Identifying and Classifying Traveler Archetypes from Google Travel Reviews

Sam Bales, Sharmin Sumy

Abstract

Grouping consumers with similar interests is important for revenue optimization. Principal components analysis, hierarchical clustering, and k-means clustering were used to identify traveler archetypes from Google travel reviews. K-nearest neighbors were used to classify the identified classes in the dataset. These prediction algorithms have high accuracy measures, but the clustering methodologies require further improvement.

1.0 Introduction

Identification of consumer subcategories is important for optimizing advertising revenue at large technology firms. Thus, the main objective of this project was to identify and classify Google users into traveler archetypes. Data used for this task were average Google reviews of various travel locations and amenities. Archetypes were then predicted using a variety of statistical learning methods based on the existing data.

The paper can be divided into three main parts. A discussion of the unsupervised statistical learning techniques used to identify traveler archetypes is presented first. The methods used to predict these archetypes are presented second. Finally, findings and further considerations for the work are detailed.

2.0 Methods

There were three main phases of the analysis. First, an exploratory analysis was conducted to understand the dataset and detect any relationships that may prove useful later. Second, unsupervised learning methods were used to understand the structure of the data and assign traveler archetypes. This phase of the project is the most critical to its success as the results of this analysis were used in later methods. Finally, supervised learning methods were built on the constructed class structures. The specific learning methods employed in each portion of the project are presented below.

2.1 Data

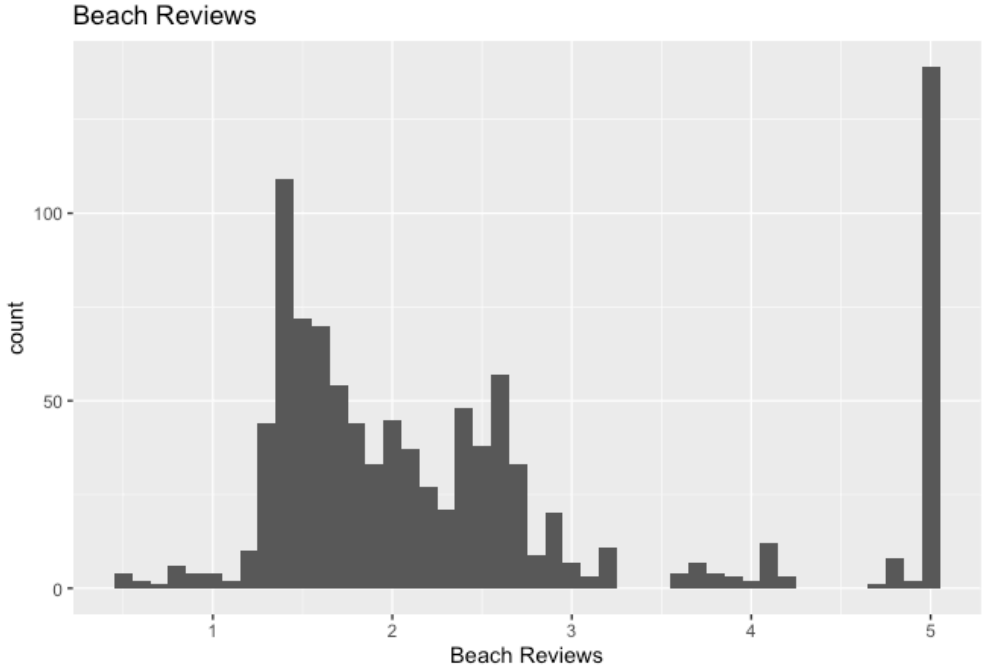
The dataset consists of average Google review from 5454 distinct individuals across 24 distinct travel amenities and leisure businesses. Each review is on an ascending scale of 0 to 5.

Table 1 in the appendix contains a complete description of variables contained in the dataset. This data was obtained from the University of California, Irvine's Machine Learning Dataset Repository.

