**Cluster Analysis**

In this part, Ward, K-means and DBscan clustering methods are used to detect special patterns and clusters for collected flight dataset. After analyzing of each attribute in dataset, there are several attributes may contribute to flight delay, such as: wind speed, gust speed, visibility and air speed. These attributes are selected to do clustering analysis in the following step.

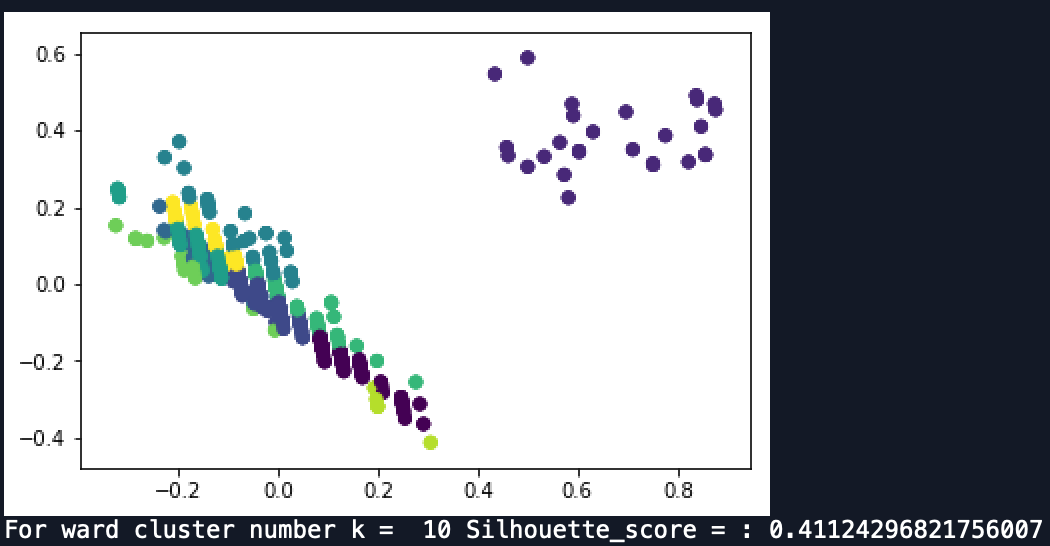
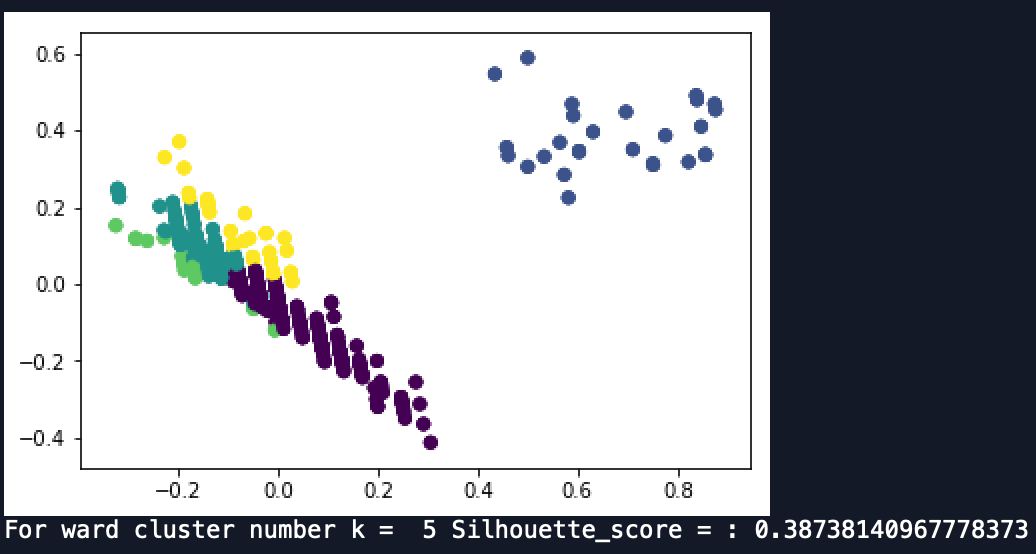
Based on outputs and PCA plots generated by Python, we could find there is obvious pattern for collected dataset even we didn’t do clustering analysis. Most of points are located on the left side of plot as a line shape and several points spread on the right side. To get precise cluster analysis, Ward, K-means, and DBscan methods need to be processed.

