Project 3

SAXA 3

Mike Johnson | Kesh Kamani | Ryan Mathis | Khushi Patel | Andrew Singh

1. Executive Summary

The marketing and product management team at Travelbiz has engaged with our data analytics team to identify consumer segments based on existing Google user ratings data of various travel destinations that can be leveraged in targeted travel packages for customers. Using the data provided by Travelbiz, the data analytics team identified four distinct consumer segments:

- **Shopper / Socializer** - This consumer tends to enjoy malls in conjunction with a night out at a restaurant or bar. They also like a trip to the zoo.
- **Cultural Enthusiast** - Theatres and museums generally rated favorable by these consumers. They also enjoy parks.
- **The Critic** - If this user had a negative experience at a destination, they can be expected to let their opinion be heard with a low user rating.
- **Juicer** - This consumer stands out for their generally high ratings for juice bars. This user tends to enjoy hotels and the art gallery, and opts for burgers and pizza versus sitting down in a restaurant.

Using the consumer segments described above, it is recommended that Travelbiz develop and implement targeted travel packages that align with each consumer segment's higher ratings. Targeted ads for these travel packages can be purchased and delivered through platforms such as Google (the source of the data analyzed) and Meta. Through a targeted marketing and product strategy developed around these consumer segments, the data analytics team is confident that Travelbiz would experience increased customer engagement and satisfaction with their travel packages.

2. Introduction

The key focus of any marketing and product management team is developing, communicating, and delivering products and services that consumers value. In the context of this engagement, the focus is to answer the following question: *What should a package look like and where should these be advertised?*

3. Methodology

3.1 Data Source

Travelbiz provided a dataset of 5,455 users and their average ratings for 24 different travel destinations. Note that the user count differs from 5,456 claimed by Travelbiz. This suggests that one user record is missing. Given that the missing record accounts for a very small proportion (<0.0%) of the data, it should not have a material impact on the analysis of this engagement.

3.2 Analytical Approach

To accomplish the main objectives of this exercise, cluster analysis was performed using K-Means Clustering - an unsupervised learning algorithm used for clustering data points into distinct groups based on their feature similarity. This algorithm partitions the data into clusters (k), where each data point is assigned based on the nearest mean.

Given that cluster analysis was performed 24 different dimensions, Principal Component Analysis (PCA) was applied. PCA is a technique used for reducing the number of variables in a dataset while preserving as much of the information as possible. It's useful for simplifying the data and reducing meaningless data that can distort analysis results.

3.3 Assumptions

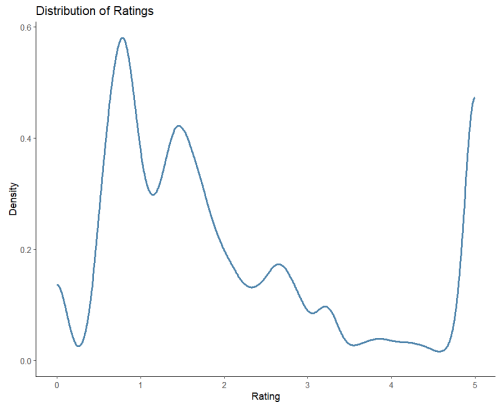
Google Review ratings range from 1 to 5. The dataset shared does include average ratings for some users that are 0 or less than 1. This could mean that the user had "not visited" or "has no interest in submitting a review". For this exercise, the latter was assumed because the intent of this exercise was to leverage what is **known** about the types of destinations the consumer values. Note that there was one user with no average score in Gardens. In this case, it was assumed that the user had no interest in this type of destination, and a value of 0 was used.

4. Data Analysis

4.1 Exploratory Analysis

An exploratory analysis was performed prior to any clustering analysis to understand some of the overall patterns for user reviews.

The density plot below shows the distribution of ratings across all destination categories. There are three distinct peaks at 1, 1.5, and 5; suggesting that users will typically leave a review if they had a strong negative or positive experience.



4.2 Model Development

This engagement focuses on identifying distinct consumer segments based on user data. To accomplish this, an unsupervised machine learning technique called K-Means Clustering. Additionally, Principal Component Analysis (PCA) was applied to address the high dimensionality of the data.

