Behavioral Patterns of Travelers

Determining Significant Factors Driving Tourism

Authors: Saxa Team #9 - Kassandra Sellers, Clark Necciai, Mabel Barba, Benny Huang, Tony Campoverde

2024-2-27

Table of Contents

[1 Executive Summary and Problem Statement 2](#_Toc159502165)

[2 Methodology 2](#_Toc159502166)

[2.1 Data Pre-Processing 2](#_Toc159502167)

[2.1.1 Variable Inspection 2](#_Toc159502168)

[2.1.2 Outlier Evaluation 3](#_Toc159502169)

[2.1.3 Correlation 3](#_Toc159502170)

[2.2 Distance Measure/Clustering Procedure Selection 4](#_Toc159502171)

[3 Unsupervised Modeling Approach 4](#_Toc159502172)

[3.1 Principal Component Analysis 4](#_Toc159502173)

[3.2 K-Means Clustering 6](#_Toc159502174)

[4 Recommendations 7](#_Toc159502175)

[5 Conclusion 7](#_Toc159502176)

[6 Appendix 8](#_Toc159502177)

# Executive Summary and Problem Statement

Our analytics team was approached by the travel organization, Travelbiz, to analyze data captured by Google reviews in 2018 regarding travel and tourism. By leveraging our key findings, their goal is to build and market optimal travel packages to Europe. Travelbiz’s primary interest is not only in identifying specific target traveler groups, but also in developing and marketing travel packages for each group.

To accomplish this goal, we leveraged two different types of analysis through the unsupervised machine learning techniques of Principal Component Analysis (PCA) and K-Means Clustering. These models guided us in determining that three distinct traveler groups exist of which to market packages. Through an extensive and in-depth analysis, our team ultimately segmented our users with cluster labels, subsequently finding insightful travel aspects to market distinct travel packages. Our primary tasks in this analysis were twofold:

* Utilize PCA and Clustering techniques to identify and describe distinct travel groups.
* Uncover, evaluate, and recommend travel packages guided by how travel groups uniquely rate locations and activities.

# Methodology

## Data Pre-Processing

We began with an introductory analysis, in which we evaluated variable distributions. Our variables were solely continuous. We combed through all 25 variables and 5455 user ratings to inspect for missing data, value integrity, and outliers. The *Gardens* variable, with a skewed distribution, contained the only missing value in the entire dataset. Our team decided that in order to maintain the maximum number of useful observations, we should impute the missing value. Due to the median being the better measure of central tendency for skewed distributions, the *Garden* variable’s median value was imputed for its missing value.

### Variable Inspection

A notable aspect of this dataset that our team had to address was the issue of user ratings of zero (0) for any given tourist destination. According to the provided data dictionary, ratings of zero indicated that the tourist destination was either not visited or was not given a review. Our team first considered dropping these reviews due to their value missingness and the effect that inclusion of these irrelevant reviews would have in skewing our unsupervised methods of variable reduction and clustering. We decided against this arbitrary approach in dropping these zero-value reviews. We instead imputed each variable’s median in place of zero valued reviews. Twenty percent (20%) of our rows contained such values and losing these rows altogether would cost a significant loss of useful information from other variables. Our decided-upon approach enabled us to retain all observations. Additionally, imputation of the median maintained the integrity of our distributions given they were all skewed.

### Outlier Evaluation

Knowing that PCA and clustering methods can be highly susceptible to outliers, we created and ran our analysis over a secondary dataset consisting of non-outlier values. Failure to approach this analysis in this additional way might have negatively affected our final interpretation of the created principal components and clustering outcomes. In conjunction with our full dataset consisting of 5455 observations, the secondary reduced dataset’s dimensionality consisted of a mere 1104 observations. Aside from the notable reduction in the number of observations, another issue with arbitrarily removing these ‘outlier’ observations was that a vast majority of the distributions no longer had high ratings. Despite our reduced dataset performing better in terms of variance captured during PCA, our team suspected and later confirmed that this approach did not yield enough distinguishable information to find disparate clusters from which to build recommendations. Evaluating the models with each dataset gave us the ability to see if the sensitivity toward outliers would affect overall performance, finding that the full dataset resulted in reliable clusters insensitive to these assumed outliers. We ultimately decided on utilizing the results of PCA and clustering resulting from use of the full dataset for these reasons. Throughout this report, we will occasionally note the disparity in performance between each dataset as a practice in thoroughness within each section but will primarily focus on results yielded by use of our full dataset.