Ward method:

Ward method used variance of clusters as standard to separate data points. Minimizing variance of clusters is goal of ward method. During ward clustering process, cluster number k is set to 2, 5, and 10. Among these 3 different cluster numbers, Silhouette score of ward method is the highest one and equals to 0.6588 when cluster number equals to 2. The PCA of plot demonstrates that cluster number should be 2 which give the best clustering result.

图片包含 屏幕截图



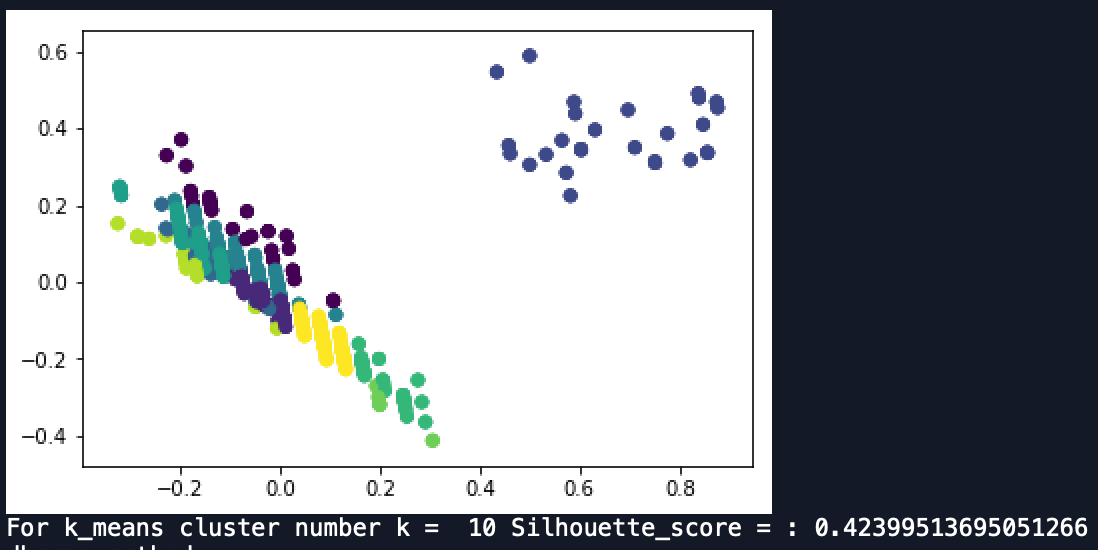
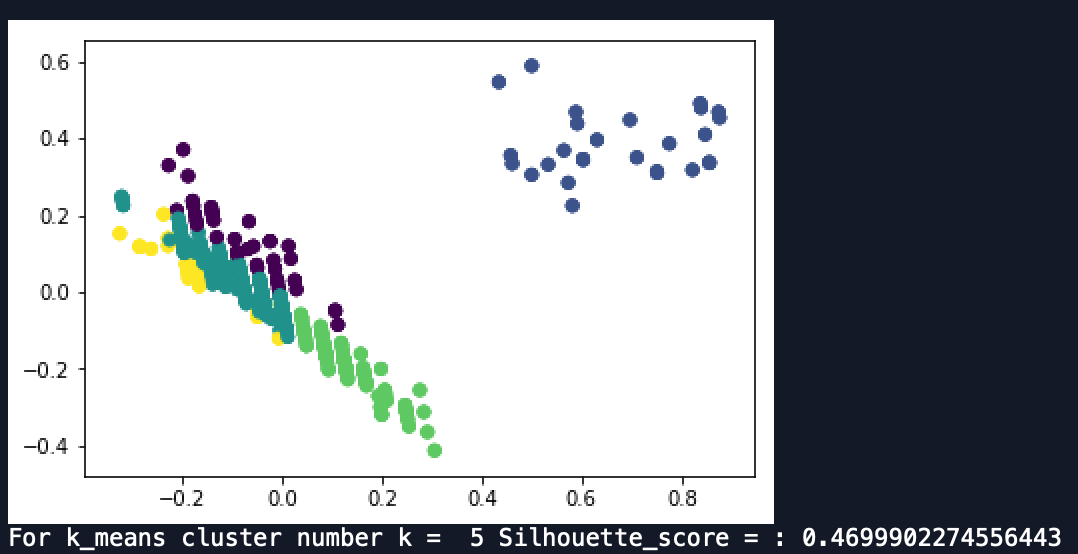
自动生成的说明

K-means method:

K-means method used distance between points and centroid to separate data points into clusters. During K-means clustering process, cluster number k is also set to 2, 5, and 10 which is same as Ward method’s cluster numbers. Among these 3 different cluster numbers, Silhouette score of K-means method is the highest one and equals to 0.6588 when cluster number equals to 2. The PCA plot of K-means method also demonstrates that cluster number should be 2 which give the best clustering result.

