Practicum (Course code 833559901)

# Exploring Israel’s Tourism Using Unsupervised Learning Methods

By

Ravid Motorin 201314374

Renana Sanders 311500805

Department of Information Science

Bar-Ilan University

Course Lecturer: Dr. Ariel Reosenfeld

Revised version- October 2023

# Table of Contents

## Background

### The Global Tourism Industry: Significance and Growth

Over the past seven decades, tourism has experienced practically uninterrupted growth and diversification, becoming one of the world's foremost economic sectors. International tourist arrivals rose from 25 million globally in 1950 to nearly 1.5 billion in 2019 (UNWTO, 2020). Tourism generated $1.7 trillion in export earnings and created 1 in 10 jobs worldwide prior to the COVID-19 pandemic (UNWTO, 2019). The UN World Tourism Organization (UNWTO) projects international arrivals to reach 1.8 billion by 2030, with tourism accounting for 11% of global GDP and 380 million jobs (UNWTO, 2019). This sustained expansion reflects tourism's immense value for stimulating economic growth, employment, exports, infrastructure development, and socioeconomic advancement worldwide (UNWTO, 2018).

### Key Drivers and Considerations for Tourism Destinations

A complex array of factors shape tourism flows and the attractiveness of destinations. Fundamental elements include the natural and cultural endowments of a place that motivate travel (Gelbman & Timothy, 2011). Accessibility, transportation networks, tourism infrastructure, and hospitality services also play an essential role (Mak, 2017). Prices, accommodation facilities, restaurants, cleanliness, safety, attractions, and weather conditions influence travelers' experiences and satisfaction (Jin & Wang, 2016). Tourism demand varies across source markets and tourist segments based on demographics, cultural norms, and the purpose of travel, whether for leisure, family, religious, educational, business or health reasons (Chang, 2018). As the tastes and expectations of travelers evolve, destinations must continually enhance their offerings, while balancing environmental, social and economic sustainability (Milano et al., 2019).

### Israel's Inbound Tourism Industry

Tourism has become a vibrant and growing component of Israel's service economy. International arrivals more than doubled over the past decade, from 2.7 million in 2010 to over 4 million in 2019 (OECD, 2020). Tourism directly accounted for 2.8% of Israel's GDP and 3.6% of employment in 2018, while also stimulating indirect impacts and supply chain activities (OECD, 2020). Top source markets prior to the pandemic included the United States, France, Russia, Germany, and the United Kingdom (Israel Ministry of Tourism, 2020). However, the COVID-19 crisis led to a massive 73% year-over-year decline in foreign arrivals in 2020 compared to the record numbers in 2019 (Israel Ministry of Tourism, 2021). As Israel rebuilds its tourism sector in the post-pandemic era, enhancing global competitiveness and resilience will be key priorities.

Applying Innovative Machine Learning Techniques to Uncover Insights

This pioneering research leverages machine learning algorithms to analyze data from Israel’s 2016-2017 Inbound Tourism Survey for insights to guide strategy and planning. The survey collected detailed information on tourist profiles, trips, activities, and satisfaction from nearly 3000 visitors (Israel Ministry of Tourism, 2017). We implemented three leading unsupervised learning approaches – association rule mining, clustering, and anomaly detection – to uncover hidden patterns and relationships within this rich dataset without defining any categories a priori. This represents the first known application of advanced artificial intelligence to derive strategic value from the survey data.

### Association Rule Mining for Pattern Discovery

Association rule mining identifies interesting associations and correlations between variables in transactional databases like purchases, website activities, or tourism choices (Agrawal et al., 1993). It enables market basket analysis, cross-marketing recommendations, decision support systems, and other applications (Hahsler et al., 2019). Efficient algorithms like Apriori and FP-growth are used to find frequent itemsets from which high-confidence rules are generated (Agrawal & Srikant, 1994; Han et al., 2004).

### Clustering Techniques for Segmentation

Clustering algorithms group data points based on similarity, allowing the segmentation of customers, documents, images, or other entities into categories (Xu & Wunsch, 2005). Popular approaches include k-means, hierarchical, density-based, and spectral clustering, using distance, density, graph connectivity or other metrics (Jain, 2010). Clustering supports customer relationship management, targeted marketing, recommendation systems, and other applications across sectors (Aggarwal & Reddy, 2013).

### Anomaly Detection for Identifying Outliers

Anomaly detection involves identifying rare items, events or observations that differ significantly from the majority of data points (Chandola et al., 2009). Statistical, proximity-based, clustering-based, classification-based, information theoretic and other techniques can be applied (Aggarwal, 2017). Anomaly detection aids in detecting fraud, network intrusions, faults, health conditions and other critical issues (Chandola et al., 2009). It also facilitates analyzing skewed data distributions and concept drift (Gupta et al., 2014).

### Implications for Tourism Strategy and Practice

The patterns, associations, segments, and outliers revealed by these algorithms provide data-driven insights to enhance tourism marketing, service design, resource allocation, and visitor satisfaction. Our techniques establish a foundation for ongoing innovation in leveraging analytics and artificial intelligence to strengthen Israel's tourism ecosystem. The knowledge generated also has relevance for tourism organizations globally working to optimize management practices and policies.

## Research Question, Hypotheses and Relevant Methodologies

Initial visualization performed during the data exploration stage, showed clear patterns in site preferences (in terms of visitation patterns) based on attributes like demographics, trip purpose, religion etc. This suggests that we might be able to define tourist "types" based on common behavioral traits according to their travel itineraries

- **Research Question: Is it possible to characterize tourist types based on their travel itineraries and visited sites.**
  We will examine this question as part of the main stage of our research when we will apply unsupervised ML including clustering algorithms.

- **Research Hypotheses:**

The following 6 hypotheses were checked using the appropriate statistical tests (Anova-F-test and 2 sample T-test using spicy library). Setting the confidence level at 0.05, all of the relations described in these hypothese were proven to be significant.
In terms of methodology, it is worth mentioning that in order to perform the statistical tests we needed to create a variable that includes a single numeric value for each tourist (up to this point, we only had a vector with 246 binary values).
We decided to use the amount of times that the value "1" appears in a tourist's vector as the tourist's variable for the purpose of the statistical tests. This calculation equals to the number of activities the tourist participated in (from the total revised sum of 246 activities as explained in the previous section).

- ➢ Hypothesis #1: There is a relationship between a tourist's **religion** and their travel itinerary

Data exploration shows that christian tourists strongly prefer visiting the Old City of Jerusalem in which several christian holy places are located while tourists with no stated religious affiliation or of other faiths tend to prefer more secular sites like Tel Aviv Beach and urban centers.

This suggests religious tourists plan itineraries specifically to visit holy sites tied to their faith. Their religion directly influences their travel goals and site selection.

- ➢ Hypothesis #2: There is a relationship between a tourist's **age** and their travel itinerary, specifically looking at groups with very young or very old tourists.

Data exploration shows that very young or old tourists in a group, tend to skew the group's preferences in terms of site choices.

- ➢ Hypothesis #3: There is a relationship between a tourist's **country of origin** and their travel itinerary

Data exploration suggests that tourists from different countries have varied interests and cultural backgrounds that influence where they choose to visit. We will look specifically into the top 5 visiting countries.

- ➢ Hypothesis #4: There is a relationship between a visit being a tourist's **first / repeating visit** and a tourist's travel itinerary.

Data exploration shows that first-time visitors tend to focus more on the most famous cultural sites like the Old City of Jerusalem, while repeat visitors distribute their time among a wider range of locations.

- ➢ Hypothesis #5: There is a relationship between the **length of the visit** and the tourist's travel itinerary.

Data exploration shows that tourists who stay for a longer time do tend to visit more unique locations. However, the relationship is nonlinear - the peak number of unique sites is reached at 8-11 days, after which more days result in diminishing additional locations. This suggests that most new sites are visited early in the trip, with repeating/revisiting locations thereafter.

