Customizing Travel Based on User Ratings

Corinne Medeiros
July 25, 2021
https://corinnemedeiros.github.io/

Executive Summary

This study examines TripAdvisor travelers' ratings within a range of different categories of attractions throughout East Asia and aims to provide insight for customized travel itineraries and potential travel social groups. These data come from the travel and tourism domain and can help provide planning support for tour companies and travel marketing. A travel agency could adjust daily agendas based on who is traveling and what the most popular attractions are. Additionally, ratings can help in putting together future trips in other areas of the world based on similar activities available. All in all, with users' interests in mind, more informed decisions can be made when creating travel packages and advertising trip options.

To analyze the data, I explored the relationships between ratings by category and across all users with visualizations in Python, R, and Tableau. Through machine learning analysis, specifically clustering, I was able to group users into three clusters based on shared preferences.

Intro and Background of the Problem

With recommendations from friends and family, travel video bloggers, and social media, travelers are faced with seemingly endless decisions. On top of the logistics of transportation, accommodations, and timing, people have to decide where to go, what to see, what to eat, and what activities to do. To add to that, there are countless travel agencies and tour companies advertising packages and prepared agendas to ease some of that pressure, but the variety of companies out there still presents travelers with even more decisions to make. From a travel agency perspective, the competition is high, and travelers are

looking for more personalization when it comes to experiences and planning. The American Express Travel Survey in 2019 found that “92% of leisure travelers are overwhelmed by the options presented during the trip booking process” and “more than half say they would pay extra for a personalized itinerary” (Peterson, 2019). Generating insights from user data is incredibly valuable and using those insights to shape business plans and stand out among other businesses can make all the difference in becoming an option on the table for people researching travel.

The dataset from the UCI Machine Learning Repository contains one csv file with data from 2018 TripAdvisor.com reviews on destinations within East Asia (Renjith, 2018). There are 980 observations, and each user has average ratings in 10 categories including art galleries, dance clubs, juice bars, restaurants, museums, resorts, parks, beaches, theaters, and religious institutions. Ratings are on a scale of Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0).

In this analysis I am exploring the following research questions:

Which users enjoy similar attractions?

Which users dislike similar attractions?

Which users could have similar travel itineraries?

Which categories are most popular to include on all itineraries?

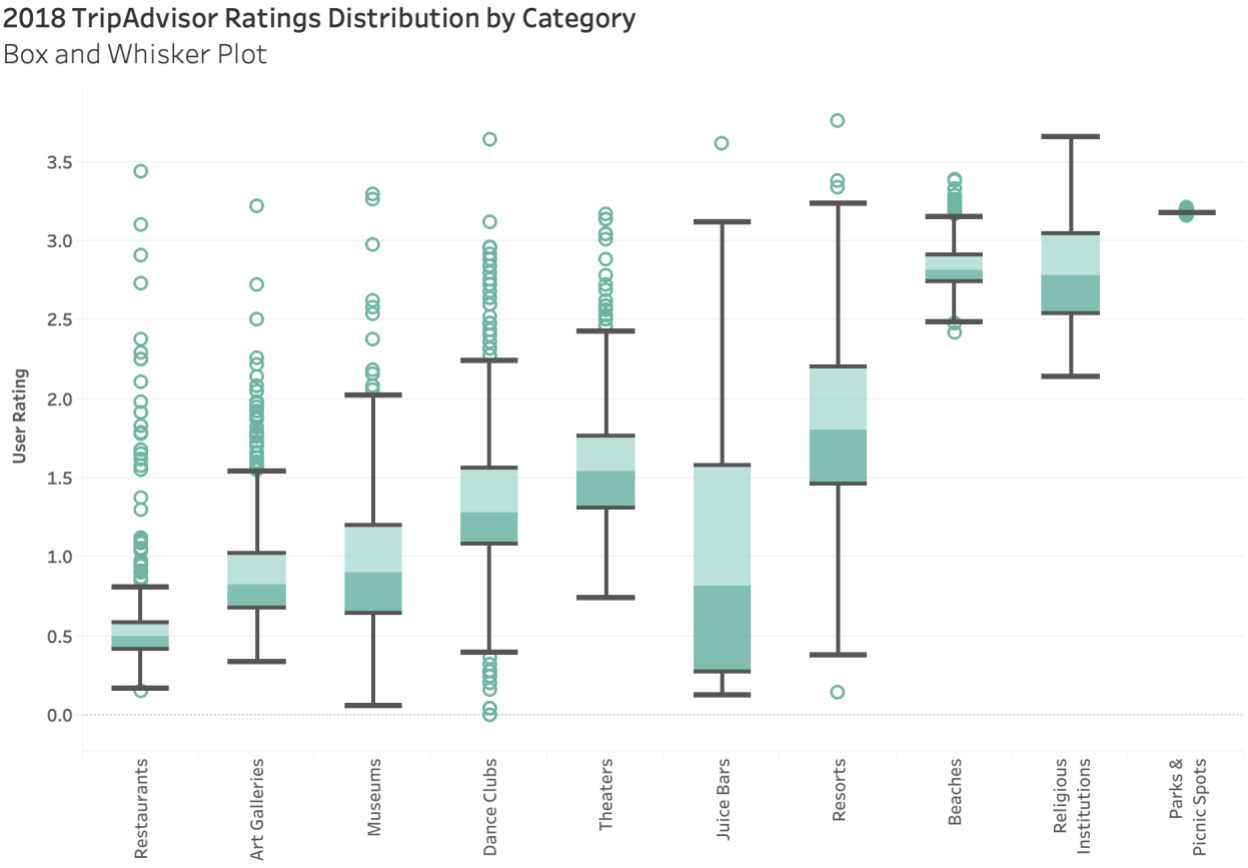
Which categories, if any, should not be included on itineraries?

