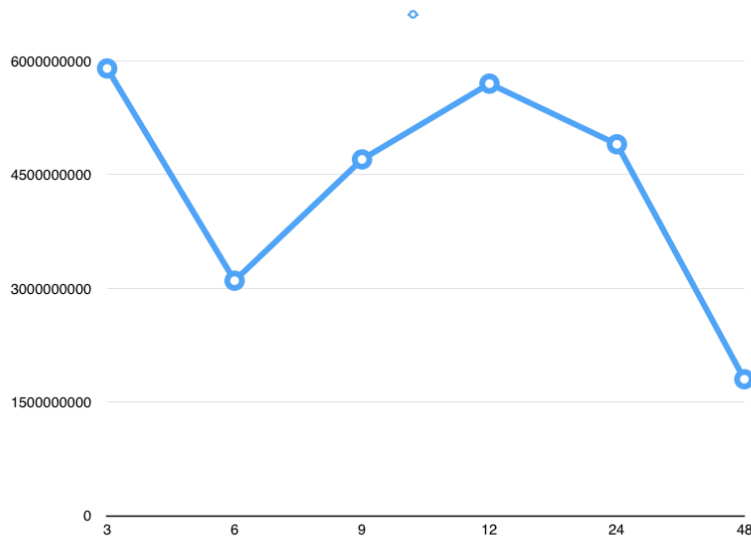


2.

(i)

- Plot the within-cluster sum of squares (wc) as a function of K.

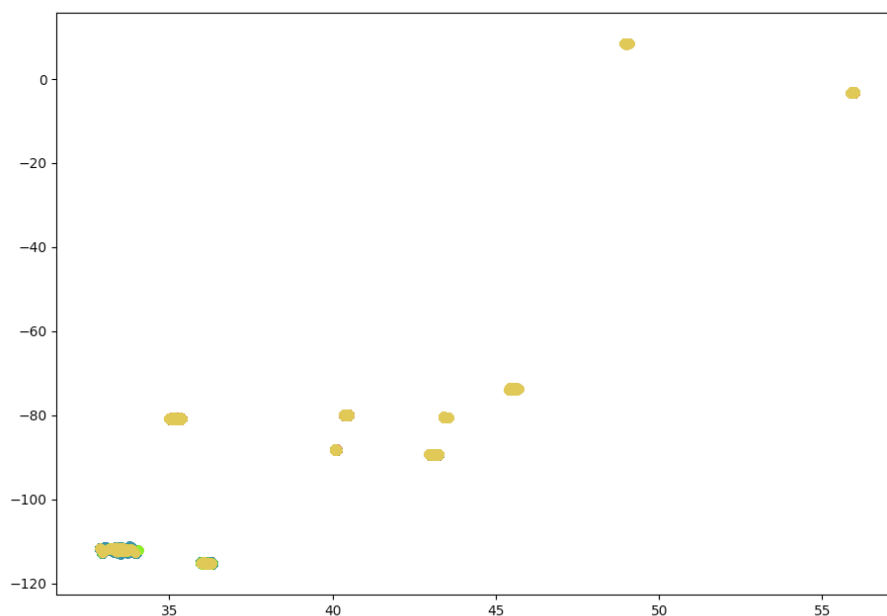


- Choose an appropriate K from the plot and argue why you choose this particular K.

I would choose K equal to **6** since this is the knee point of the graph

- For the chosen value of K, plot the clusters with their centroids in two ways: first using latitude vs. longitude and second using reviewCount, checkins. Discuss whether any patterns are visible.

