Raman Kahlon
PUID: 00291-51712

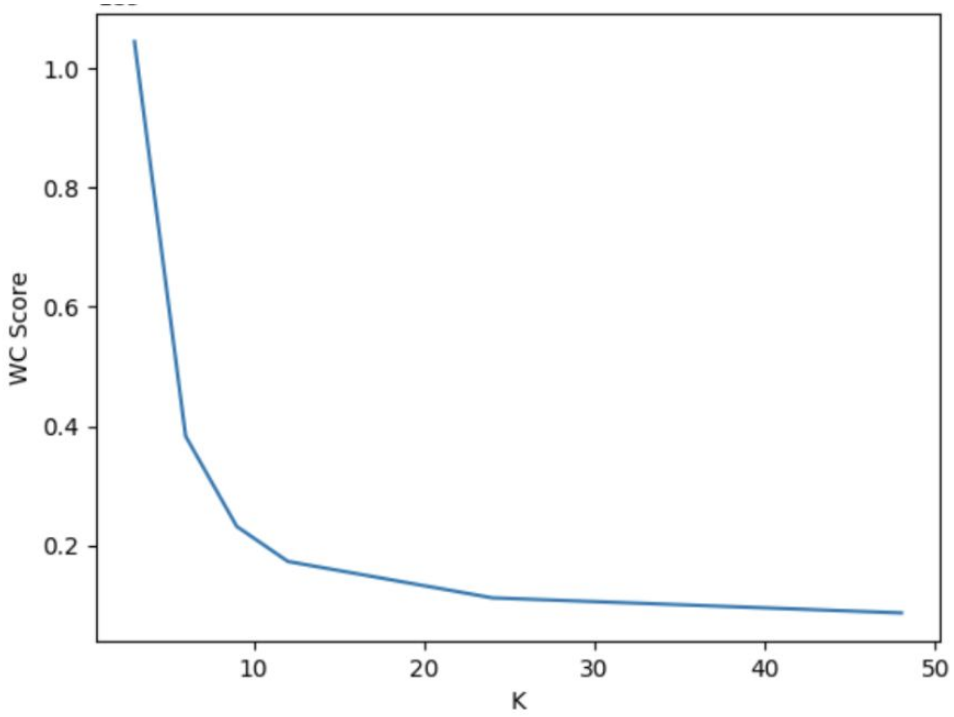
HW 5 Report

Part 2.1 Theory

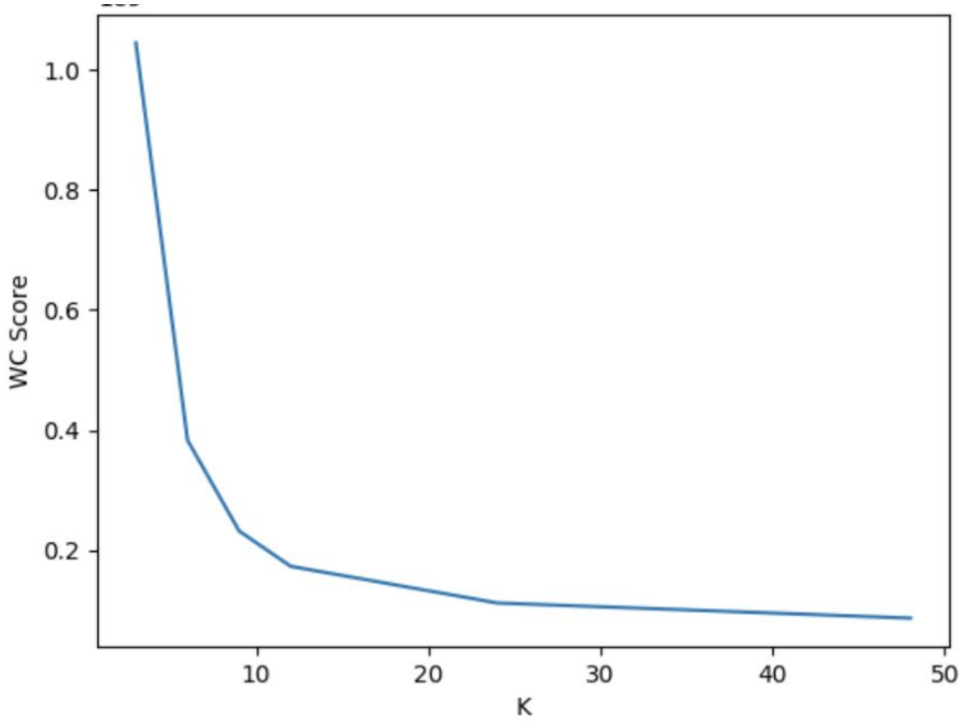
For large variables, K-means runs computationally faster than other algorithms if K is small. K-means also produces tighter clusters than hierarchical clustering, especially if the clusters are globular. The disadvantages include: it is difficult to predict the K value, the algorithm wouldn't work well with global clusters, different initial partitions can result in different final clusters, and K-means doesn't work well with clusters of different size and different density.