➢ Hypothesis #6: There is a relationship between **group size** and travel itinerary.
Data exploration revealed that large groups visited the Old City of Jerusalem less frequently while solo tourists went to Tel Aviv beach and Jerusalem city center more often and displayed more variation in site choices compared to couples and groups.

## Data Review and Preprocessing Applications

## Our Data

Our study is based on a survey conducted by the Ministry of Tourism in collaboration with Prof. Noam Shovel from the Hebrew University of Jerusalem. The survey involved face-to-face interviews and GPS tracking of tourists in Israel between December 2015 and November 2017.

We focused on individual tourists who planned their visits independently, using two main files: "Individuals" (2,453 tourists, 96 columns) and "Activities" (173,488 activities, 112 columns).

Key challenges in our preprocessing phase included (1) data cleaning, standardization, and feature selection (2) reindexing the "Activities" dataset to create "activities per tourist" information (3) Qualitative reduction of the number of activities to focus on the most common ones. We then created binary activity vectors per tourist for machine learning analysis.

## Preprocessing Applications

1. **Cleaning standardizing and feature selection**

Main goal: “Cleaning” our datasets and extracting the relevant columns for further analysis

a. Manual review of  the “Activities” dataset: Before coding, we carefully reviewed a dataset with 170,000+ rows of free-text activities, uncovering issues like spelling errors, non-standard formats, and Hebrew characters needing cleaning. We also found location data errors and tagging inconsistencies, such as categorizing taking a cab as an accommodation or tourism activity when it should be transportation.
   Our manual review helped define what qualifies as a "tourist site," leading to the decision not to include food-related activities but recognizing markets as tourist sites. We also

determined whether to group specific locations within broader geographic categories. These decisions were guided by factors like geography and activity frequency.

b. Defining rules for the “Activities” dataset cleaning and standardizing:
Based on the manual review, we defined a set of rules, categories and assumptions to clean the free text data. For example, identifying work/study related trips as irrelevant.

c. Removing irrelevant rows from “Activities” dataset: we defined a function “Unneeded_Row” to identify activities that should be removed, such as work-related trips, airport arrivals/departures, etc. This function filtered the dataframe and dropped unnecessary rows.

d. Standardizing activity names in “Activities” dataset: we defined functions to standardize the activity names using the tourist area, municipal area and other columns. For example, activities in the "Old City" were standardized to "Old City Jerusalem". This made the activity names more consistent.

e. Feature selection: we only kept relevant columns that we thought might have an influence on tourists' choice of sites: visit_purpose_category, Country_of_Residence, Summer_Winter, first_visit, main_purpose, How_many_people_in_group, Number_of_women, Number_of_men, number_of_days, age groups.

f. Harvesting the revised data set : Finally, the revised “Activities” data set was exported for further analysis.

**2. Reindexing Activities dataset**

Main goal: Extracting activity data for each tourist from the "Activities" dataset and incorporate it into the "Individuals" dataset as new binary value columns.

Initially, we added all to the "Individuals" dataframe, all initialized with "0" values.

To achieve this, we created a dictionary (ind_act_dict) that links tourist IDs to lists of activities they participated in. We used this dictionary to set the value "1" in the relevant activity columns for tourists who performed those activities, leaving "0" for those who didn't. This process was computationally intensive as it required updating the "Individuals" dataframe for each tourist.

To optimize the workflow and avoid repeating this time-consuming process, we generated a new "Individuals" Excel file with all the added columns.

**3. Removing rare activities and creating a binary activity vector per tourist**

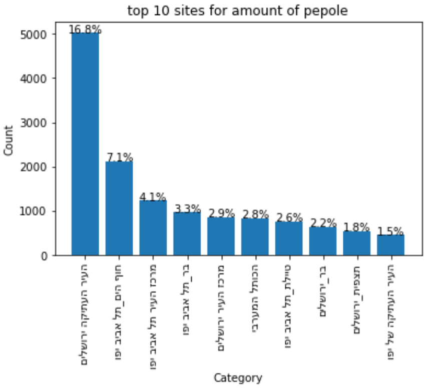
Main goal: Reducing the number of activities for machine learning modeling.

Initially, we had 2,237 activities that were imported from the “Activities” dataset to the “Individuals” dataset. Those activities included many rare ones. In order to focus on more common activities, we decided to only consider activities visited by at least 10 tourists. This reduced the activities count to 223.

We counted each activity, once, per tourist (even if they participated in the same activity multiple times). Following, we created a binary activity vector for each tourist, representing their participation (1) or non-participation (0), in the selected activities, this formed a uniform-length feature for all tourists, serving as our independent variable for machine learning modeling.

## Preprocessing Exploratory Visualizations

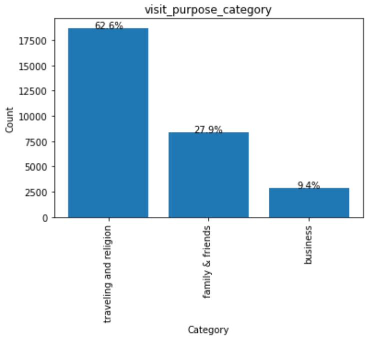
Our initial analysis focused on exploring relationships between activity sites and various categorical variables like visit purpose, religion, etc. Below are some of the interesting findings that we gained in this phase of our research (more can be found within the appendix).

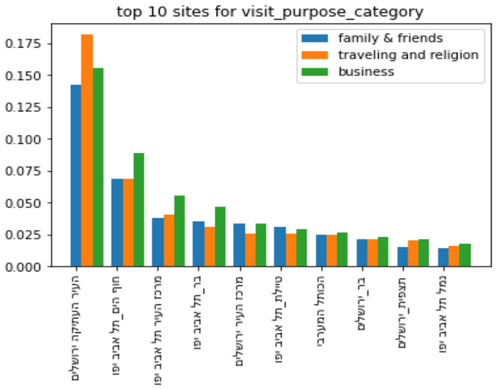


- **All 10 top sites are located  in Jerusalem or Tel-aviv**

- **62.6% of the tourists arrived for traveling / religious purposes, 27.9% for visiting friends and relatives and 9.4% for business.**

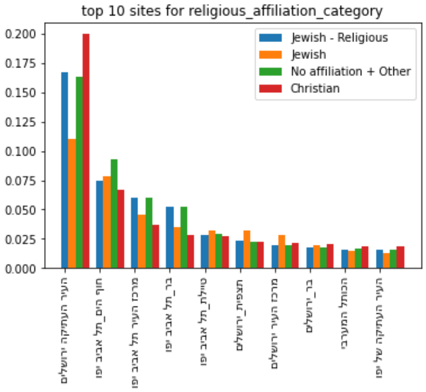
Visitors with different visit purposes seem to have different site perforations, for example the most prevalent visit purpose among old city visitors is traveling and religion, following is business and last is family & friends, while in many sites in Tel-Aviv the business tourists are more prevalent.

