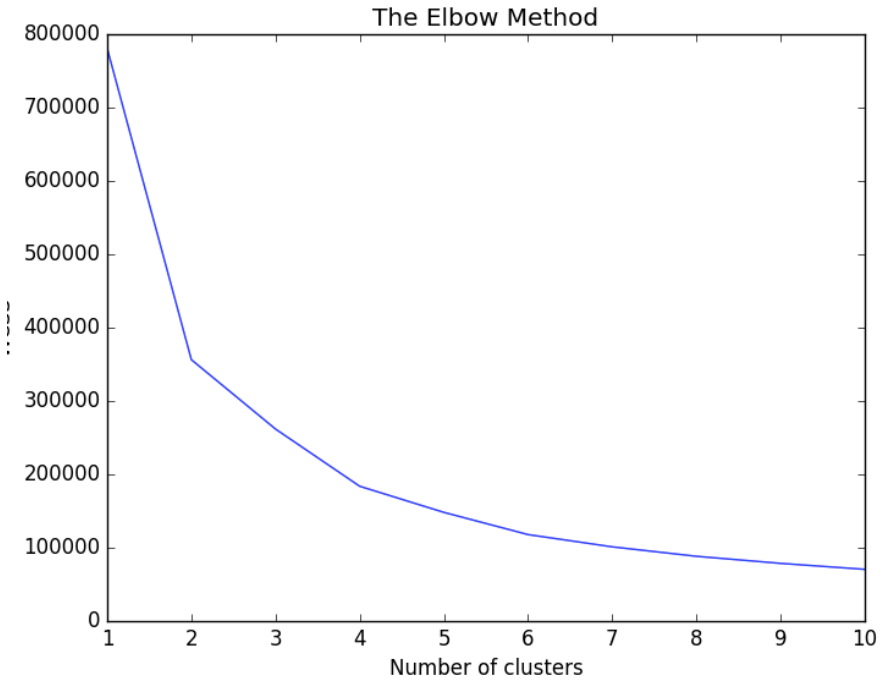
Dataset:
The dataset is a 249 x 7 .csv file describing the number of reviews each user did. Every row has a specific user ID (249 users), every column reflects a different field of review. This dataset was specifically chosen from the UCI Machine Learning Repository to match the review section of our app, so that, for example, Nature and Picnic could represent hunting and fishing (or similar activities).

Out[3]:

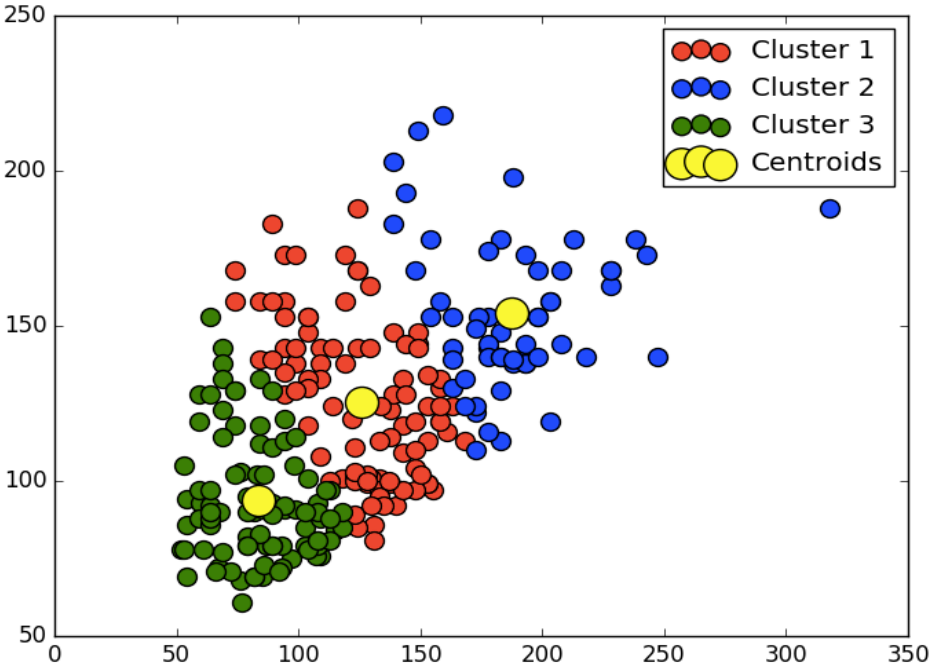
	User Id	Sports	Religious	Nature	Theatre	Shopping	Picnic
0	User 1	2	77	79	69	68	95
1	User 2	2	62	76	76	69	68
2	User 3	2	50	97	87	50	75
3	User 4	2	68	77	95	76	61
4	User 5	2	98	54	59	95	86
5	User 6	3	52	109	93	52	76
6	User 7	3	64	85	82	73	69
7	User 8	3	54	107	92	54	76
8	User 9	3	64	108	64	54	93
9	User 10	3	86	76	74	74	103
10	User 11	3	107	54	64	103	94
11	User 12	3	103	60	63	102	93
12	User 13	3	64	82	82	75	69
13	User 14	3	93	54	74	103	69
14	User 15	3	63	82	81	78	69
15	User 16	3	82	79	75	75	82
16	User 17	5	59	131	103	54	86
17	User 18	5	56	124	108	56	85

Clustering Nature – Picnic:

Elbow (N. 3 cluster was chosen)



Scatter Plot: Nature – Pic nic

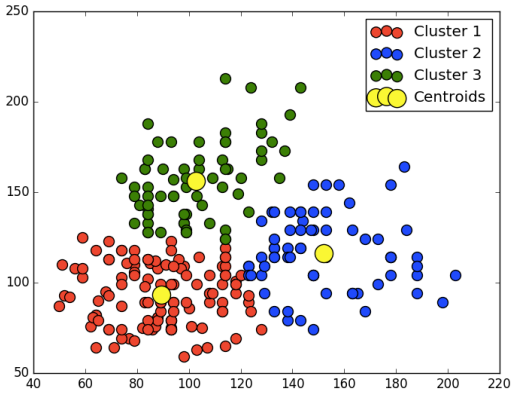


Consideration:

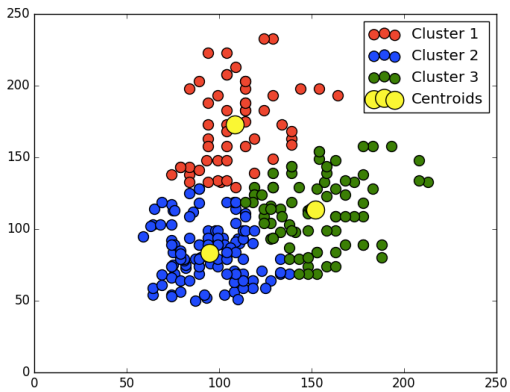
Given the elbow, we are able to identify 3 clusters of users (50 – 120, 120 – 170, 170 – 250), depending on the number of reviews they did. Regarding nature and pic nic, it's clear the presence of a linear incremental association, since in every cluster there's almost a 1:1 relation between them. As matter of fact, it's also understandable since a person intrigued by nature will likely have pic nic in his free time.

Proof:

We also tried to plot other features to verify this intuition. The output was a different, non linear association between features depending on parallel clusters, totally different from the regular one we found.



Religion – Theatre



Theatre - Shopping