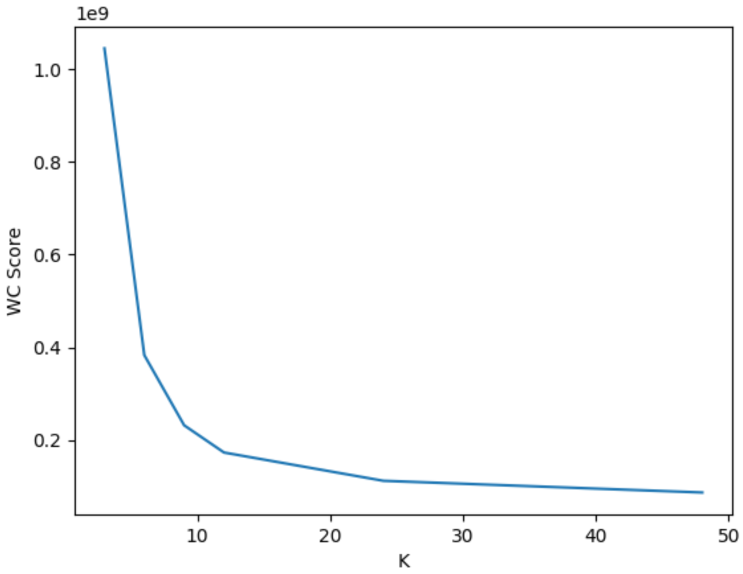
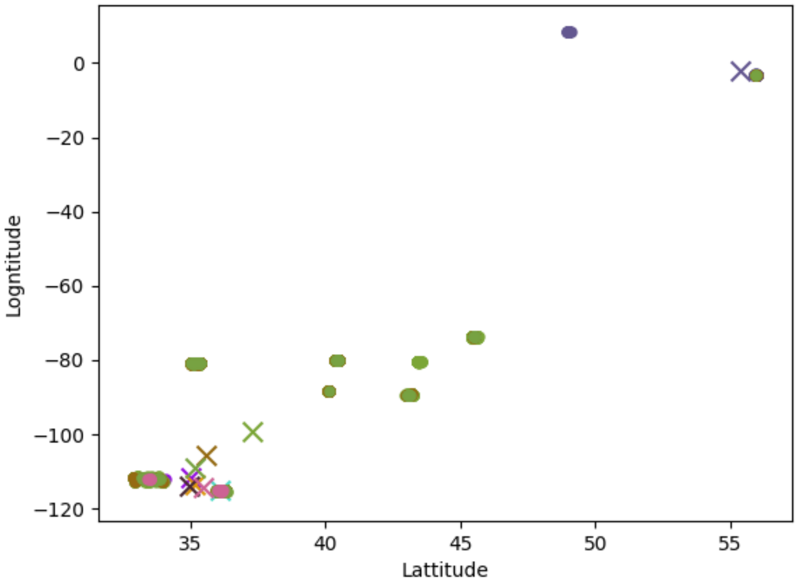
Muhammed Onus

