Finding the right RV campsite for ME!

Hendrik Henning

June 14, 2021

Abstract

Introduction

Finding the perfect RV camping spot can be a challenge for travelers with over 10 000 campsites to choose from in the US alone. This work would usually require a lot of research and planning, and often then you only see the camp site when you arrive. Perhaps you don't like a busy neighborhood or you prefer to have restaurants close by. Your unique preferences are hard to describe and this piles on top of the research to make it a daunting task to find that perfect spot.

RVing has become popular over the years as in addition to holiday goers some have opted in for an RV lifestyle to avoid mortgages, live simply, and allow for some traveling! For those that have overcome the challenges of small spaces and earning a consistent living while on the move this has become a dream come true. This project aims to help both holiday goers and permanent RVers a way of easily finding campsites that are similar to the ones they have found they liked.

Data and Methods

Data science offers tools that can categorize any group of objects into distinct groups. The tools will take several descriptive variables and find objects that are most similar and group them. There are various tools to use that group objects in different ways with different approaches. In this project, I will compare two different approaches, namely, Kmeans and Hierarchical clustering.

The Kmeans algorithm will take a predetermined amount of groups and fit the objects in the analysis to those number of groups as best it can. The measure that this algorithm uses to determine the best fit is by reducing the distance of objects in the cluster to the center of the cluster. This is done iteratively with starting points usually chosen at random and stops when the centers of the groups stop moving. The boundaries between groups tend to be straight lines and planes. Some strengths of the algorithm are that it is easy to visualize and utilize. Some drawbacks include fragmenting of groups through assuming that groups are well defined and easily separable.

The agglomerative method of building hierarchical models uses an approach that includes uniting two objects or groups that are most similar. Every single object starts as its own group and then gets paired as the algorithm moves through its iterative cycles. The algorithm ends when there is only one group after all groups have joined. A distance cut-off can be chosen and the various groups at that cutoff will be labeled separately. Some benefits of this method are that irregularly shaped clusters are seen as one cluster, another is that it is easy to see and understand how the clusters form through dendrograms.

The two algorithms will be evaluated by looking at the comparison of the different clusters generated.

The campsite data will be found online at the poi-factory which holds a community-generated CSV list of campsites with GPS coordinates for GPS software. The campsite identity will be generated by using the

Foursquare API to access data of surrounding venues, parks, and natural habitat data such as lakes and beaches. A radius of 4 kilometers will be chosen for the required data to include a wide variety of data points.

Results and Discussion

A total of 16 976 unique campsites in North America was identified and fed into the Foursquare API to yield venue data. A total of 146 388 venues were identified with 600 different categories of venues. 8112 campsites had no venues in a 4-kilometer radius. The remaining campsites and venues were then taken to generate models.

The Kmeans model was evaluated first and the Elbow method found a distinct drop in the means squared plot when the number of Kmeans clusters reached 6 as seen in Figure 1. We see that both the Kmeans and the hierarchical models found similar groups, with similar numbers that have similar venues close by. Visual inspection of the Folium maps generated showed clear trends that were not explained by looking at the numbers and top 4 venues alone. This indicates that for most groups a large number of venues played significant roles in clustering the groups.

