Problem Statement 1

Improved Apriori Techniques (Sampling and Partitioning)

Take Dataset from Kaggle
https://www.kaggle.com/datasets/newshuntkannada/dataset-for-apriori-and-fp-growth-algorithm

Apply Sampling and Partitioning techniques of Improved Apriori. Propose your own technique to select the number of samples and number of partitions with proper justification. Show performance analysis with different number of samples and partitions with different support and confidence to find frequent items. Write your conclusion based on it.
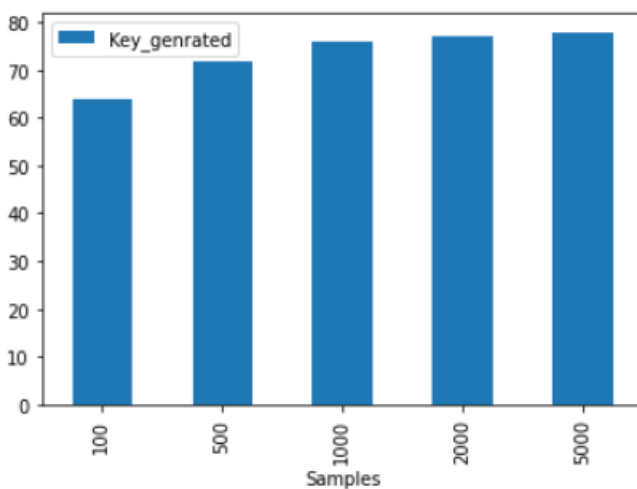
Sampling techniques of Improved Apriori:

Performance analysis with different number of samples for confidence level=0.30 and min support=0.10
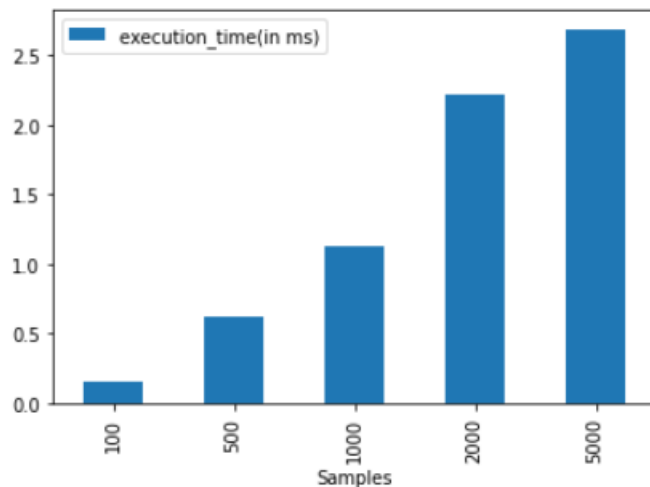
Execution Time:
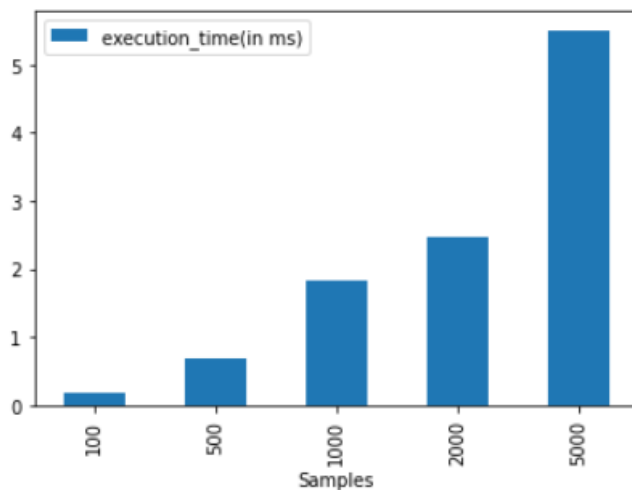
Key Generated:

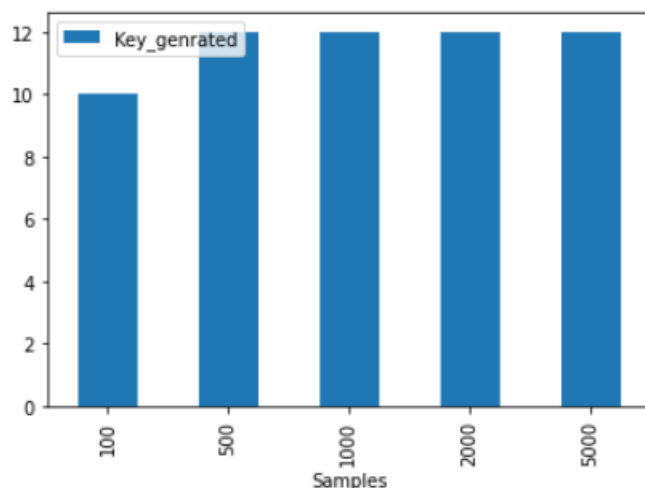Performance analysis with different number of samples for confidence level=0.30 and min support=0.30

