Trip Hacking with Google Reviews

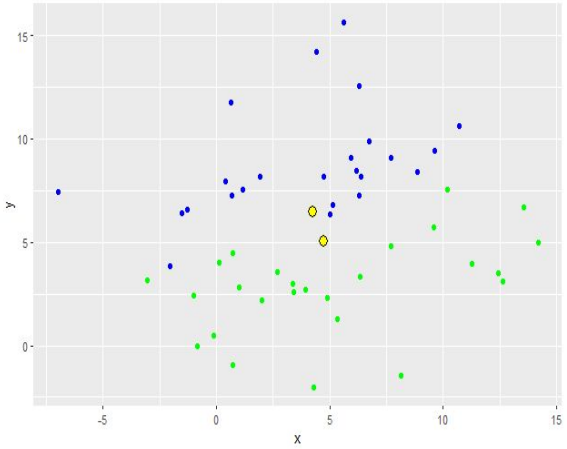
Sean Shen
Jabbir Ahmed
Abrar Ahmed

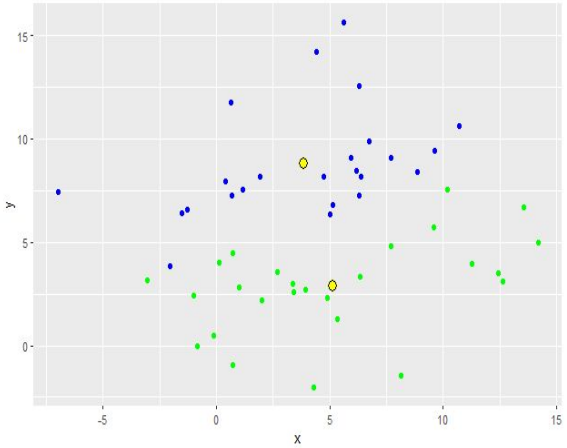
Social Science: Our Goals

- **Leverage Google Reviews** to analyze travel preferences across different categories (e.g., restaurants, parks, museums).
- **Apply K-means clustering** to segment travelers based on their preferences.
- **Categorize travelers** into distinct groups to tailor personalized travel itineraries.
- **Use data-driven insights** to predict consumer preferences for future trip recommendations.
- **Enhance travel planning** by aligning itineraries with user ratings and reviews.

K-Means Clustering

- Classifies points into clusters by their geometric distance to a centroid
- Centroid: central point of a cluster
- Iteration: group data points by centroids, new centroid from averaging points in a cluster, reassign data points again
- Final: centroids stabilize
- Silhouette score: ratio between average distance of points with current cluster and max of that and distance of points with the nearest cluster





churches	resorts	beaches	parks	theatres	museums	malls	zoo	restaurants	pubs.bars	local.services	burger.pizza.shops
1.572862129	2.253211	3.258830716	4.36280977	4.065986038	3.245777	2.585148	1.99148342	2.444555	2.5195288	2.265410122	1.424101222
1.127966752	3.121432	2.276560102	2.14611253	2.406649616	3.393159	4.686266	3.33537084	4.5085422	3.4309335	2.281445013	1.971751918
2.485465426	2.651396	2.400718085	2.14761968	1.978710106	1.790957	1.780971	1.49295213	1.6173803	1.51485372	1.527553191	1.368218085
1.190554855	1.54203	2.669230769	3.22934426	4.014174023	4.235826	4.53309	2.44498108	2.8459773	2.10175284	1.993392182	1.822383354
1.549589041	1.789909	1.965114155	2.01388128	1.941826484	1.903333	3.141826	2.11876712	2.5207763	2.33703196	2.270456621	2.256986301
0.6524	0.91562	1.59894	1.58512	1.60116	1.62136	2.931	1.95448	2.75586	2.951	3.32822	4.28952
1.826903073	2.407305	2.53141844	3.89539007	4.233995272	3.013712	3.315887	2.95416076	3.2047045	3.11028369	3.45179669	2.813120567
1.229709794	1.558549	1.698754534	2.15899637	2.244195889	2.56717	3.702624	3.69565901	4.5597098	4.53887545	4.127363966	2.030205562
1.51765812	4.355214	3.932	3.8197094	4.071880342	3.30988	2.702085	2.22471795	2.7284957	2.49635897	1.810615385	1.701982906
hotels.other.lodgings	juice.bars	art.galleries	dance.clubs	swimming.pools	gyms	bakeries	beautyspas	cafes	view.points	monuments	gardens
1.091797557	1.047784	1.186492147	1.26375218	0.859790576	0.646492	0.610175	1.07551483	1.1313962	4.80762653	2.280383944	1.743211169
2.020319693	3.169795	3.935984655	0.72943734	0.650869565	0.656458	0.690448	0.81893862	0.826087	0.87644501	0.891662404	1.023222506
1.480505319	1.793045	2.117513298	1.47069149	1.547393617	1.710412	2.426662	2.32696809	2.0932447	2.966875	2.735465426	2.852792553
1.790542245	1.726293	1.403972257	1.00238335	0.678095839	0.496003	0.597629	0.60041614	0.6981967	0.8268348	1.157187894	1.281929382
2.311780822	3.324612	4.130684932	4.08826484	3.774429224	2.559041	1.589315	0.68141553	1.1613242	1.52118721	1.517899543	1.549452055
4.53438	4.9108	3.88002	0.7918	0.80278	0.93534	1.35624	1.08302	0.58454	0.5652	0.54742	0.65346
4.05250591	2.194137	1.373002364	0.90692671	0.926122931	0.749905	0.754988	0.72995272	0.7562648	2.55328605	3.012576832	2.78427896
1.729032648	1.363906	1.35185006	1.09041112	0.644256348	0.360508	0.329045	0.4314994	0.5778476	1.20623942	1.025405079	1.186287787
1.603247863	1.556513	1.752786325	1.25300855	0.550769231	0.47735	0.821026	1.0477094	0.8515385	0.89003419	1.099538462	1.240324786

