there is a final significant improvement when changing the model from five to six clusters. The *flexmix* package provides a function which automatically chooses the best model and confirms the hypothesis by stating that a six-cluster model fits best. The cluster sizes range from 66 to 276 observations.

## 5.1 Issue with Data Dimensionality

Before continuing with the results, a short interjection might be in order. As stated in chapter 4, clustering in general can be understood as trying to find clusters of observations in an $n$-dimensional space. This is pretty straightforward in one, two and three dimensions, yet increases in complexity with an increasing number of variables. As the dataset consists of 45 different variables, it is apparent that whatever algorithm is chosen, it tries to find clusters in a 45-dimensional space. A particular issue in this regard that has been pointed out by Dolnicar et al. (2014) in their paper "Required Sample Sizes for Data-Driven Market Segmentation Analyses in Tourism".

Fundamentally, they state that "the validity of data-driven market segmentation analyses depends on having available a sample of adequate size" and that there is general rule on how such an "adequate size" might look like (Dolnicar et al., 2014). That is why the authors addressed this issue by

using a simulation study to find a general guideline for data-driven market segmentation. The conclude that "a sample size of 70 times the number of variables proves to be adequate" (Dolnicar et al., 2014). Having in mind that the analysed dataset of vacation activities has 45 different variables at a sample size of 1003 observations, it is easy to see that this criteria is not fulfilled. That is why I will perform a second, distance-based approach on the dataset to gain more robust results. As covered in section 4, there are many approaches to clustering, especially in a distance-based setting. One approach, that has been covered in greater detail is the *bagged clustering* approach. Such an additional analysis can be found in figure 12 in Appendix A and will be compared to the main results in the upcoming chapter.

# 6 Results

Finally, it is time to see what insights market segmentation and cluster analysis can generate by analysing the results of the empirical analysis and following the 10 step procedure of Dolnicar et al. (2018) (See page 7 for more). Having extracted the segments using a *Finite Mixture Model of Bernoulli Distributions*, it is time to profile and describe the gathered segments. Due to the vastness of the clusters and its variables (45 per cluster), the clusters will be shown individually. An overview of all clusters can be found in Appendix A.