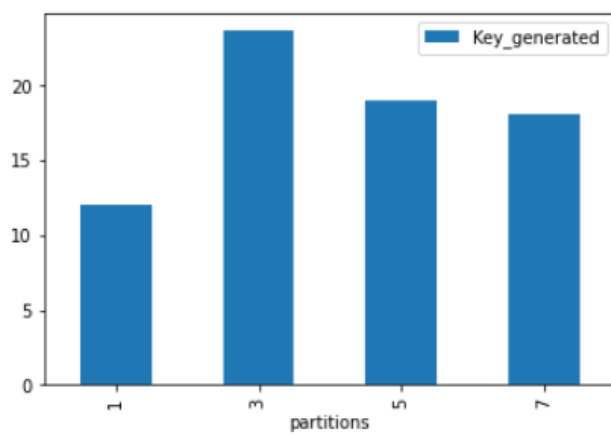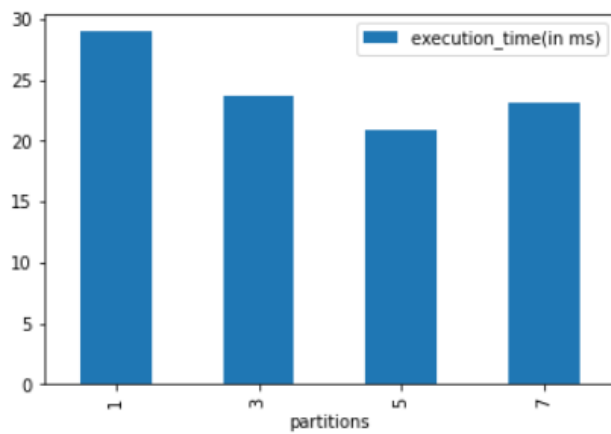***Execution Time:***

Key Generated:

Conclusion:

After observing the above output, we can see that as the sample number is increasing the execution time and key generated are increasing and after particular number like in this case 1000 sample key generated are appr. Constant, so this can be the benefit of sampling as at 1000 samples the execution time is comparatively very low with respect to overall dataset's execution time. We can use 1000 sample for apriori algorithm with almost desired output and less time complexity.
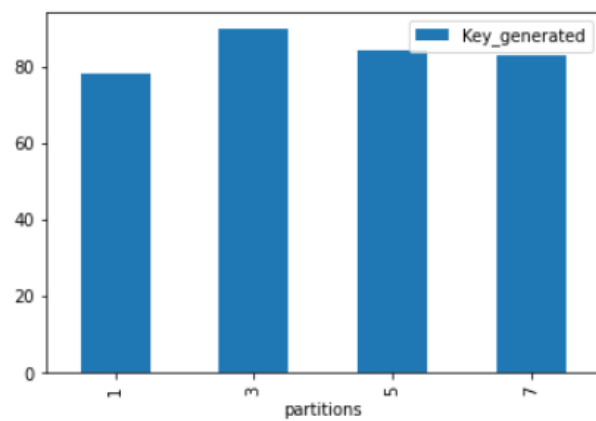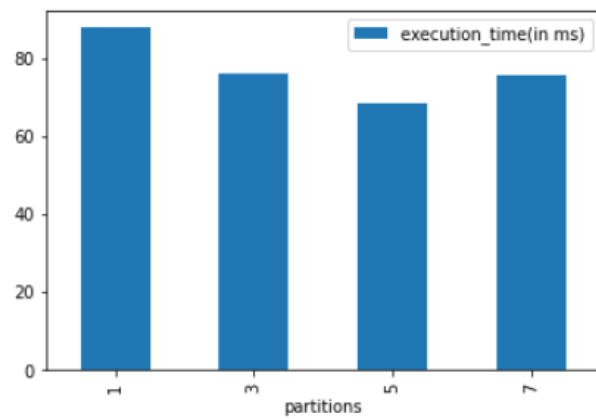
Choosing sample number: As per observation we can take sample number as total transactions divide by total items. Like in this case 8732/7= 1247.

Partitioning techniques of Improved Apriori:

Performance analysis with different partitions with min support 0.30:

Performance analysis with different partitions with min support 0.10:

Conclusion: After observing the above output, we can see that as the partitions are increasing the execution time and key generated are mix. But for 5 Partitions the algorithm giving proper outcomes;