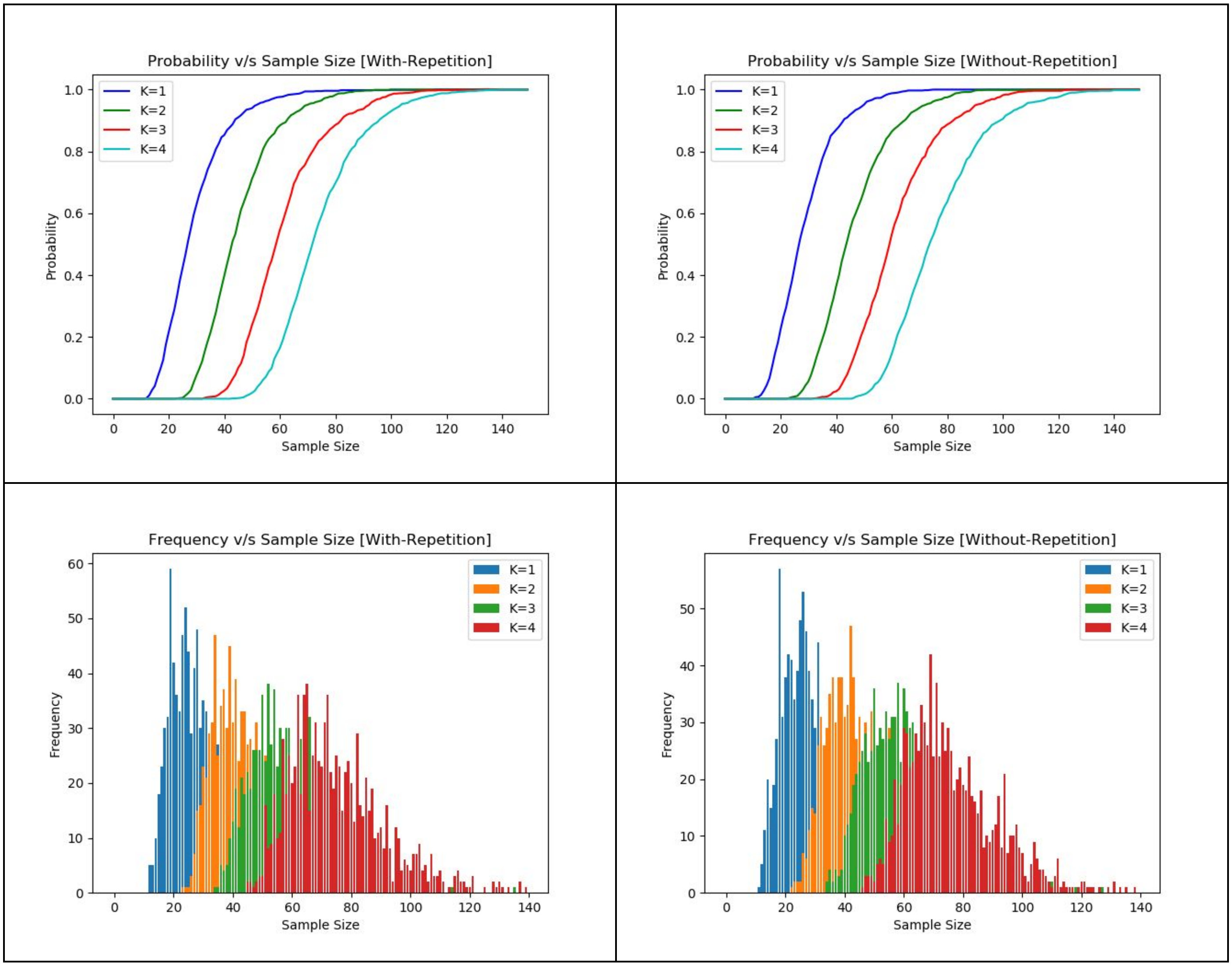
Data Mining

Report A1

Question 1
Report, Inference attached with each Graph in Tableau Tool.