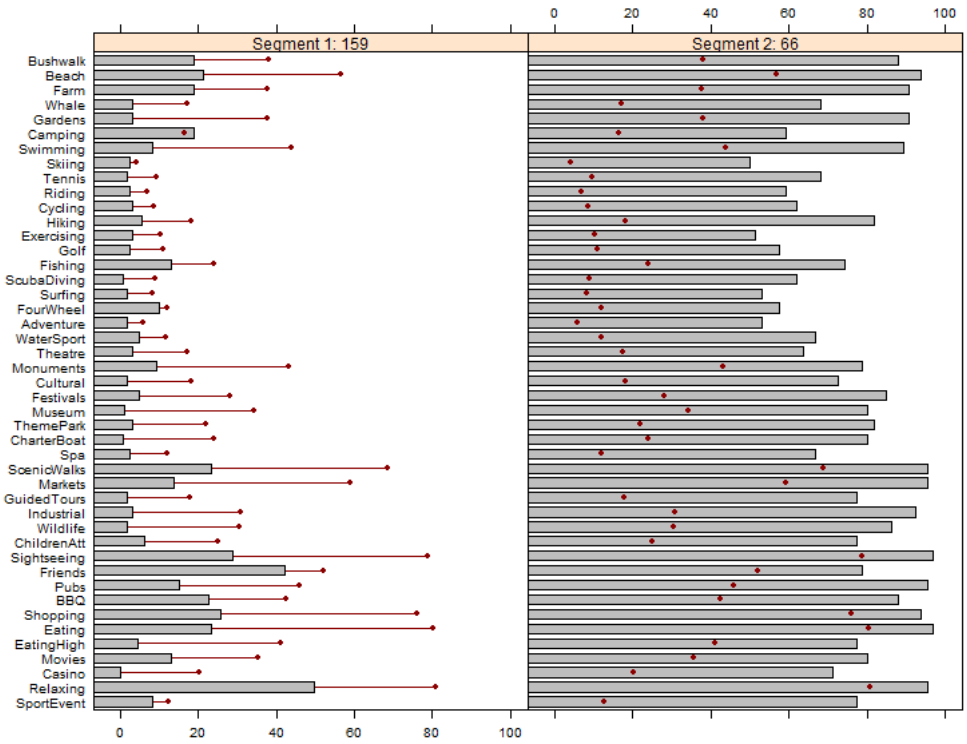


Figure 8: Cluster 1 and 2 of the six-cluster finite mixture model

The two most distinct and straightforward clusters are *cluster 1* and *cluster 2*, which we can see in figure 8. The grey bars indicate the probabilities

of a person in the respective cluster to have undertaken this activity. The red dots represent the marginal probabilities of the entire sample. In the left panel, it can be seen that people in segment 1 have a low probability of pursuing any activity. Nearly every activity has a probability below its sample mean. The only activities with a probability higher than 40% are "Relaxing" and "Friends", however it is likely that these activities profit from a distinct high sample mean of roughly 80%. Nevertheless, it fits the picture of a non-active, relaxing and passive cluster, which is why - for practical reasons - it will be called the **Breakfast-In-Bed-Cluster**. Cluster 2, as seen in the right panel of figure 8, shows a contrary picture. Nearly every activity shows a high probability of being undertaken and without any exception, a higher level than in the sample, indicated by the red points all being within the grey bars. This market segment of 66 people shows a high readiness of doing as much as possible. Generally common activities like "Relaxing", "Shopping", "Eating" and "Sightseeing" nearly reach a probability of 100% and even rare activities (see figure 5 for reference) like "Skiing" and "Riding" reach quite a decent probability with roughly 50% and 60%, respectively. Due to this overall picture, this segment will be called the **Off-We-Go-Cluster**.

The remaining extracted segments are not as straightforward as the first two clusters and distinctive features might only appear when comparing two different segments. Figure 9 shows the third and fourth cluster in the same manner as the first ones. Segment 3 has some significant characteristics when looking at the activities "Bushwalk", "Beach", "Swimming", "Scenic Walks" as well as the common activities "Relaxing" and "Sightseeing". All of those variables have a probability that is higher than the sample mean. Additionally, "Fishing", "Hiking", "Whale", "Camping" and other outdoor related activities show probabilities that exceed the sample mean by nearly a 100% (e.g. "Hiking" has a cluster probability of roughly 43% and a sample probability of 20%). Seeing that it is apparent that this cluster focuses on outdoor activities, whether they are sports-related like "Hiking" or sightseeing-related like whale watching ("Whale"). Due to these characteristics, I am going to call this cluster the **Bear-Grylls-Cluster** (Named after the famous outdoor-

adventurer *Edward Michael Grylls*).