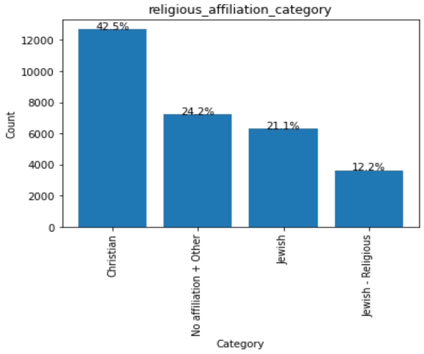




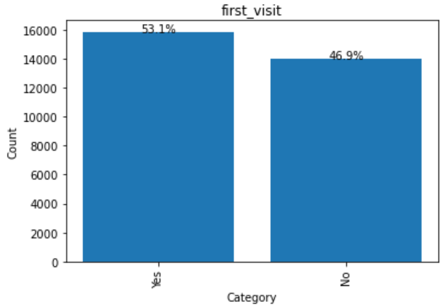
- **42.5% of the visitors are christian, 33.3% are Jewish and 24.2% had other / no religious affiliation .**

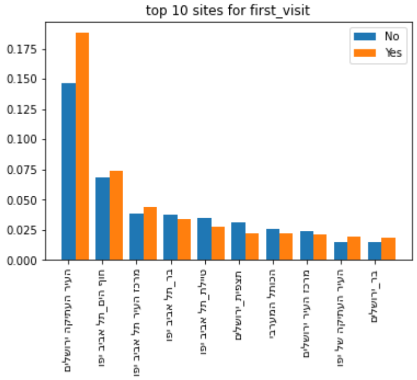
There are some differences in site preference by religion, for example: christians visit old city Jerusalem the most while non-religious jews visit it the least.



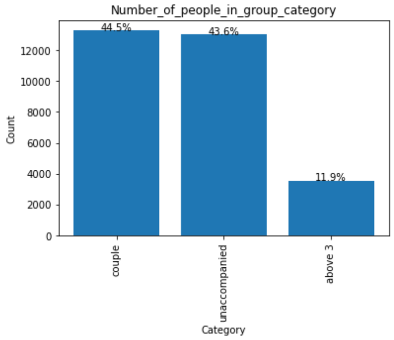


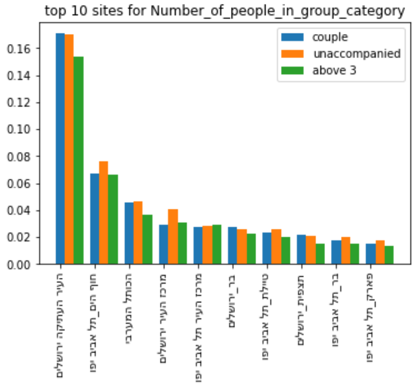
- **53.1% of the tourists were first timers which visited the old city more than returning visitors**





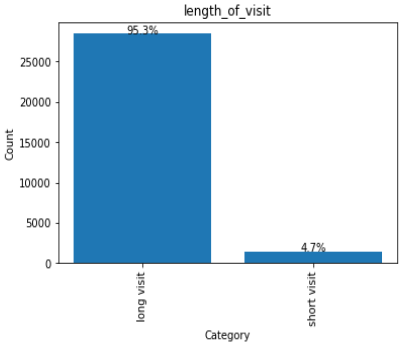
- **43.5% of the tourists arrived solo (by themselves), 44.5% arrived as couples, and 11.9% in a group of 3 persons and above.**

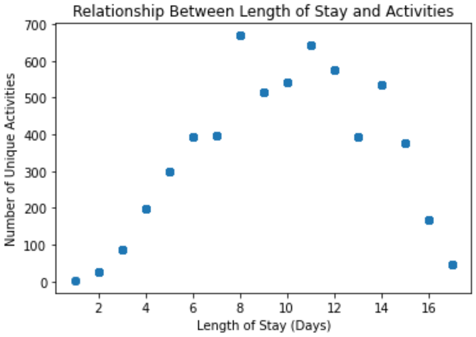




Large groups are more prone to visit the old city of Jerusalem while solo tourists are more prone to Tel-Aviv beach and the Center of Jerusalem. Solo travelers displayed more differences in site choice compared to couples and larger groups.

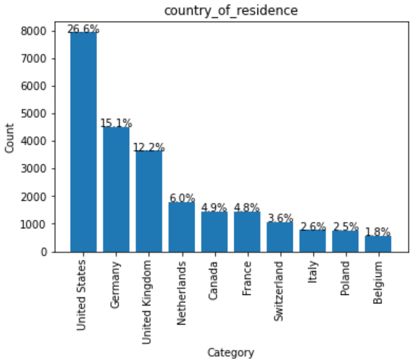
- **95.3% of tourists stay in Israel for a long visit (more than 3 nights).**

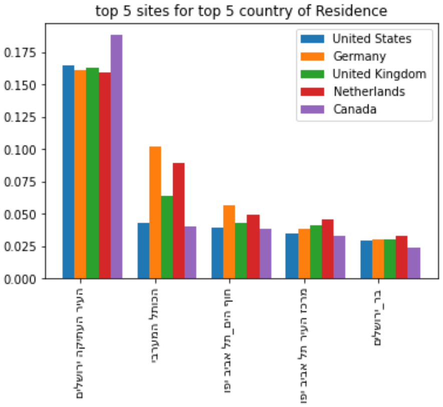




Tourists who come for a short visit, visit the top 5 sites more than the tourists who come for a long visit (refer to the appendix for relevant figures). The Correlation coefficient for the relationship between length of stay and unique activities is: 0.285 (moderately correlative) yet the relationship is not linear (there is a peak around 8-11 days of visit).

- **The vast majority of tourists came from European and North American countries with the United states and Germany being the most prominent.**

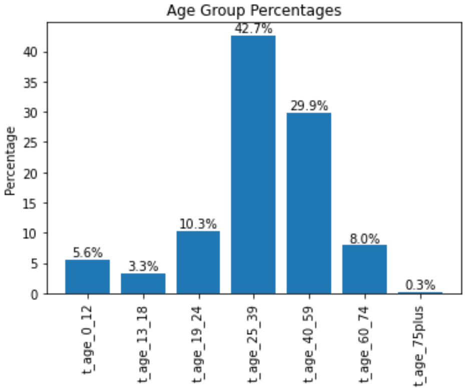




Canadian tourists tend to visit the old city of Jerusalem more than their counterparts. The Western Wall is notably visited less by Candians and Americans and more by tourists from Germany and the Netherlands. Notably the Jerusalem bar is in the top 5 but not the Tel Aviv bar.

- **72.6% of tourists are between the ages 25 and 59.**

Top 5 sites for each age group are entirely different, nonetheless, in terms of geographic areas, the Center of Tel Aviv, the old city of Jerusalem, is in the top 5 sites for 3



age groups.

## Data Processing - Machine Learning Modeling

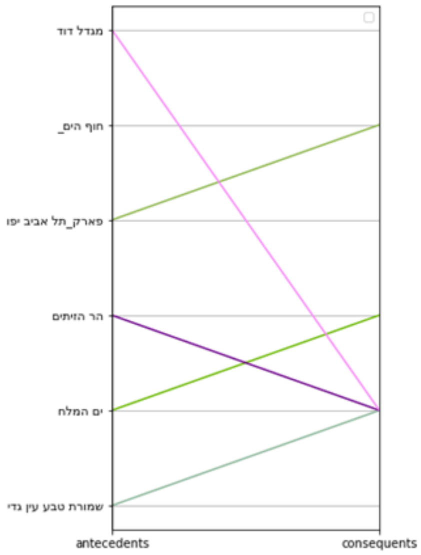
This analysis explores patterns in tourism activity data from Israel using unsupervised machine learning techniques. The goal is to gain insights into tourist behaviors based on site visits without defining categories a priori. Both descriptive models like association rule mining and clustering, as well as anomaly detection algorithms are implemented.

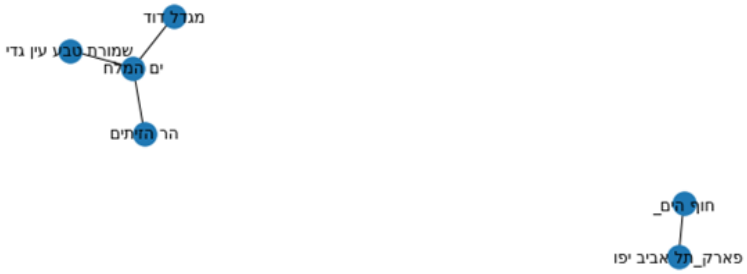
### Association Rule Mining for Pattern Discovery

Association rule mining identifies interesting associations and correlations between variables in transactional databases like purchases or website activities (Agrawal et al., 1993). Here it uncovers co-visitation patterns across tourist sites.