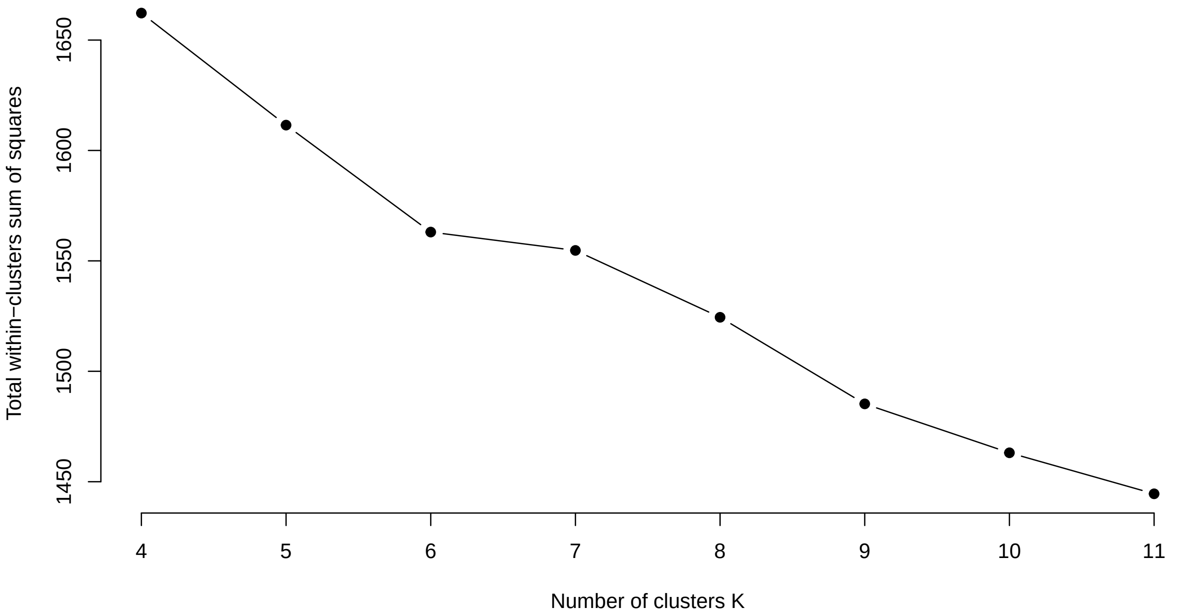

Figure 1: A figure showing how the number of clusters was determined using the Elbow method and Kmeans clustering.

It must be noted that the Kmeans square mean distance kept dropping significantly for the larger cluster numbers indicating that there are possible sub groups or that the data set is a large dispersion of data that is loosely connected. Building the hierarchical model allowed for visual inspection of the clusters as seen in Figure 3. There are two larger groups of clusters that then break down 4 with 2 of those containing the majority of campsites.These 2 larger groups then break down to 4 sub groups, 2 each.Breaking up the hierarchical model into 6 groups generates groups that are significantly distant from one another while still retaining significant group similarity. The result of 6 clusters seemed to give enough seperation of the data set while still not getting too many sub clusters that would in essence look very similar to one another.