2.1.1 Exploratory Analysis

An important question during the exploratory analysis was whether the data should be centered and scaled for use in unsupervised learning methods. Examination of univariate plots of the average reviews proved useful. Figure 1.1 shows a histogram of beach reviews, which is representative of the distributions seen across features. One can see the majority of observations fall in the range of 1.5 to 3, with a large peak at 5. A similar phenomenon can be seen in Figure 1.2. The median review for swimming pools is approximately 1.5, and the distribution has a long tail with an outlier at 5.

Figure 1: Histogram of Beach Reviews



Feature distributions are only part of the story, and bivariate plots revealed important insights. The scatterplot presented in Figure 1.3 plots reviews of juice bars versus reviews od gyms, and the relationship is similar for most features. A mild linear relationship between can be seen, but the most noteworthy structure can be seen from the margins of the plot. Note the alpha shading makes clusters of observations look darker. Thus, there are several reviewers that have an average review of 5 for juice bars, but have never submitted a review for a gym. Similarly, there are many individuals that gave a review for one amenity and not

the other as seen by the lines at the bottom and left-hand side of the plot. This does not warrant special consideration for the performance of the algorithms, but portends well-separated clusters based on missing and non-missing reviews.

Figure 2: Boxplot of Swimming Reviews

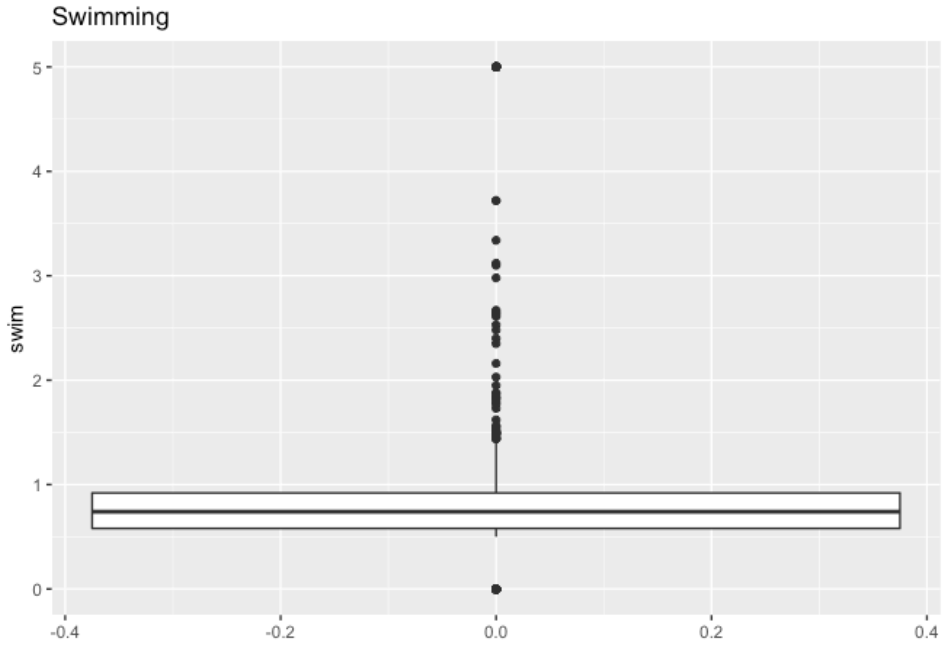
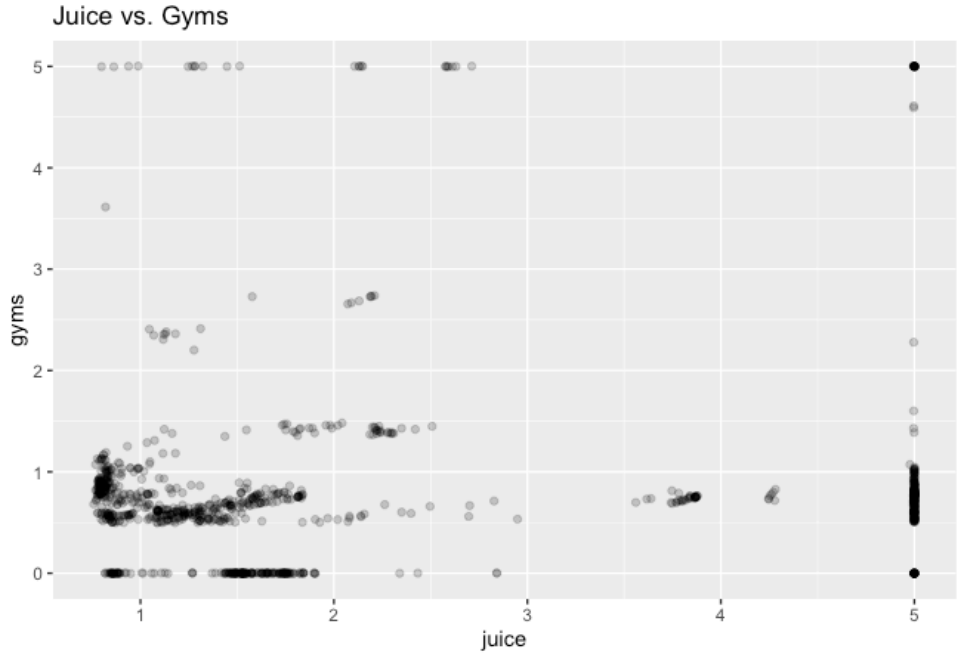


