RITWIK SRIVASTAVA - 16303400

MACHINE LEARNING: LAB 4 - CLUSTERING

Challenges Encountered

Coding Challenges Encountered: During the implementation, an error popped up while computing cluster averages, triggering a warning for an "Empty Slice Mean" and encountering invalid values during division. This problem, initially puzzling, was traced back to similar variable names, 'dataset' and 'datasets'.

Use of Yellowbrick: Before installing Yellowbrick, I experimented with various kernels in VSCode, a method previously useful for sorting imports. I was considering an alternate approach but it deviated from lab-4 requirements. I confirmed the necessity of installing Yellowbrick from a friend and proceeded.

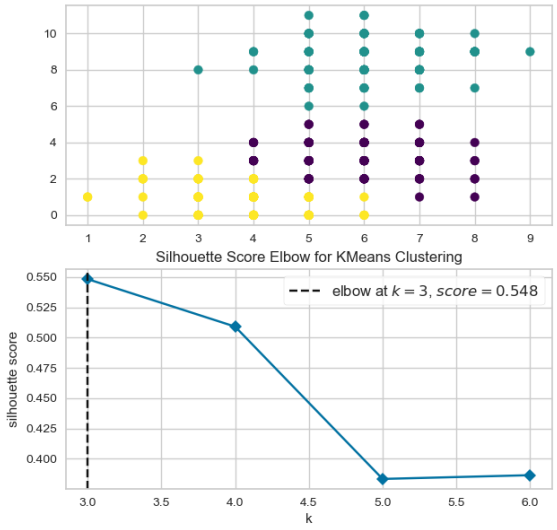
Algorithm Analysis

The implemented K-Means Algorithm customizes clustering according to the given rubrics. Initially cluster centers are allocated randomly and then refined to determine optimal cluster assignments. It begins by setting a maximum iteration limit = 50. Then, distances between data points are calculated and cluster centers to assign each point to the nearest cluster. Arbitrary allocation of clusters for equidistant points occurs randomly. The algorithm updates cluster centers based on the mean of data points assigned to each cluster until convergence or reaching the maximum iterations. This iterative process ensures accurate clustering by adjusting cluster centers to better fit the data distribution.

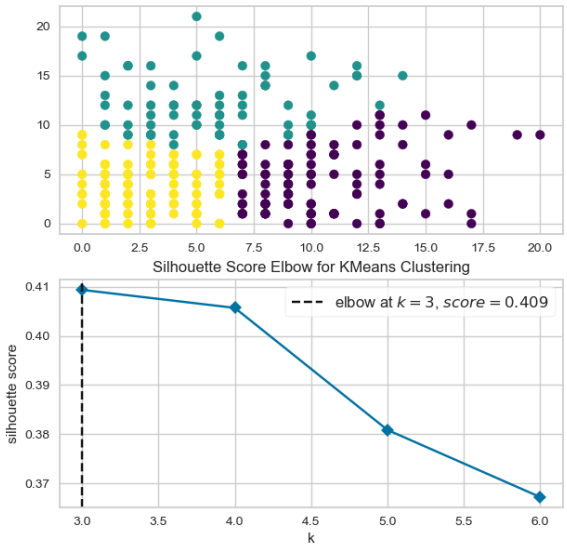
Visualizations and Results

The algorithm visualizes clustered data for different k values between 3 and 6 using K-Means through scatter plots. It also uses silhouette scores, consistently highlighting 3 clusters as optimal, to assess how well-defined and separate these clusters are. This visualization approach provides valuable insights into cluster formations, crucial for evaluating the algorithm's performance. More analysis can be found with the diagrams.

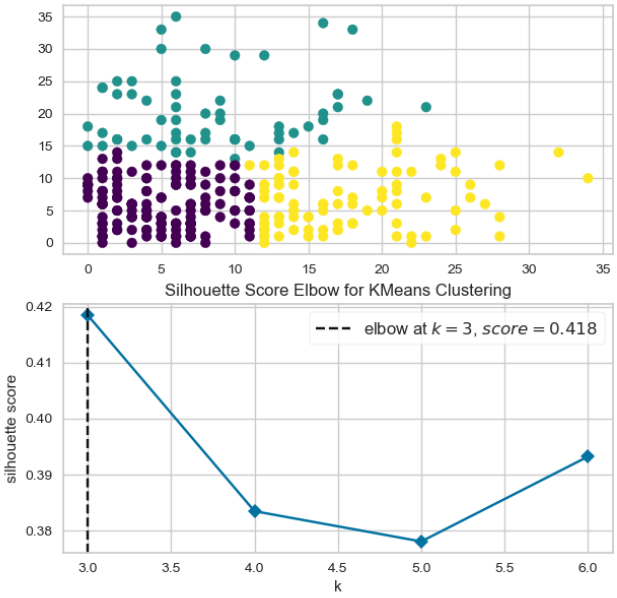
Plotted Figures for the 5 generated datasets of 300 values and 2 features.