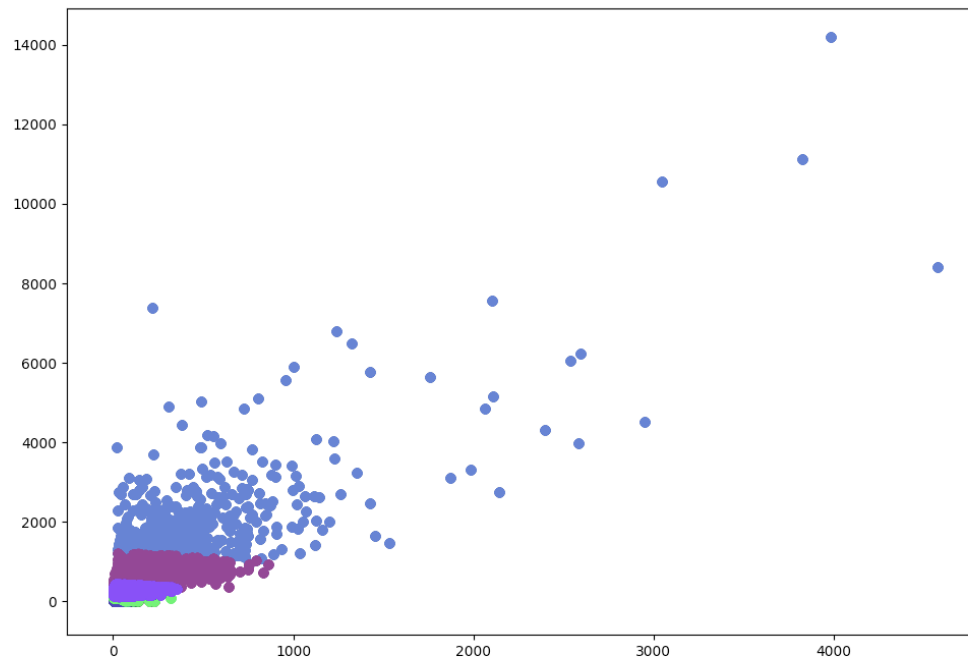
Latitude vs. longitude:



For this graph, the data itself is very well divided. The data has several places that has very high density. I think that is because there is a high concentration of restaurants in certain areas. For

example, one dense spot maybe Chicago, and others might be San Francisco or New York. However, our k-means clustering method doesn't do a very good job because the data is already kind of clustered by itself.

ReviewCount vs. checksins:



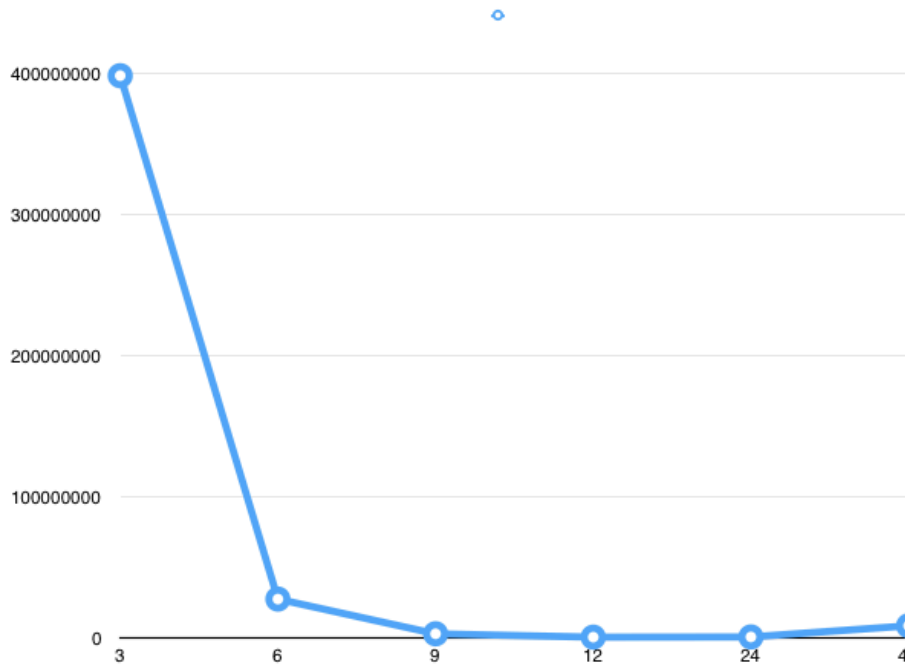
For this graph, the color difference is more obvious than the longitude and latitude one. We can see that the cluster is divided as the points go to the upper right corner. I would say this is because the data here is less already clustered than the previous one. K-means does a better job in combination of x and y.

(ii)

- Describe how you expect the transformation to change the clustering results.

I would expect the transformation to make the clustering more ideal since log transformation remove the skewness in the data.

- Plot the within-cluster sum of squares (wc) as a function of K.