Figure 3: Scatterplot of Juice Bars vs. Gyms



2.2 Unsupervised Learning Methods

The use of unsupervised learning methods was the cornerstone of the analysis, and a variety of methods were employed. Principal component analysis (PCA) was used to understand the structure of the data, while hierarchical and k-means clustering were used to examine and classify the different traveler archetypes. Each methodology revealed distinct insights about the data.

2.2.1 Principal Component Analysis

PCA performs a singular value decomposition on the data matrix to yield a rotation matrix that contains the eigenvectors of the covariance matrix and their associated eigenvalues. The eigenvectors can be interpreted as the directions that account for the most variability in the data, which is measured by the eigenvalues. Most importantly, the matrix containing a subset of the eigenvectors can be used to construct a lower dimensional representation of the original data through an outer product with the data matrix. Dimensionality reduction is the main benefit of PCA for this problem, since it allows one to visualize the entirety of the data in two dimensions.

PCA is sensitive the variability of features and standardization is regularly used. However, the reviews are on a fixed scale of 0 to 5, and the exploratory analysis showed the variability was mostly homogenous across features indicating standardization was not necessary in this case. PCA was not used to construct traveler archetypes, but the insights were used in the clustering methods

2.2.2 Hierarchical Clustering

Agglomerative hierarchical clustering assigns observations in a dataset to a cluster based on how the observation is being linked to a cluster and the metric used to calculate the distance between the point and the cluster. Hierarchical clustering is a very flexible and customizable technique, and allowed for the exploration of many methods to create traveler archetypes, and the different methods were evaluated using dendrograms.

The following distance metrics were considered while using hierarchical clustering: Euclidean, Pearson, and Manhattan. Each distance metric has advantages and disadvantages. Euclidean distance operates as a quadratic penalty function, thereby generating very tight and distinct clusters that are geometrically “close”. Similarly, Manhattan distance calculates the absolute value of the distance between observations, and thus functions as a linear penalty function. Pearson distance considers the correlation between observations and is not a function of geometric proximity. Pearson distance will be small for vectors that are nearly parallel and large for vectors that are nearly orthogonal, regardless of their relationship in the data space.