Three algorithms were tested on the tourism data: Apriori, FPGrowth and FPMax. Apriori uses a bottom up approach, generating candidate itemsets to test against the data. FPGrowth stores compressed dataset information in a tree structure to efficiently find frequent itemsets. FPMax starts with maximal itemsets and prunes into smaller sets. Parameters like minimum support and confidence thresholds are specified to filter results.

FPGrowth identified the most interesting rules at minimum support of 2% and confidence of 20%. It revealed strong bidirectional associations between geographically proximate sites like Dead Sea-Ein Gedi Nature Reserve, indicating visitors to one often also go to the other. Local attractions like Tel Aviv's beach and park were frequently visited together. Dead Sea visitors also commonly went to Jerusalem sites like the Mount of Olives and David's Tower. However, there were limited links bridging Dead Sea/Jerusalem and Tel Aviv attractions.Nightlife venues only showed minor one-way associations with tourist sites.

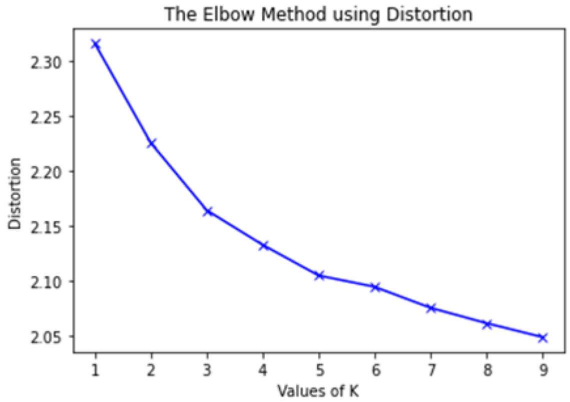


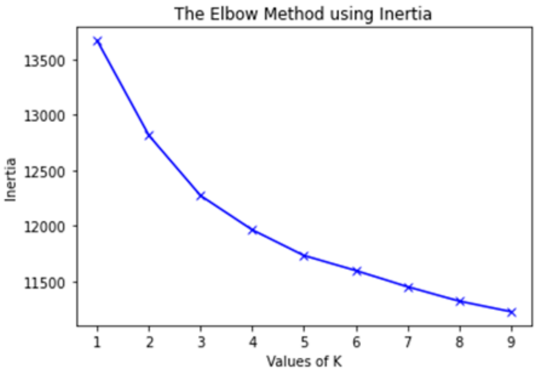


In summary, association rule mining uncovered that tourist site preferences interconnect more strongly within geographic regions than between regions. It also found that natural and cultural heritage attractions drive visitation more than nightlife. This helped us gain a base understanding of how travel itineraries work in Israel.

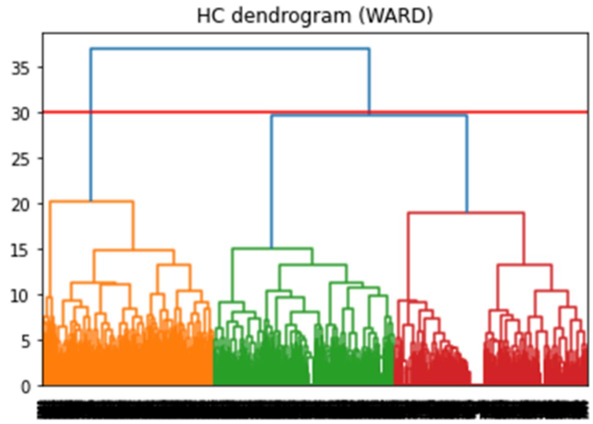
### Clustering for Segmentation

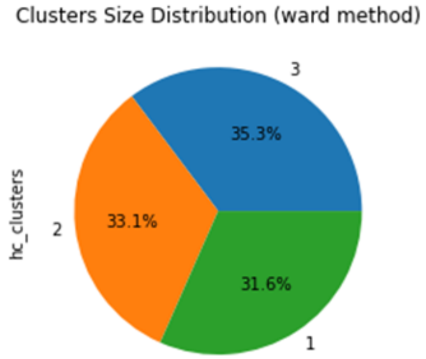
Clustering groups similar data points together based on a measure of similarity, revealing intrinsic segments (Jain, 2010). Here it clusters tourists based on their activity patterns. K-means clustering did not uncover clear groupings, as distance metrics had no elbow point.





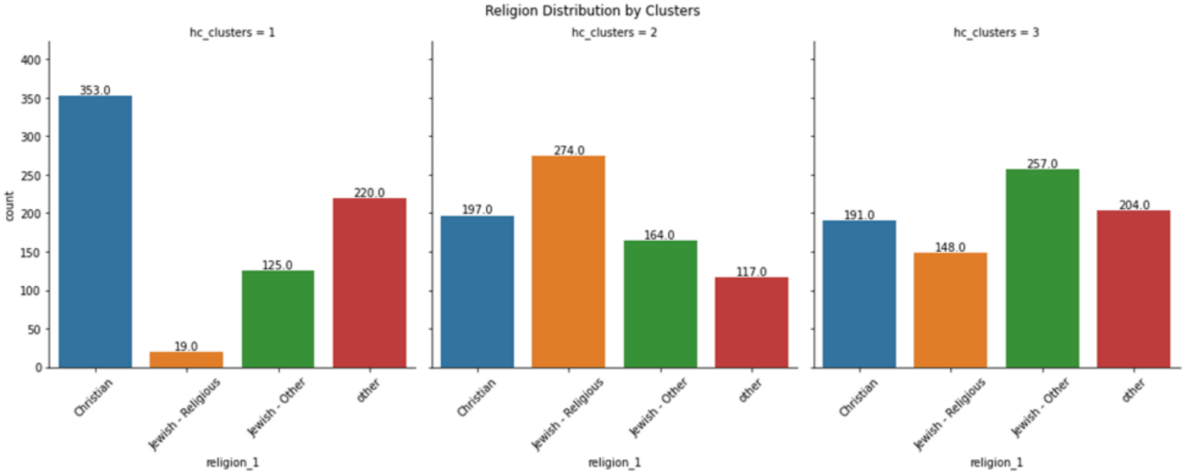
However, hierarchical clustering using Ward's minimum variance method produced well separated, balanced clusters. Cutting the dendrogram manually gave 3 significant clusters comprising 32%, 33% and 35% of the data.





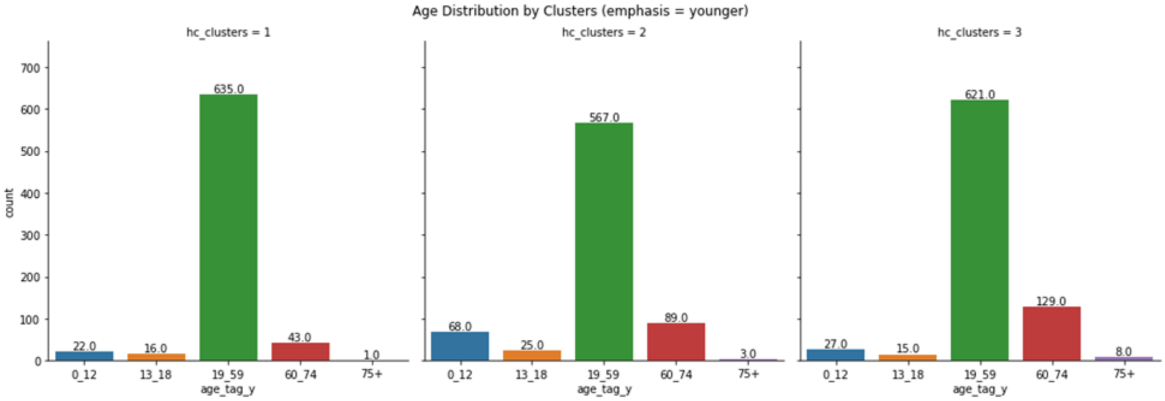
**Exploring cluster characteristics to identify tourist profiling:**