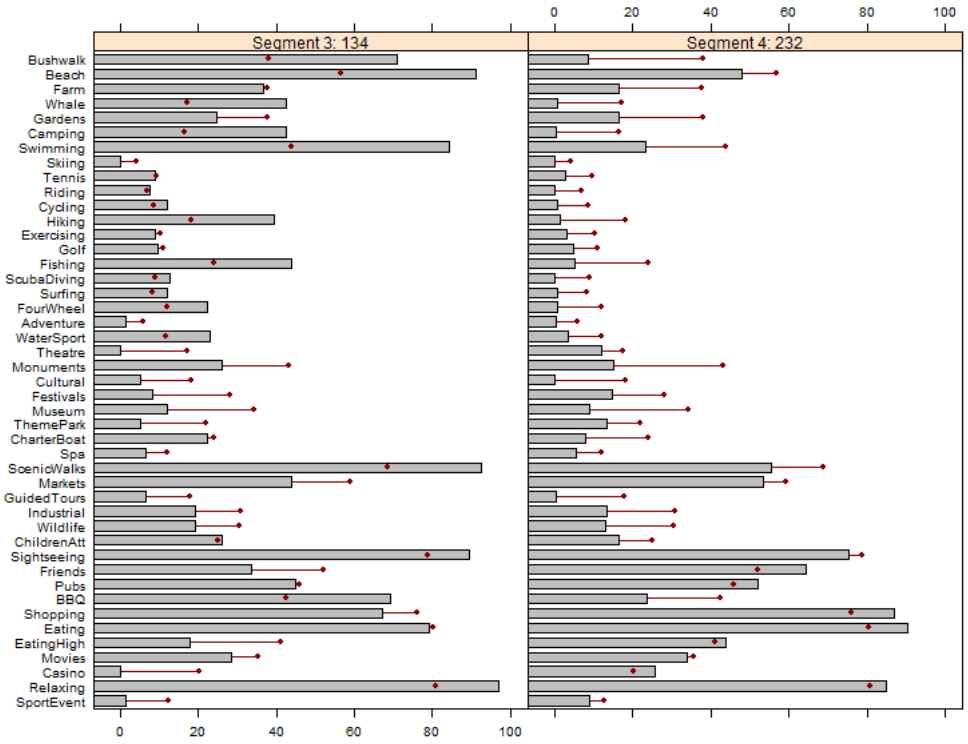


Figure 9: Cluster 3 and 4 of the six-cluster Finite Mixture Model

The presence of a high probability of an activity within a cluster is not the only way to describe and profile a segment. The absence of an activity can also contain valuable information on the key characteristics of a market segment. Segment 4, for example has a similar structure to segment 3, yet some significant differences are present. At first, it can be seen that the sports- and outdoor-affiliated activities, especially the activities ranging from "Bushwalk" to "Charter Boat", show distinctively lower probabilities in segment 4 than in segment 3. Nevertheless, some activities show increased levels of probabilities in segment 4. For instance, "Casino", "Movies, "Eating High", "Pubs" and "Friends" as well as "Shopping" and "Eating" have cluster-probabilities higher than or roughly equal to the sample mean. Common activities like "Relaxing" and "Sightseeing" are present as always. These

occurrences indicate that this cluster/segment captures tourists that enjoy partying, meeting friends, eating out and having a drink with friends. Due to these characteristics, I am going to name this segment the **Saturday-Night-Fever-Cluster**.