In addition to the distance metrics, complete, average, and single linkages between observations were considered. since they have different operating characteristics. Complete

linkage creates distinct and well-separated clusters by considering the distance between a new point and the furthest point in an existing cluster and thus can be considered a dissimilarity metric. Single linkage creates sprawling, long-trailing clusters by considering the distance between a new point and the closest point in a cluster. Average linkage measures the average distance between a new point and all the points in a cluster, and creates clusters that are more sprawling than complete linkage, but less long-tailed than single linkage.

2.2.3 K-means Clustering

The second clustering method used was k-means. Unlike hierarchical clustering where the number of clusters is determined post hoc, the number of clusters to create using k-means is assigned a priori. Additionally, the performance of k-means can be monitored through within-cluster sum-of-squares, giving the user an objective measure of accuracy. For these reasons, K-means was used to identify user archetypes and assign clusters for the supervised learning phase.

2.3 Supervised Learning Methods

The identification of traveler archetypes has limited application for solving business problems. Correctly classifying individuals into the identified archetypes is critical for translating the data into actionable insights. To identify and classify Google users into traveler archetypes, a number of different supervised learning methods were investigated. These methods include Linear Discrimination Analysis (LDA), Quadratic Discrimination (QDA), K-Nearest Neighbors (KNN) and Tree-based Classification methods.

2.3.1 K-Nearest Neighbors

K-NN is a widely used and high performing non-parametric classification algorithm. The KNN classifier bases class predictions of an observation on *k* adjacent points. Adjacency—or nearness—is computed most commonly as Euclidean distance. A majority vote of the *k* neighbors is then used to assign the class. Much like PCA, hierarchical clustering, and k-means, k-NN is sensitive to the scale of the data. Any variables that are on a relatively larger scale will have a greater effect on the distance between the observations, and hence on the KNN classifier, than variables that are on a small scale. K-NN was anticipated to perform the best out of all supervised learning methods due to its similarity to the unsupervised methods used to determine traveler archetypes.

2.3.2 Linear Discriminant Analysis

LDA models the distribution of the predictors separately in each of the response classes, and then uses Bayes' theorem to flip them around into estimates. For more than one predictor, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate gaussian distribution which has a class specific mean and common variance. Class means and common variance must be estimated from the data, and once obtained are then used to create linear decision boundaries in the data. LDA then simply classifies an observation according to the region in which it is located.

2.3.3 Quadratic Discriminant Analysis

QDA is very similar to LDA, but does not assume constant variance across classes. Heterogenous class variances change the decision boundaries from a linear to quadratic, thus changing the behavior of the classifier. However, this additional complexity comes at cost. LDA is a simpler model with higher bias but less variation. QDA is a more flexible model that has lower bias but higher variance. LDA will outperform QDA when the decreases in bias are outweighed by the increases in variance.

2.3.4 Tree-based Methods

Classification trees are a highly flexible statistical learning method that can capture non-linearities and interactions present in the data. Tree based methods begin by constructing decision trees through binary recursive splitting of the data. At each split, a cut point is chosen to minimize the heterogeneity of the resulting nodes. Single decision trees are notorious for relatively poor performance. Two methods—bagging and random forests—were used to improve the performance of tree-based methods in this project.

2.3.4.1 Bagging

Bootstrap aggregation, or bagging, is a general-purpose procedure for increasing the performance of a machine learning algorithm. Bagging involves bootstrap sampling training datasets from the original dataset, fitting the statistical learning methods, and aggregating the predictions appropriately. For this project, bagging was used on classification trees, so a majority vote of the predicted classes from the fitted trees was used as the prediction.

2.3.4.2 Random Forests

Bagging can be sensitive to strong predictors that repeatedly result in the sample split regardless of the bootstrap sample, which can lead to highly correlated trees and poor performance. Random forests remedy this situation by taking a random sample[1] of the predictors at each point. Many trees are fit on bootstrapped samples and predictions are aggregated across all the models, very similar to bagging. The result is a collection of uncorrelated trees, and random forests often perform better than bagged trees.