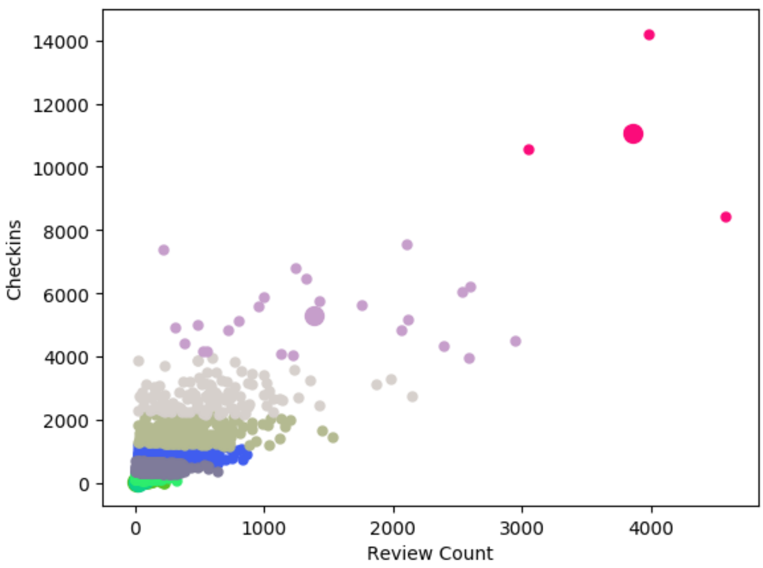
**Q2)**

**i)**



This figure is the main figure where algorithm uses given K values. I choose 9 as my K because it is obvious that after 9 line becomes steep and increasing K stops being efficient when we think how much time it takes.

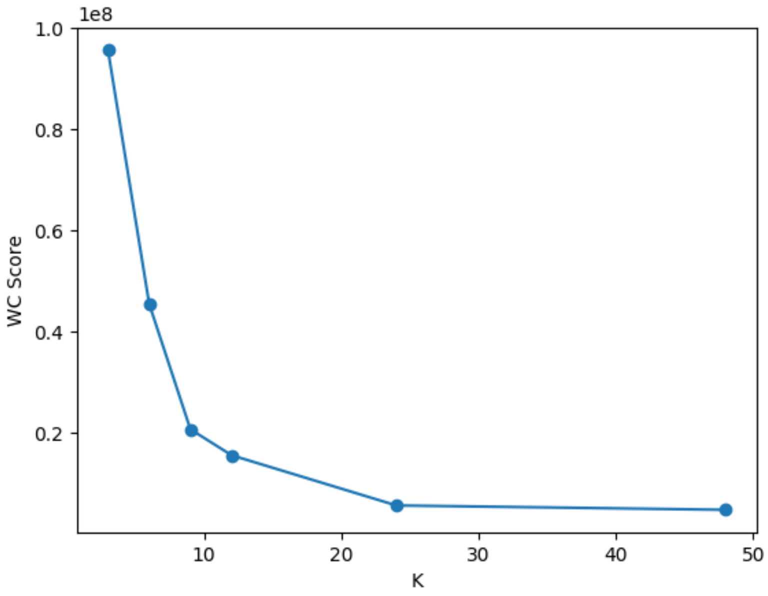




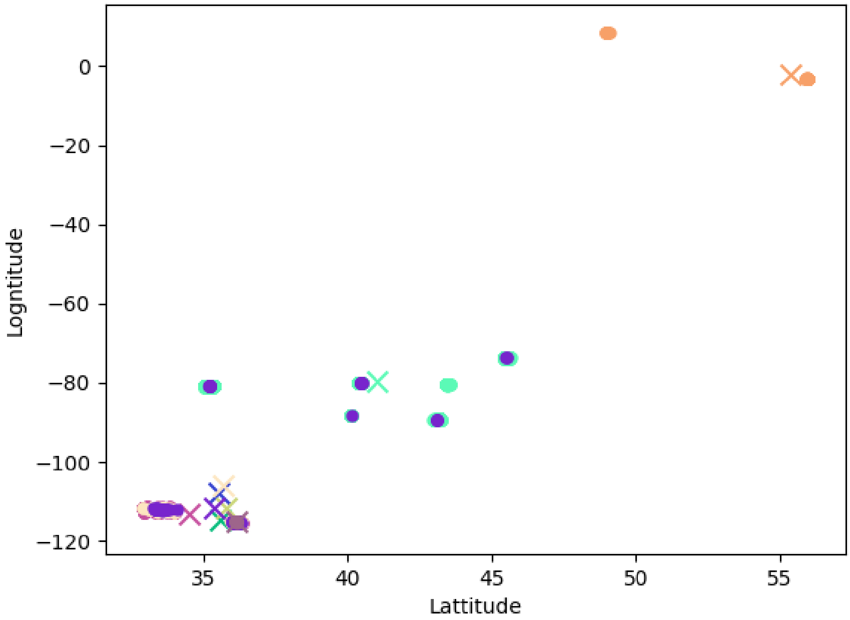
It is obvious that review count and checkins are correlated but we don't see such thing in latitude v longitude figure because they are all in different cities and a correlation wouldn't make any sense.