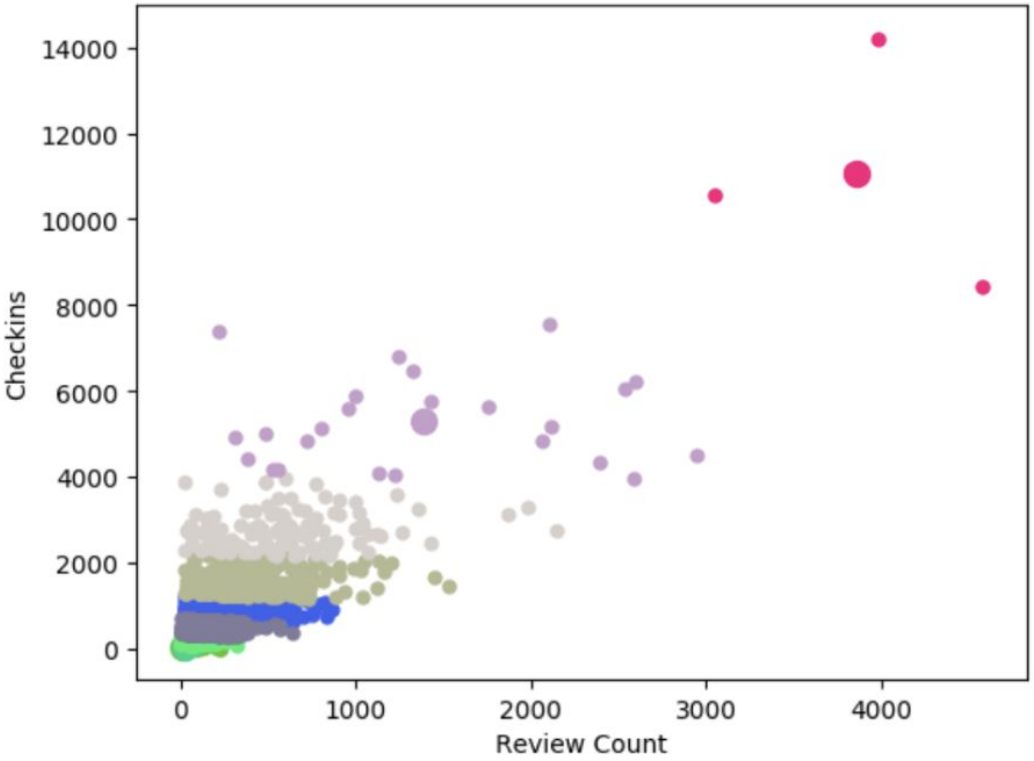
Part 3 Analysis

1.



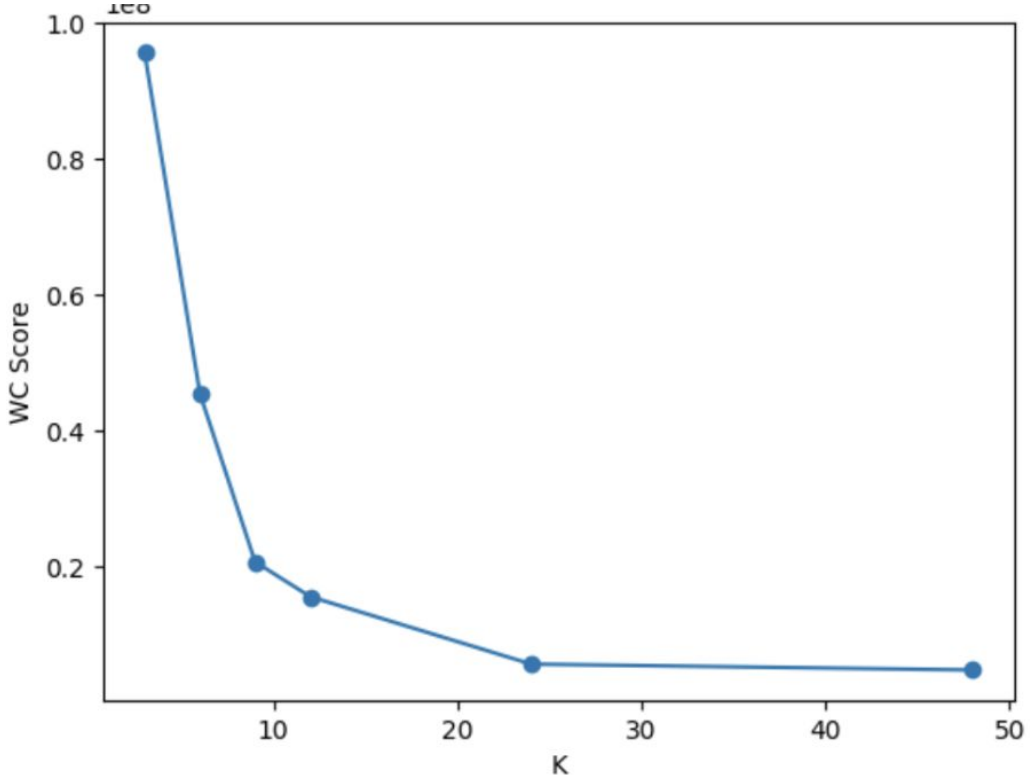
This is the plot that is formed when the algorithm uses given K values. I chose K = 9 because, as we can observe, the slope begins to become flat after K = 9. As we increase K after K = 9, the algorithm decreases in efficiency as it takes a longer time.