2.3.5 Cross validation (CV)

Cross validation was used to estimate the test accuracy rate of each supervised learning model. CV involves randomly dividing the set of observations into K "folds" of approximately equal size. The first fold is treated as a validation set, and the model is trained on the remaining $K - 1$ folds of data. This trained model is then used to predict the target in the Kth fold, and an accuracy metric, ACC_1, is computed. This procedure is repeated K times where a new validation set is used during each iteration. This process results in K estimates of the test error: ACC, $ACC_2, \dots, ACC_k$. The K fold CV estimate is computed by averaging these values,

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{n} ACC_i$$

This project required classifying observations into one of several categories; therefore, accuracy, misclassification rate, and the kappa statistics were used as accuracy measures in CV.

2.3.6 Kappa Statistic

Accuracy can be a misleading metric since it is possible to make correct classifications simply by chance alone. Kappa was the preferred accuracy metric for this project, since it statistic adjusts accuracy by accounting for the possibility of a correct prediction by "guessing". The following is the formula for calculating the kappa statistic:

$$k = \frac{\Pr(\mathrm{a}) - \Pr(\mathrm{e})}{1 - \Pr(\mathrm{e})}$$

In this formula, $\Pr(a)$ refers to the proportion of actual agreement and $\Pr(e)$ refers the probability of making a correct classification purely by chance. Kappa values range from 0 to a

[1] $\sqrt{p}$ variables when building a random forest of classification trees, where p is the total number of predictors.

maximum of 1 with a value of 1 indicates perfect agreement, a value of 0 indicating no agreement, and values between 0 and 1 indicating varying degrees of agreement. Depending on how a model is to be used, the interpretation of the kappa statistic might vary. Values above 0.8 were considered acceptable for this project. Traditional metrics such as precision, recall, and specificity can still be calculated with multiple classes, but the objective of this analysis was overall accuracy, not a specific error rate.

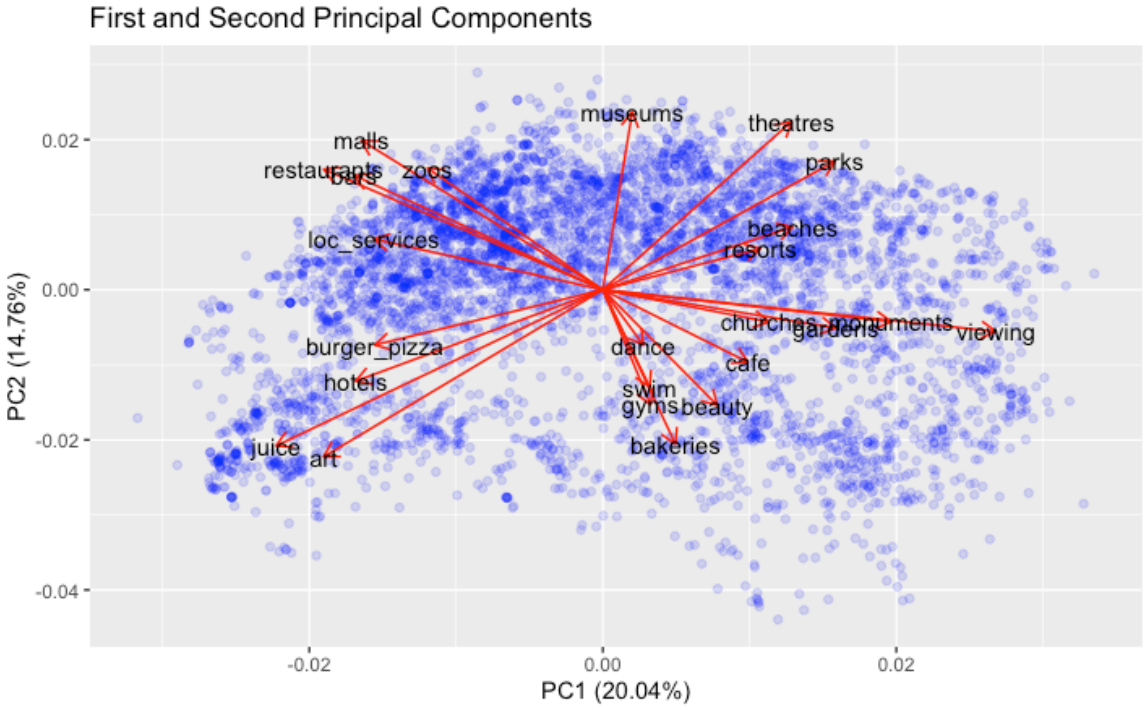
3.0 Results (classes assigned, accuracy, Kappa)