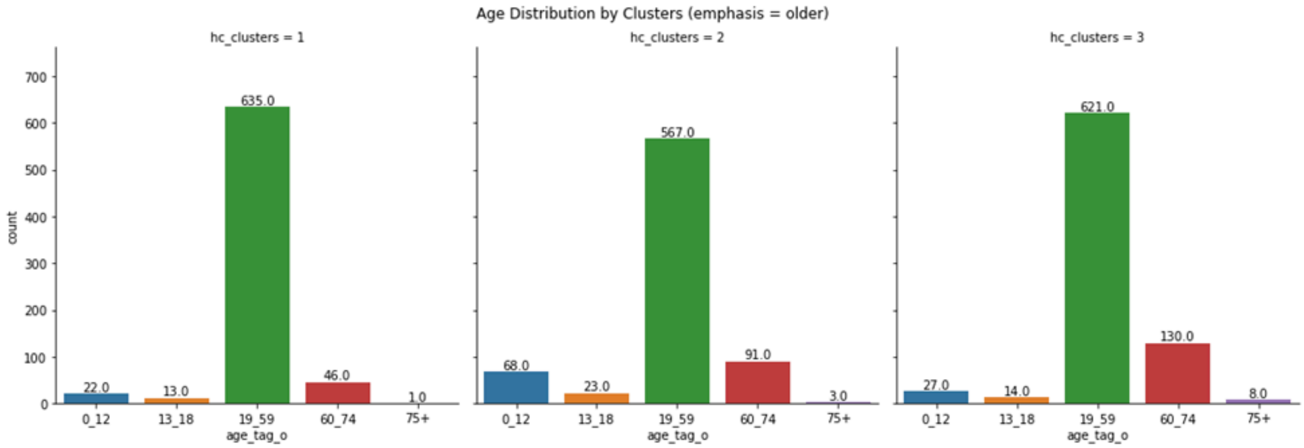
<u>Revisiting Hypothesis #1</u>: (connection between tourist's religion and their activities vector)



cluster #1 is the most common among Christian tourists (~47%), cluster #2 is the most common among religious jewish tourists (~62%) and cluster #3 is the most common among other jewish tourists (~47%).

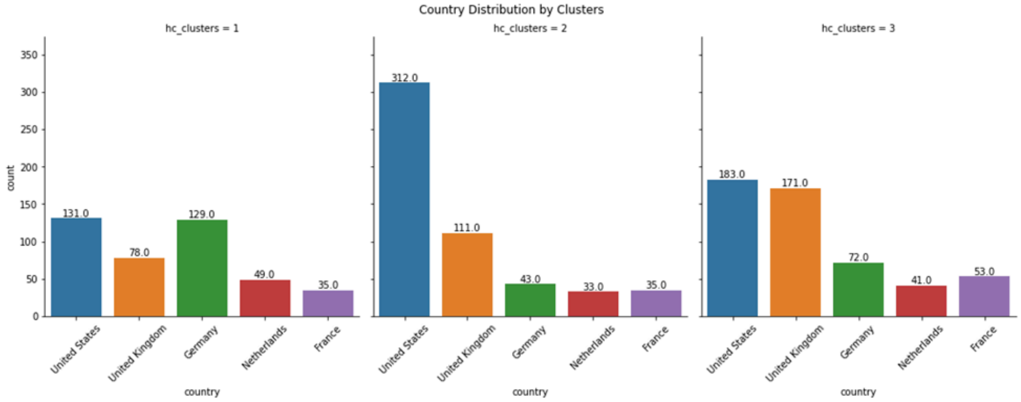
<u>Revisiting Hypothesis #2</u>: (connection between tourist's age and their activities vector)





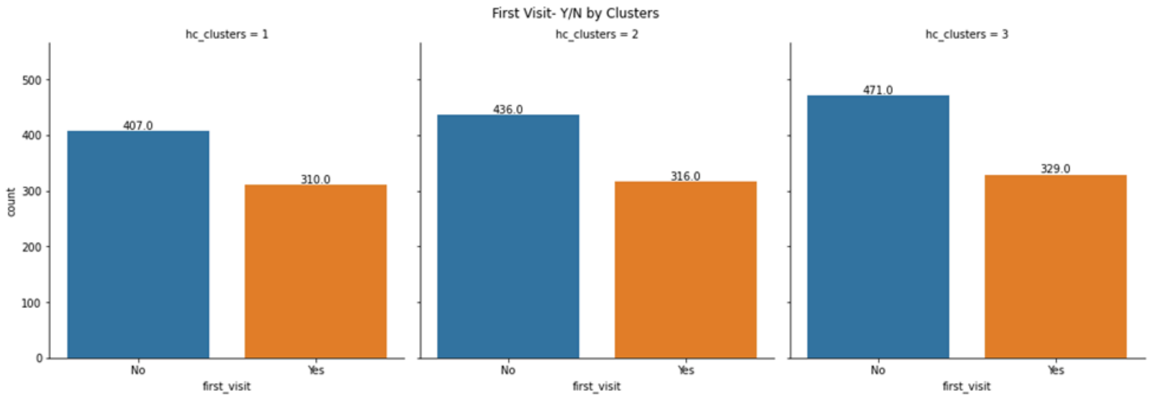
there is no significant difference between the two catplots, hence the emphasis on younger / older participants does not really make a difference.

Revisiting Hypothesis #3: (connection between tourist's country of origin and their activities vector)



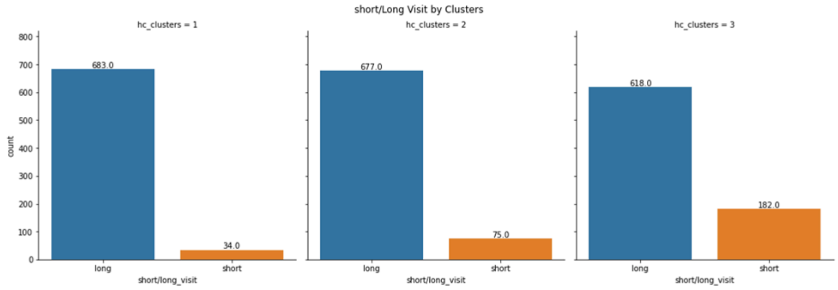
Approximately 50% of tourists from the United States are in cluster #2.
Regarding European tourists, the distribution between the clusters is quite balanced (~34% in cluster #1, ~26% in cluster #2 and ~39% in cluster #3.
These results seem to be too random to be indicative.

Revisiting Hypothesis #4: (connection between tourist's visit being first / recurring and their activities vector)



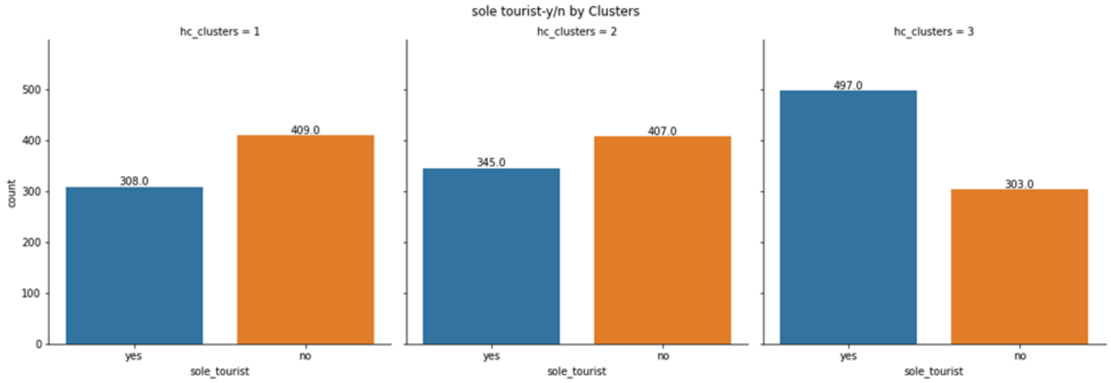
The division between first and recurring visits is quite consistent throughout all clusters without an unusual visible result.

