Nichollette Acosta 800829090 ITCS 3162 - Assignment 2

I. **Get 40 more CNN news articles online (try to find them in different categories, e.g., sports and finance). Record the category of each article.**

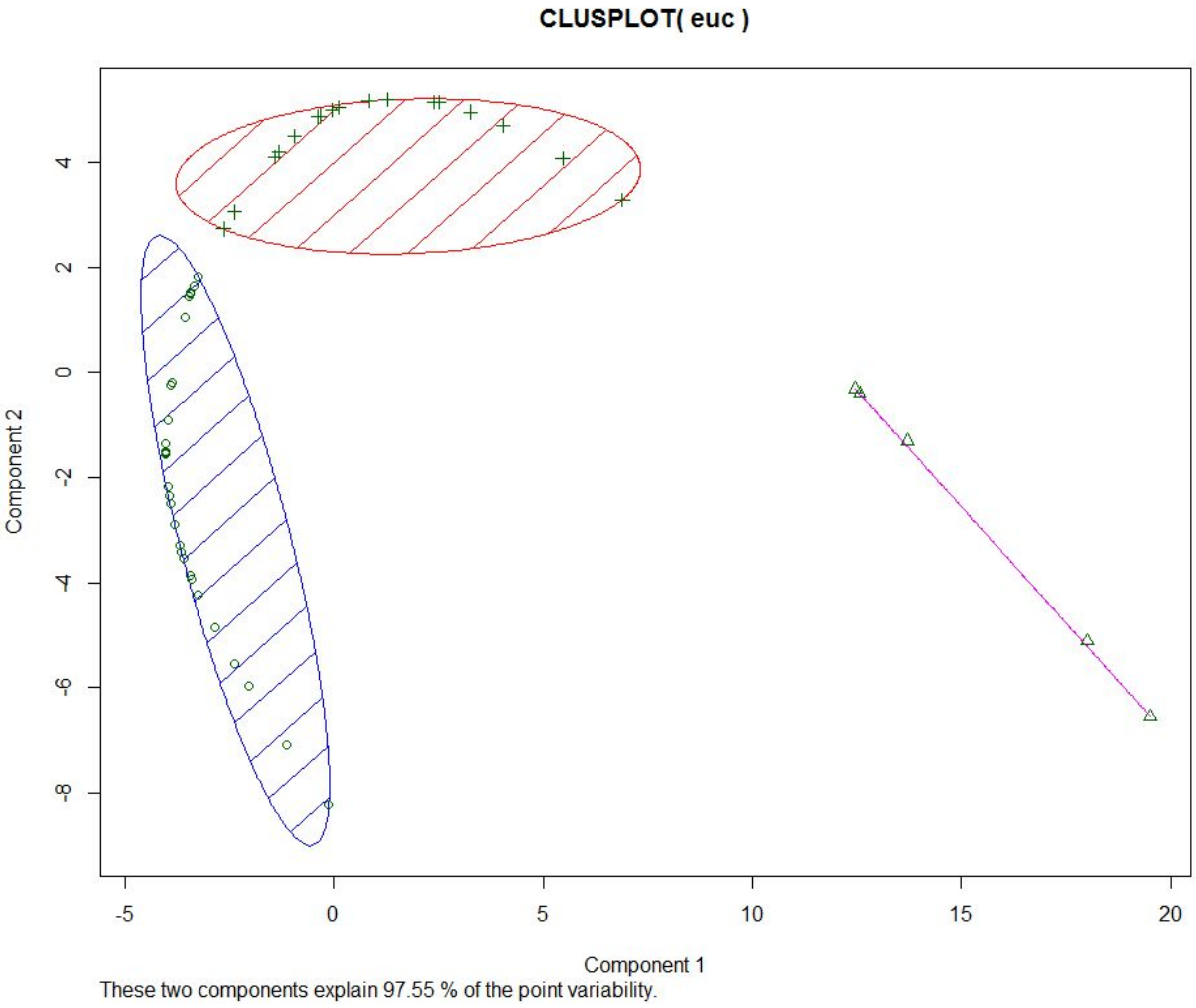
II. **Convert all 50 CNN news articles to data matrix (each row is an article and each column is a unique term)**

.