To improve the performance and ensure consistency across variables, the data was standardized by subtracting the mean value of each variable and dividing by the standard deviation. This standardization process ensures that the mean is 0 and the standard deviation is 1 for each variable in the dataset.

4.2.1 Initial K-means Clustering

Prior to the Principal Component Analysis, an initial K-Means Clustering was performed. The results of the analysis were not meaningful: 1) the optimal number of clusters (k) determined was 2 (Figure 1) and 2) contained a silhouette width of 0.15, the two

clusters are not well separated (Figure 2). It was concluded that the lack of meaningful clusters were largely due to the high dimensionality of the data, with 24 different variables being evaluated.

4.2.2 Principal Component Analysis (PCA)

To address the high dimensionality, Principal Component Analysis was performed. Based on this analysis, it was determined that three Principal Components explained 42% of the variation in the data (Figure 4).

When plotting each variable on the first two components, the contribution of each variable and their correlation can be visualized (Figure 5). A review of the contributions of variables to each principal component indicates that Restaurants and Pubs/Bars are high contributors on the first and third dimensions. In the second dimension, Theatres, Parks, and Museums have high contributions.

4.2.3 Initial K-means Clustering With Principal Component Analysis (PCA)

A second K-Means cluster analysis was performed using the first three principal components from the PCA. This model had more meaningful results: 1) the optimal number of clusters (k) determined was 4 (Figure 6) and showed an improved average silhouette width at 0.39 (Figure 7) and 3) a cluster plot showing some distinct clusters with some overlap between them (Figure 8). With this cluster analysis, distinct consumer groups were identified.

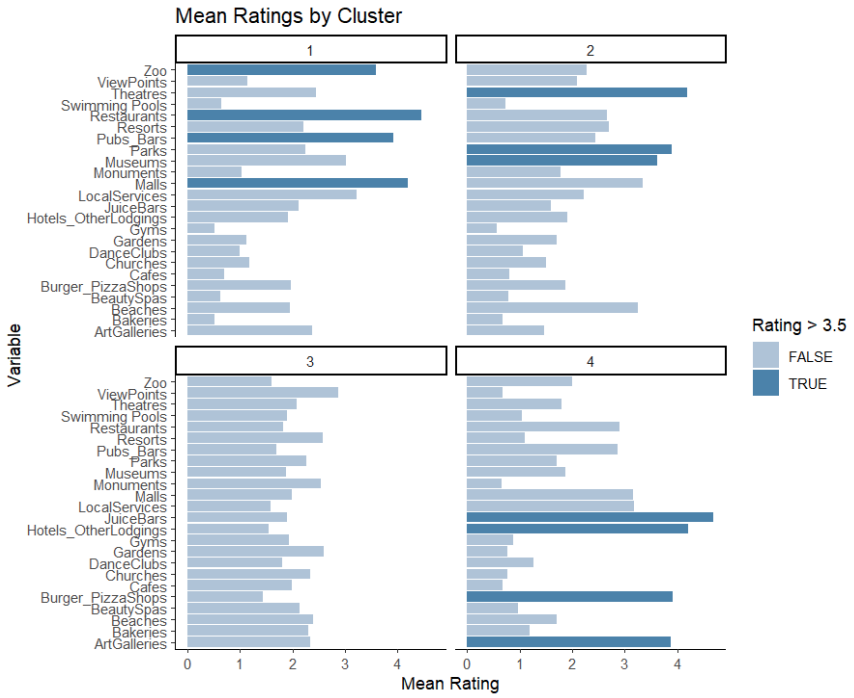
5. Results

The K-Means Clustering with Principal Component Analysis resulted in four distinct consumer segments identified:

- **Shopper / Socializer (Cluster 1, n = 1,761)** - This consumer tends to enjoy malls in conjunction with a night out at a restaurant or bar. They also like a trip to the zoo.
- **Cultural Enthusiast (Cluster 2, n = 2,058)** - Theatres and museums generally rated favorable by these consumers. They also enjoy parks.