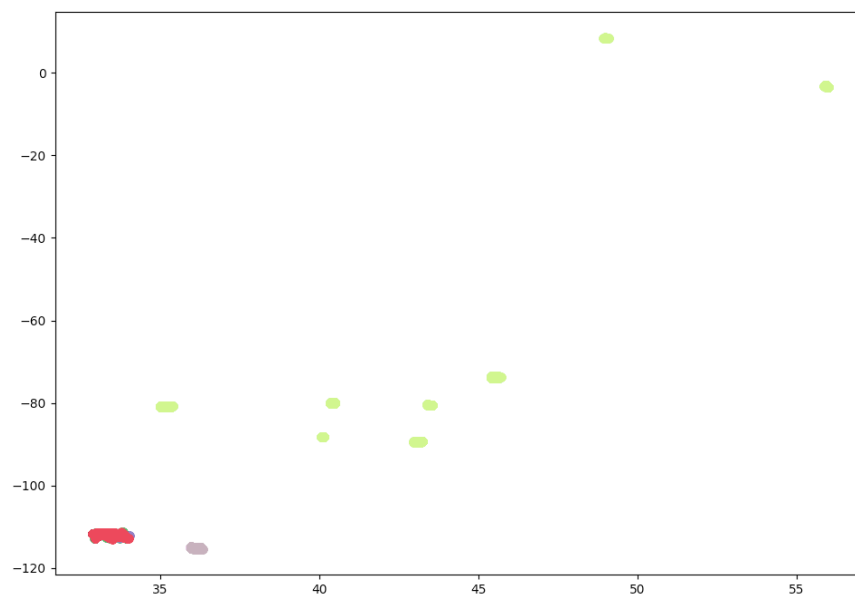


- Choose an appropriate K from the plot and argue why you choose this particular K.

I would choose K equal to 6 because this is the knee of the graph.

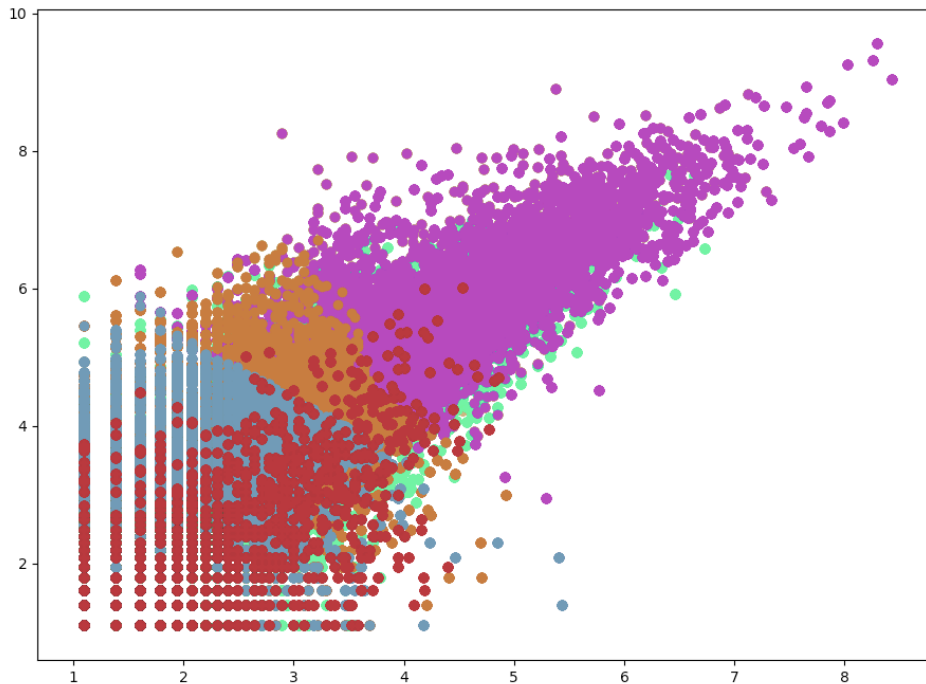
- For the chosen value of K, plot the clusters with their centroids in two ways: first using latitude vs. longitude and second using reviewCount, checkins. Discuss whether any patterns are visible.

Latitude vs. longitude:



There is minimal difference between this one and (i) because we didn't change the data for those two attributes at all.