III. **III.Run K-means clustering with Euclidean, Cosine and Jarcard similarity. (Specify K as the number of categories of your 50 CNN news articles)**

<u>EUCLIDEAN:</u>

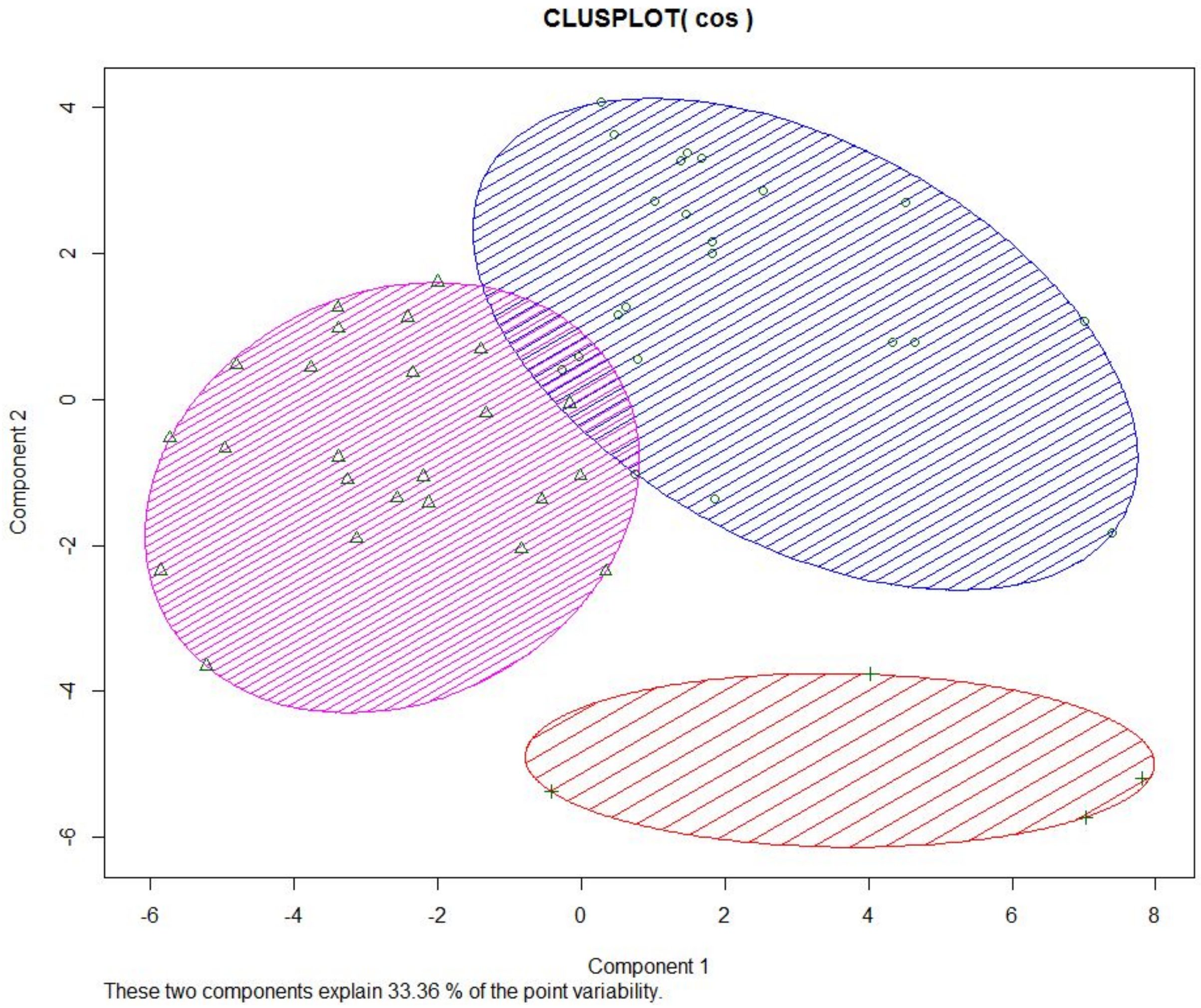
-K-means clustering with 3 clusters of sizes 28, 5, 17