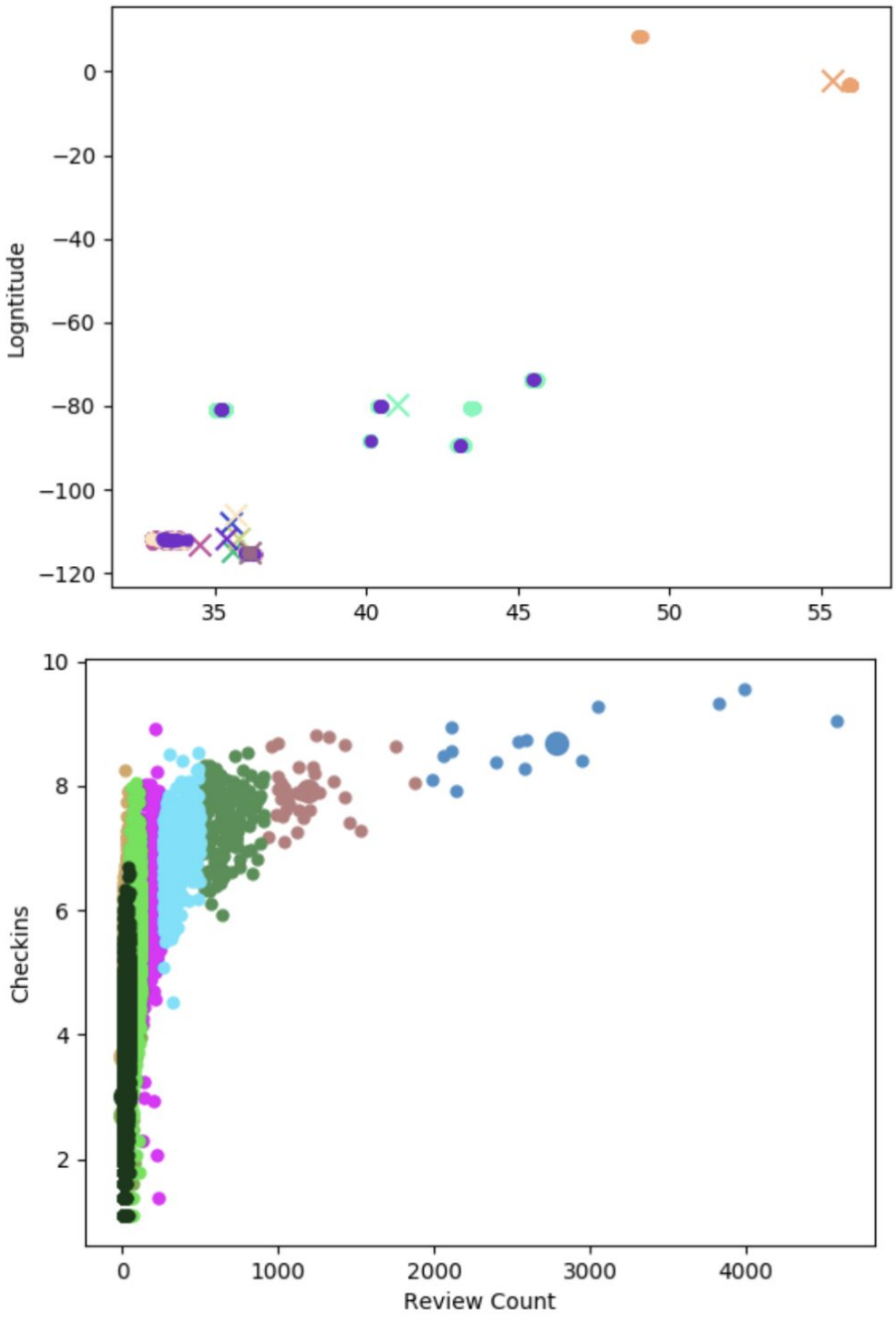


The attributes reviewCount and checkins are clearly correlated, as we can observe on the plot above. In the latitude vs. longitude plot, however, they are not because since the restaurants are in different cities, we wouldn't expect there to be a correlation.