- **The Critic (Cluster 3, n = 969)** - If this user had a negative experience at a destination, they can be expected to let their opinion be heard with a low user rating.
- **Juicer (Cluster 4, n = 667)** - This consumer stands out for their generally high ratings for juice bars. This user tends to enjoy hotels and the art gallery, and opts for burgers and pizza versus sitting down in a restaurant.

These categories were determined by measuring the mean ratings for the destination category and identifying mean ratings greater than 3.5 in each cluster.



6. Discussion

The consumer segments identified can ultimately be used to develop travel packages that align with the segment's higher rated destinations categories. For example, a travel package for cluster 1 would focus on destinations like malls, restaurants, and bar offerings. As a result, Travelbiz could capitalize by using targeted ad campaigns managed by Google and Meta, targeting user interests which are aligned with the previously discussed segments.

A limitation of the data is the lack of descriptive characteristics of individual users such as age, gender, and income which would support a deeper cluster analysis of users. Additional information about each destination location per user rating could also improve the analysis.

7. Recommendations

Based on the cluster analysis performed, it is recommended that TravelBiz develop travel packages tailored to the Shopper / Socializer, Cultural Enthusiast, and Juicer. These travel packages should align with the destination categories that they generally rated higher. Travelbiz should then develop targeted user ad campaigns using Google and Meta ad-campaign platforms. However, the Critic segment is a group of users that TravelBiz should not be invested heavily into given that their average rating for all destination categories are low. This is a group of consumers that would likely not engage with TravelBiz's offerings and/or would not enjoy the packages offered.

In the long-term, TravelBiz should invest in this user review data to identify and refine key consumer segments. Additional demographic information, and contextual details around user ratings would be helpful to inform their product and marketing strategies.

8. Conclusion

The cluster analysis revealed four distinct consumer segments. By understanding the preferences and behaviors of these segments, Travelbiz can develop tailored travel packages that align with each segment's interests, enhancing customer engagement and satisfaction. Additionally, targeted marketing campaigns on platforms like Google and Meta can effectively reach these consumers, driving increased interest in the company's offerings. While the Critic segment may not warrant significant investment, focusing on the other segments can lead to significant benefits. Moving forward, Travelbiz should continue to refine its consumer segments with more detailed demographic and contextual data to further enhance its marketing and product strategies.

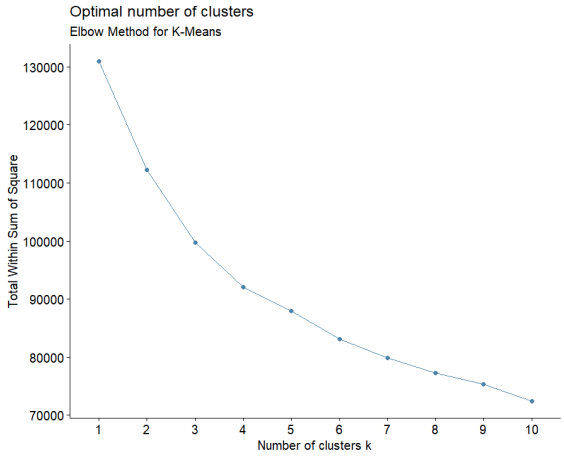
9. Appendix

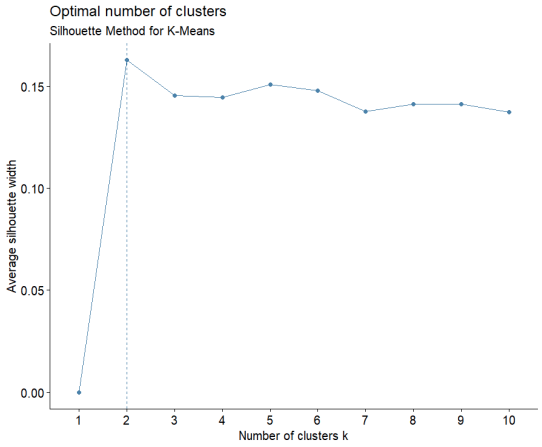
Figure 1: Optimal Number of Clusters (k) Without PCA

The plots below show the output of the optimal number of clusters for K-Means without principal component analysis. Both plots suggest 2 as the optimal number of clusters.