<u>Revisiting Hypothesis #5</u>: (connection between visit duration (short / long) and activities vector)



the majority of tourists who arrive for a short visit are in cluster #3 (~62%).

<u>Revisiting Hypothesis #6</u>: (connection between group size and activities vector)

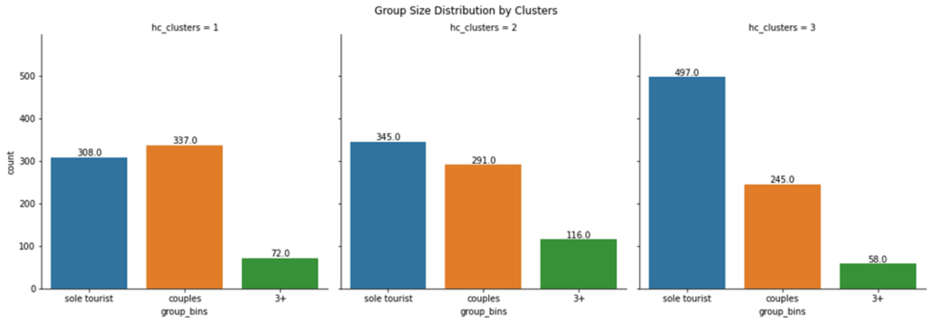


As explained above, we checked this hypothesis in two ways:

(1) Binary group size (sole tourist vs tourist with companion)

While in clusters 1 and 2 there is a similar distribution between the two groups, with a larger number of companion tourists compared to single tourists, the distribution in cluster 3 is in the opposite direction with ~62% single tourists to ~37% companioned.

(2) Sole tourist vs couple's vs tourists with more than one companion.



When examining the group size feature through a wider prism, we get more detailed information. We can see that ~43% of single tourists are in cluster #3 (The remaining single tourists are similarly divided between the other clusters). In addition, ~47% of tourists with more than one companion (the 3+ category) are in cluster #2 (the remaining single tourists are similarly divided between the other clusters). In terms of the couple's category, there is no substantial difference between the clusters.

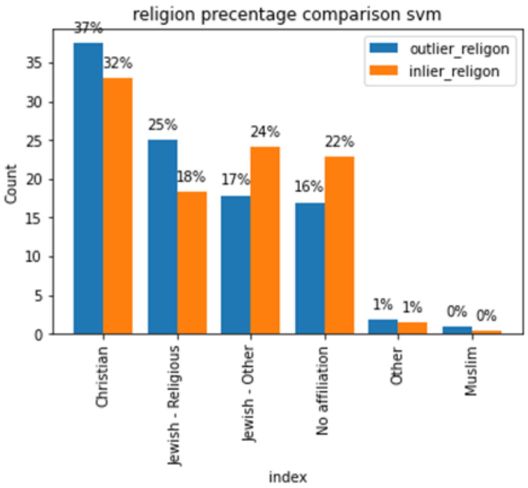
| **Hypothesis** | **Cluster #1** | **Cluster #2** | **Cluster #3** |
|---|---|---|---|
| Hypothesis #1 | Christians | Jews- Religious | Jews- Other |
| Hypothesis #2 | | Younger (0-18) participants | Older (60-75+) participants |
| Hypothesis #3 | | United States | |
| Hypothesis #4 | No specific indication | | |
| Hypothesis #5 | Short visits | | |
| Hypothesis #6 | Single tourists | More than one companion (3+) | |

### Anomaly Detection for Identifying Outliers

Anomaly detection identifies rare occurrences differing significantly from most data (Chandola et al., 2009). It has applications in detecting fraud, system faults, network intrusions and more. Here it flags atypical activity patterns.

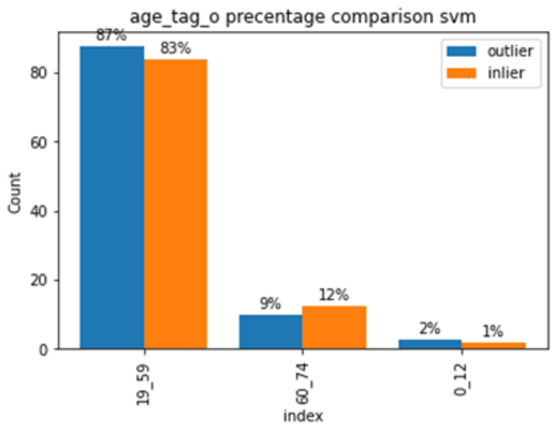
Local Outlier Factor (LOF) quantified outlier scores based on local density differences. It flagged 10% of tourists as outliers. Support Vector Machine (SVM) modeled normal points to classify outliers, identifying 5%. Isolation Forest was ineffective, finding minimal outliers. Association rule mining compared inliers and outliers. While both showed connections between proximal sites, inliers had more links between major attractions but outliers had more localized focus

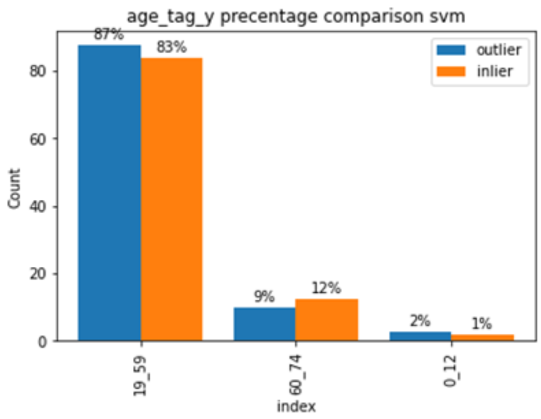
Revisiting Hypothesis #1: (connection between tourist’s religion and their activities vector)



The outliers have more Christians but less jews than the inliers.

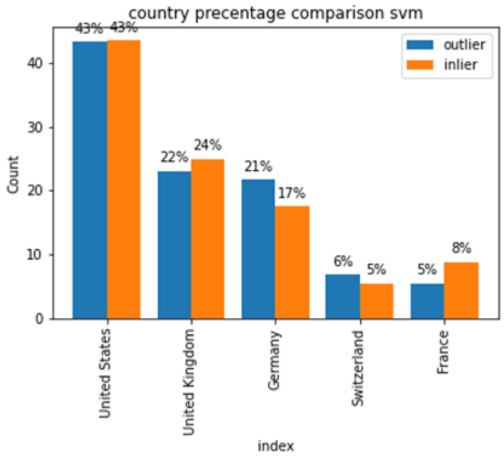
Revisiting Hypothesis #2: (connection between tourist’s age and their activities vector)We checked the age factor once with an emphasis on the youngest group participants and again with an emphasis on the oldest group participants.





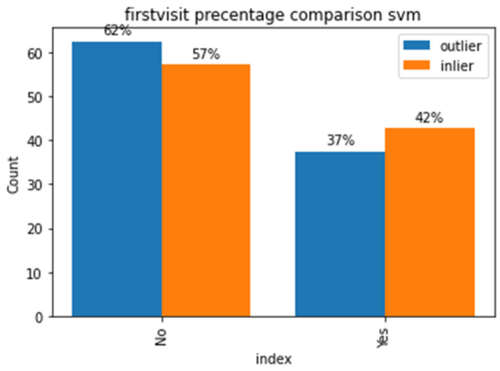
Inliers and outliers look similar in both visualizations.