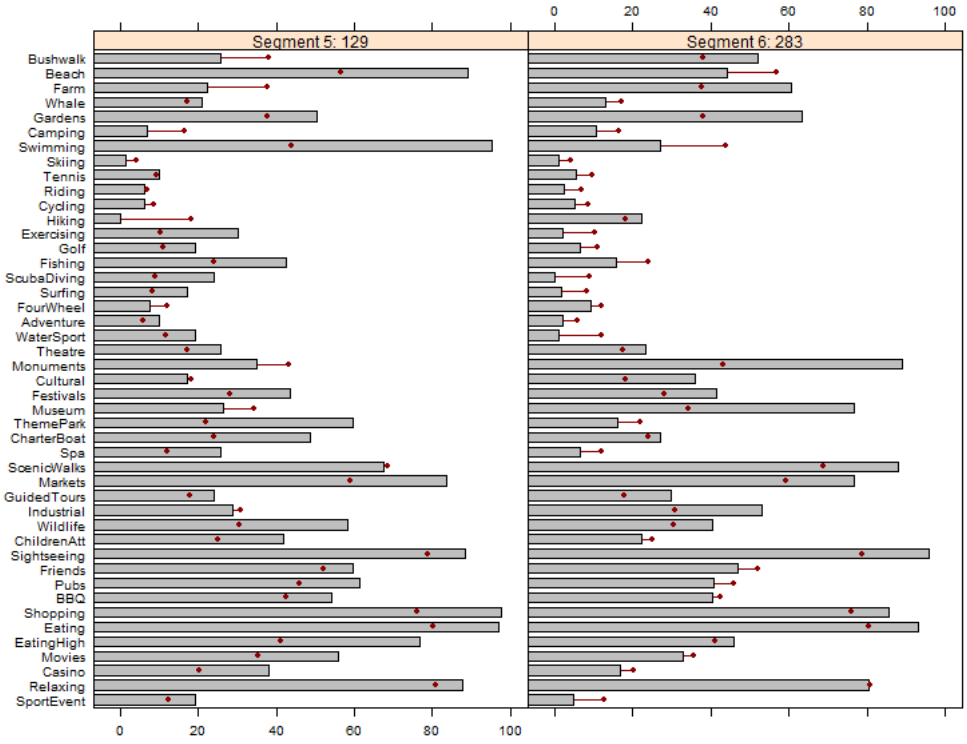


Figure 10: Cluster 5 and 6 of the six-cluster Finite Mixture Model

The last two clusters show an entirely new picture, as they are the first clusters - aside from the previously mentioned **Off-We-Go-Cluster** - that show above-mean probabilities in the arts and culture section. For the sake of convenience, let us start with cluster 6 this time. Typically, the common activities like "Relaxing", "Shopping" etc. have high probabilities of being undertaken. The distinctive feature of this cluster, however, can be seen in the activities "Monuments", "Cultural", "Festivals", "Museum", "Theatre" and "Gardens". In contrast to the previous clusters, which focused on sports-

or social-related activities, this cluster has a significant manifestation of high probabilities in the culture section of the activities. With 283 observations, it is also the biggest cluster in absolute terms. Aside from that, the cluster also shows average to high probabilities in the social section of the activities in the lower third of the right panel in figure 10. These key characteristics suggest the name **Culture-cluster** for this segment.

The only cluster remaining is segment 5 (see the left panel of figure 10), which contains 129 observations. This extracted segment is the only one without clear key characteristics. There are certainly some similarities between segment 5 and segment 6, however the significant arts and culture block of segment 6 does not seem to be present here. Some activities that show above-mean probabilities are "Wildlife", "ChildrenAtt", "ThemePark" and "CharterBoat". These feature would indicate that it might be some sort of "family-related" market segment. In addition, activities like "Beach" and "Swimming" are nearly at a 100% which supports this initial hypothesis. The remaining variables exhibit average or slightly above-average probabilities, which is why I will name this segment the **Family-Fun-Cluster**.

## 6.1 Robustness check using bagged clustering

All these clusters intuitively make sense and each cover at least some part of common human behaviour, however a robustness check might be in order. That is why I performed a bagged clustering analysis on the data set (see chapter 4.1.3 for further discussion). An overview of the cluster probabilities, similar to the ones in figures 8, 9 and 10 can be found in Appendix A. Clearly, the *Bed-In-Breakfast-Cluster* can be found in Cluster 2 and the *Off-We-Go-Cluster* in cluster 1. The *Culture-cluster* can be found in cluster 5 and the *Saturday-Night-Fever-Cluster* can be found in cluster 6. Some features of the *Bear-Grylls-Cluster* can be found in Cluster 3, however some main features (like a high probability in *Hiking*) do not seem to be present. The *Family-Fun-Cluster* is not as present in this method as in the model-based approach, however even in the model-based approach it was quite a weak relationship.

Keeping in mind that, like Dolnicar et al. (2018) repeatedly mention, market segmentation is inherently exploratory and that that the cluster results often depend on the chosen methods, the robustness check is an indicator that the extracted structure by the model-based approach can indeed be viewed as valid to some degree.

## 6.2 Application of results

When dealing with subjects like market segmentation, the question "why bother?" necessarily arises. The results gathered in the previous sections are most certainly interesting from a theoretical perspective, however also provide meaningful insights on a practical, real-world level. As this thesis primarily deals with the theoretical and empirical analysis of a market with a model-based clustering approach and not policy implications of a market segmentation analysis, I will only provide some examples of real-world implications based on the previous results.

Touristic areas that already have a lot of pubs and restaurants might benefit from opening a cinema or a casino, as it may attract new customers from the *Saturday-Night-Fever-Cluster*. Similarly, regions with a strong historical foundation and several monuments can profit from having a lot of museums and theatres. Tourists from the *Culture-Clusters* might then have incentive to travel to these destinations. A third boost could be gained from the *Family-Fun-Cluster*. The simultaneous presence of children attractions, wildlife parks and theme parks might attract families with children in a more efficient way.