Results are divided into two parts. First, the constructed traveler archetypes are discussed. Second, the results from the supervised learning methods are presented.

3.1 Traveler Archetypes

Figure 4 shows a biplot of the two first principal components of the unstandardized reviews dataset with an overlay of the original feature vectors. Four or five distinct groups of the original data vectors can be seen in the plot.

Figure 4: Biplot of First and Second Principal Components



A summary of the average values for each of the five classes assigned by the k-means algorithm is presented in Table 2. Some variables have a wide diversity across classes, while others have less separation. For example, the average rating for zoos is similar across classes, whereas restaurants have a wide dispersion across classes. Similarly, Figure 5 shows a biplot of

Table 2: Average Archetype Values

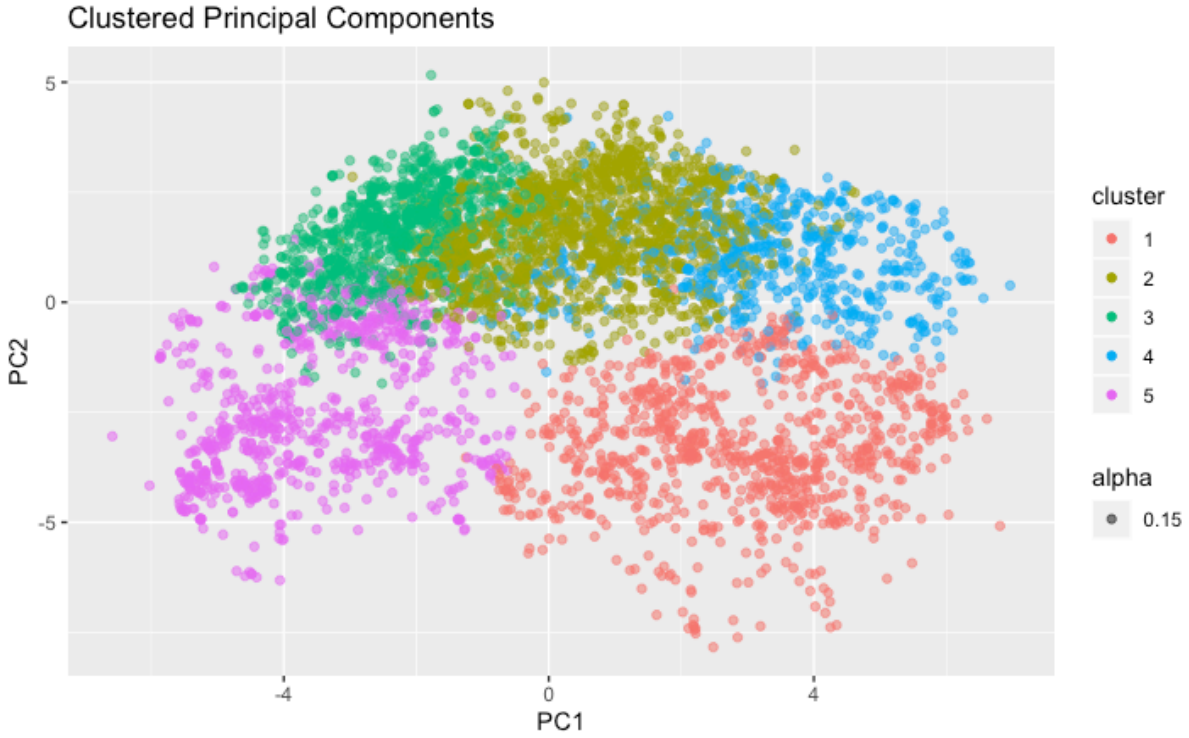
Archetype	Churches	Resorts	Beaches	Parks	Theatres	Museums	Malls	Zoos
1	1.36	2.81	2.99	3.44	4.05	3.81	3.88	2.53
2	1.17	1.96	1.77	2.05	2.16	2.71	3.99	3.71
3	1.64	2.23	3.11	4.29	4.07	3.18	2.72	2.19
4	2.34	2.70	2.50	2.23	2.05	1.88	1.94	1.57
5	0.95	1.56	1.98	1.95	2.01	2.27	3.52	2.33

