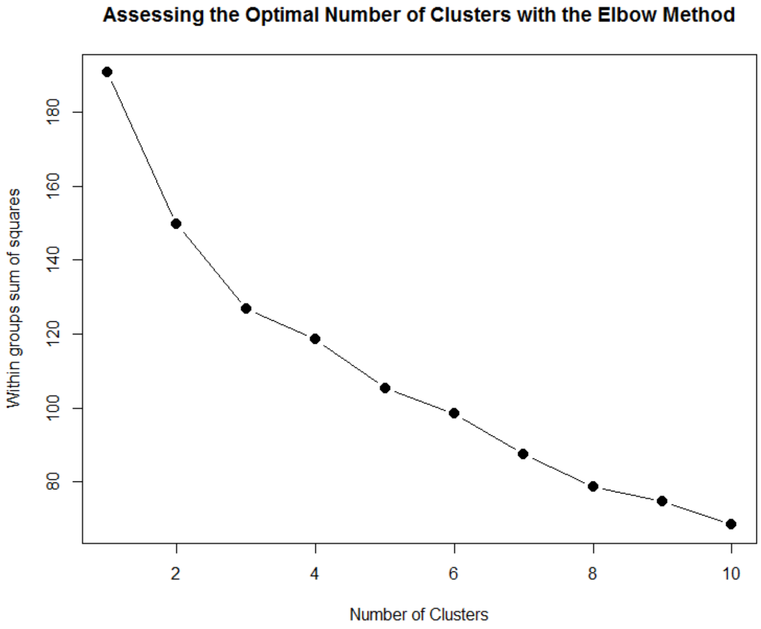
Assignment # 3: Clustering Assignment

An online shopping site has the following primary pages or sections: Home, Products, Search, Prod_A, Prod_B, Prod_C, Cart, Purchase. A user may browse from "Home" to "Products" and then to one of the individual products. The user may also search for a specific product by using the "Search" function. A visit to "Cart" implies that the user has placed an item in the shopping cart, and "Purchase" indicates the user completed the purchase of items in the shopping cart. The site has collect the hypothetical session data for 100 sessions. This data is available in CSV format, Sessions.CSV, on course website. Use K-means clustering algorithm to cluster these user sessions into segments.Try different clustering runs with various numbers of clusters (e.g., between 4 and 8), and select the result set(s) that seem to best answer as many of the following questions as possible.

First look at the data and run for 4 cluster only and check the cluster centroids:

Cluster#	Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
1	1	0.333333	0.904762	0.238095	0.333333	0.428571	0.428571	0.142857
2	0.555556	0.75	0.333333	0.694444	0.527778	0.444444	1	1
3	0.904762	0.857143	0.190476	0.952381	0.333333	0.428571	0.52381	0
4	0	0.909091	0.363636	0.136364	1	0.5	0.227273	0

One solution often used to identify the optimal number of clusters is called the **Elbow method** and it involves observing a set of possible numbers of clusters relative to how they minimize the within-cluster sum of squares. In other words, the Elbow method examines the within-cluster dissimilarity as a function of the number of clusters.



Between 4 and 8, from the example above, we can say that after 6 clusters the observed difference in the within-cluster dissimilarity is not substantial.

- Question 1A - If a new user is observed to access the following pages:
  Home => Search => Prod_B,
  according to your clusters, what other product should be recommended to this user?

Cluster	Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
1	1	0.5	1	1	1	0	1	1
2	1	0	1	1	1	1	1	0.5
3	1	1	1	0	1	0	0.5	0
4	1	0	1	0	1	0	0	0
5	1	1	1	0	1	1	0	0
6	1	0.666667	1	1	1	1	0	0

- Answer 1A – Based on the given routine and cluster analysis above, **"Prod_A"** should be recommended to this user which would result high percentage of purchase rate than other options.

- Question 1B - Explain your answer based on your clustering results. What if the new user has accessed the following sequence instead: Products => Prod_C ?

Cluster	Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
1	0.625	1	1	0.25	0.875	1	0.125	0
2	0	1	0	0	1	1	0.125	0
3	1	1	0.833333	0.666667	0	1	1	0.833333
4	0	1	1	1	0	1	1	0.5
5	0	1	0.125	0	1	1	1	1
6	1	1	0	1	0.5	1	0.833333	0

- Answer 1B – Based on the given routine and cluster analysis above, **"Prod_B"** should be recommended to this user which would result high percentage of purchase rate than other options.

- Question 2 - any clustering help us identify casual browsers ("window shoppers"), focused browsers (those who seem to know what products they are looking for), and searchers (those using the search function to find items they want)? If so, are any of these groups show a higher or lower propensity to make a purchase?

Cluster	Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
1	0	1	1	0	1	0.375	0.5	0
2	1	1	1	0.333333	0.833333	0.833333	0	0
3	0	1	1	1	0	1	1	0.5
4	0.5	1	1	0	0.5	1	1	1
5	1	1	1	0.25	0.25	0.25	1	0
6	1	1	1	1	0.2	0.6	1	1

- Answer 2: when a user search and check Product C and B at the same time, he/she less like to purchase. On the other hand, when a user check Product C and A, they will more likely to purchase

- (Optional Questions) Question_3 - Do any of the segments show particular interest in one or more products, and if so, can we identify any special characteristics about their navigational behavior or their purchase propensity?

Cluster	Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
1	0	0.111111	0.111111	1	0.444444	0	0.777778	0.666667
2	0	1	0.310345	0.034483	1	0.655172	0.482759	0.310345
3	0.941177	0.764706	0.647059	0.764706	0.529412	1	0.529412	0
4	1	0.9	0.05	0.9	0.35	0	0.65	0.45
5	1	0.25	0.833333	0.083333	0.25	0	0.416667	0.166667
6	0.923077	0.615385	0.846154	0.846154	0.230769	0.692308	1	1

- Answer_3: Based on Cluster # 6 as shown above, following navigational behavior has highest change of purchasing the product :

Home => Search = > Prod_A

- Question_4 - If we know that, during the time of data collection, independent banner ads had been placed on some popular sites pointing to products A and B, can we identify segments corresponding to visitors that respond to the ads? If so, can we determine if either of these promotional campaigns are having any success?

Cluster	Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
1	1	0.666667	1	1	1	1	0	0
2	0	0	0	1	1	0	0.5	0.5
3	1	1	0	1	1	0.75	1	0
4	0.5	1	0	1	1	0	0	0
5	1	1	0.2	1	1	0	1	1
6	1	0	1	1	1	0.666667	1	0.666667

- Answer_4: Based on Cluster #2 as shown above, from the ones who followed the promotional campaign link directly without opening any other websites but Prod_A and Prod_B, half of them ended up purchasing the product. In other words, the campaign had 50% success.