Methods

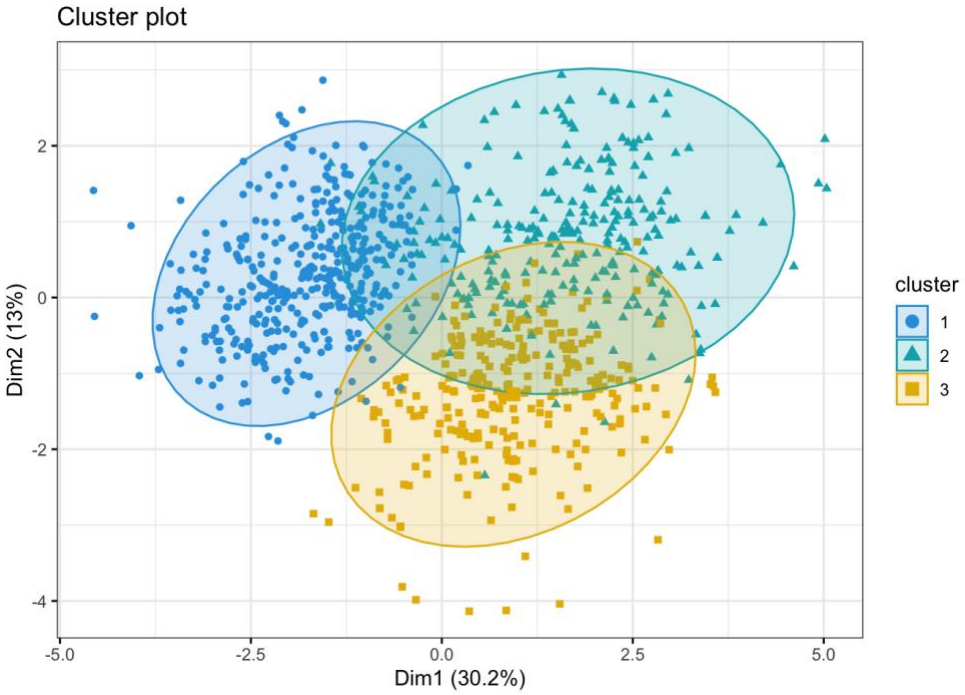
Data cleaning, processing, and initial visualizations took place with Python in Jupyter Notebook. By looking at printed data summaries I could tell that the dataset was pretty clean and didn’t have any missing values or outliers such as ratings outside of the 0-4 scale. I changed the column names to be more descriptive and set the User ID column as the index. To get a sense of patterns and trends in the data, I generated a few heatmaps and a pairs plot. The heatmaps provided some insight on category popularity, specifically that fans of dance clubs were also fans of parks, beaches, and religious institutions but not

museums and art galleries. Restaurant lovers generally had lower ratings for dance clubs. Ratings for parks, beaches, and religious institutions were high across the board. My pairs plot revealed very little significance in relationships between variables, aside from a positive relationship between resort ratings and museum ratings.