ReviewCount vs. checkins:

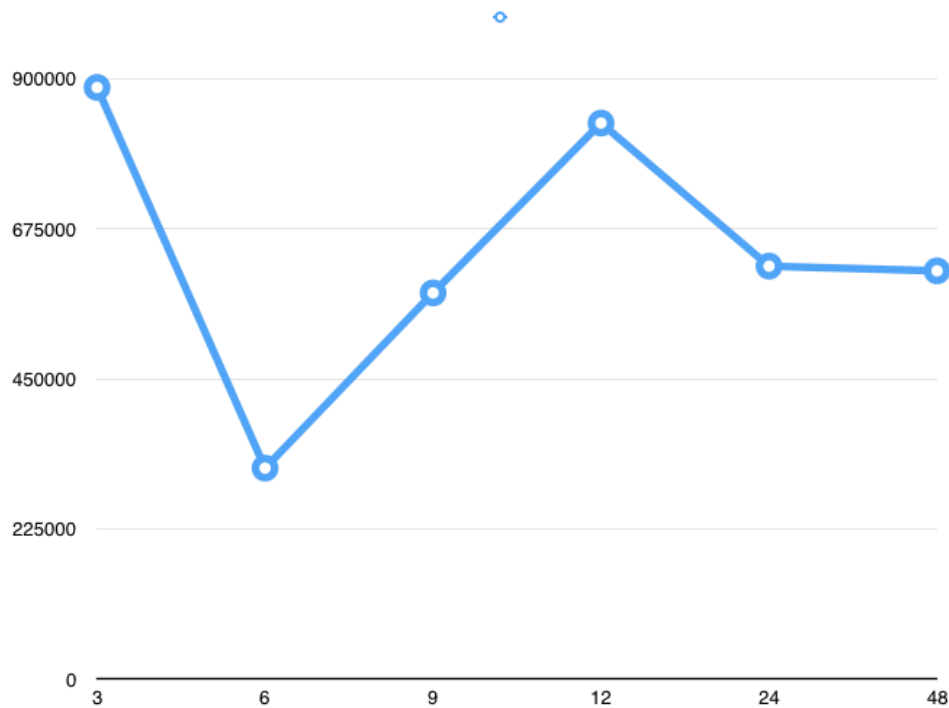


After the log transformation, the data has a clearer pattern because we eliminated the skewness and outliers by performing a log transformation. The clusters are very clearly colored. K-means does a good job clustering data here.

(iii)

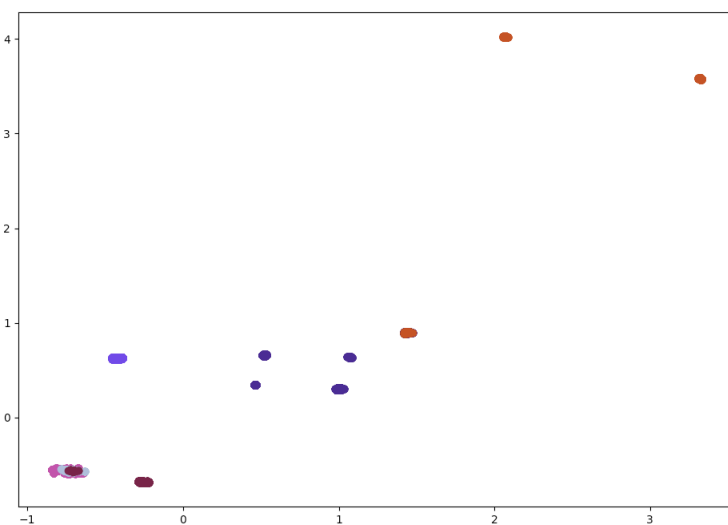
- Describe how you expect the transformation to change the clustering results. By normalizing the data, it should improve the result.

- Plot the within-cluster sum of squares (wc) as a function of K.