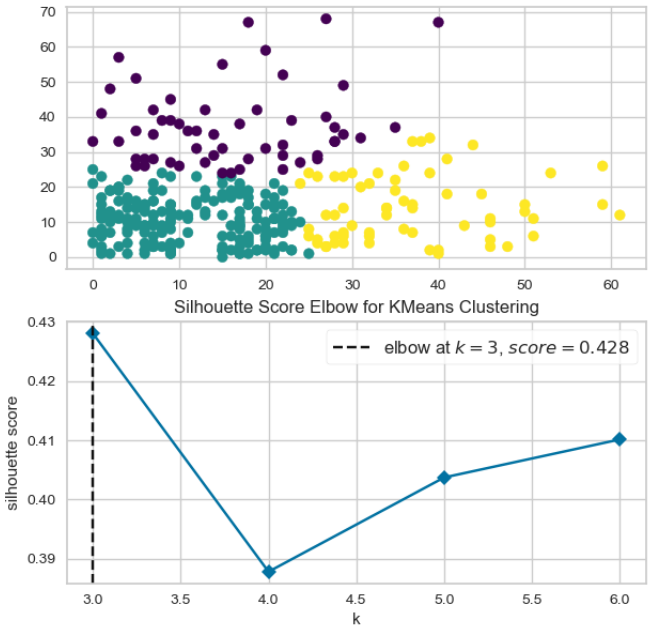
Standard deviation 1



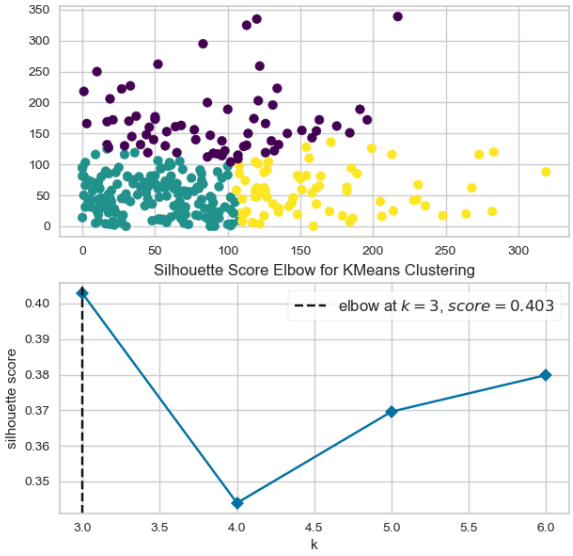
Standard deviation 5



Standard deviation 10

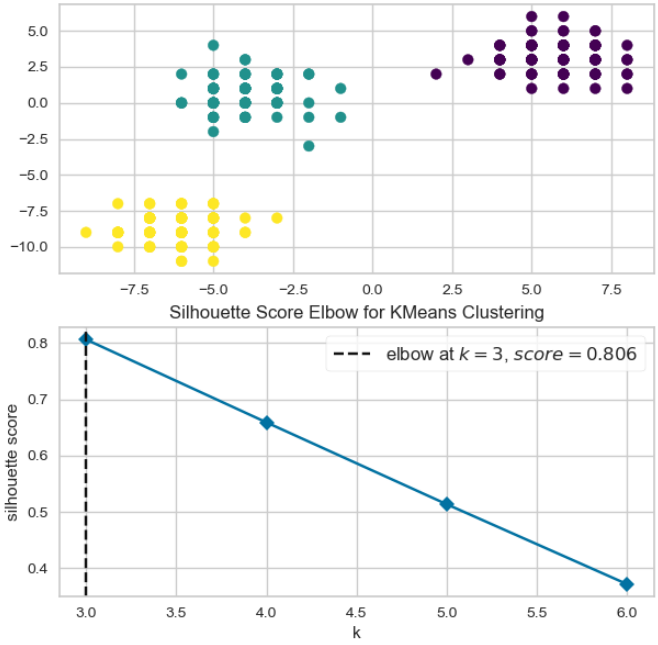


Standard deviation 20

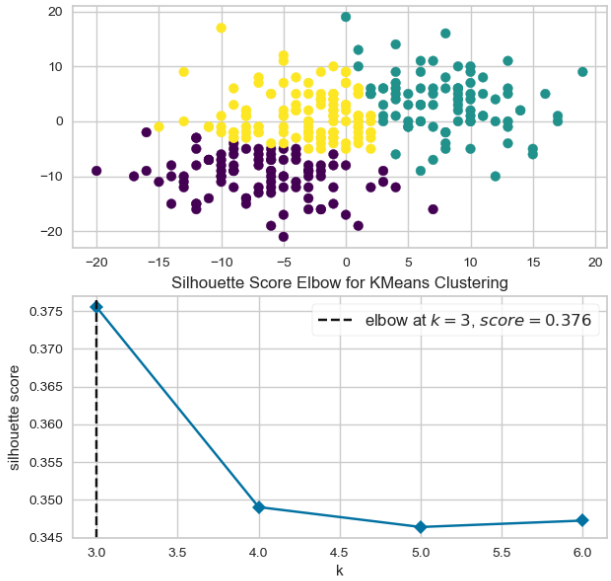


Standard deviation 100

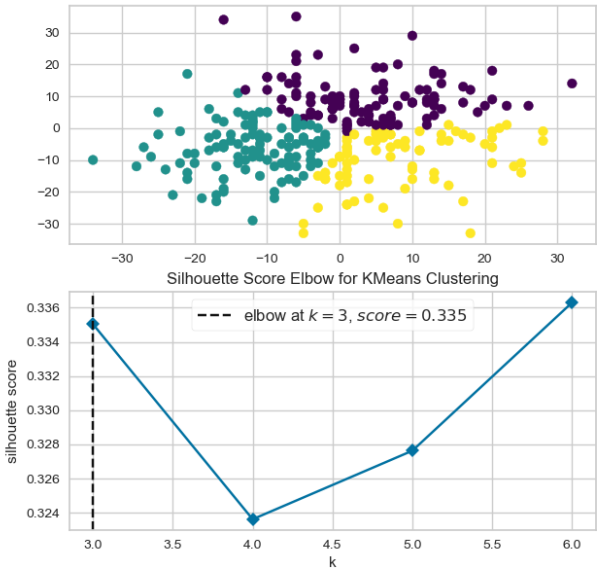