### Correlation

In reviewing our correlation matrix visualizations, we found that the dataset with all observations did not show strong correlations among variables [Appendix-Figure#1]. Some stronger correlations that appeared within the provided correlation matrix did provide some foresight into which variables may contribute better to PCA. However, given that the majority of our variables displayed weak correlations and because PCA relies on highly correlated variables from which to build principal components, we suspected that PCA over our full dataset may not yield favorable results. However, our dataset with outliers removed highlighted many variables with strong correlations which is generally more preferred when using the PCA method.

## Distance Measure/Clustering Procedure Selection

Each of our values denotes average ratings of tourist destinations/attractions and are all continuously scaled between values greater than zero up-to and including five. When attempting to assess the similarity between two or more observations, we utilized the Euclidean distance metric. The Euclidean distance can be used to assess similarity in our analysis due to our variable data types being all continuous and having the same scale of values.

When deciding on an appropriate clustering approach, our team deliberated between two popular unsupervised learning techniques: K-Means and Hierarchical Clustering. We ultimately decided on utilizing K-Means as our choice of clustering method. Given our relatively large dataset, K-Means stood out as being the preferred method being that it scales better to larger datasets as opposed to Hierarchical Clustering. Despite the drawback that K-Means requires a predetermined *K* number of clusters, our team bounded our desired maximum number of clusters. It was limited to five to avoid having too many differing numbers of travel packages. Because of this, our search space over the optimal number of *K* clusters was reasonable. Lastly, in conjunction with K-Means, we can leverage our principal components from PCA to reduce the feature space and therefore defend against noisy clustering which K-Means is susceptible to in high-dimensional feature spaces.

# Unsupervised Modeling Approach

## Principal Component Analysis

Given our original dataset across twenty-five variables, attempting to segment our data into separate clusters may prove difficult due to the “curse of dimensionality.” This refers to cases in which disadvantages such as overfitting, data sparsity, and performance degradation occur when working with large numbers of features relative to the number of observations. Our team opted to investigate the potential benefits of applying PCA to our dataset, the main objective of which is dimensionality reduction through creation of orthogonal, weighted linear combinations of our original variables. PCA does so by identifying principal components which reflect the directions in which our original data varies the most. Applying PCA over both the full and reduced datasets yielded results in-line with our assumptions garnered from observing our dataset’s correlation matrix. Given that we saw a large degree of weak correlations among our full dataset and that PCA performs better in highly correlated input spaces, it came as no surprise when the proportion of variance explained to number of principal components was found to be sub-optimal.

We observed our principal components’ coordinates which denotes our original variables’ correlations to a given principal component. The correlational magnitude each principal component has with a variable alludes to the degree of influence each variable had in the creation of that principal component [Appendix-Figure#2]. Notably, for example, we can see that from our coordinate plot, the first principal component relates most to variables such as Restaurants, Malls, Zoos, Pubs/Bars, ViewPoints, Monuments, Gardens, Churches, and Cafes. The same type of notable relationships among variables and the *second* principal component can be inferred in the same way when observing the coordinate plot. Referring back to how our variables were correlated from our correlation matrix, we see these same relationships reappear during the formation of the corresponding principal components.

The first principal component captures the maximum amount of variation in our original input features and contains contributions from all features to varying degrees. When we observe how each variable contributed to each principal component, we find a vast majority of low-level contributions to our principal components [Appendix-Figure#3]. When we observed the contributions to our first principal component, the highest contributing variables were the same as mentioned previously, yet still had relatively low levels of impact on the creation of our principal component. This low-level of contribution across all created principal components was reflected in the less-than-desirable amount of variance captured per principal component [Appendix-Figure#4]. Being that our goal is dimensionality reduction, our hope is that our first few principal components can capture high amounts of variance. However, we found that ultimately, our outcome from PCA yielded principal components that failed to capture a large majority of the variance with a low dimensionality. After deliberation on how to proceed, we determined that we still required a greater amount of variation in the data in order for our clustering analysis to be valid. We would, however, need to be considerate as to the number of retained principal components to avoid high dimensionality. We found a fair balance of variance to dimensionality which resulted in the retention of eleven principal components which captured 75% of our data’s variance.

## K-Means Clustering

In order to effectively identify and describe segments of users based on their ratings of attributes, our team leveraged K-Means clustering as an approach to find distinct groupings of similar observations. In short, K-Means’ method of finding optimal clusters given a distinct *K* number of clusters consists of iteratively assigning observations to clusters and subsequently updating centroid positions based on mean values of observations. This continues until the positions of the centroids stabilize. This clustering technique explicitly assumes spherical clusters and may underperform in high-dimensional feature spaces because of this. Our team believes that our efforts in variable reduction through PCA will aid in attaining reliable clusters relative to the original number of features.