Each row is a centroid of a cluster in a 24-Dimensional Space

Average rating profile- overall preferences of users to different attractions, archetypes of users

Cluster 1: Culture Seekers

Key Preferences:

- Museums: 3.24
- Parks: 4.36
- Theatres: 4.06

Ideal Itinerary:

- Tours of historical sites
- Famous parks
- Live theatre performances

Cluster 5: Nightlife & Fun Seekers

Key Preferences:

- Dances Clubs: 4.08
- Swimming Pools: 3.77
- Spas: 2.33

Ideal Itinerary:

- Nightlife-focused trips
- Clubs
- Beach parties
- Recreational pools and spas

Cluster 7: City Sightseers

Key Preferences:

- Theatres: 4.23
- Parks: 3.90
- Restaurants: 3.20
- Monuments: 3.01

Ideal Itinerary:

- A city-focused tour
- Visit famous landmarks, monuments, cultural centers, and local restaurants

Cluster 9: Outdoor Adventurers

Key Preferences:

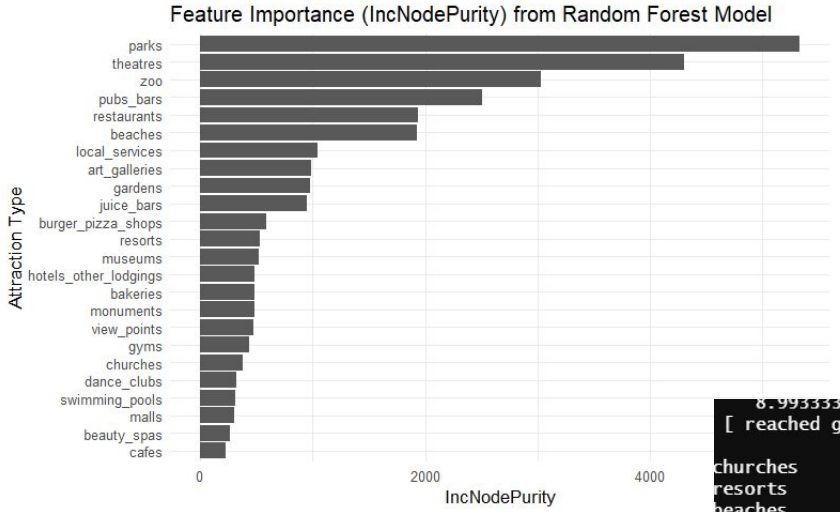
- Resorts: 4.35
- Beaches: 3.93
- Parks: 3.82

Ideal Itinerary:

- A nature and adventure-based itinerary
- Stays at resorts
- Beach activities
- Visits to parks