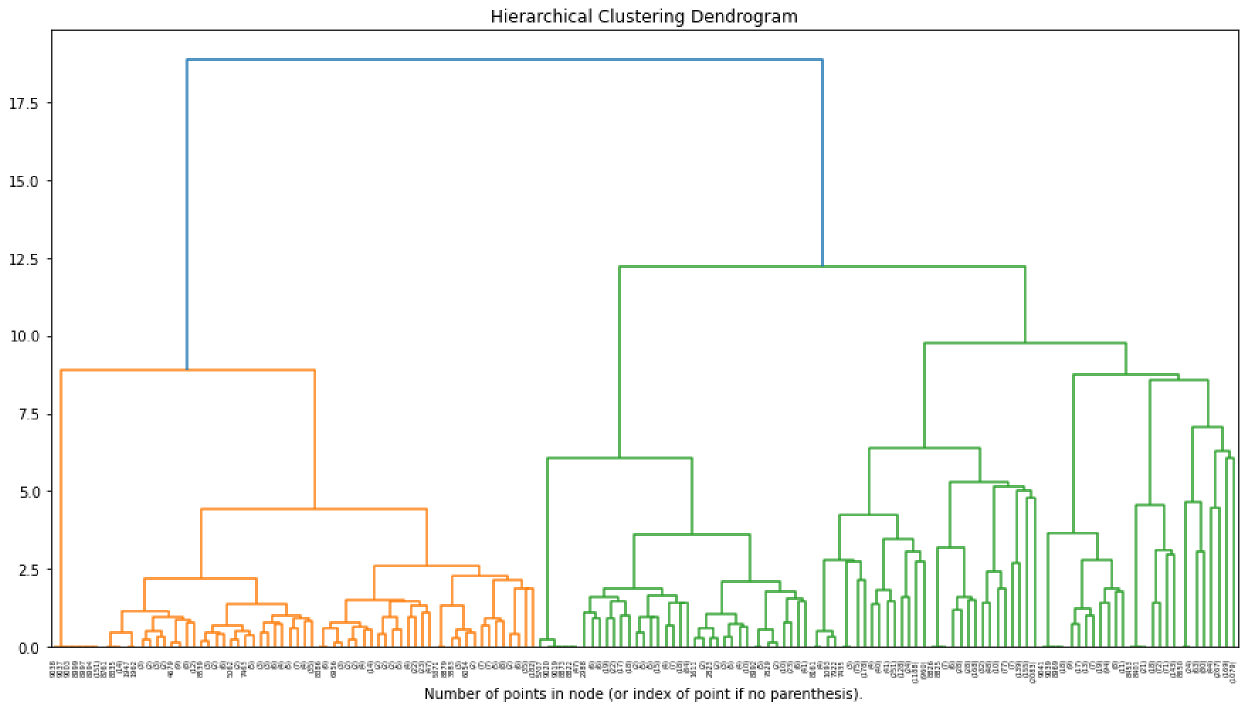

Figure 2: A firgure displaying the dendogram of the hierarichal model.

Comparing the 2 models was not straight forward as there was no ground truth, as is the general problem with clustering.

Comparing the groups to see relative numbers of the clusters and common campsites gave the results shown in Table 1. The Agreed number was calculated on the number of campsites that were in both models in the same group. The Total numbers are the total amount of campsites in each respective model. The agreed score was calculated by taking the number of agreed campsites and deciding by the model's total campsites for that cluster. This result will show how strongly each model agrees with the other.

$$AgreedScore = Agreed/TotalSites\%$$

Table 1: A Table showing some comparative results for the 2 models developed

	Agreed	Total Hierarchical	Total Kmeans	Agreed Score (Hier.)	Agreed Score (Kmeans)
Cluster 0	531	2063	1194	25.7	44.4
Cluster 1	5262	5738	6366	91.7	82.6
Cluster 2	266	401	351	66.3	75.7
Cluster 3	4	547	507	0.7	0.7
Cluster 4	160	160	260	100.0	61.5
Cluster 5	182	202	433	90.0	42.0

The results show that the models agree with the clusters 1, 2 and 4 and that cluster 3 had almost no similarities. For cluster 4 and 5 the hierarchical model had a large percentage of the cluster where the Kmeans model had a significantly smaller population. This suggests that the Kmeans model took in a larger