| Segment name | Number of observations | Distinctive features |
|---|---|---|
| Bed-In-Breakfast-Clusters | 159 | Low in-cluster probabilities in all categories;<br><br>Only generally common activities have high probabilities |
| Off-We-Go-Cluster | 66 | High in-cluster probabilities in all categories; |
| Bear-Grylls-Cluster | 134 | Focus on outdoor and sports-related activities;<br><br>*Key Activities:*<br>Hiking, Scenic Walks, Bushwalk, Swimming |
| Saturday-Night-Fever-Cluster | 232 | Focus on social and consumption-related activities;<br><br>*Key Activities:*<br>Pub, Eating, Movies, Casino |
| Family-Fun-Cluster | 129 | Focus on family-friendly activities;<br><br>*Key Activities:*<br>Children Attractions, Theme Park, Wildlife |
| Culture-cluster | 283 | Focus on arts and culture;<br><br>*Key Activities:*<br>Museum, Monuments, Theatre |

Table 4: Overview of the six extracted segments along with its distinctive features

# 7 Conclusion

As always, the question at the end of a thesis is: Did it work and is it useful? I have gathered literature on market segmentation and analysed the data set from a descriptive point of view. Afterwards, I set up an empirical framework with a subsequent empirical analysis. The gathered results using the Bernoulli mixture model show a strong six-cluster pattern in the Australian tourism data, which has also been supported by the bagged clustering approach. The *Breakfast-In-Bed* refers to a consumer group of passive, hardly active people who distinguish themselves by having low probabilities of engaging in any activity, whereas the *Off-We-Go* defines itself following the contrary approach. Consumers from the *Bear-Grylls* cluster prefer to engage in outdoor, sports-related activities, whilst the *Saturday-Night-Fever* cluster chooses social and consumption-related activity patters. The *Family-Fun* market segment display high probabilities of signing up for family-friendly attractions. Ultimately, consumers from the *Culture* cluster focus on attractions like museums, theatres and monuments. The resulting cluster features and key characteristics can be used to efficiently target consumer groups, subsequently giving institutions, governments and companies a framework for the optimal allocation of monetary resources. Market segmentation in general, especially in combination with cluster analysis, can be a powerful tool for governments, agencies and companies to target consumers based on their needs, desires and behaviour. The recent advancements in machine learning and statistical analysis have created powerful new methods for classification and clustering, which surely will improve the ways how we as economists analyse data and derive implications. It is an exciting field and I am sure that it will continue to thrive in the upcoming decades. I would like to end this thesis with a quote by G.K. Chesterton, which could not be more fitting:

"The traveler sees what he sees. The tourist sees what he has come to see." - G.K. Chesterton

# A References

# References

Aldenderfer, M., & Blashfield, R. K. (1984). *Cluster analysis* (1. print..). Sage Publ. https://permalink.obvsg.at/wuw/AC02410264

Andersen, H. C. (1975). *The fairy tale of my life : An autobiography*. Paddington Press.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725. https://doi.org/10.1109/34.865189

Bouveyron, C., Celeux, G., Murphy, T., & Raftery, A. (2019). *Model-based clustering and classification for data science: With applications in r*. Cambridge University Press. https://books.google.at/books?id=KACjDwAAQBAJ

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Camilleri, M. A. (2018). Market segmentation, targeting and positioning. In M. A. Camilleri (Ed.), *Travel marketing, tourism economics and the airline product: An introduction to theory and practice* (pp. 69–83). Springer International Publishing. https://doi.org/10.1007/978-3-319-49849-2_4

Deza, M. M., & Deza, E. (2009). *Encyclopedia of distances*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00234-2_1

Dolnicar, S. (2020). Market segmentation analysis in tourism: A perspective paper. *Tourism Review*, *75*(1), 45–48. https://doi.org/10.1108/TR-02-2019-0041

Dolnicar, S. (2021). Market segmentation for e-tourism. Springer International Publishing. https://doi.org/10.1007/978-3-030-05324-6_53-1

Dolnicar, S., Grün, B., & Leisch, F. (2018). *Market segmentation analysis: Understanding it, doing it, and making it useful*. Springer Nature. http://dx.doi.org/10.1007/978-981-10-8818-6

Dolnicar, S., Grün, B., Leisch, F., & Schmidt, K. (2014). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research*, *53*(3), 296–306. https://doi.org/10.1177/0047287513496475

Drasgow, F. (2014). Polychoric and polyserial correlations. *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat02493

Driver, H., & Kroeber, A. (1932). *Quantitative expression of cultural relationships*. University of California Press. https://books.google.at/books?id=kpYexQEACAAJ

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern classification and scene analysis* (Vol. 3). Wiley New York.

