# CS 565 – Project 1

Zixuan Wen
U29106598

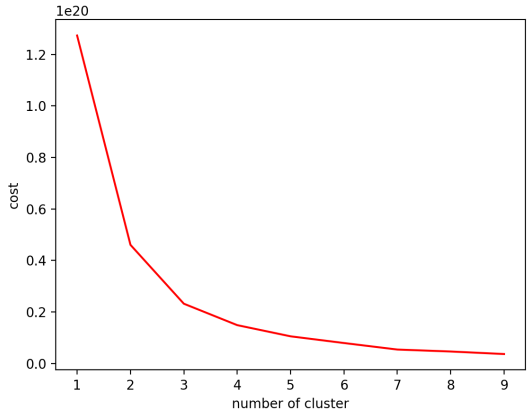
October 5, 2019

## Exercise 1

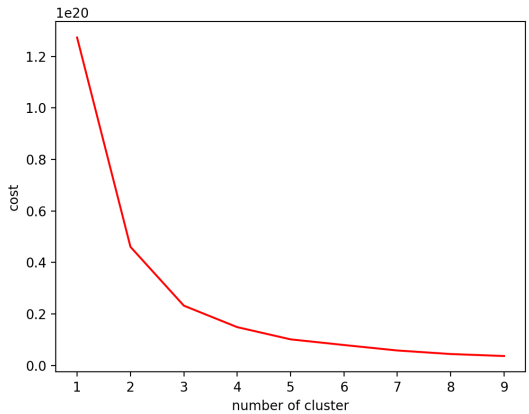
see attached source code

## Exercise 2

graph for k-means performance under different number of clusters:




From the graph above, we can see that the cluster represents the data reasonably well when the number of cluster is around 3 to 5.
graph for k-means++ performance under different number of clusters:

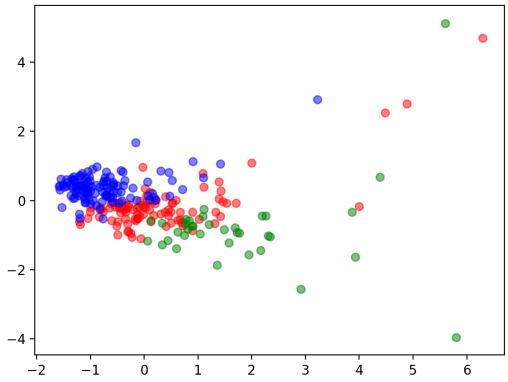



From the graph above, we can see that the cluster represents the data reasonably well when the

number of cluster is around 3 to 5.

## Exercise 3

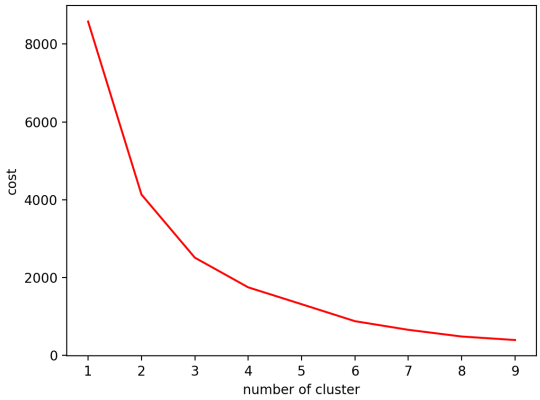
graph for cluster after PCA analyze:




From the graph above, since the data is relatively centered at the same spot due to the $total_votes$ feature, which is obtained from multiplying$vote_average$ and $vote_count$.
So the cluster tend to have a long oval shape instead of regular circle one.

## Exercise 4

graph for 1d-kmeans performance under different number of clusters:




From the graph above, we can see that the cluster represents the data reasonably well when the number of cluster is around 5 to 7.

## Exercise 5

The computed disagreement distance of cluster produced by k-means++ and cluster produced by 1d-kmeans is 2778614