Nichollette Acosta 800829090 ITCS 3162 - Assignment 2

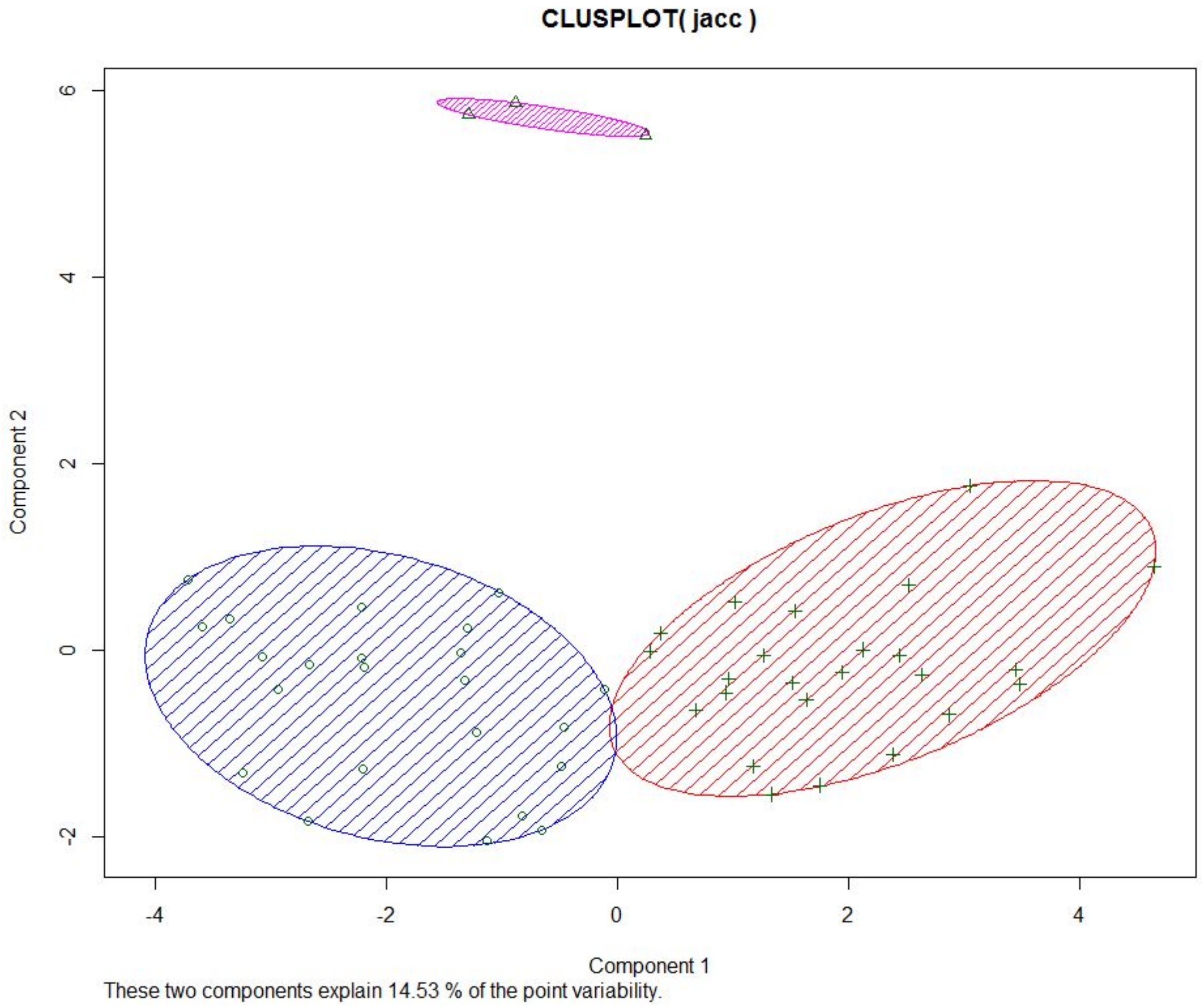
COSINE:

-K-means clustering with 3 clusters of sizes 22, 24, 4



JACCARD:

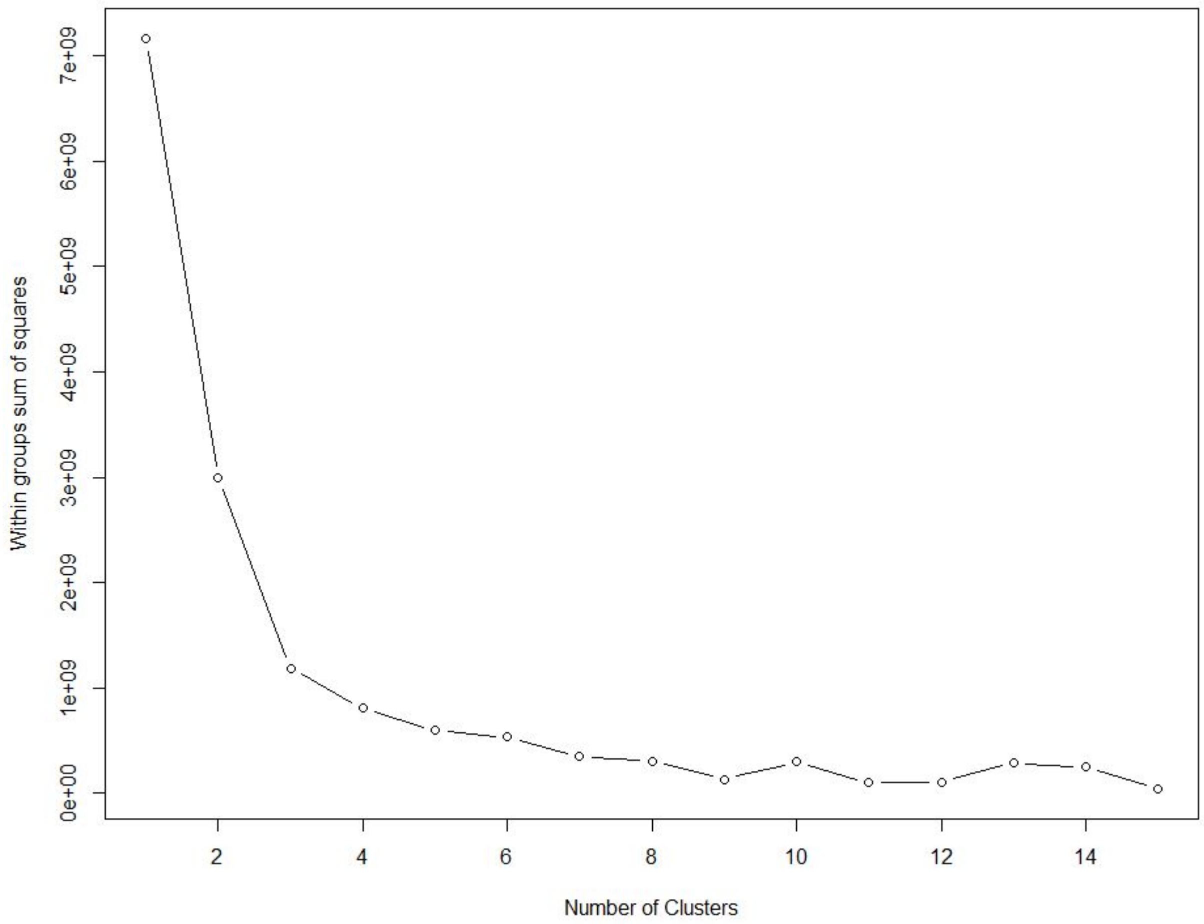
-K-means clustering with 3 clusters of sizes 3, 23, 24