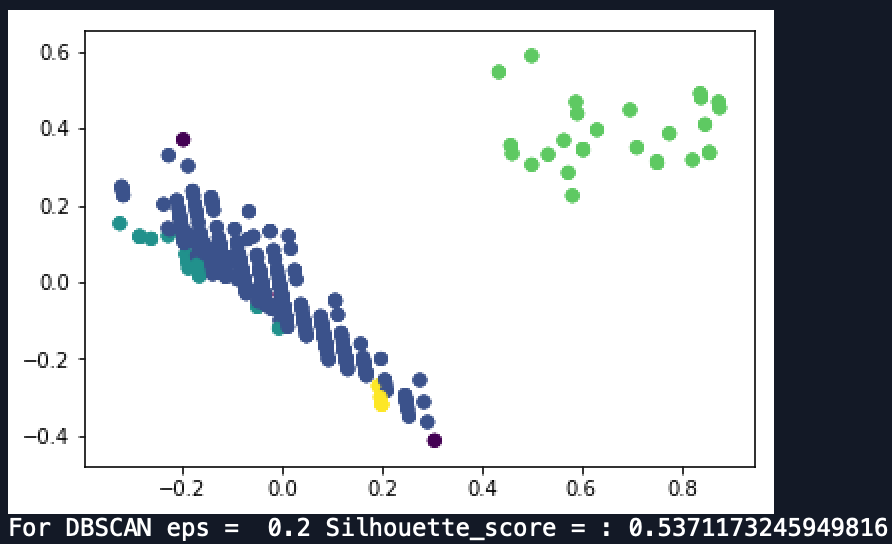
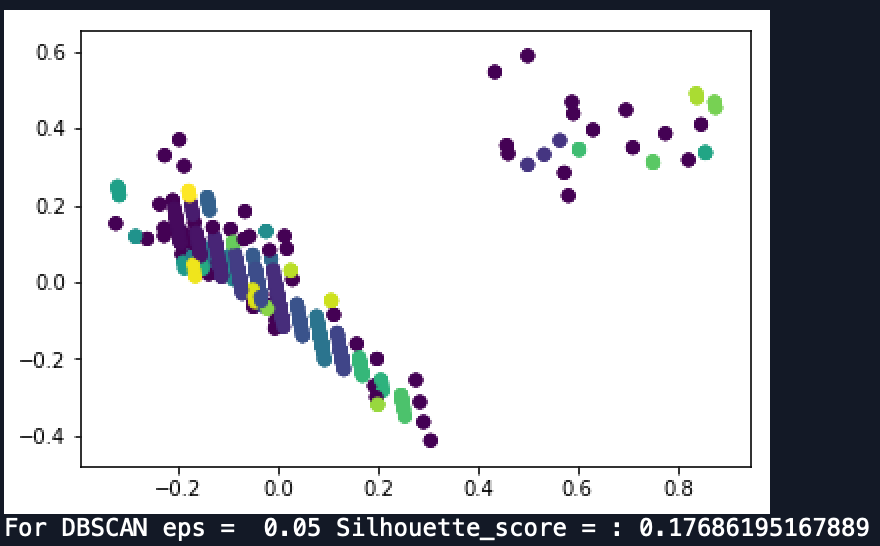
图片包含 屏幕截图



自动生成的说明

DBscan Method:

DBscan method used neighborhood distance to separate data points. During DBscan clustering process, maximum distance that to be consider as same neighborhood: eps is set to 0.05, 0.2, and 0.5. Among these 3 different eps values, Silhouette score of DBscan method is the highest one and equals to 0.6588 when eps equals to 0.5. The PCA of plot of DBscan method also shows that two cluster number gives the better clustering result.

图片包含 屏幕截图



自动生成的说明

All three clustering methods showed that Silhouette score is the highest when cluster number is two. PCA graphs and outputs of Python shows that there are distinct clusters and patterns in our data. Therefore, we could use those attributes: wind speed, gust speed, visibility and air speed to train the machine learning and do the further step.