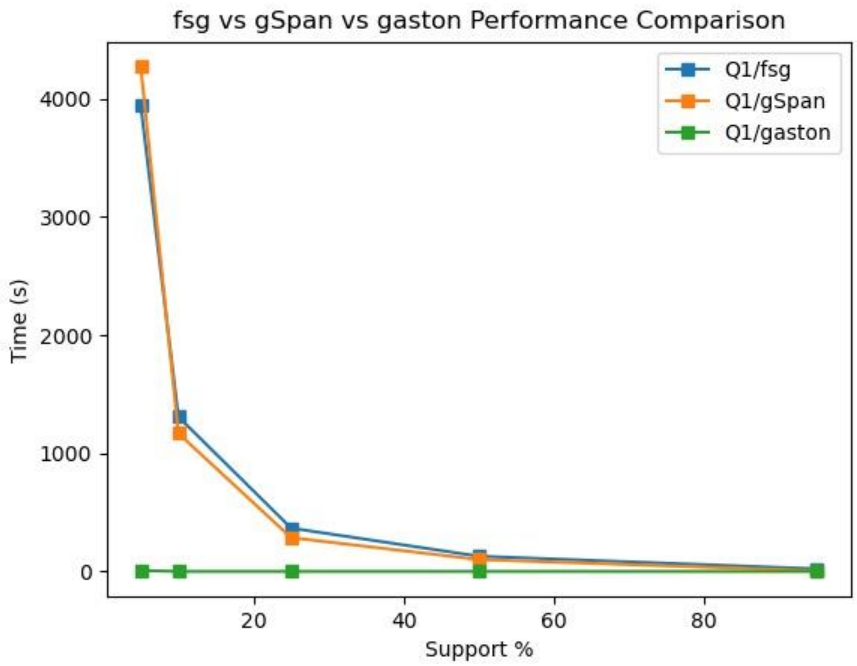
Q1.



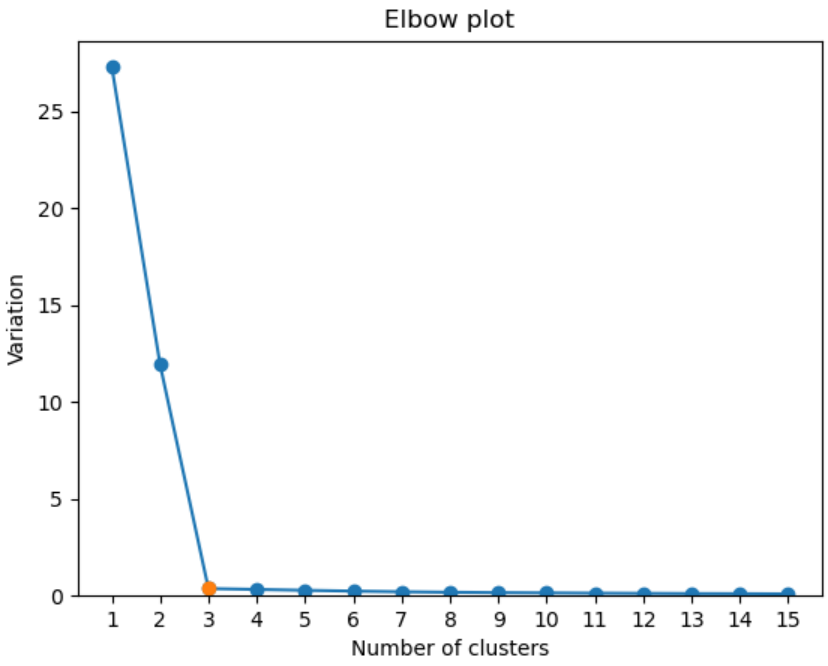
We note the following observations:

1. On increasing the minimum support threshold, the time taken by all the algorithms decreases. This is because the total number of frequent subgraphs decreases. Since these algorithms take advantage of the Apriori property, they do not generate more candidates using infrequent subgraphs, and as a result terminate quicker. On decreasing the support threshold, the growth rate (increase in time) increases. This is because the number of frequent subgraphs grows exponentially and hence, the algorithms take exponentially more time to finish. And when thresholds are very high, growth rate becomes constant.
2. FSG and gSpan take similar amounts of time in most of the support thresholds. This might be because these algorithms differ mostly on the candidate generation step and not on the subgraph isomorphism part. However, gSpan is a bit faster than FSG in support 10-50, the reason being that gSpan does not have to do candidate generation and false candidate test. At 0 support, the number of candidates is extremely high, so the benefit of false candidate generation is not helping and hence the time taken by FSG is less.
3. Gaston performs better than both FSG and gSpan. From the Gaston paper, it seems like gaston is using a special data structure called Embedding List for each candidate subgraph to make the graph frequency counting step much faster than in the other algorithms. FSG and gSpan may keep an array of identifier of graphs in the dataset for which a given candidate is a subgraph of. Embedding lists also include the identifier of the node involved in the last extension for each graph in the array. Using this structure, they might have saved time on subgraph isomorphism. And when the supports are low, these subgraph isomorphism tests will be high (because of more frequent graphs), so gaston performs extremely well there.