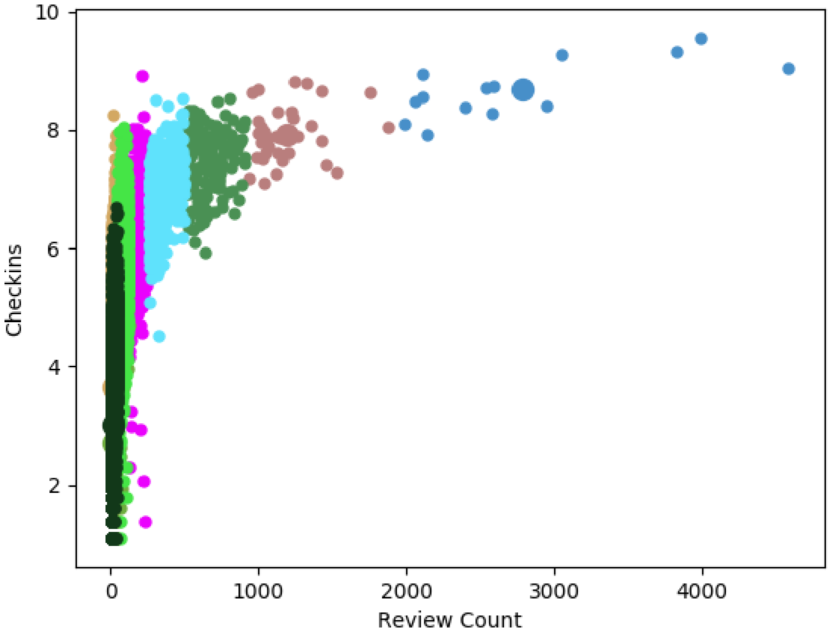
**ii)**

K vs wc would be almost the same but with lower WC score values.



For two other plots I wouldn't expect so much change about correlations but probably figure shapes will be different.