Archetype	Dance	Swim	Gyms	Bakeries	Beauty	Café	Wildlife Viewing	Monuments
1	1.03	0.66	0.53	0.59	0.73	0.77	0.85	1.35
2	1.01	0.65	0.45	0.47	0.55	0.65	1.12	1.01
3	1.21	0.87	0.67	0.62	0.97	1.04	4.75	2.52
4	1.59	1.77	1.88	2.53	2.13	1.86	2.64	2.45
5	1.30	1.08	0.87	0.99	0.92	0.75	0.85	0.78

Archetype	Restaurants	Bars	Location Services	Burger/Pizza Joints	Gardens	Juice Bars	Art	Hotels
1	2.98	2.39	2.06	2.03	1.54	1.87	1.60	1.96
2	4.64	4.47	3.67	2.02	1.11	1.56	2.03	1.75
3	2.67	2.72	2.68	1.60	1.78	1.24	1.16	1.86
4	1.76	1.57	1.57	1.43	2.62	1.84	2.35	1.54
5	3.21	2.93	2.90	3.31	0.92	4.75	4.25	3.75

the first two principal components colored by class assigned by k-means. The five groups of vectors seen in Figure 4 have their own associated class, demonstrating the structure identified during PCA was reinforced by k-means clustering.

Figure 5: PCA and Clusters



K-means clustering was favored for reasons mentioned in section 2. Additionally, class assignments were different, but the overall structure was similar between hierarchical clustering and k-means. Table 3 shows the number of observations assigned to each cluster using hierarchical (Euclidean distance, complete linkage) and k-means clustering.