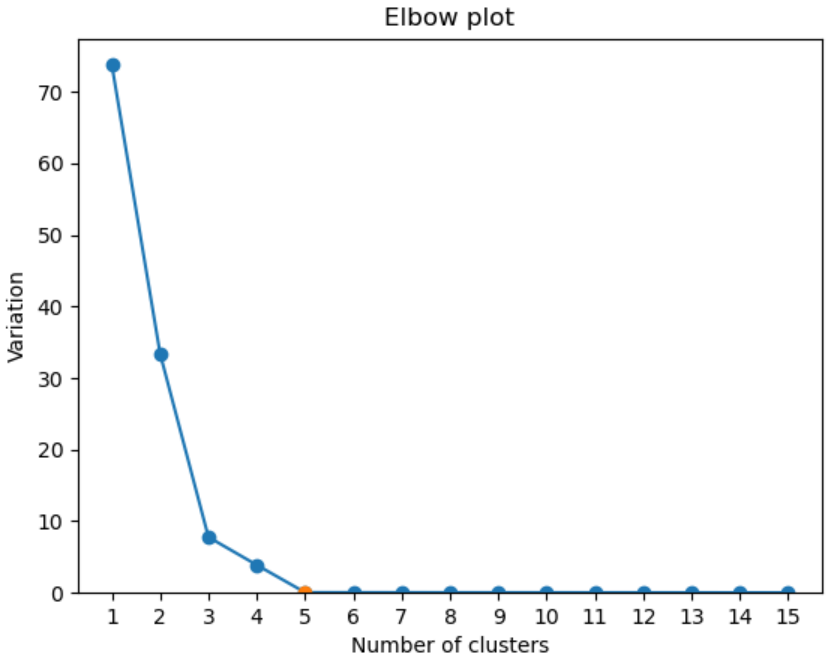
Q3

In the following plots, the red dot represents the elbow point, I.e., the number of clusters.

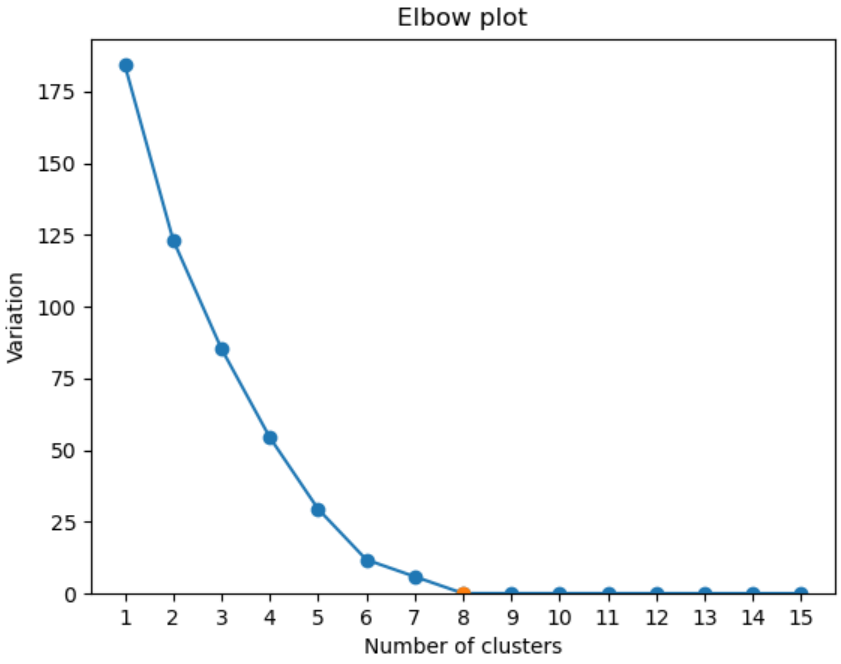
2D



3D


Elbow plot
Variation
Number of clusters
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

4D


Elbow plot
Variation
Number of clusters
0
25
50
75
100
125
150
175
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15