IV. Evaluate K-means clustering results with SSE

EUCLIDEAN:

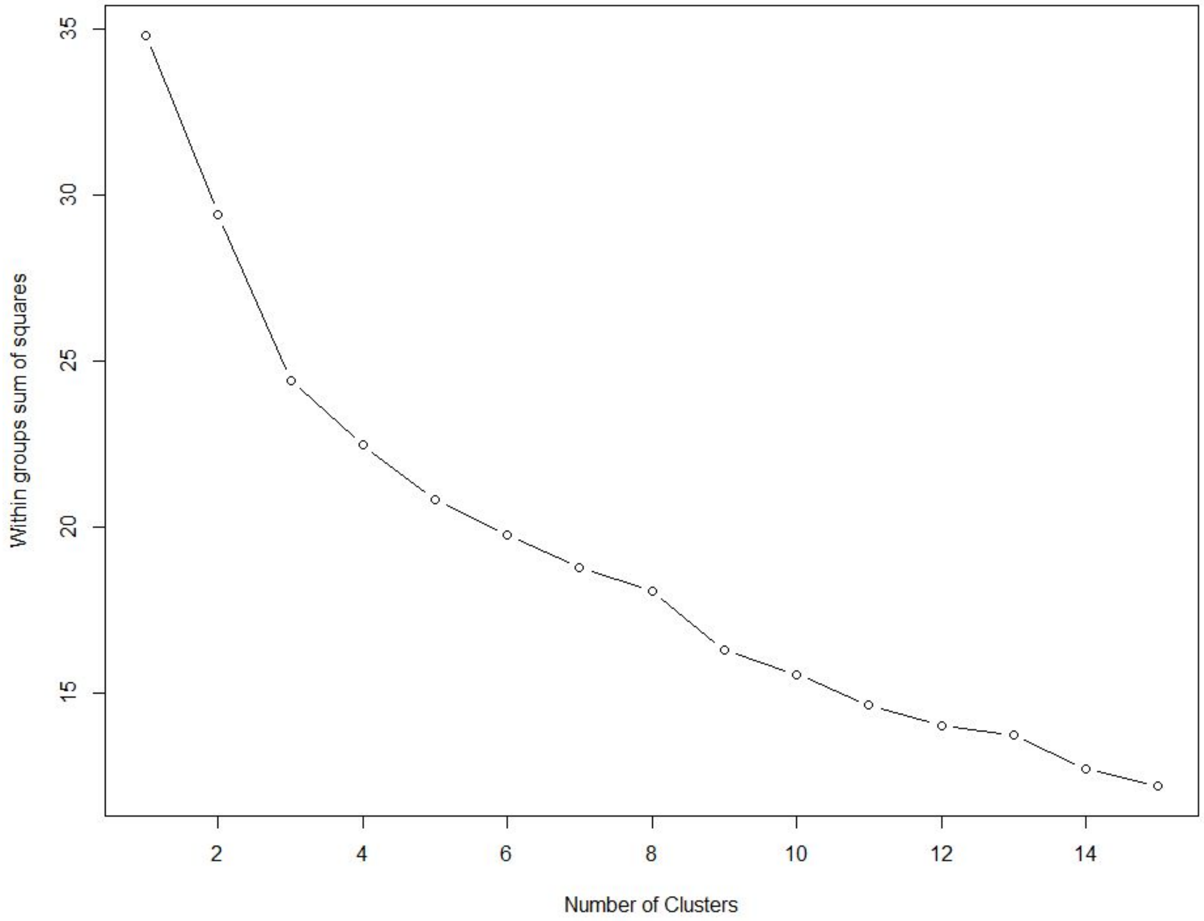
```
-Within cluster sum of squares by cluster:
 [1] 207983335 398248464 580462712
 (between_SS / total_SS =  83.4 %)
```



COSINE:

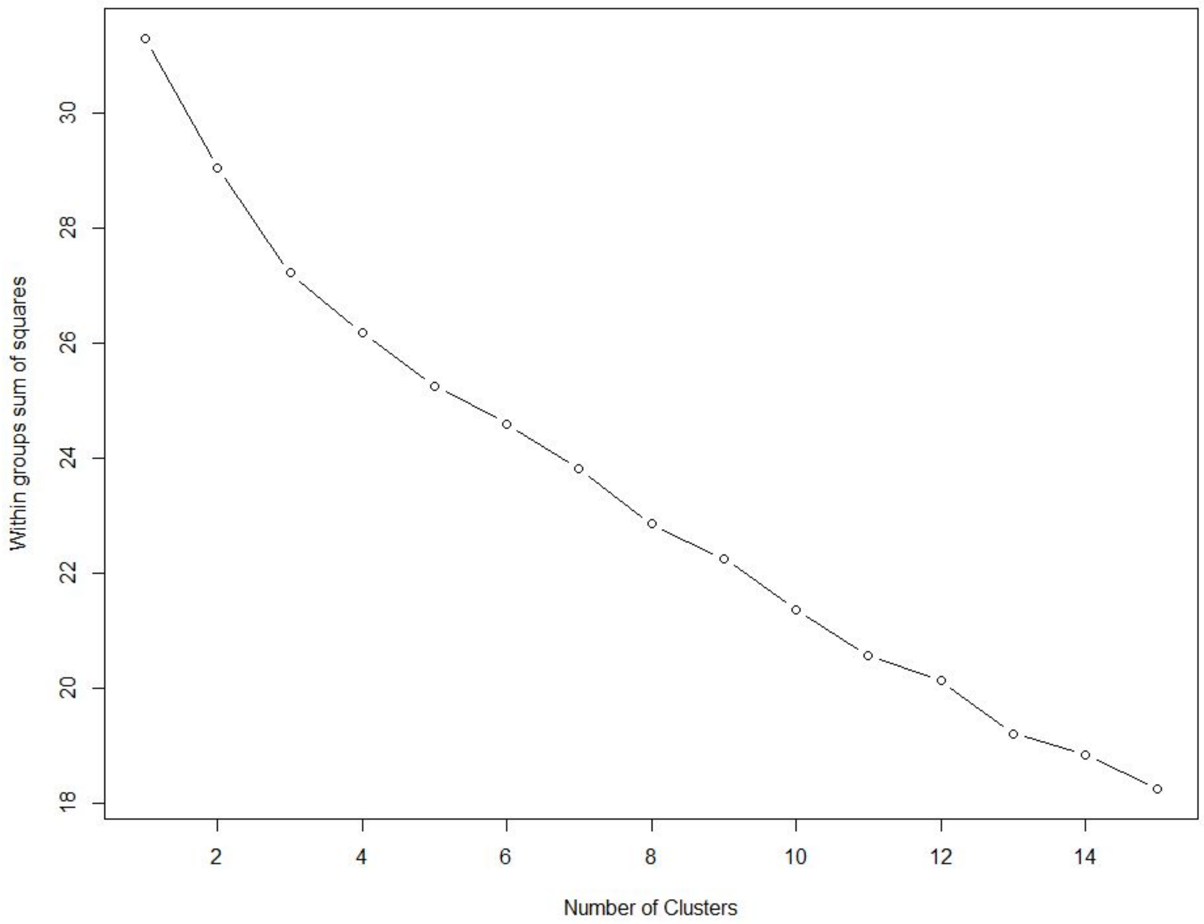
-Within cluster sum of squares by cluster:

```
 [1] 12.6744129 10.8218117  0.9139836
 (between_SS / total_SS =  29.9 %)
```



JACCARD:

```
-Within cluster sum of squares by cluster:
  [1]  0.5708887 14.2528880 12.3981739
 (between_SS / total_SS =  13.0 %)
```



Looking at the results the Euclidean distance showed the best results. The sum of squares value was the highest, implying that the Euclidean K-means created more accurate cluster results than the other two.