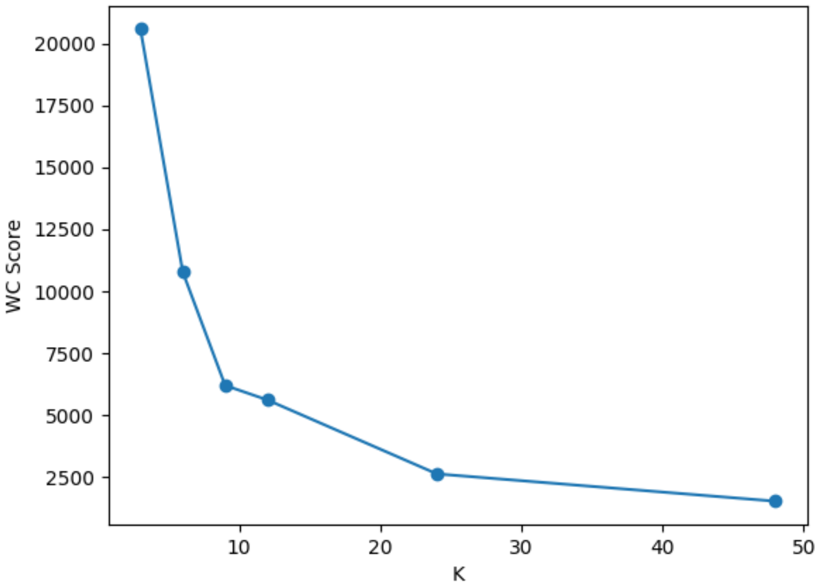


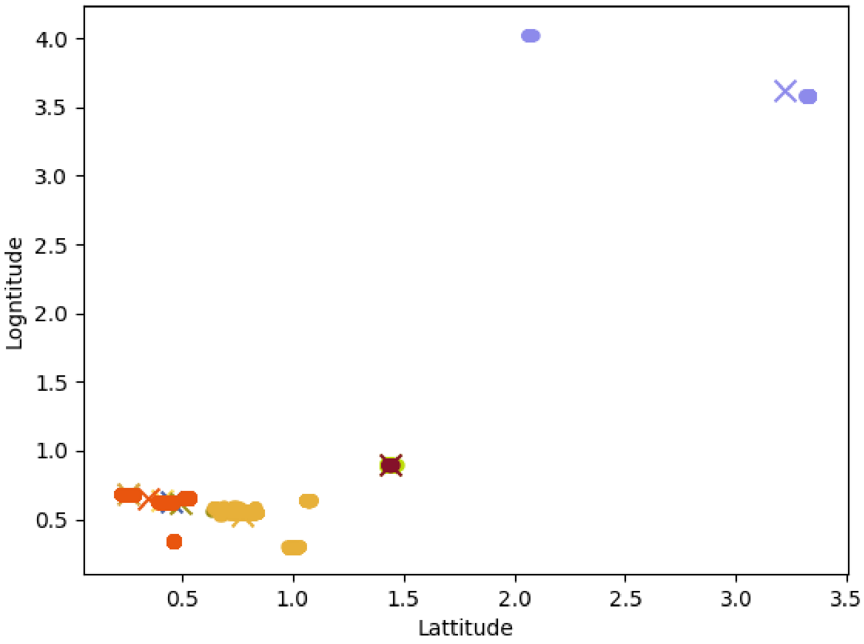


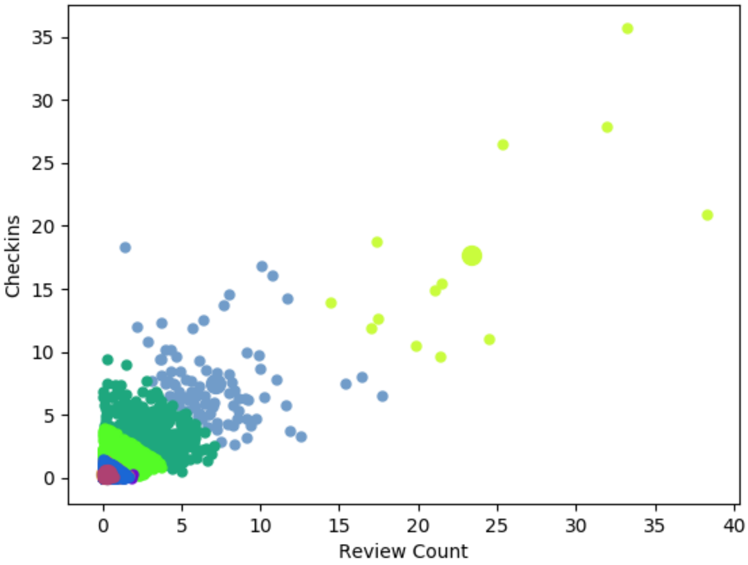
As expected coordinates still have no correlation but the second figure's shape is different because log transform lowers the value of numbers.

**iii)**

I don't expect much change because numbers are lowered in the same ratio. But WC scores would be lower.