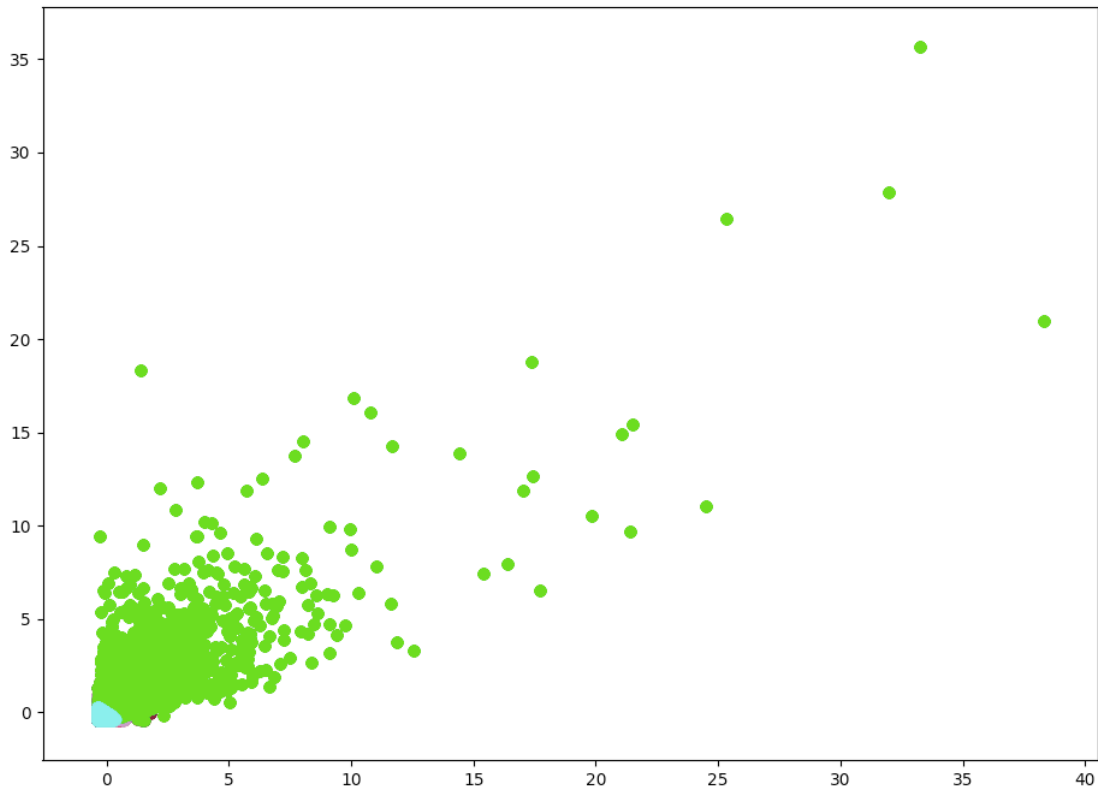
- Choose an appropriate K from the plot and argue why you choose this particular K.  
I would choose K equal **6** because it is the point that the within-cluster sum of squares is the lowest.

- For the chosen value of K, plot the clusters with their centroids in two ways: first using latitude vs. longitude and second using reviewCount, checkins. Discuss whether any patterns are visible.  
Latitude vs. longitude:



The difference is that by normalizing the result, it gives this graph a clearer cluster

ReviewCount vs. checkins:

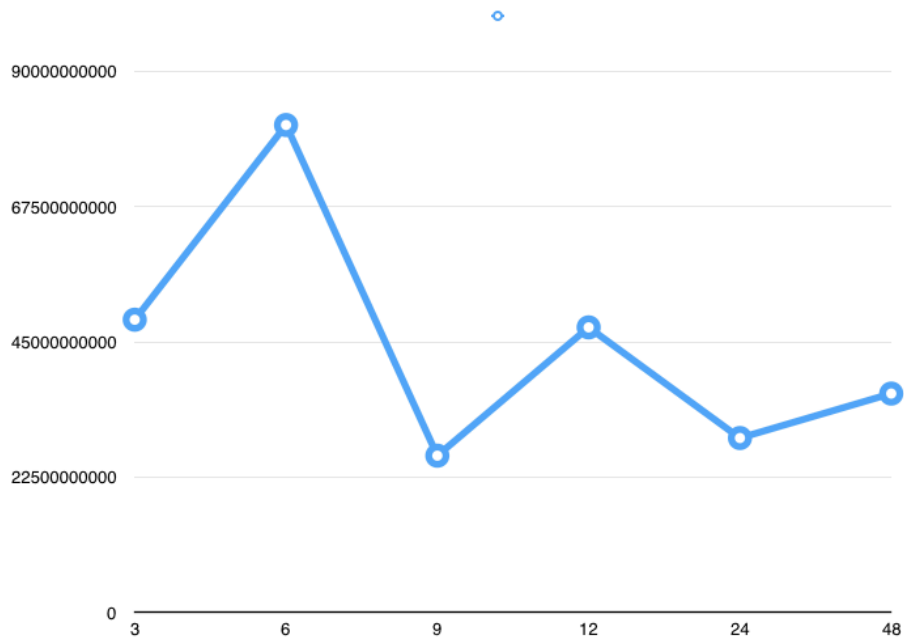


By normalizing the result, the cluster is more centralized.

(iv)

- Describe how you expect the transformation to change the clustering results.  
It will give similar result as Euclidean distance

- Plot the within-cluster sum of squares (wc) as a function of K.