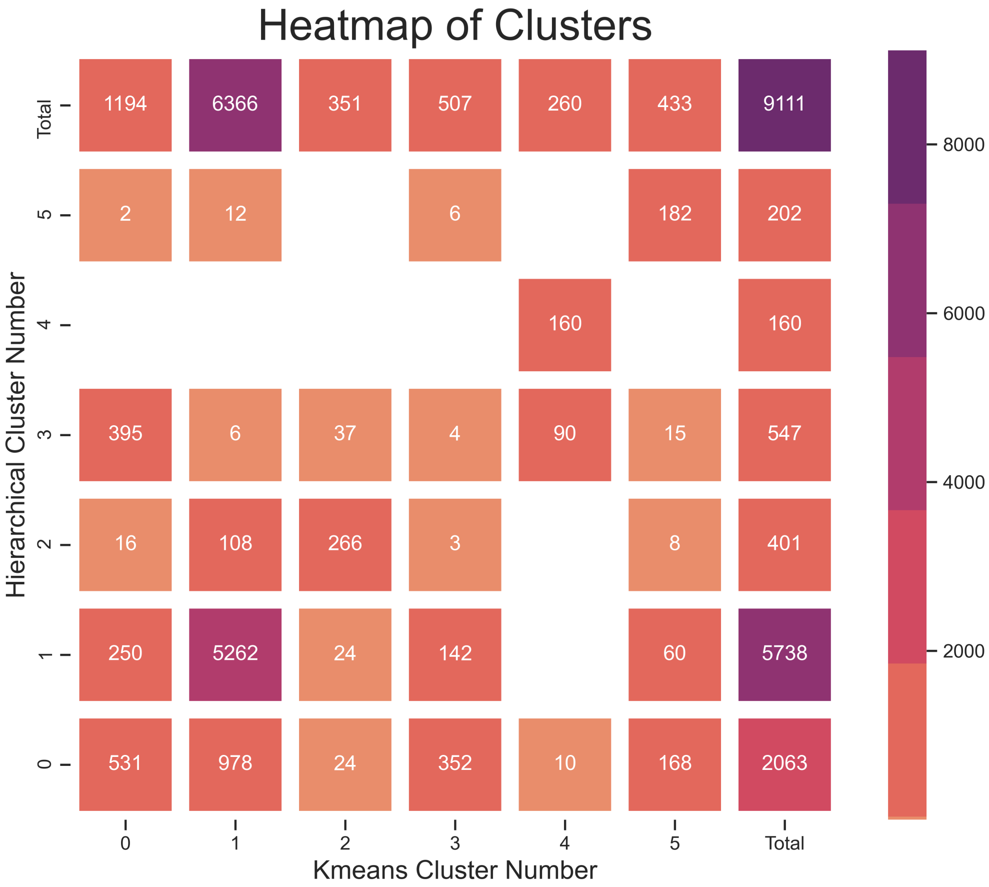

Figure 3: A Figure showing the relationship between the Kmeans and hierarchical models developed in a heat map. Totals are shown at the top and right for the corresponding cluster groups and models.

amount of campsites in that cluster from other clusters. This is due to the way Kmeans order points by individual distances to the center of the cluster and in so doing has a high probability of including campsites from other clusters that are oddly shaped and not spherical. This is the first indication that the hierarchical model is superior for the classification of campsites.

Looking at the numbers of the campsites it can be seen that the clusters 0 and 1 had a high percentage of campsites, indicating a high concentration of points that are distinct from the other camp sites. Looking at the cluster 1 briefly one can see that it has many campsites in urban areas and would suggest a high concentration of venues close to the campsite.

Table 2: A Table showing the top 4 sites for each Hierarchical cluster. Links go to interactive map results hosted on GitHub Pages

	1	2	3	4
Cluster 0	Campground	Construction & Landscaping	State / Provincial Park	Park
Cluster 1	American Restaurant	Fast Food Restaurant	Hotel	Convenience Store
Cluster 2	Lake	Campground	Harbor / Marina	State / Provincial Park
Cluster 3	Campground	Lake	Home Service	Trail
Cluster 4	Campground	NULL	NULL	NULL
Cluster 5	Trail	Campground	State / Provincial Park	Mountain

Table 3: A Table showing the top 4 sites for each Kmeans cluster. Links go to interactive map results hosted on GitHub Pages

	1	2	3	4
Cluster 0	Campground	Lake	Park	State / Provincial Park
Cluster 1	American Restaurant	Fast Food Restaurant	Hotel	Convenience Store
Cluster 2	Campground	Lake	Home Service	Trail
Cluster 3	Construction & Landscaping	Campground	Home Service	Lake
Cluster 4	Trail	Campground	Scenic Lookout	State / Provincial Park
Cluster 5	Lake	Campground	State / Provincial Park	Harbor / Marina

Looking at the top 4 venues close to the sites we can see some similarities between the Kmeans model and that of the hierarchical clustering. Cluster 1, as noted before showed a high degree of similarity with venues that include restaurants, hotels and convenience stores. When looking at cluster 4 which was highly contested we see that the hierarchical model included only campsites that have other campsites in its vicinity. There are no other venues included in this cluster for this model. The Kmeans cluster 4 included many other venues around the camp sites, and show that the Kmeans model could not isolate this group with a significant portion of the campsites as the hierarchical model did. We can call this strike 2, the hierarchical model is in the lead. The other clusters show some similarities in that the campsites tend to be away from towns and cities, and include a preference for certain types of nature spots. These will be discussed further below. When investigating further we see that the Kmeans top venues for the other clusters include venues that are similar in some of the other clusters, where the hierarchical model does not have this. This would suggest that the Kmeans model was not able to separate the groups significantly in order to achieve clear and distinct preferences.