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

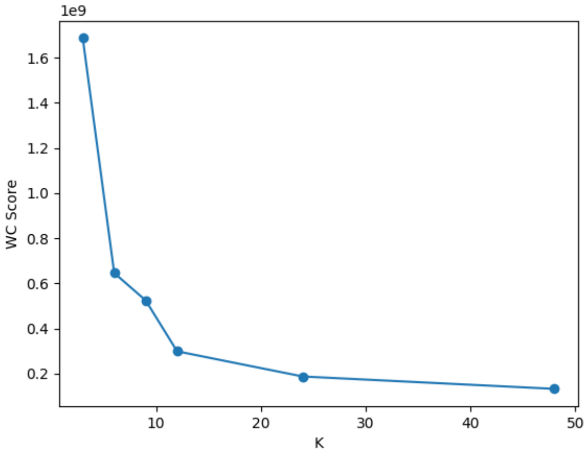

Logntitude
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Lattitude

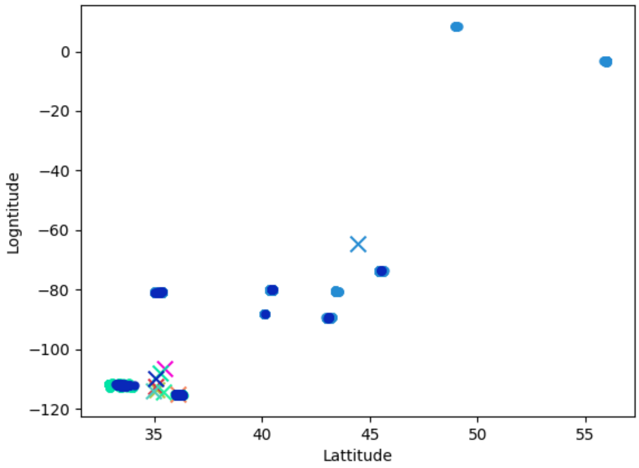


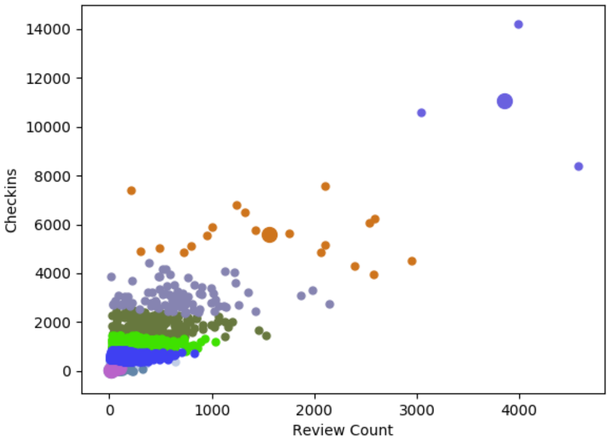
As expected there is no change in the shapes but of course numbers are much lower.

**iv)**

I don't expect so much change in the shapes but numbers would be different. However WC score looks a bit different around K=9, but in general it also looks similar to the original one.