Revisiting Hypothesis #3: (connection between tourist’s country of origin and their activities vector) When referring to tourists’ country of origin, we decided to examine only the five countries from which the most tourists come to Israel.



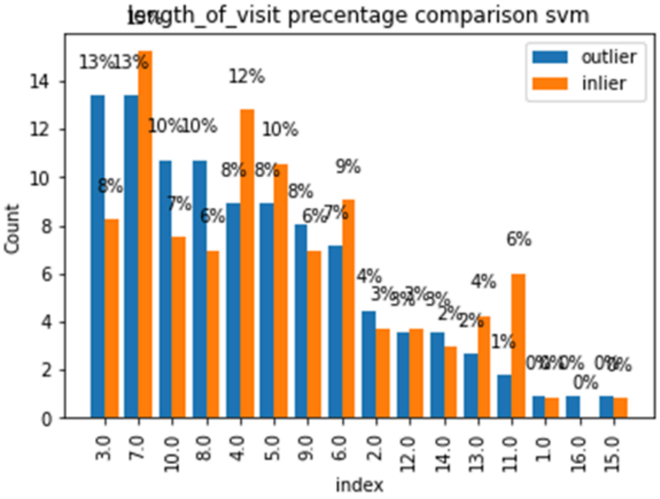
There are similar American tourists in the outlier's data and in the inliers data. The inliers have more British tourists than the outlier's data. Germany and Switzerland has more tourists in the outliers than inliers and France has more tourists in the inliers than the outliers.

Revisiting Hypothesis #4: (connection between tourist’s visit being first / recurring and their activities vector)



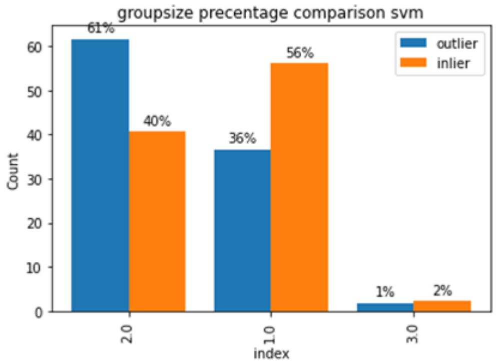
Tourists from the inliers came more for the first visit than the tourists from the outlier's data.

Revisiting Hypothesis #5: (connection between visit duration (short / long) and activities vector)

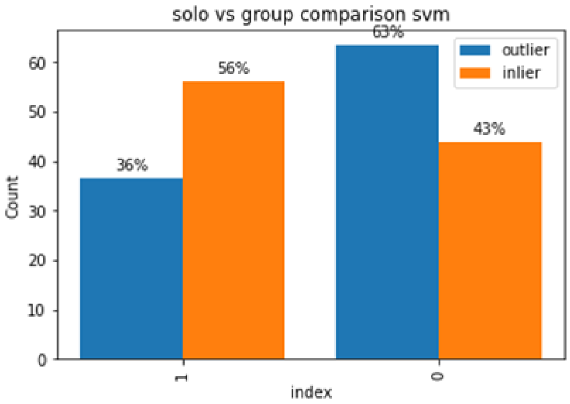


Inliers and outliers have major differences for almost any length of stay except of the outliers of 1 night and 15 or 16 nights.Tourists that stay 3 ,10,8 nights are more in the outlier's data. Tourists that stay 7,4,5,6,13,11 nights are more in the inliers data.

Revisiting Hypothesis #6: (connection between group size and activities vector)



The significant differences are in the smaller groups. There are more couples in the outliers and more sole tourists in the inliers data. There is hardly any difference between the larger groups.



Outliers have fewer solo travelers in general but have more groups. This fits with the previous findings that suggest that the outliers have more couples.

| **Hypothesis** | **Outliers** | **Inliers** |
|---|---|---|
| Hypothesis #1 | Christians | Jews- Religious and others |
| Hypothesis #2 | No difference was found | No difference was found |
| Hypothesis #3 | Germans and Swiss tourists | British, French tourists |
| Hypothesis #4 | Repeat visitors | First time visitors |
| Hypothesis #5 | Tourists that stay 3 ,10,8 nights are more in the outlier's data. | Tourists that stay 7,4,5,6,13,11 nights are more in the inliers data. |
| Hypothesis #6 | couples/groups | Solo travelers |
| Sites | Localized visits with Tel-Aviv being disconnected. | Jeusalem-Dead sea-Tel Aviv connections. |

## Discussion and Conclusions

Our study delves into the dynamic landscape of the Israeli tourism industry, leveraging unsupervised machine learning techniques such as association rule mining, hierarchical clustering, and anomaly detection. By harnessing the power of these advanced computational methods, we endeavor to discern distinct tourist types and patterns, based on the data collected by the tourism ministry in an extensive survey in 2016-2017. Through this innovative approach, we attempt to unlock valuable insights that can be utilized to inform targeted strategies, optimize resource allocation, and pave the way for personalized and tailored experiences, ultimately fostering sustainable growth and development within the Israeli tourism sector.

**Conclusions after applying Frequent itemset and association rule mining:**

1. Jerusalem and the Dead sea-Masa area are strongly connected, Tel-Aviv area activities are also connected between them.
2. Tel-Aviv and Jerusalem area activities do not have a strong connection between them.
3. The connected activities are geographically close.

**Conclusions after applying clustering algorithms:**

1. K-means clustering is not suitable for our data. We were not able to locate an "elbow point" that would indicate a suitable k for this method. Therefore, we applied hierarchical clustering.
2. By applying hierarchical clustering using Ward's method, we were able to identify three tourist clusters (types). After examining the clusters in the light of our research hypotheses, we gained several insights that distinguished the clusters from one another:
   - A significant percent of tourists who identified themselves with one or more of the following features were tagged to cluster #1:
     - Christians
     - arrived alone
     - arrived for a relatively short visit (3 nights max)
   - A significant percent of tourists who identified themselves with one or more of the following features were tagged to cluster #2:

    - religious jews
    - came from the US
    - arrived with more than one companion
    - arrived with relatively younger companions (0-18)
- A significant percent of tourists who identified themselves with one or more of the following features were tagged to cluster #3:
    - Jews (other, non religious)
    - arrived with relatively older companions (60-75+)

**Conclusions after applying anomaly detection algorithms:**

After examining the inliers and outliers identified by LOF and SVM in light of our research hypotheses:

| Outliers | Inliers |
|---|---|
| ● More Christians | ● More Jews (religious and non-religious) |
| ● More American tourists | ● More British tourists |
| ● More first time visitors | ● More repeat visitors |
| ● More couples/groups | ● More solo travelers |
| ● Visited top/general locations | ● Visited specific sites |

So in summary, the outlier analysis provided insights into different behavioral patterns and associations compared to typical inlier tourists across factors like religion, nationality, repeat visitation, group size and sites visited.

The outliers seem to represent more casual or first-time tourists focused on top destinations, while the inliers capture more experienced repeat visitors with a deeper interest in specific locations.

Although the general data showed no connections between Tel-Aviv locations and Jerusalem Locations- the partitioned data to outliers and inliers showed some degree of connection between Tel-Aviv and Jerusalem locations.

This analysis demonstrates how anomaly detection can reveal variations from expected tourist patterns, informing both marketing and product development.