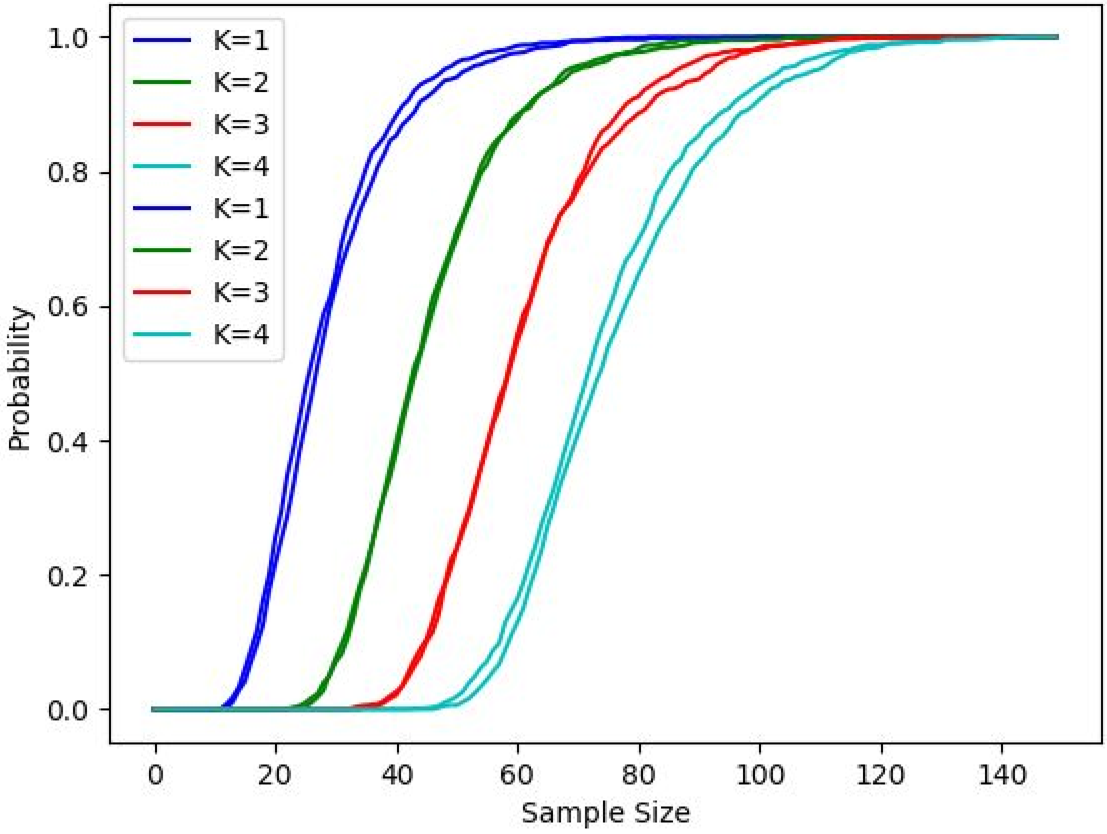
Question 2
Generated random dataset using make_blob library with 1000 samples with 2 features each of 10 classes.

To perform sampling (with and without repetitions) on the dataset I did 1000 trials for each K values and calculated the cumulative count required to satisfy the condition of having at least k elements of each class.



I observed that more trials are required to get at least K elements for each class in with repetition. Which is logically correct as samples per class decreases in without replacement case. Below graphs show that for larger K values without repetition will take more sample size than with repetition case. From the frequency graph we can see that without repetition histogram has higher frequency for large sample size than with replacement graph. Also comparing the highest peak value in both graphs, we can see that without repetition has highest peak for each K value with larger sample size.

With simulation we can see that the sample size required to get at least K elements for each class is a lot less than the worst case.



Probability vs Sample size [With and Without Repetition combined]