With the previously mentioned Euclidean Distance as our chosen distance metric, we evaluated our reduced dataset consisting of eleven principal components over a range of two to five clusters. Due in part to the nature of unsupervised clustering methodology, there is no single objective metric in determining the optimal number of clusters. Because of this, we opted to run our range of possible clusters of two to five over twenty-four separate performance indexes. Each value of *K* would be tested for each performance index to determine the optimal *K* via majority vote. Cluster assignments can be heavily influenced by the initial centroid position. Each of our cluster methods was run thirty times in order to account for this centroid randomization to attain stable results. We felt this was a significant enough number of iterations to allow for stable results. Testing revealed that a strong majority, fifty percent (50%), of the twenty-four tested indexes selected the optimal number of clusters to be three.The *Silhouette* metric*,* a commonly relied upon measurement which relays how reliable an observation’s position is within a given cluster, show that for our optimal number of clusters (three), a vast majority of our observations had a strong probability of belonging to the resulting assigned cluster.Each of the other values of *K* had an equal number of indexes (four) in their favor. Our clusters, despite being formed from relatively high-dimensionality, actually displayed incredibly strong separation of cluster labels within just two dimensions [Appendix-Figure#5]. Knowing this, alongside a strong majority of indexes supporting three being the optimal number of clusters, our team is confident that the distinct number of travel groups and corresponding number of travel packages is three.

As a practice in thoroughness to control for the possible influence of outlier values, we ran an alternative method of clustering, K-Medoids, to observe if there were significant changes in clustering assignments. As the position of the centroid changes to now be that median based, it was expected that if our K-Means clusters were heavily influenced by outliers, there would be significant shifting of cluster labels from K-Medoids. It was determined that cluster labels changed minimally, signaling additional confidence in our K-Means optimal number of clusters being three.

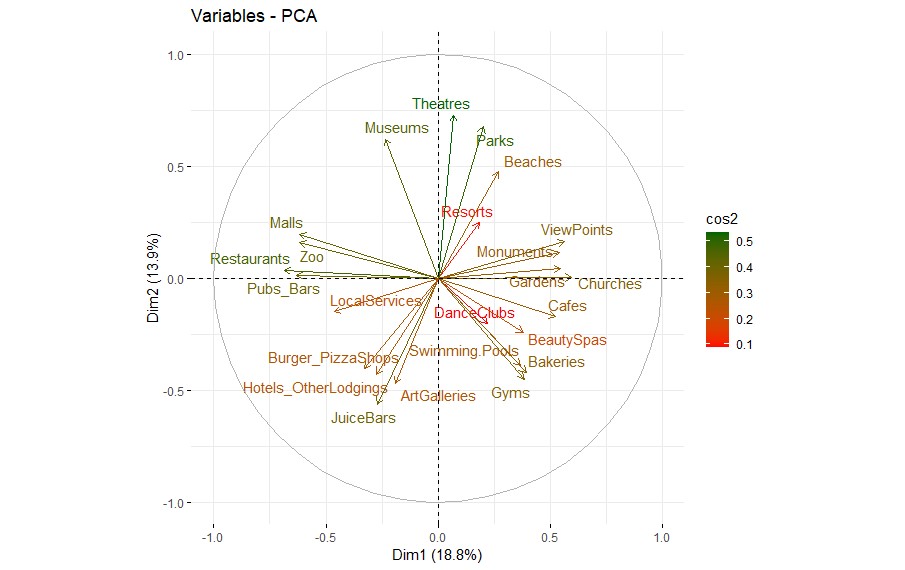
# Recommendations

* K-Means Cluster labels were mapped back to the original dataset. mapping allows for a deeper statistical understanding as to which destinations/attractions each group rates highly.