Overall the hierarchical model is superior, and will be taken as the chosen model.

Group Identities

We would like to go into more detail regarding the groups generated and give them names that are easy to remember.

Cluster 0 - A little out the way

Cluster 0 offers campsites that are close to population centers, but not in them. You can have the convenience of having stores and gas stations close by as well as being away from hustle and bustle of every day life.

Cluster 1 - City Dwellers

Cluster 1 has a very high amount of campsites with 5738 out of the total 9111. This would indicate a very dense population, and upon inspection we can see that many of the campsites are in populated areas with many venues in the 4 kilometer radius.

The RV camper that wants to be close to many different stores and have the convenience of many spots to choose from would travel to spots in this cluster. Rvers that do migratory work and part time jobs in many towns and cities would find this cluster recommended.

Cluster 2 - Wet feet

Cluster 2 contains a large amount of campsites that have water close, in addition many have State parks and show that you can have a wet nature experience in your RV.

Cluster 3 - The all rounder

Cluster 3 campsites tend to have many different venues close by including State Parks, lakes, mountains and some are even a 15 minute drive from population centers. If you're someone that can't make up your mind on the kind of experience you want to have this cluster could be for you.

Cluster 4 - Popular Isolated Nature Spots

In Cluster 4 we only see other campsites in the 4 kilometer vicinity and no other venue types. This is different from the group in which there were no venues within the same radius and would suggest that the campsites, even though isolated, would be more popular in nature due to a higher concentration of campsites in the vicinity.

The increased popularity could indicate frequent travelers, or isolated nature spots that are easily accessible. In short you are deep in nature, but not too far away from home.

Cluster 5 - Need a View?

Cluster 5 campsites have a high concentration of mountains and State Parks. Is there any need to say there is only one campsite from cluster 5 in Florida? If you need a site with a view this cluster seems to cater well, as well as hiking trails and other campsites.

Limitations of the model and study

The clustering of campsites yielded acceptable results, but there were some areas of concern. Some clusters are very large in size with little distinction between sub groups, some of the nature clusters can feel redundant and seem similar upon inspection and we had a very large group of campsites that never made it to the clustering stage because they had no venues within a 4 kilometer radius.

In order to improve on the result the study should be modified to include a very large area venues. This radius size can be optimized in order to ensure that every single camp site has at least one entry. Further, a scoring system can be developed where nearer venues and venues would score higher so that the campsites with a high concentration of very close venues would still be grouped together.

It might also be very beneficial to perform the clustering and generate fewer groups, let's say 3 or 4 - a number that can be optimized, and then perform clustering a second round on the generated clusters in order to separate them into sub groups which would ideally give more predictable results.

The model and descriptions are at present only in report form and maps that are hard to access, and therefore a website with easily accessible descriptions and maps would allow the model to be accessed by the public, which could benefit from the results.

Conclusion

The clustering of campsites was achieved with moderate success and a successful hierarchical model was generated. The hierarchical model proved superior to the Kmeans model that gave similar numbers of clusters, but was found to to mix the groups and give less distinction to each group. This was particularly pronounced with smaller groups.

Further work could be done to improve this model and expand on its capabilities, namely scoring venues according to distance and allowing a larger radius for venues to be recorded and modeled. There is also the possibility of assigning sub groups if they are distinct enough. The project shows some promise to be used as an app or website for RVers to find suitable campsites that suit them personally.