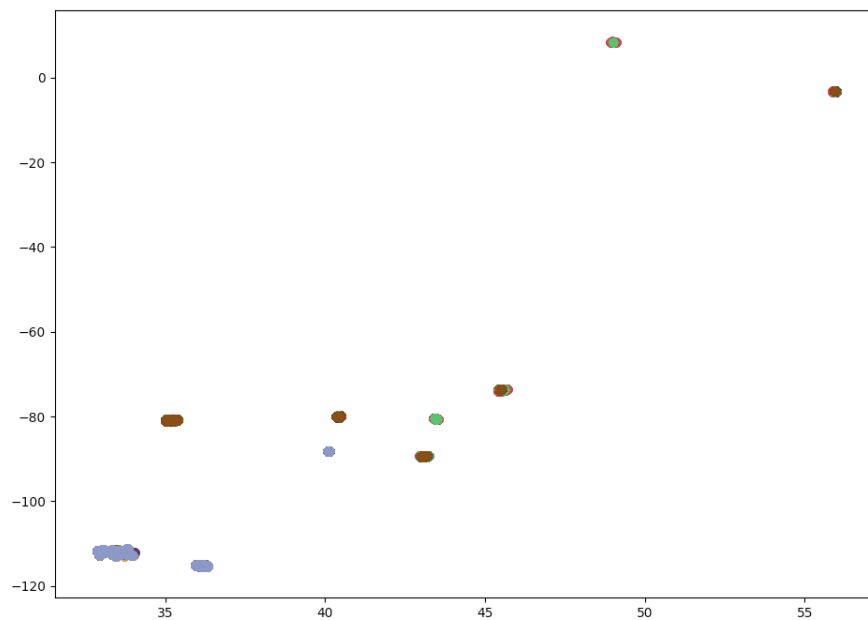


- Choose an appropriate K from the plot and argue why you choose this particular K.

I would choose K equal to 9 because it is the lowest point

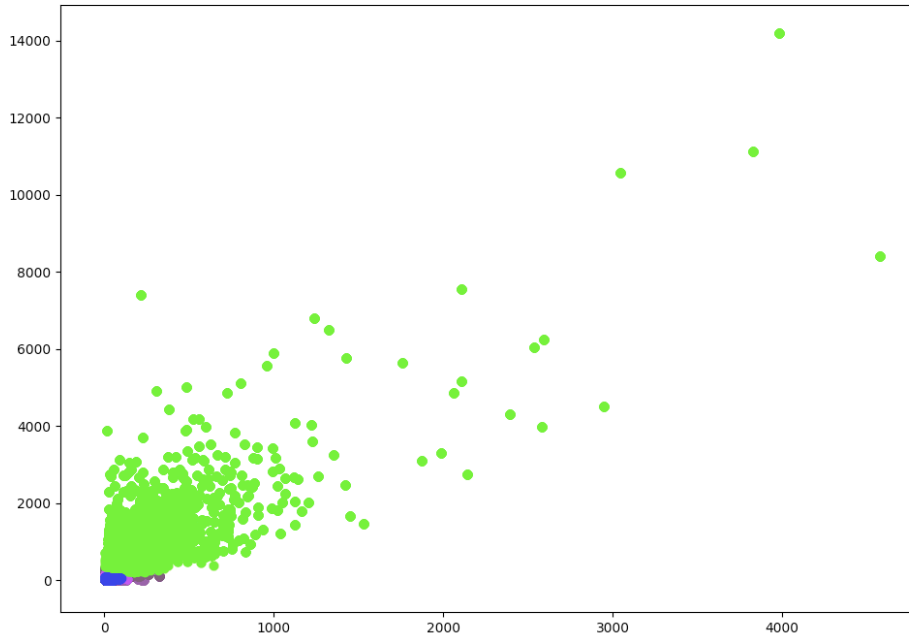
- For the chosen value of K, plot the clusters with their centroids in two ways: first using latitude vs. longitude and second using reviewCount, checkins. Discuss whether any patterns are visible.

Latitude vs. longitude:



The cluster is clearer than part i.