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *SIGKDD Explor. Newsl.*, *4*(1), 65–75. https://doi.org/10.1145/568574.568575

Frühwirth-Schnatter, S., Celeux, G., & Robert, C. P. (2019). *Handbook of mixture analysis*. CRC Press, Taylor & Francis Group. https://permalink.obvsg.at/wuw/AC15319610

Géron, A. (2020). *Hands-on machine learning with scikit-learn, keras, and tensorflow : Concepts, tools, and techniques to build intelligent systems* (Second edition, seventh release.). O'Reilly. https://permalink.obvsg.at/wuw/AC16308631

Goryushkina, N. E., Gaifutdinova, T. V., Logvina, E. V., Redkin, A. G., Kudryavtsev, V. V., & Shol, Y. N. (2019). Basic principles of tourist services market segmentation. *International Journal of Economics and Business Administration*, *7*(Issue 2), 139–150. https://doi.org/10.35808/ijeba/222

Grün, B., Leisch, F., et al. (2007). Applications of finite mixtures of regression models. *URL: http://cran. r-project. org/web/packages/flexmix/vignettes/regression-examples. pdf*.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304. https://doi.org/10.1023/A:1009769707641

International Labour Organization (ILO). (2017). Tourism at a glance. https://www.ilo.org/global/docs/WCMS_544196/lang--en/index.htm

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytologist*, *11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Kamthania, D., Pawa, A., & Madhavan, S. (2018). Market segmentation analysis and visualization using k-mode clustering algorithm for e-commerce business. *Journal of Computing and Information Technology*, *26*(1), 57–68. https://doi.org/10.20532/cit.2018.1003863

Leisch, F. (1999). Bagged clustering. *51*. https://epub.wu.ac.at/id/eprint/1272

Leisch, F. (2003). FlexMix: A general framework for finite mixture models and latent class regression in R. *Faculty of Business - Papers (Archive)*. https://ro.uow.edu.au/buspapers/487/

NETSTEL Software. (2022). *New york - chicago - flight time*. https://flight-time.org/flight/new-york-2-chicago-1586

Nielsen, F. (2016). Hierarchical clustering. In *Introduction to hpc with mpi for data science* (pp. 195–211). Springer International Publishing. https://doi.org/10.1007/978-3-319-21903-5_8

Paullin, C. O. (1932). *Atlas of the historical geography of the united states*. Carnegie Inst. of Washington American Geographical Society.

Preud'homme, G., Duarte, K., Dalleau, K., Lacomblez, C., Bresso, E., Smaïl-Tabbone, M., Couceiro, M., Devignes, M.-D., Kobayashi, M., Huttin, O., Ferreira, J. P., Zannad, F., Rossignol, P., & Girerd, N. (2021). Head-to-head comparison of clustering methods for heterogeneous data: A simulation-driven benchmark. *Scientific Reports*, *11*(1), 4202. https://doi.org/10.1038/s41598-021-83340-8

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

The World Bank Group. (2022a). International tourism, expenditures (% of total imports). https://data.worldbank.org/indicator/ST.INT.XPND.MP.ZS

The World Bank Group. (2022b). International tourism, expenditures (current USD). https://data.worldbank.org/indicator/ST.INT.XPND.CD

Tourism Research Australia. (2019). *Tourism satellite account 2018–19: Summary of key results*. https://www.tra.gov.au/ArticleDocuments/185/Tourism%20Satellite%20Account%202018-19.pdf.aspx

Towner, J. (1985). The grand tour. *Annals of Tourism Research*, *12*(3), 297–333. https://doi.org/10.1016/0160-7383(85)90002-7