All the generated values are positive integers. This was done to generate cleaner plots for the lab. However, if this were to go against the requirements, I have attached diagrams for integers - positive and negative both, below.



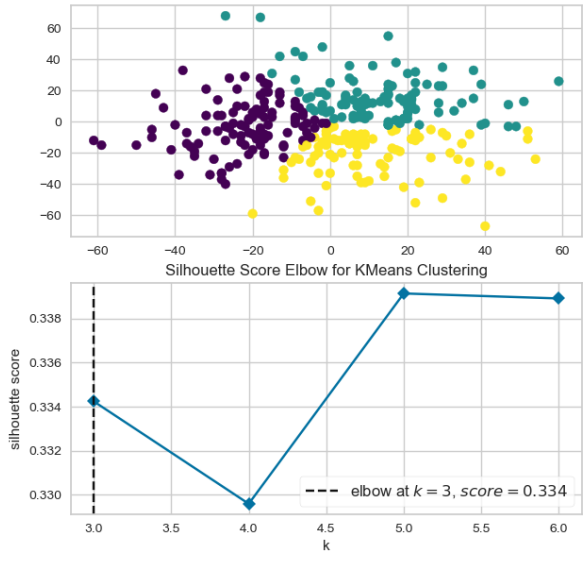
Standard deviation 1



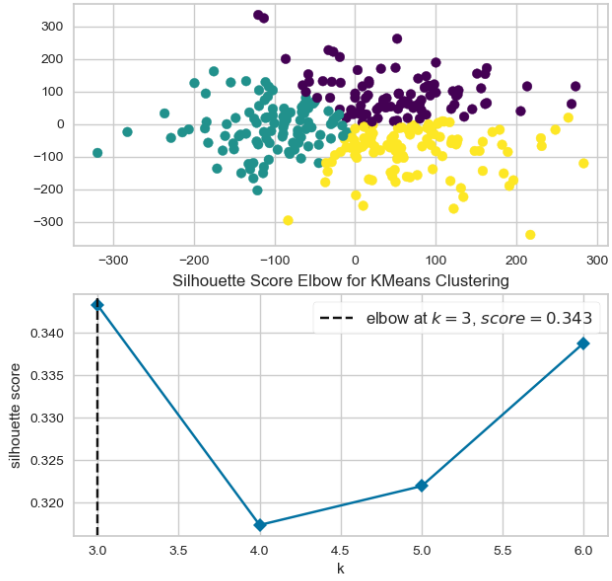
Standard deviation 5



Standard deviation 10



Standard deviation 20



Standard deviation 100