Random Forest

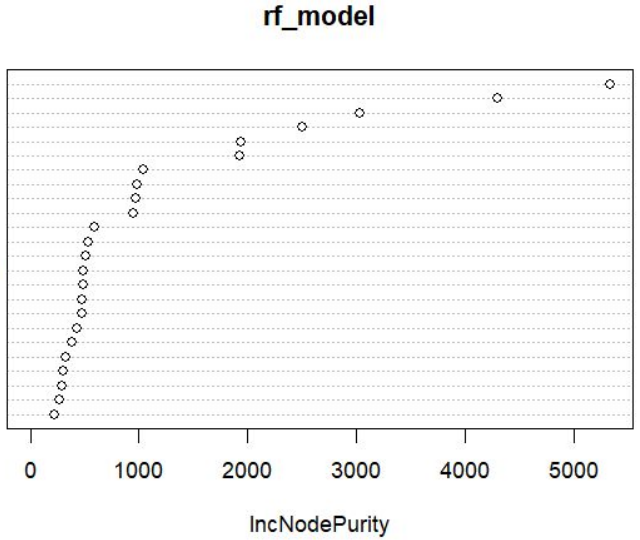
- **Random Forest is an Ensemble Method**: Random Forest combines multiple decision trees to create a more accurate and stable model for making predictions. Each tree is built on a different subset of the data, allowing the model to learn from diverse perspectives and reducing the risk of overfitting, which can occur when relying on a single tree.
- **Mechanism of Operation**: The algorithm operates by using a technique called bootstrapping, where random samples of the data are taken to build individual decision trees. Each tree provides a prediction, and the final output is determined by averaging the results (for regression) or majority voting (for classification), thus providing a robust decision-making process.
- **Importance of Feature Selection**: One of the key advantages of using Random Forest is its ability to assess the importance of different features or variables in predicting outcomes. In our project, this helps identify which types of attractions (like parks, nightlife, or museums) significantly influence user ratings and preferences, allowing us to focus on what matters most to visitors.
- **Understanding IncNodePurity**: IncNodePurity is a critical metric derived from the Random Forest model that quantifies the importance of each feature by measuring how much each feature contributes to the model's predictive accuracy. Higher values indicate that a particular category of attractions has a greater impact on user decisions, helping us to prioritize our focus on the most influential factors



Graph 2

RF Model

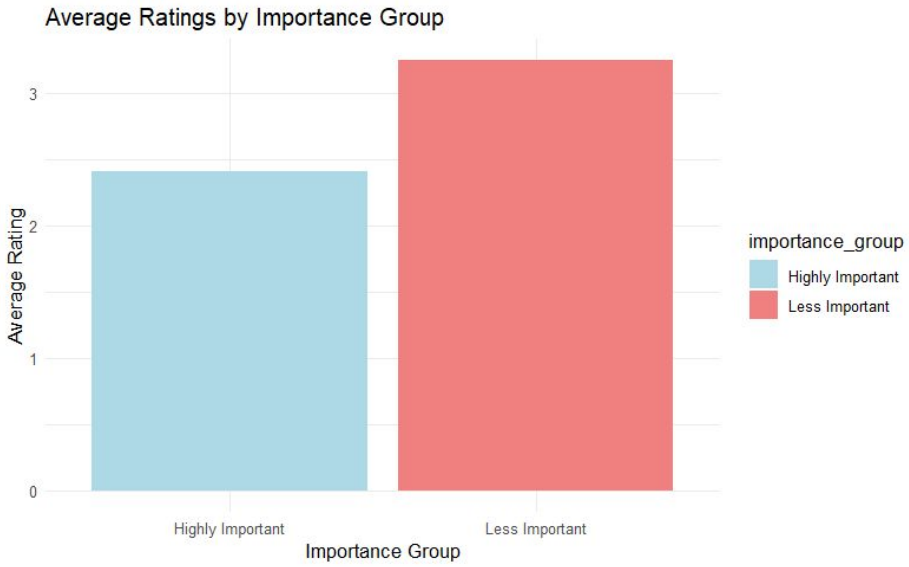

```
 [ reached getOption("max.print") --
                          IncNodePurity
churches                        381.6302
resorts                         530.1468
beaches                        1929.6702
parks                          5334.6368
theatres                       4304.3620
museums                         516.2777
malls                           297.1897
zoo                            3032.7217
restaurants                    1934.8890
pubs_bars                      2507.0687
local_services                 1041.5066
burger_pizza_shops              590.2555
hotels_other_lodgings           485.8014
juice_bars                      953.0920
art_galleries                   985.3714
dance_clubs                     322.5823
swimming_pools                  309.0884
gyms                            431.7453
bakeries                        484.9232
beauty_spas                     265.4240
cafes                           224.7806
view_points                     474.6157
monuments                       481.0322
gardens                         974.8897
```