ReviewCount vs. checkins:



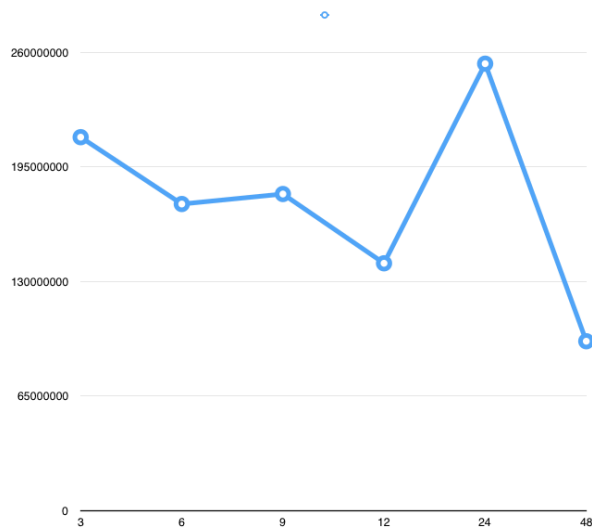
Because the lower left corner has more data, there is a lot of clusters in that region. The result is similar to normalizing the data.

(v)

- Describe how you expect the down-sampling to change the clustering results.

I would expect there is less points in the plot. Plot will be less dense.

- Plot the within-cluster sum of squares (wc) as a function of K.



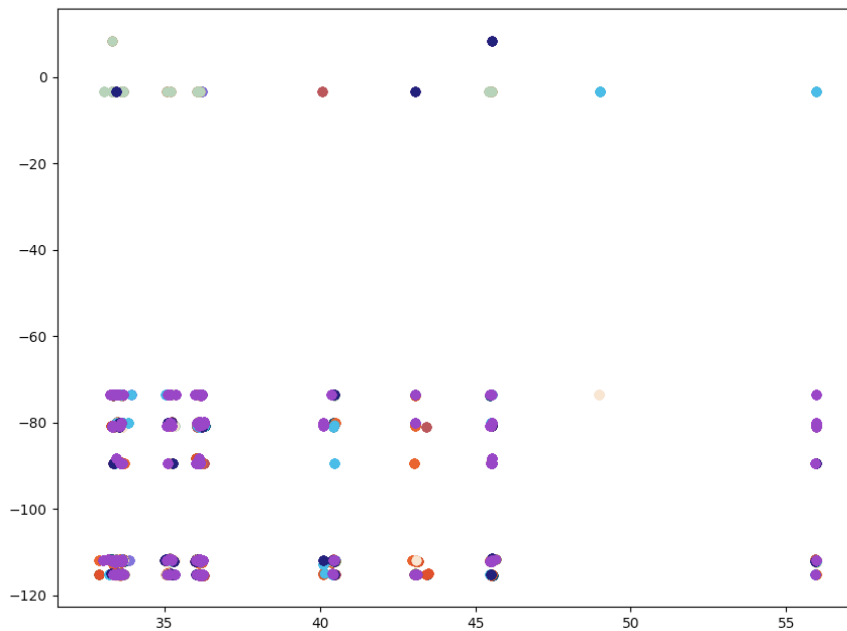


- Choose an appropriate K from the plot and argue why you choose this particular K.

I would choose K equal to **12** because it is the knee point

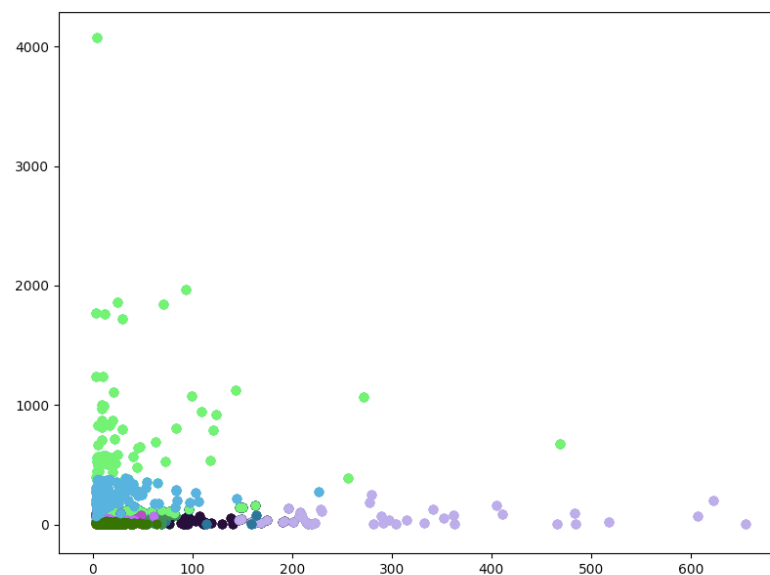
- For the chosen value of K, plot the clusters with their centroids in two ways: first using latitude vs. longitude and second using reviewCount, checkins. Discuss whether any patterns are visible.