**Further Recommendations (to the Israeli Ministry of Tourism)**

- Develop targeted marketing campaigns toward the key tourist segments identified, such as campaigns tailored to Christians, Jewish religious groups, Jewish non-religious groups, and repeat visitors.
- For first-time and casual tourists (outliers), focus marketing on major top destinations like Jerusalem, Tel Aviv, and the Dead Sea. Highlight "must see" locations.
- For repeat and engaged visitors (inliers), promote more niche and specific sites and experiences based on interests. Develop loyalty programs.
- Encourage development of tailored tour products and experiences catering to the needs of the different clusters and segments based on factors like religion, age, and group size.
- Invest in improving connectivity and joint packages between Tel Aviv and Jerusalem to capture multi-city visitors.
- For U.S. visitors, boost marketing and ease of transportation between key sites like Jerusalem and the Dead Sea area.
- Develop personalized trip planning tools and apps to provide tailored recommendations matching tourist profiles and interests.
- Implement a CRM system to nurture relationships with repeat visitors and identify emerging needs and opportunities.
- Recommendations for further exploration:
    1. Adding to the clustering and anomaly detection algorithms seasonality, gender and visit purpose as additional points to check.
    2. Conduct the same research on internal tourism data
    3. Conduct surveys with tourists after trips to gain direct feedback on experiences, satisfaction, pain points etc. Incorporate post-trip data.
    4. Apply time series analysis on visitation data to identify seasonal patterns and trends over time. Monitor for changes.
    5. Build predictive models using supervised learning techniques to forecast future tourist demand, especially for repeat and loyal visitors.
    6. Enrich data with third-party data sources on tourism trends, market conditions, disruptions etc. to provide greater context.

7. Incorporate  review analysis on social media to add semantic dimensions about perceptions, sentiments, interests etc.

The advanced analytics provides a powerful lens into tourist segments, motivations and behaviors. Leveraging these insights via targeted strategies and offerings will enable sustainable growth as a destination.

# Bibliography

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (pp. 207-216).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB (Vol. 1215, pp. 487-499).

Alpaydin, E. (2020). Introduction to machine learning. MIT Press.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-58. https://doi.org/10.1145/1541880.1541882

Chandola, V., Banerjee, A., & Kumar, V. (2012). Anomaly detection for discrete sequences: A survey. IEEE Transactions on Knowledge and Data Engineering, 24(5), 823-839. https://doi.org/10.1109/TKDE.2010.235

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD (Vol. 96, No. 34, pp. 226-231).

Gelbman, A., & Timothy, D. (2011). Border complexity, tourism and international exclaves: A case study. Annals of Tourism Research, 110–131. https://doi.org/10.1016/j.annals.2011.05.023

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS One, 11(4), e0152173. https://doi.org/10.1371/journal.pone.0152173

Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and Data Engineering, 26(9), 2250-2267. https://doi.org/10.1109/TKDE.2013.184

Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2011). The arules R-package ecosystem: Analyzing interesting patterns from large transaction datasets. Journal of Machine Learning Research, 12(Jun), 1977-1981.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation. Data Mining and Knowledge Discovery, 8(1), 53-87. https://doi.org/10.1023/B:DAMI.0000005258.31418.83

Israel Ministry of Tourism. (2019). Inbound tourism survey annual report 2019. https://www.gov.il/BlobFolder/reports/inbound-tourism-survey-2/he/9259\\_ENG-web%20report%202019\\_v4.pdf

Israel Ministry of Tourism. (n.d.). Incoming tourism surveys. https://www.gov.il/en/departments/general/incoming-tourism-surveys-en#:~:text=Covid%2D19%20has%20caused%20a,to%20renew%20when%20tourism%20recovers

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM

Computing Surveys, 31(3), 264-323. https://doi.org/10.1145/331499.331504

OECD. (2020). OECD tourism trends and policies 2020. https://doi.org/10.1787/6b47b985-en

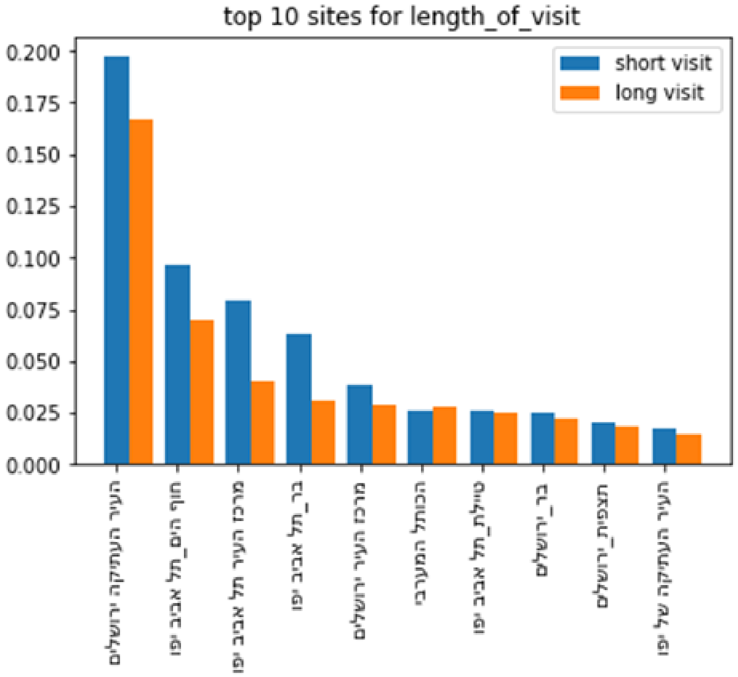
Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., ... & Kloft, M. (2018). Deep one-class classification. In International Conference on Machine Learning (pp. 4393-4402). PMLR.

Tan, P. N., Kumar, V., & Srivastava, J. (2005). Selecting the right objective measure for association analysis. Information Systems, 29(4), 293-313. https://doi.org/10.1016/j.is.2004.01.005

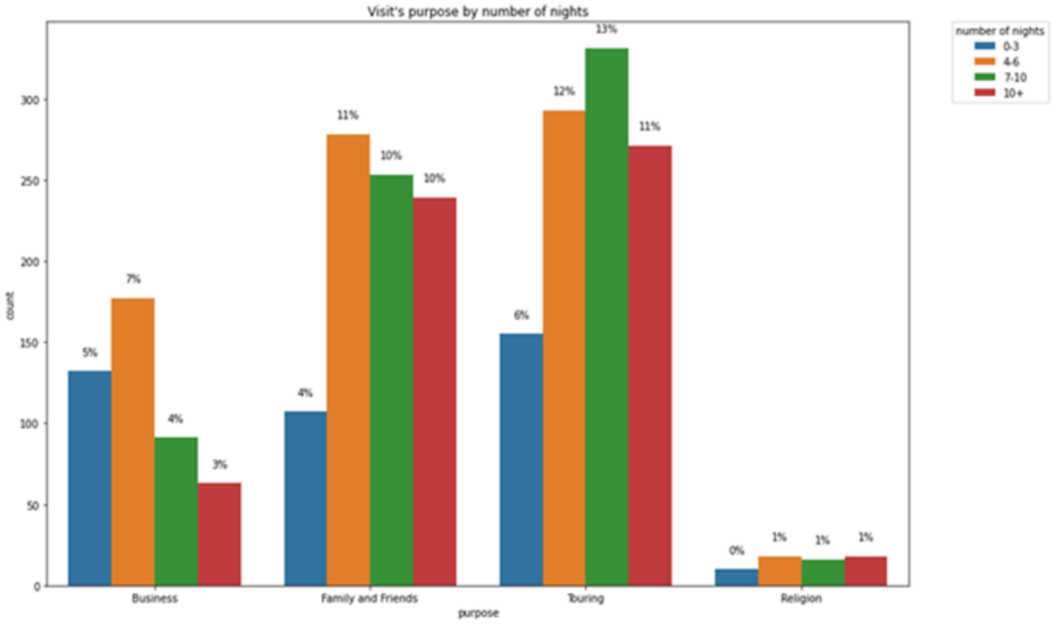
Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In International Conference on Learning Representations. https://openreview.net/forum?id=BJJLHbb0-

## APPENDIX

Length of visit frequency and Top sites by visit length



Tourists who come for a short visit visit the top 5 more than the tourists who come for a long visit.



It can be seen that longer tourism (7-10 days or more) is mostly held for tourism / visiting and family purposes.