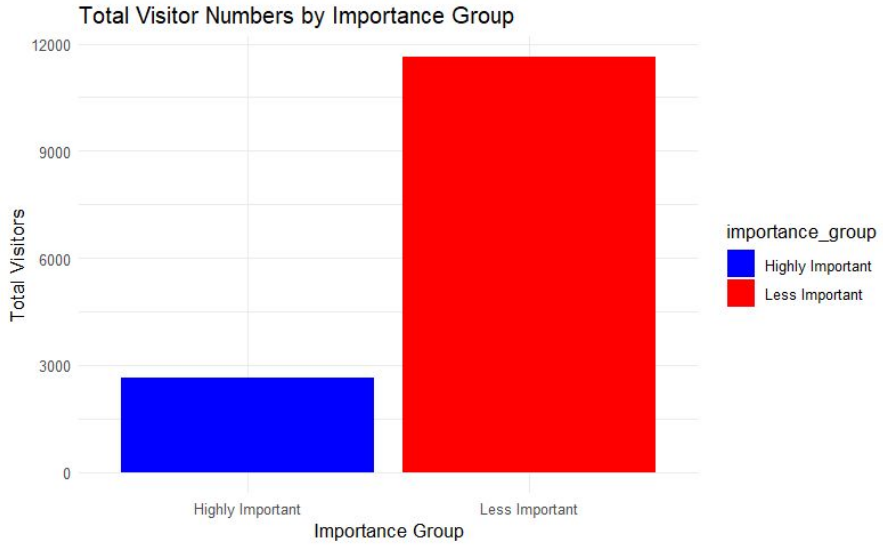


Output RF

Analyzing Travel Attraction Data

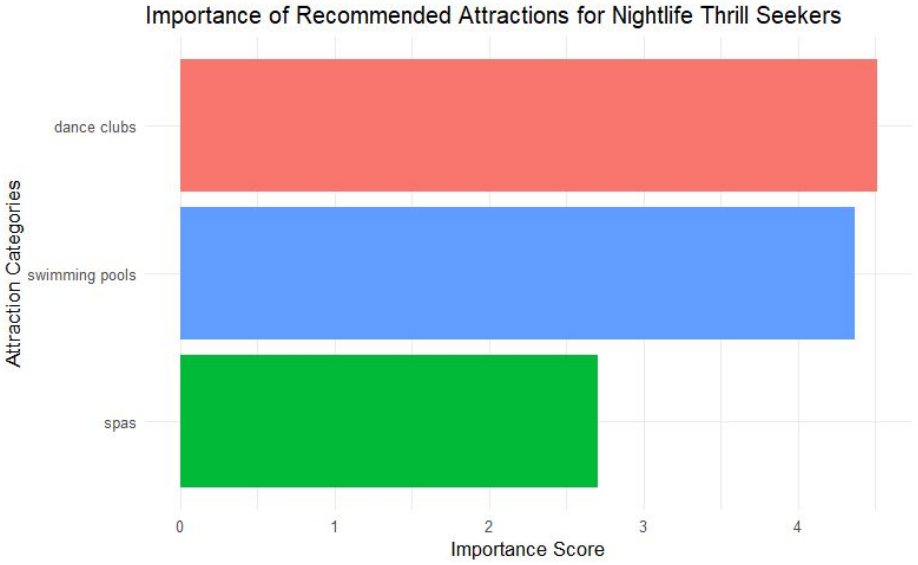
- **Data Overview**: We created a dataset containing 24 categories of travel attractions in Europe, such as parks, beaches, and museums. Each category has an average rating (from 1 to 5) representing visitor satisfaction, and a number indicating how many visitors each attraction received.
- **Categorizing Importance**: To better understand which attractions matter most to travelers, we classified them into two groups: "Highly Important" (like parks and restaurants) and "Less Important" (like bakeries). This helps us focus on the attractions that have a greater influence on visitors' experiences.
- **Summary of Findings**: By analyzing the data, we calculated the average ratings and total visitor numbers for each importance group. This summary reveals which types of attractions are more popular and have higher ratings, guiding businesses and tourism boards in their marketing efforts.
- **Visualizing Insights**: We created bar charts to visually represent our findings. The first chart shows the average ratings for highly important versus less important attractions, while the second chart displays the total visitor numbers for each group. These visuals make it easier to understand the key trends and differences in visitor preferences.


Average Ratings by Importance Group
Average Rating
0
1
2
3
Highly Important
Less Important
Importance Group
importance_group
Highly Important
Less Important


Total Visitor Numbers by Importance Group
Total Visitors
0
3000
6000
9000
12000
Highly Important
Less Important
Importance Group
importance_group
Highly Important
Less Important

Example of Output

- **User Input**: When users input preferences like **"dance club, swimming pools "** the system displays their selected interests, allowing them to see exactly what cultural experiences they wish to explore during their travels in Europe.
- **Relevant Categories Found**: The output identifies and lists categories matching the user's preferences, such as museums and cafes, alongside their attributes, including average ratings and **IncNodePurity** scores, which indicate the importance of each category in influencing user choices.
- **Recommended Attractions**: The system then presents tailored suggestions for the user, highlighting the most relevant and highly-rated attractions within their chosen categories, ensuring they receive the best recommendations for cultural experiences as a **Nightlife thrill seeker**.
- **Conclusion**: This recommendation system serves as a valuable tool for tourists, enabling them to receive personalized suggestions based on their interests, thereby enhancing their travel experience and ensuring they find attractions that align with their preferences.




```
What do you enjoy? (Select all that apply, comma-separated)
1. Nightlife
2. Relaxation
3. Socializing

Recommended Attractions:
```

category <chr>	importance_score <dbl>
dance clubs	4.510389
swimming pools	4.367149
spas	2.705720

THANK YOU!

QUESTIONS?