2. The K vs. wc plot would be almost the same as the earlier plot, but it will have decreased (lower) wc score values.

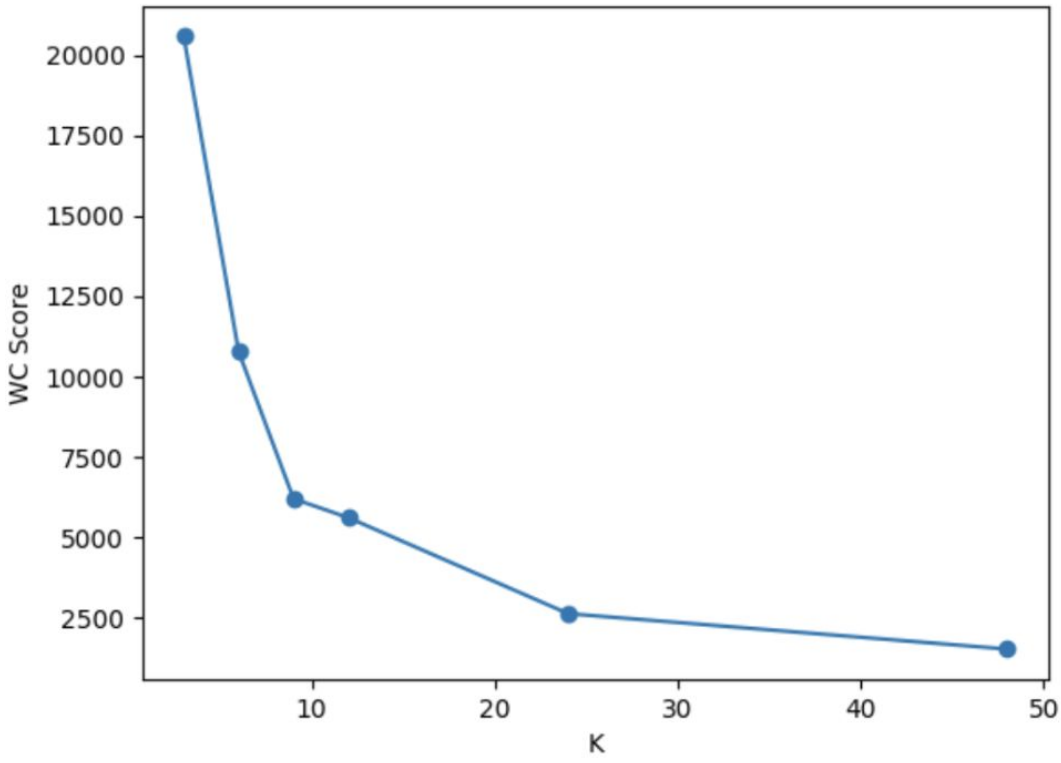


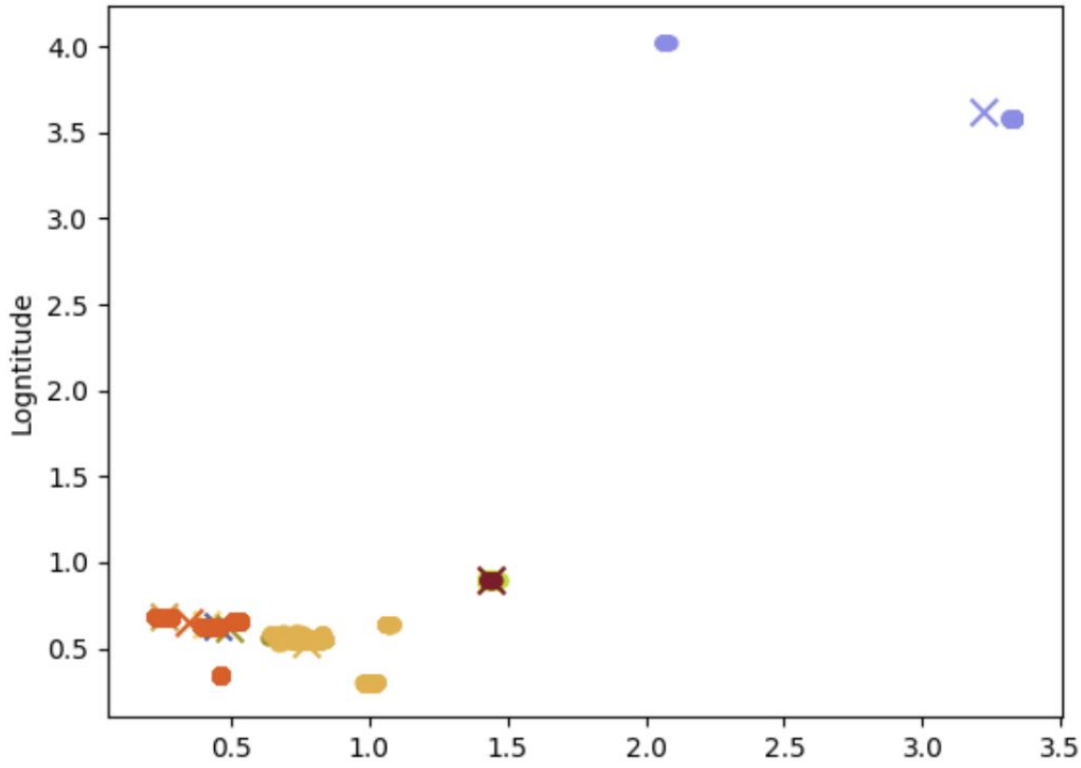
For the two other plots, I wouldn't expect a lot of change among the correlations. The figures might still have different shapes, though.