I further explored the data using Tableau to create histograms and a box plot. The box plot below illustrates the distribution trends of user ratings across all categories. Parks and picnic spots are consistently the highest, with all user ratings above 3. Beaches and religious institutions are the next most popular among users. Restaurants have the lowest average ratings of all the categories, with art galleries close behind. This would suggest that parks and picnic spots should be included in all itineraries and advertising, while restaurants and art galleries should be left off.

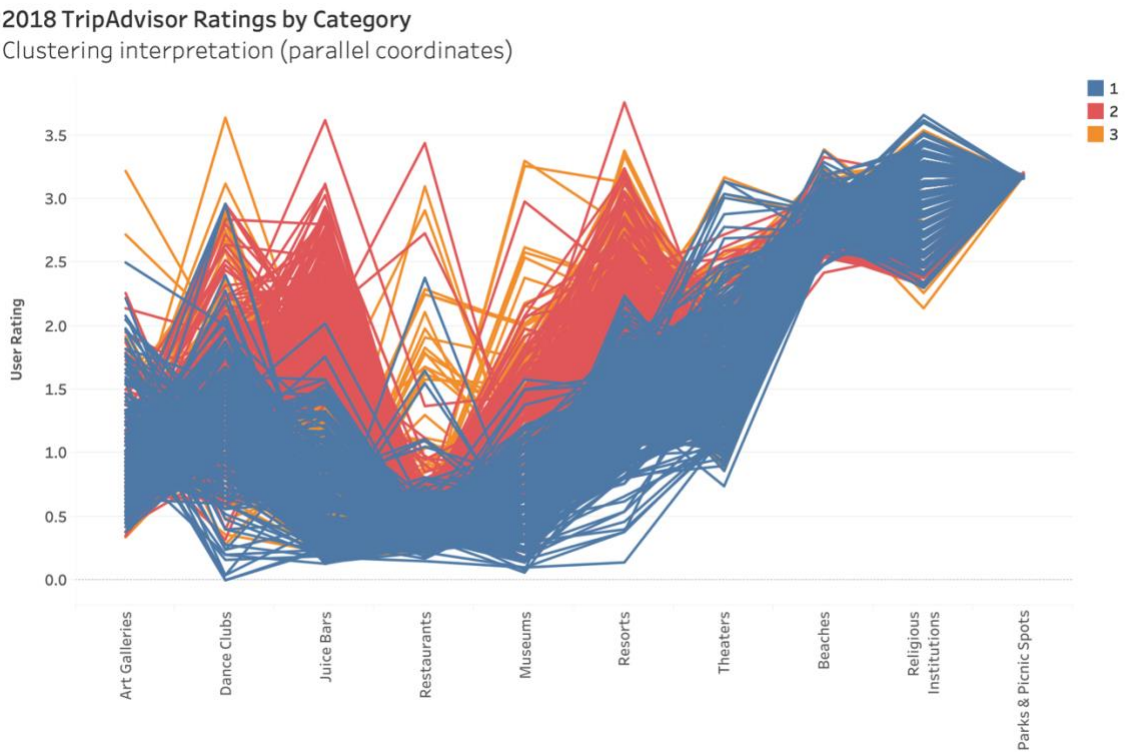


With a cleaned dataset and a better understanding of the data, I moved into R for machine learning analysis. Applying the k-means clustering algorithm to the data allowed me to find similar groups of users based on ratings per category (see Appendix for a more detailed definition of k-means). To set up the algorithm, I calculated the optimal number of clusters and visualized the results using the factoextra package. The optimal number of clusters ended up being three. In the cluster plot below, we can see the general grouping of three clusters of users, who are represented by the three differently shaped points. The percentages on the axes reveal that the most variance any of our categories accounted for was about 30 percent. This would explain the low significance originally displayed in the pairs plot. The compactness percentage, which is how similar users in each cluster are to each other, came out to 39.4%.



After studying the output of the model and clustering visualizations, I evaluated my clustering with silhouette widths and plotted them to see the average width per cluster (see Appendix for a more detailed explanation of silhouette widths). Finally, I interpreted my clusters using a parallel coordinate plot, which helped me understand the preferences of users per cluster. For final visualizations, I took my

dataset with included clustering results from R into Tableau. Below we can see the parallel coordinate plot of the three clusters.