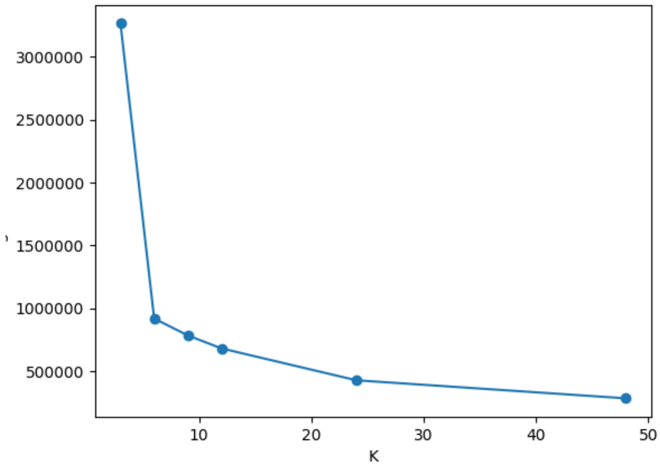




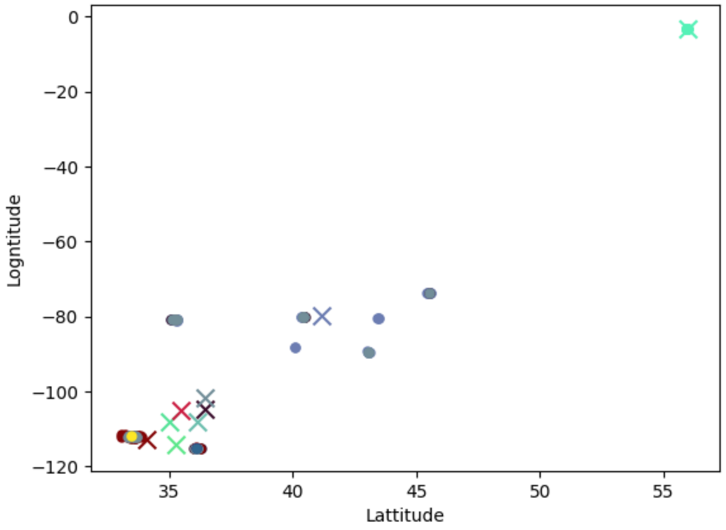


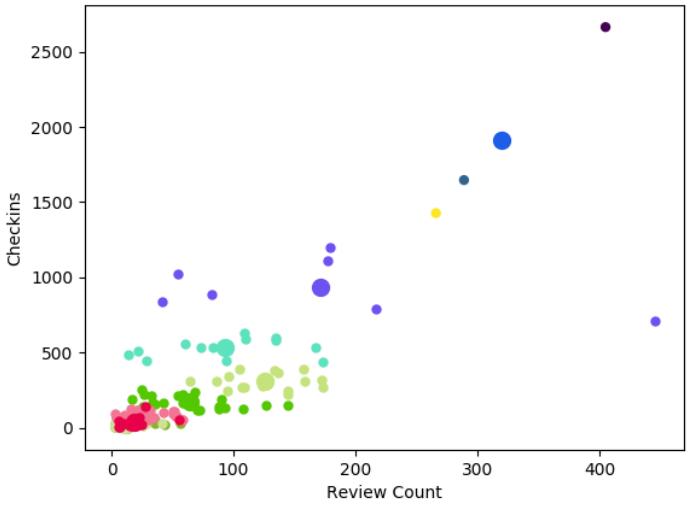
**v)**

wc vs K graph won’t change because random function is unbiased. But I would expect a more scattered figures for lat vs long and review vs checkin even though conclusions won’t change. Here is average wc vs. K figure:



Centroids from the closest trial:



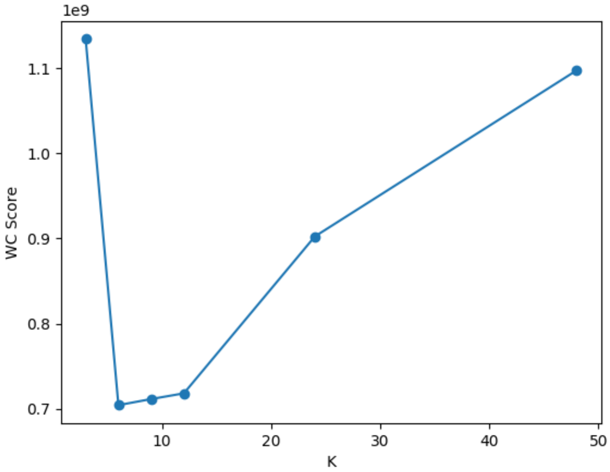


**Q3)**

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

$$where\ M_k\ is\ the\ number\ of\ points\ C_k\ have$$

Basically, we use WC and squared distance of a centroid from other centroids multiplied by the number of members it has so that bigger clusters weigh more. This is more accurate than only WC because it considers weight of every cluster along with the distances between clusters.



Scoring function shows that after 12 and before 6 error score is too high. So, we should choose 6, 9 or 12 as K.