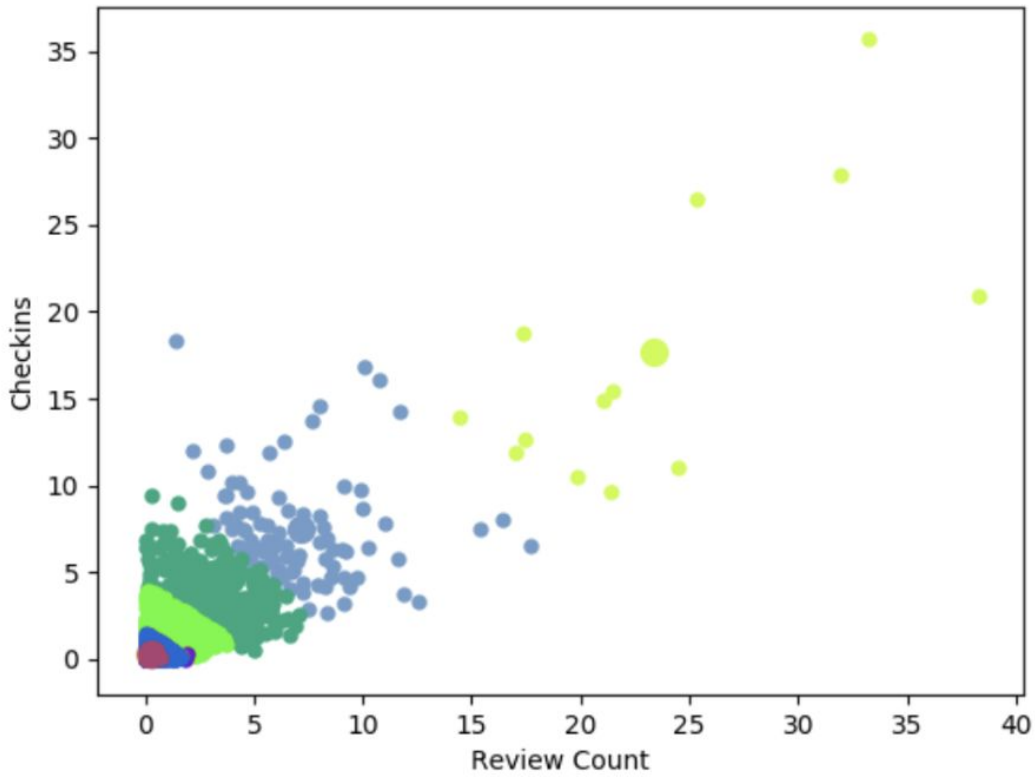


The coordinates still don't have any correlation. The second plot's shape is different than the first because of the log transformation. This would lower the number values.

3. I wouldn't expect a lot of change since the values are decreased by the same rate, however, the I would expect the wc scores to be lower.