Results

The main criteria I based my conclusions on are the compactness of clusters, and visualizing trends of the clustering as well as the variables themselves. The higher the percentage of compactness, the more similarity users in each cluster have to each other (Fonseca, 2019). The percentage produced by the k-means model was 39.4%, which isn't particularly high. The silhouette plot revealed that the average silhouette width is 0.4, which isn't very strong either. Clustering in three groups is somewhat good since there are no negative silhouette widths, and most of the values are getting close to 0.5. Ideally however, they should be bigger than 0.5 for best results (Fonseca, 2019). Looking at the parallel coordinate plot, we're able to confirm that there aren't strong distinctions between user preferences in each cluster, but there are general insights. For example, Cluster 1 users enjoy religious institutions but not dance clubs,

museums, or resorts. Users in Cluster 2 enjoy juice bars and resorts, and users in Cluster 3 generally enjoy dance clubs and religious institutions but not juice bars.

Based on the information given, I had to make some assumptions about the data. Content-wise, the categories lack specifics to help understand context. In this way, it's difficult to make detailed insights. Ratings per category could be based on many different attractions since there are many different types of museums and parks for instance. It's hard to generalize all restaurants and art galleries too. I assumed the restaurants category did not include street food since it had a lower than expected rating across all users. Based on my experience and research, I assume many travelers enjoy the street food, so that could boost the ratings if included. With no specific types of pulls mentioned regarding the data collection method, I assumed the travelers are a diverse group in terms of age, gender, and ethnicity.

With this information, a travel company could make three different itineraries with different activity recommendations for the three groups of users defined by the k-means clustering model. Travelers could also have the option of meeting and booking activities with other travelers who have similar preferences. Overall, knowing that beaches, picnics, and parks are popular across all users is important because those activities can be marketed to all potential customers. On the other hand, highlighting art galleries and museums can generally be limited to minimize risk of disinterest from travelers since those have lower ratings from everyone. Having these insights from the analysis ultimately leads to better engagement and more prepared planning.

Discussion and Conclusion

Challenges with this project include limited details on the data, and clustering accuracy. The ratings are specific to places in East Asia, but there isn't information on what the individual attractions are within each category, or which country they are in. This additional information would be ideal to have for context. The clustering algorithm was not very effective in terms of high similarity between users, so that

makes it hard to confidently cater to individual user preferences. Yet some insight is much better than none.

To take this project to the next level, I could try out different parameters for clustering, such as different values of k, and different values for nstart to see if that would improve performance. Furthermore, I could split the data into testing and training sets to experiment with a recommendation style model. If we could accurately predict which cluster a user belongs to, we could let that user know which travel agenda might align the best with their preferences, or market that particular itinerary to them to increase the chance of them getting interested in a trip right away.

With recommendation systems popping up everywhere in our daily lives, the expectation for customized experiences is growing and all industries need to keep up. Travel is no exception. People crave personalization and respond well to research results and plans that are organized with their interests taken into account. Travelers are more likely to return to a service if they had a good experience, which makes the personalization factor competitive and critical in keeping up with other companies and apps. One of the major challenges in the travel and tourism industry is the fact that there are too many choices for travelers to make. Integrating machine learning techniques with user data can help solve this problem. Overall, simplifying travel decisions for customers using personalization is the best way to meet current demands and grab travelers' attentions.

References

Choe, J. & O'Regan, M. (2015). Case Study 2: Religious tourism experiences in South East Asia. In Griffin, K and Raj, R., (ed.) *Religious Tourism and Pilgrimage Management: An international perspective*, 2nd edition, Oxfordshire: CAB International, pp191-204.