# Conclusion

# Appendix

A graph of a bar chart

Description automatically generated with medium confidence

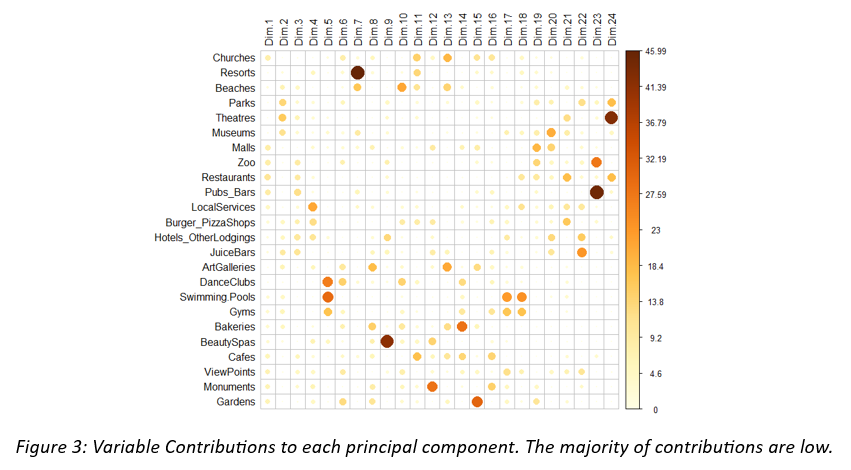


Figure 3: Variable contributions toward principal component creation generally found to be low especially in the case of the first three principal components.

Figure 1: Correlation Matrix Displaying generally weak correlations which indicates that PCA may underperform.

Figure 2: X/Y-Axis reflects correlational magnitude. Arrows/Variables with similar sign and magnitude represent the degree of correlation/relationship between them. Squared Cosine(cos2) represents variance explained for a given variable by a principal component.

A graph showing the amount of the amount of a number of individuals

Description automatically generated with medium confidence

Figure 4: Scree-Plot displaying variance contributed by each Principal Component. To retain at least 75% of the variance requires at least eleven of the total twenty-four created Principal Components.

A diagram of a cluster of red and yellow dots

Description automatically generated

Figure 5: K-means resulting in distinct clusters with minimal observational overlap due to the previous eleven Principal Component dimensions being used during clustering reduced to lower dimensional space.

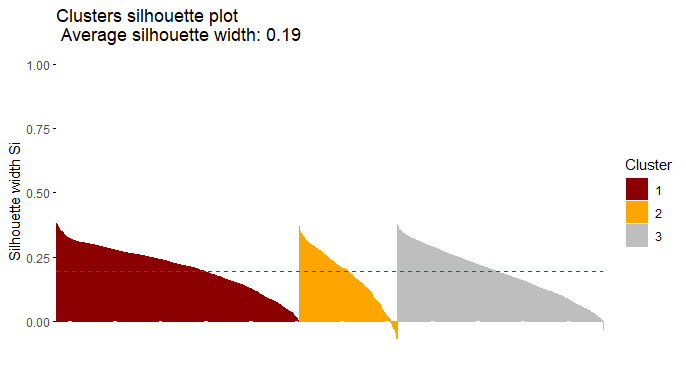


Figure 7: Silhouette metric among three clusters shows that a strong majority of observations have appropriate cluster labels.

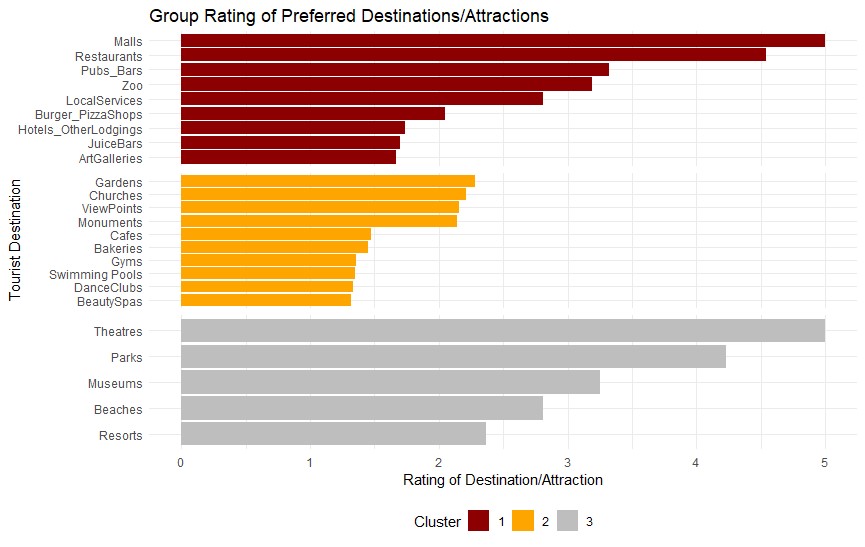


Figure 8: Bar-plot displaying highest median ratings with the corresponding target group.