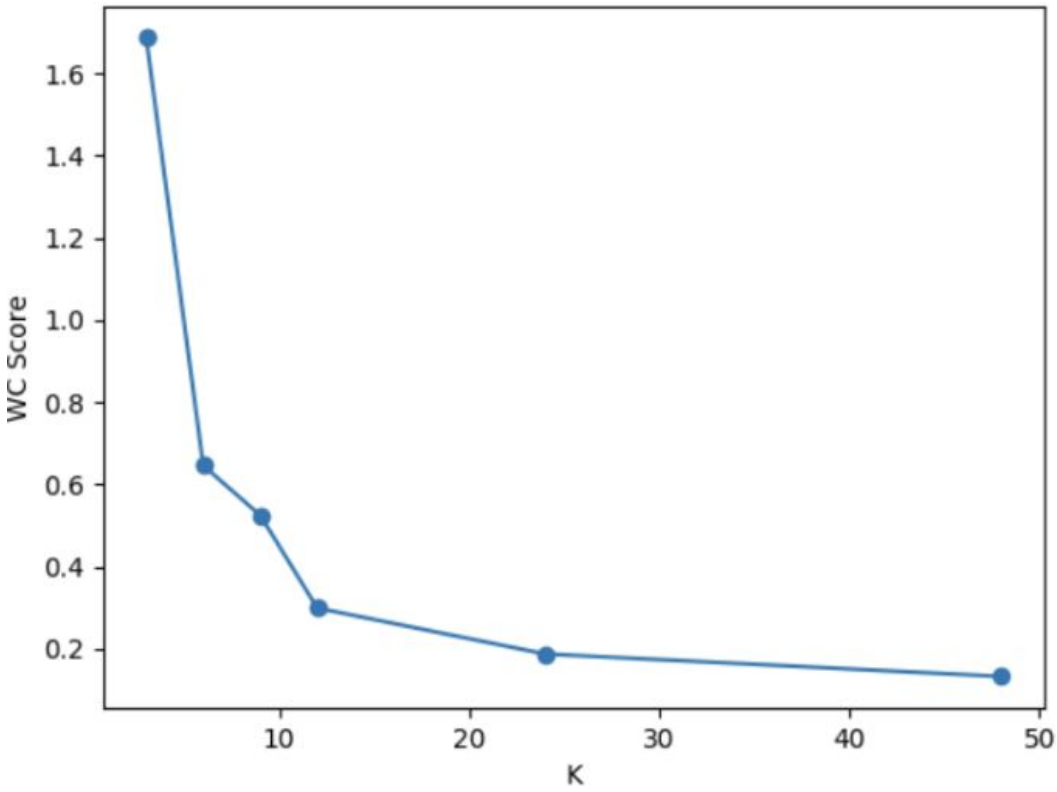

WC Score
20000
17500
15000
12500
10000
7500
5000
2500
10
20
30
40
50
K

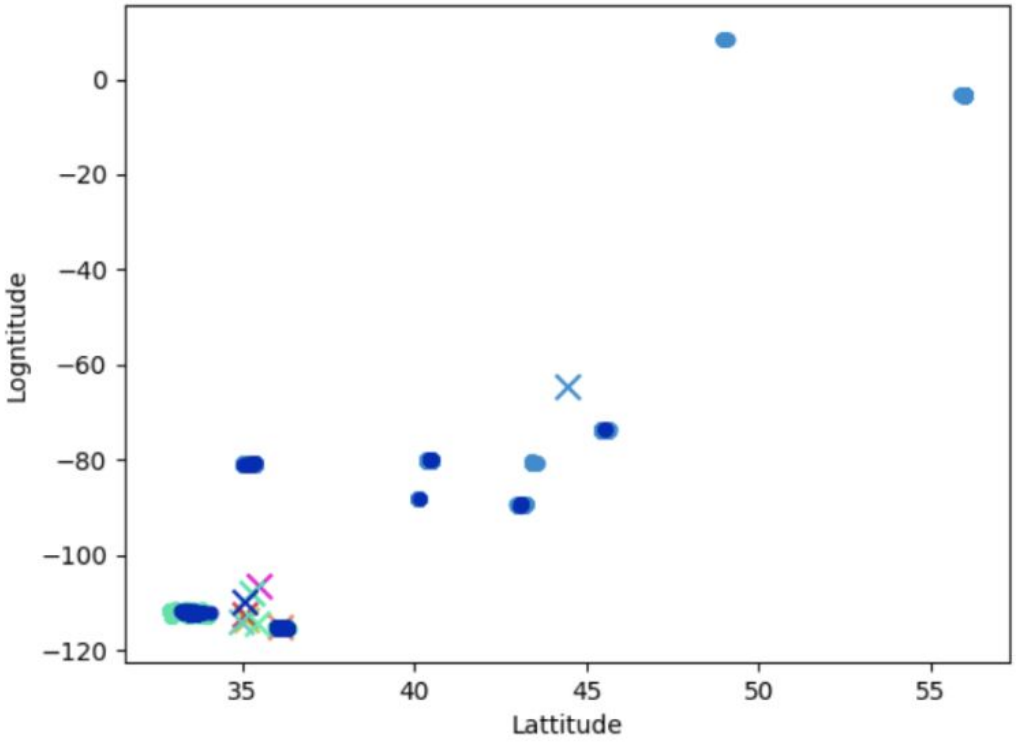

Logntitude
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.5
1.0
1.5
2.0
2.5
3.0
3.5

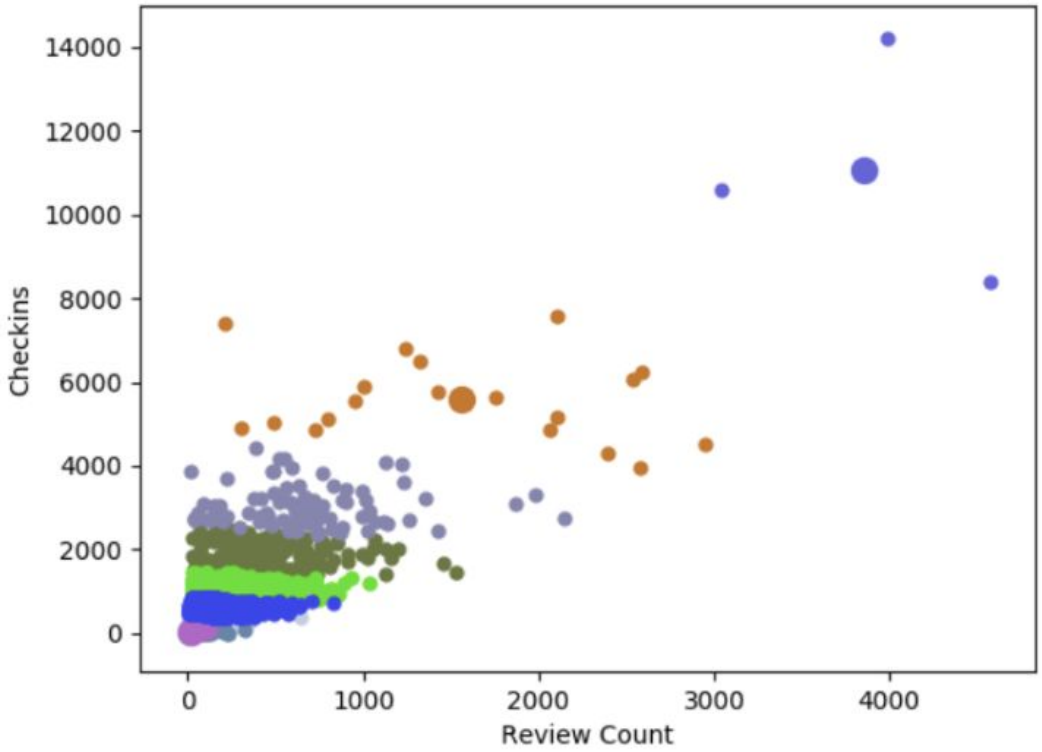


The shapes of the plots remain the same but the numbers have decreased drastically.

4. I wouldn't expect much of a difference in the shapes of the plots, but the values would clearly be different. The wc score would differ, especially near K = 9 when the slope begins to flatten, but the plot generally looks like the original plot.