DataFlair. (2021). Clustering in Tableau – Learn the steps to perform it easily. Retrieved from https://data-flair.training/blogs/clustering-in-tableau/

Dietz, L.W., Sen, A., Roy, R. et al. Mining trips from location-based social networks for clustering travelers and destinations. Inf Technol Tourism 22, 131–166 (2020). https://doi.org/10.1007/s40558-020-00170-6

Finch, S. (2021). 7 tourism marketing challenges and how to overcome them. Hearst Bay Area. Retrieved from https://marketing.sfgate.com/blog/tourism-marketing-challenges

Fonseca, L. (2019, August 15). Clustering Analysis in R using K-means. Medium. Retrieved from https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967

Goyal, A. (2021, April 25). Introduction to K-means Clustering. MarkTechPost. Retrieved from https://www.marktechpost.com/2021/04/25/introduction-to-k-means-clustering/

Kabacoff, R. (2017). Cluster Analysis. Quick-R. Retrieved from https://www.statmethods.net/advstats/cluster.html

Kassambara, A. (n.d.). fviz_nbclust: Determining and visualizing the optimal number of clusters. Retrieved from https://rdrr.io/cran/factoextra/man/fviz_nbclust.html

Kassambara, A. (2018, June 2). K-Means clustering visualization in R: Step by step guide. Retrieved from https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

Koehrsen, W. (2018, April 6). Visualizing Data with Pairs Plots in Python. Medium. Retrieved from https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166

Koksal, I. (2020, April 4). How travel apps are using AI to personalize the experience. Forbes. Retrieved from https://www.forbes.com/sites/ilkerkoksal/2020/04/04/how-travel-apps-are-using-ai-to-personalize-the-experience/?sh=5258aa07f216

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., & Li, X. (2019). Efficient user profiling based intelligent travel recommender system for individual and group of users. Mobile Networks & Applications, 24(3), 1018–1033. https://doi-org.ezproxy.bellevue.edu/10.1007/s11036-018-1059-2

Peterson, B. (2019, December 2). More than half of travelers would pay more for personalized itineraries. Travel Market Report. Retrieved from https://www.travelmarketreport.com/articles/More-Than-Half-of-Travelers-Would-Pay-More-for-Personalized-Itineraries

Renjith, S. (2018, December 19). Travel Reviews Data Set. UCI Machine Learning Repository. [https://archive.ics.uci.edu/ml/datasets/Travel+Reviews#]. Irvine, CA: University of California, School of Information and Computer Science.

Renjith, S., Sreekumar, A., & Jathavedan, M. (2018). Evaluation of partitioning clustering algorithms for processing social media data in tourism domain. 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 127-131, doi: 10.1109/RAICS.2018.8635080.

Schroer, A. (2020, April 6). Personalized hotel booking and money for delayed flights: 9 reasons why AI is a traveler's best friend. Retrieved from https://builtin.com/artificial-intelligence/ai-travel-tech

Shukla, Y., & Jyoti, J. (2017). State of Art Survey of Travel based Recommendation System. International Journal of Advanced Research in Computer Science, 8(3).

Shvili, J. (2020, August 16). Which countries are part of East Asia? Retrieved from https://www.worldatlas.com/articles/which-countries-are-part-of-east-asia.html

Tsaih, R. & Hsu, C. (2018). Artificial Intelligence in smart tourism: A conceptual framework. International Conference on Electronic Business. Retrieved from https://pdfs.semanticscholar.org/24e9/507f17e1866bb38abaa57f7e3cde1f64be58.pdf

Uçar, T. (2021). Benchmarking data mining approaches for traveler segmentation. International Journal of Electrical & Computer Engineering (2088-8708), 11(1), 409–415.

Appendix

k-means clustering:

K-means clustering analysis a partitioning style machine learning method that uses a chosen number of clusters (k) and compares the distance of each data point to the means of each cluster in order to assign them to a group (Goyal, 2021).

silhouette widths:

The silhouette coefficient, also known as the silhouette width, is calculated by the following steps. First, the average dissimilarity between an observation and the other points in the same cluster is calculated, then the average dissimilarity between all other clusters is calculated to find the lowest value, which is the dissimilarity of the cluster closest to the observation after its own cluster. The final number comes from the difference between these dissimilarities divided by the maximum values of the dissimilarities (Fonseca, 2019).

Questions

1. What can we learn about past and future travelers using these data?
2. What further data details would help make this analysis more specific?
3. Will these insights be useful outside of East Asia?
4. What potential do these data provide in terms of future predictions?
5. What other companies are using customized travel itineraries?
6. Can this analysis be tailored to specific age groups?
7. Can we implement this analysis on newer data?
8. What further visualizations and details can you provide?
9. How can travelers find other travelers who have similar interests?
10. What kind of recommendations to users can be made with this analysis?