Latitude vs. longitude:



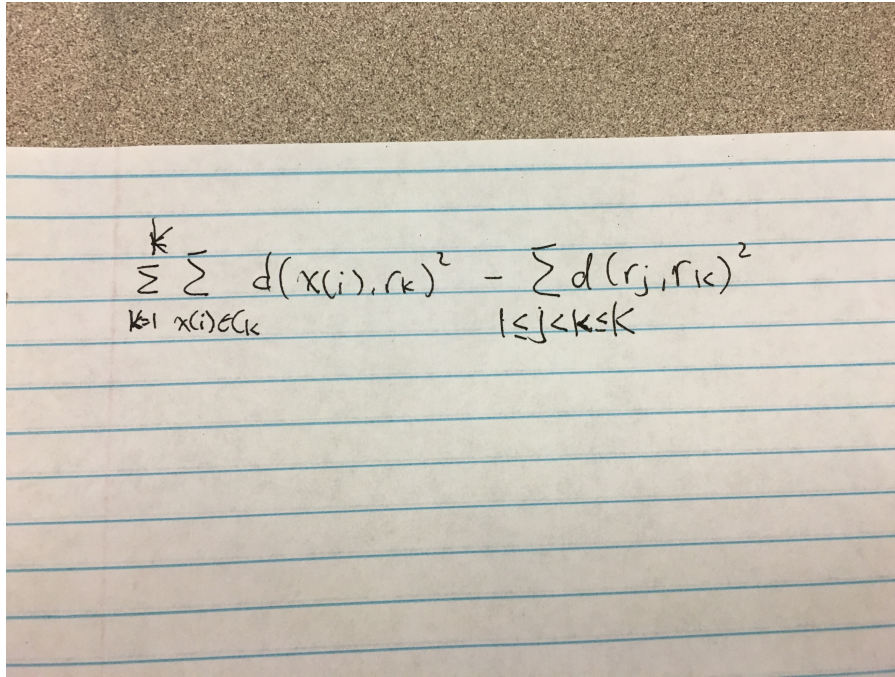
The data is more evenly distributed because the amount of data is less.

ReviewCount vs. checkins:



The data is less dense but still has a clear cluster pattern

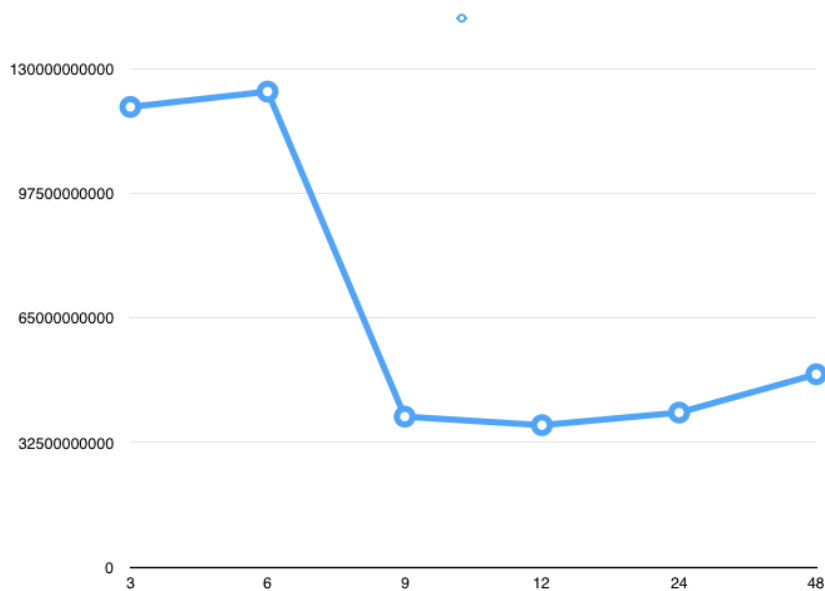
3.



Handwritten formula on lined paper:

$$\sum_{k=1}^K \sum_{x(i) \in C_k} d(x(i), r_k)^2 - \sum_{1 \leq j < k \leq K} d(r_j, r_k)^2$$

My scoring function also takes in account of the between-cluster distance. I subtract the between cluster distance from the within-cluster distance because the farther away, the better the cluster is.



The general shape looks almost the same, and therefore, the choice of K is the same.