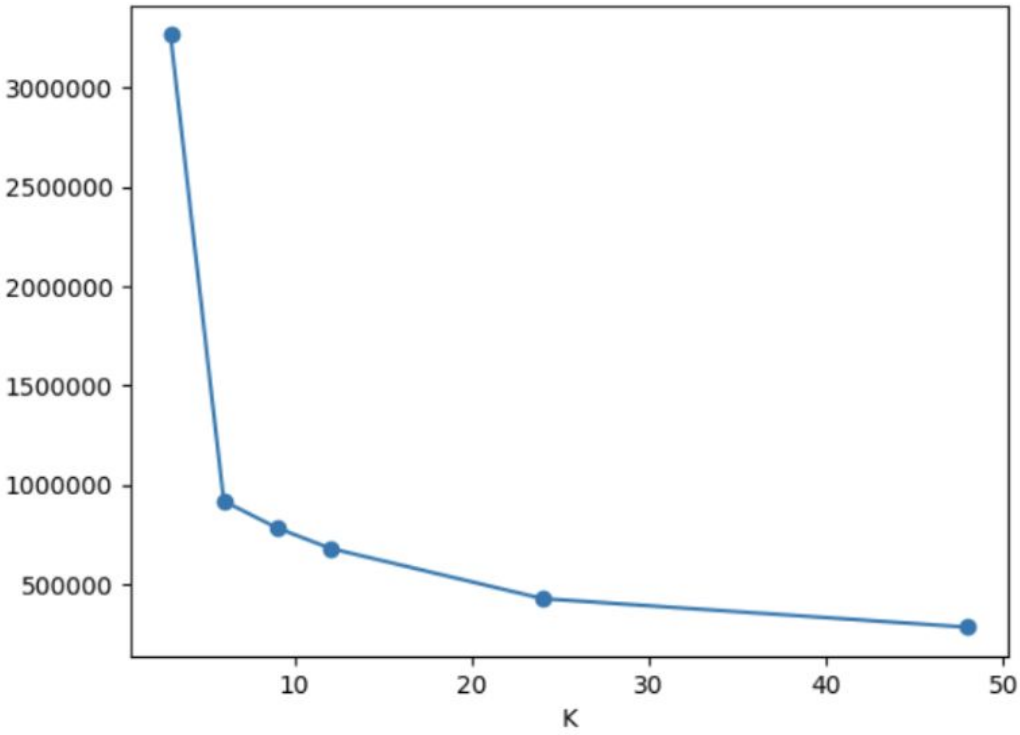

WC Score
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
10
20
30
40
50
K


Logntitude
0
−20
−40
−60
−80
−100
−120
35
40
45
50
55
Lattitude

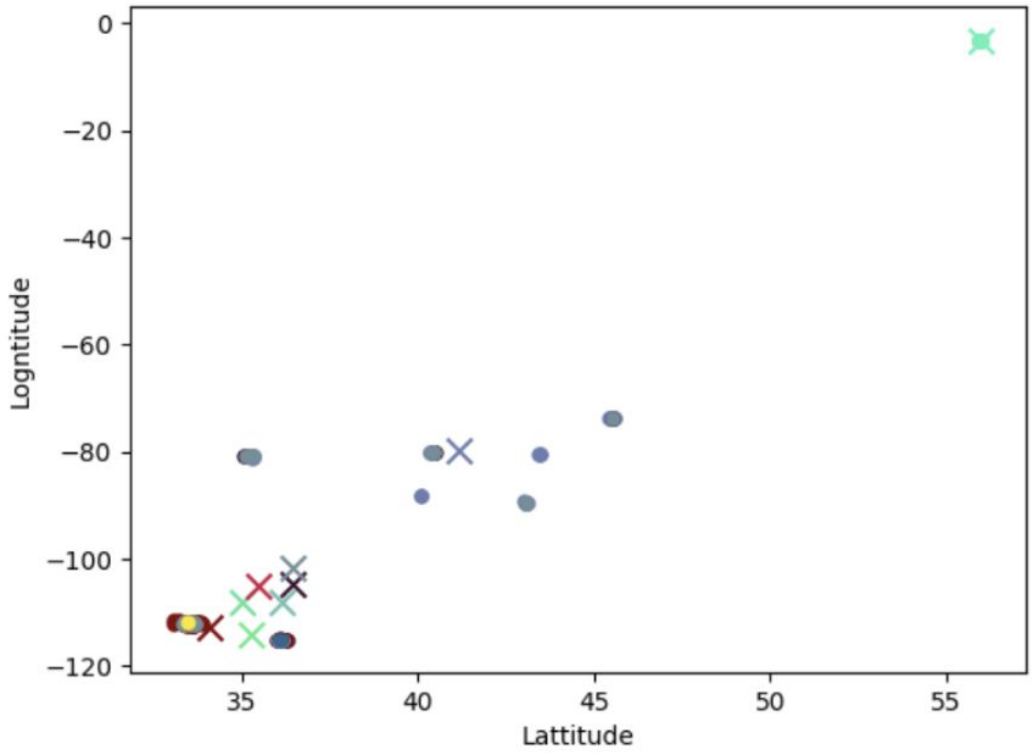


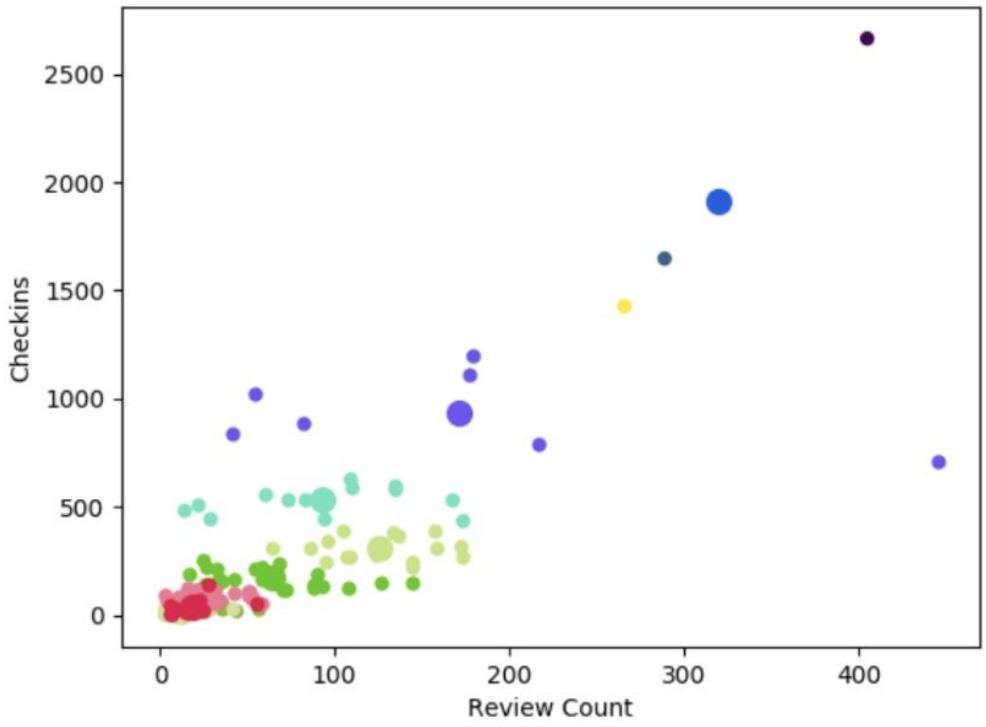
5. The wc vs. K graph wouldn't change since the random function produces an unbiased value. I would expect more scattered figures for reviewCount vs. checkins and latitude vs. longitude, despite the fact that the conclusions wouldn't change.

   Average wc vs. K plot:



Centroids from closes trial:



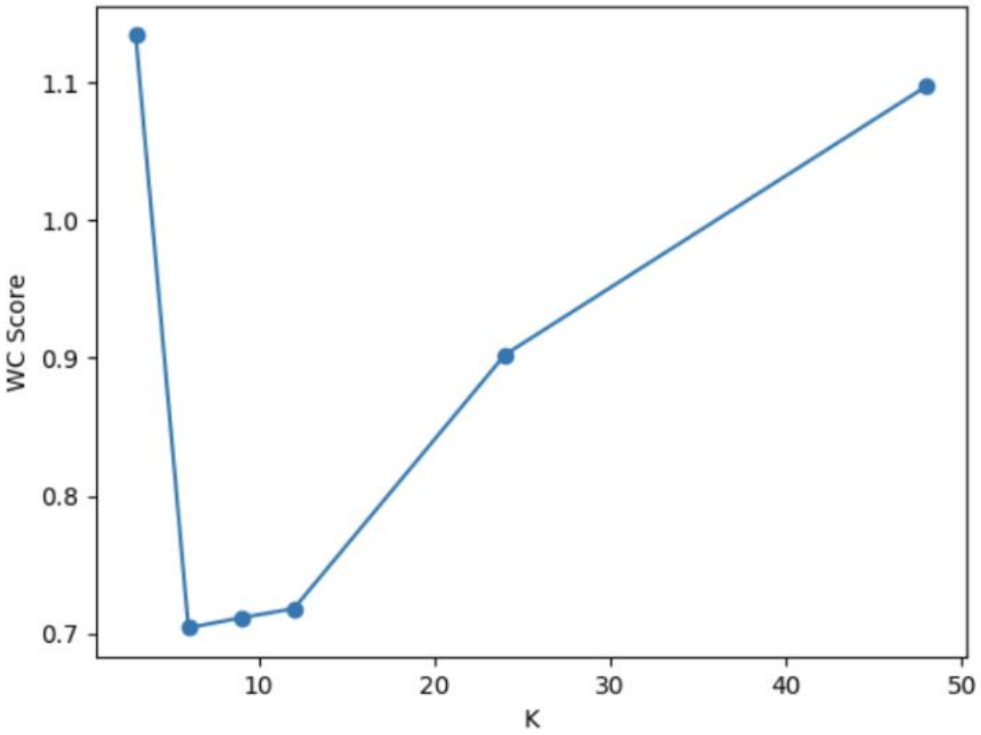


6.

$$Score(C) = \sum_{k=1}^{K} \left(\sum_{x(i) \epsilon C_k} (d(x(i), r_k)^2 + \sum_{C_m \in C} d(r_m, r_k)^2 * M_k) \right)$$

where M_k is the number of points C_k have

We need to use within-cluster sum of squares and the squared distance of a centroid from all other centroids multiplied by the number of members it has so that bigger clusters would have a greater weight. This would be a more accurate approximation than using only within-cluster sum of squares since it would take into account the weight of every cluster and the relative distances between the clusters.



The scoring function indicates that the error score is too high between K = 6 and K = 12. We should choose K = 6, 9, or 12.