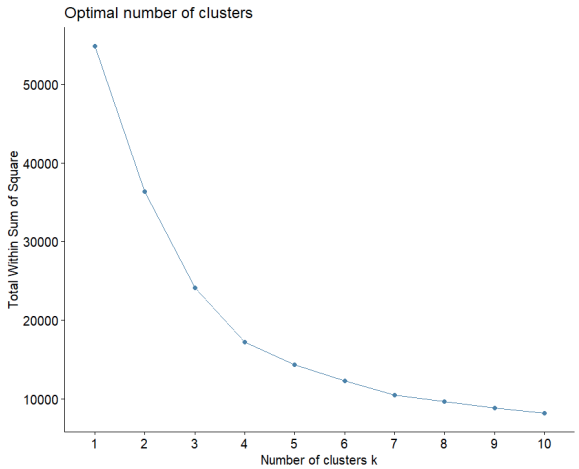
The first plot uses the Elbow Method, where the visual indicator of an “elbow” would be the optimal number of clusters. In this case, the elbow is approximately at 2.The second plot uses the Silhouette Method, where the highest average silhouette width indicates the optimal number of clusters. In this case, which is 2.

Clustering Analysis Without PCA





Clustering Analysis With PCA



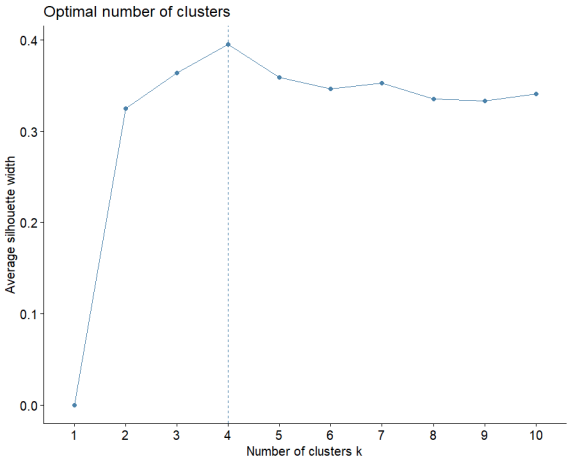
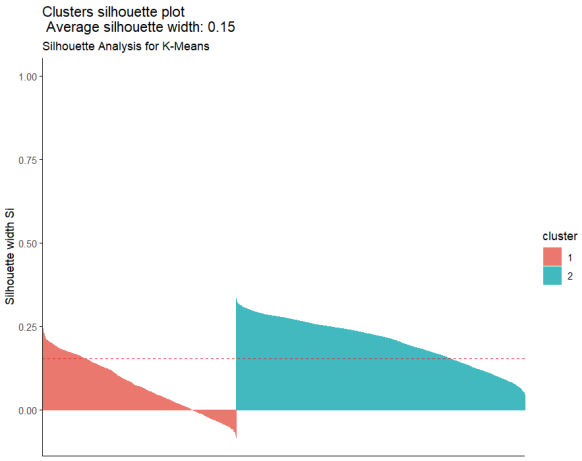