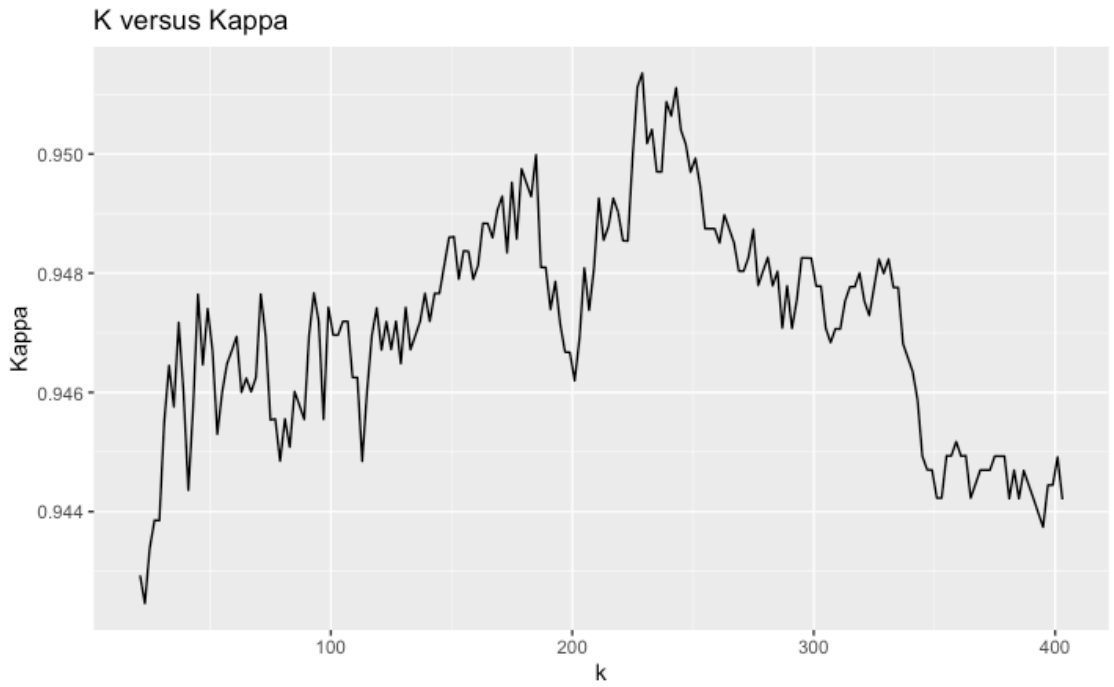
Table 3: Class Assignments

	1	2	3	4	5
Hierarchical	1805	560	1734	820	535
K-Means	1678	943	1184	740	909

3.2 Predicting Archetypes

K-nearest neighbors was the first statistical learning technique used. Figure 6 shows different values of the Kappa statistic for different values of K. The optimal number of neighbors is 229 achieving a kappa statistic of approximately 95%. This is an extraordinary level of accuracy for a classification algorithm, and does not bode well for the success of the problem.

Figure 6: K-NN Performance for Different Values of K



Linear and Quadratic Discriminant analysis were both used to predict consumer archetypes. Table 6 presents Kappa and accuracy rates for LDA, QDA and tree-based methods. Similarly to K-NN, all the included supervised learning methods have exceptional accuracy rates with kappa statistics well above 90%. Once again, this is an unbelievable level of accuracy, and the results from these methods are suspect.

Table 6: LDA and QDA Summary and Tree-based Methods,

Algorithm	**Cross Validation Kappa Statistic**	**Accuracy**
LDA	0.9478	94.4%
QDA	0.9251	94.4%
Bagging	0.951	96.2%
Random Forest	0.954	96.4%

4.0 Discussion

The success of this project relies on correctly identifying the archetype structure in the data. Alignment between PCA, hierarchical, and k-means clustering suggests the methods employed approximate the true structure. many different clustering methodologies were used to obtain classifications with similar results. Additionally, different numbers of groups were also tested without a change in the accuracy metrics. However, the accuracy rates of the supervised learning methods present several questions about the integrity of the results.

It is highly unlikely to obtain an accuracy rate above 90%, and even less likely for all tested methods to have a similar level of performance. For these reasons, the authors are skeptical that the identified traveler archetypes are robust. Additional clustering methods could be employed to test the robustness of the conclusions, but this is beyond the scope of the paper.

The classes identified above should be check by a domain expert for reasonableness. Since this in unlabeled data, it is not possible to test the model on new data. Ultimately, this model could be deployed on a small subset of customers and data could be collected on the performance of business metrics. For example, the models discussed above could be used to recommend advertisements for one group, and click-through rates could be compared to a similarly sized control group. This could be a robust methodology to validate the models.

5.0 Conclusion

Identifying consumer subgroups is essential for revenue optimization at large organizations. PCA was used to understand the underlying structure of the Google travel reviews dataset. Hierarchical and k-means clustering were used to identify consumer archetypes. K-NN, LDA, QDA, and tree-based methods were used to predict traveler classes. All supervised statistical learning methods performed extremely well; however, it is unlikely the assigned classes are correct. The models should be evaluated via an A/B test before putting them into full production.

6.0 References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer 2013. DOI 10.1007/975-1-4614-7138-7.

Travel Review Ratings Data Set. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Tarvel+Review+Ratings