World Tourism Organization. (2022). Why tourism? https://www.unwto.org/why-tourism

# B List of Tables

## List of Tables

# C List of Figures

## List of Figures



[heading=none]

# D Appendix A

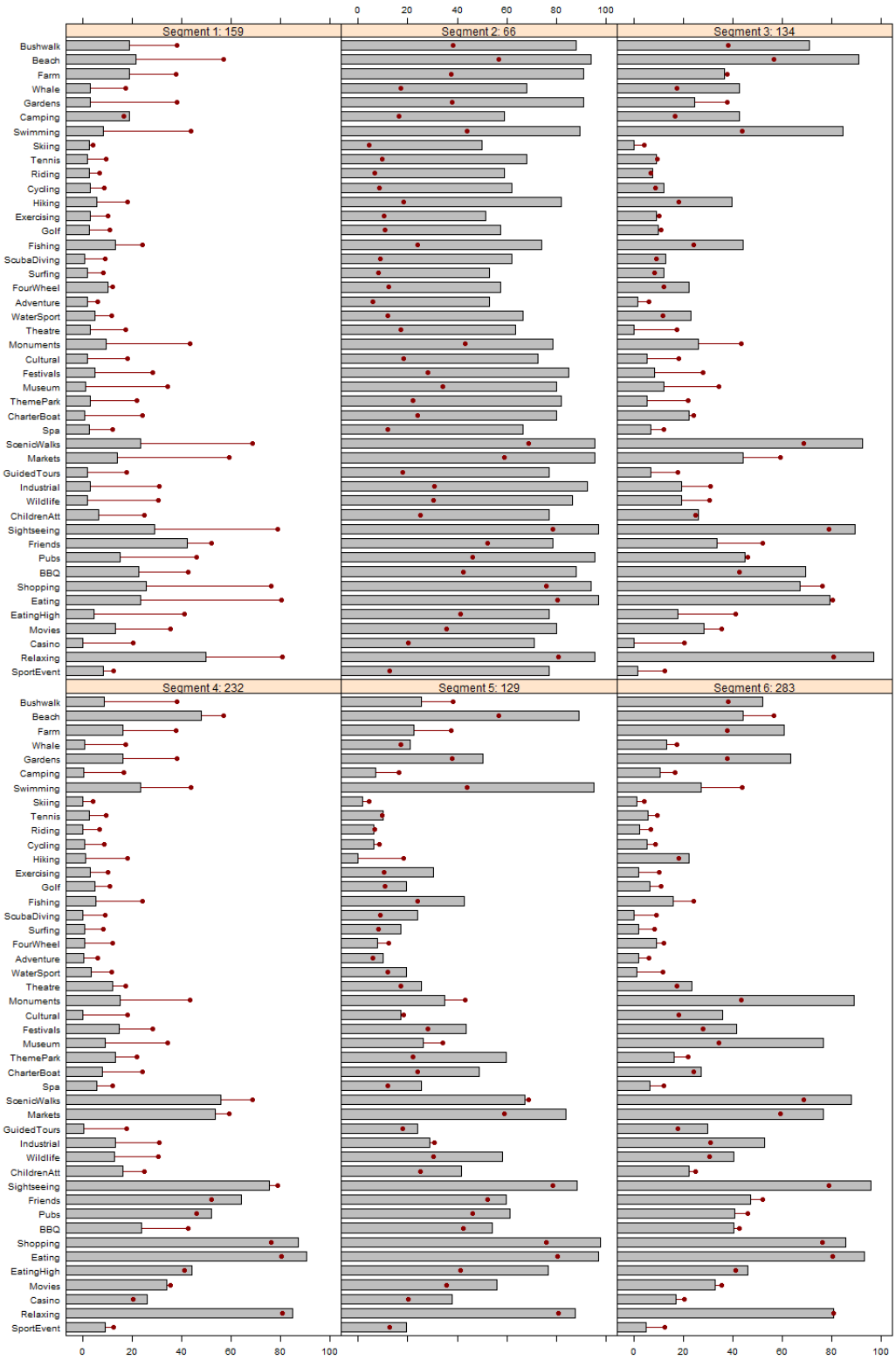


Figure 11: Finite Mixture of Bernoulli - six Cluster solution

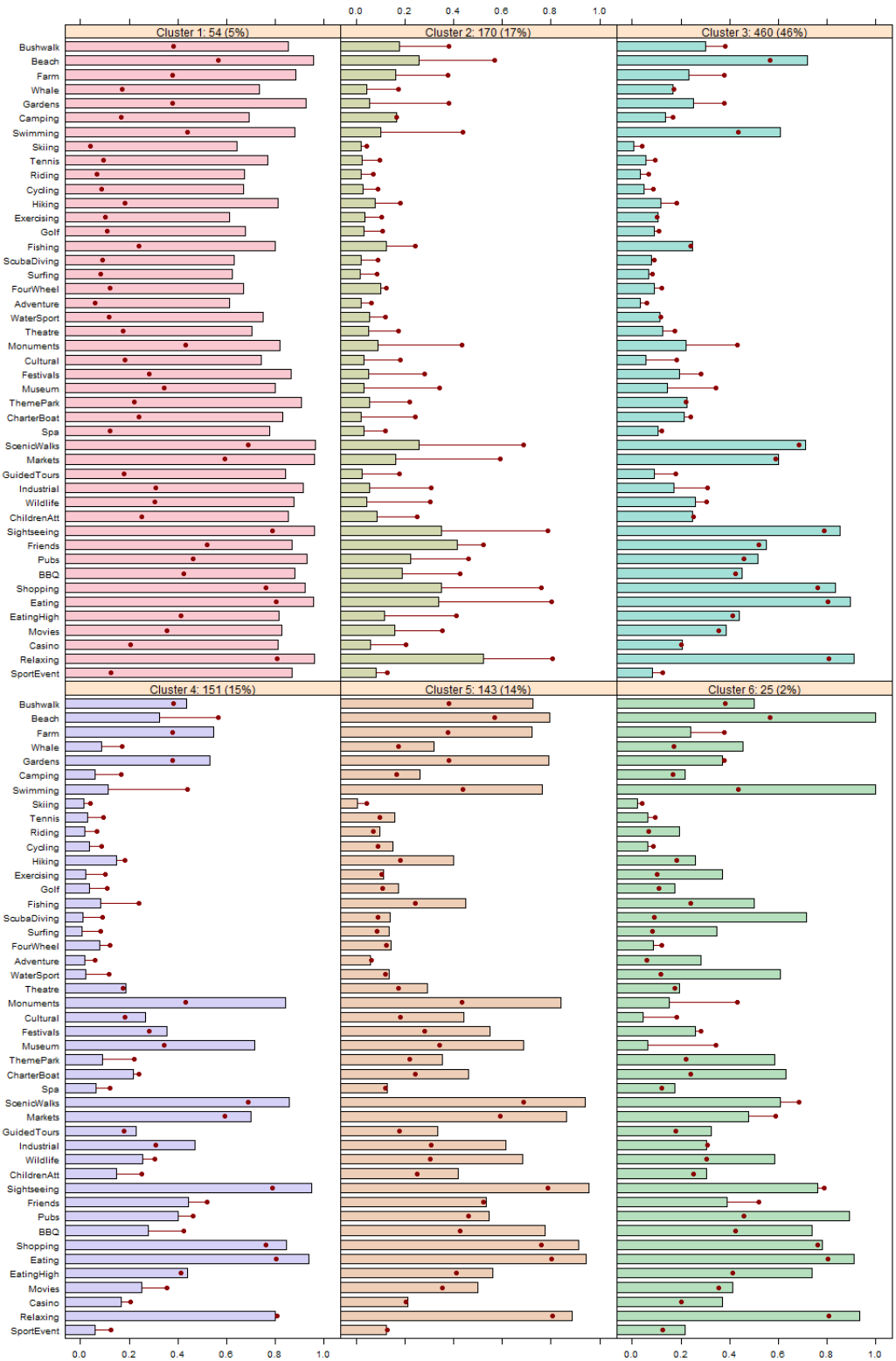


Figure 12: Bagged Clustering - six Cluster solution