Figure 2: Clusters Silhouette Plot Comparisons

Clustering Silhouette Plot without PCA



Clustering Silhouette Plot with PCA

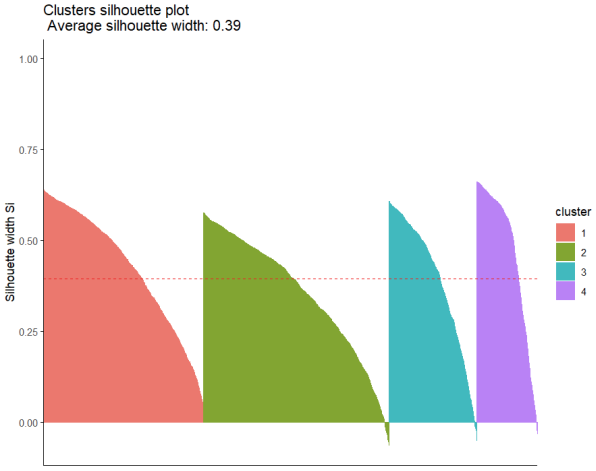
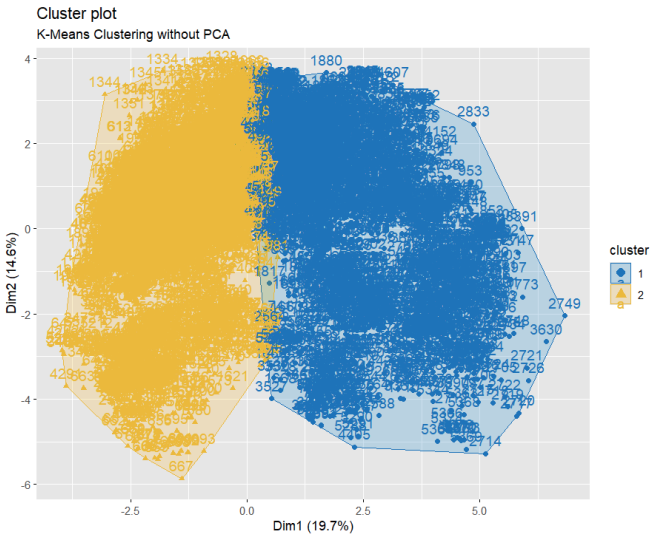


Figure 3: K-Means Clusters Plot Comparison

K-Means Clustering without PCA



K-Means Clustering With PCA

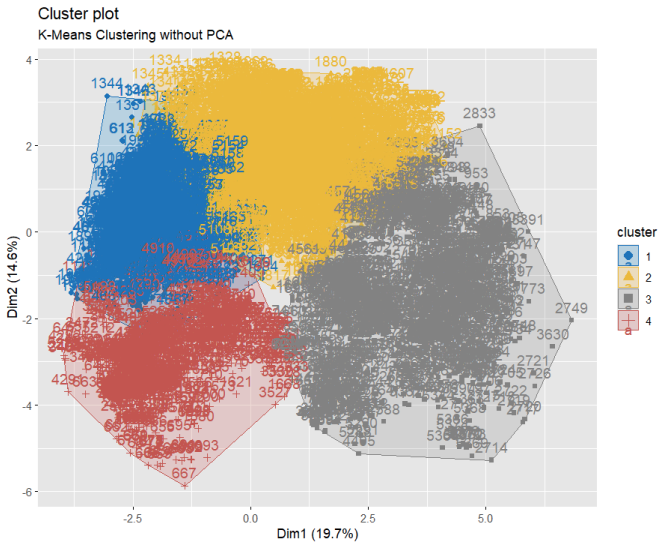


Figure 4: Principal Component Analysis - Scree Plot

The plot below is a Scree Plot from the PCA, showing the percentage of variance explained by each dimension. A visual indicator of an “elbow” indicates the optimal number of principal

components. In this case, the “elbow” is at roughly the third dimension - explaining 43% of variance.

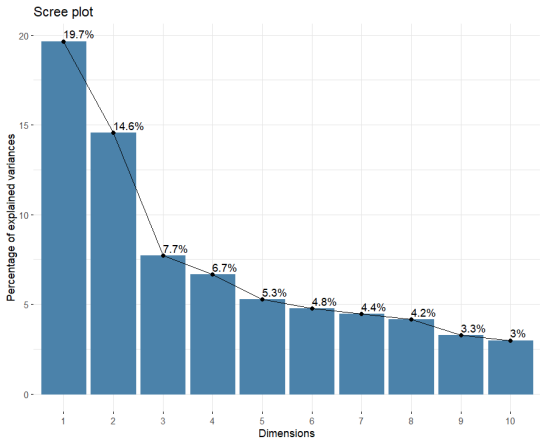


Figure 5: Variables Mapped Over Their First Two Components

Variables with a longer line indicate a higher contribution. Lines that overlap are variables that are correlated. For example, Restaurants and Pubs/Bars overlap each other, suggesting they are related.