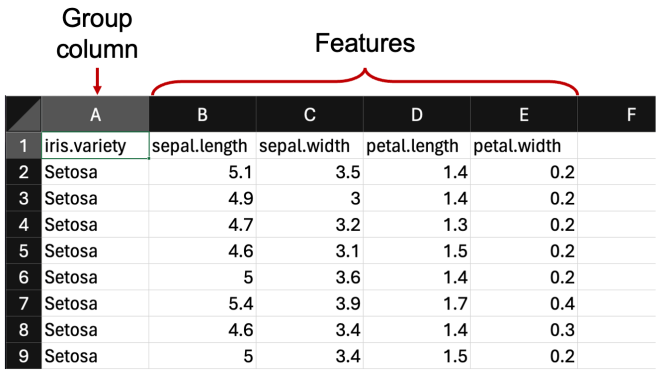
Data Visualization for Cluster Analysis - Running the App
June 11, 2024
Instructions by Ross Brancati

Part 1: Structuring your data

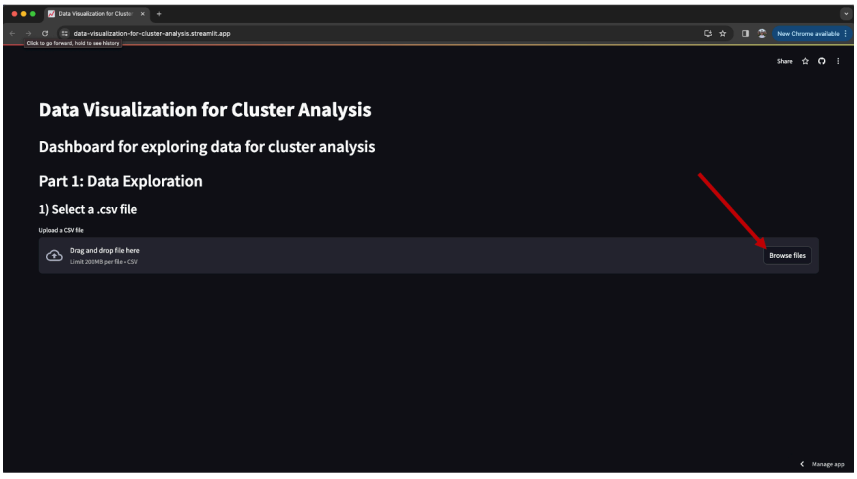
1. The data must be in .csv format with one column indicating the group of the observation, and the consecutive columns as features. In this example, the group is the variety of the iris (Setosa, Versicolor, Virginica)
    a. *Note: you can have more than one column as a group indicator. For example, if your data consists of multiple types of flowers, you can have one column indicating the type of flower, and another column indicating the variety within the type.*



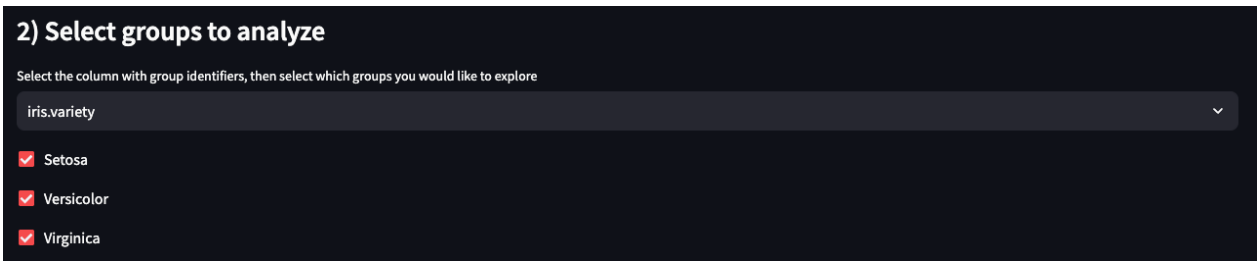
	A	B	C	D	E	F
1	iris.variety	sepal.length	sepal.width	petal.length	petal.width	
2	Setosa	5.1	3.5	1.4	0.2	
3	Setosa	4.9	3	1.4	0.2	
4	Setosa	4.7	3.2	1.3	0.2	
5	Setosa	4.6	3.1	1.5	0.2	
6	Setosa	5	3.6	1.4	0.2	
7	Setosa	5.4	3.9	1.7	0.4	
8	Setosa	4.6	3.4	1.4	0.3	
9	Setosa	5	3.4	1.5	0.2	

Part 2: Running the app

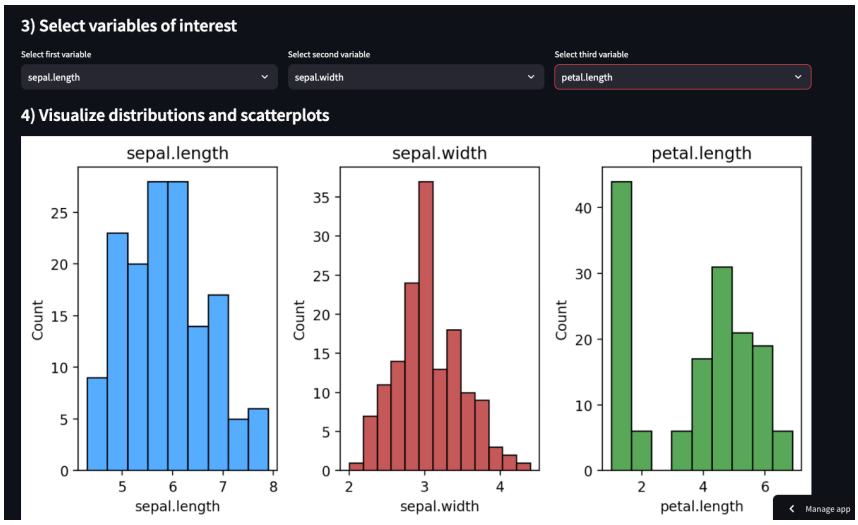
1. Click on the link to the application: https://data-visualization-for-cluster-analysis.streamlit.app/
2. Click browse files and navigate to your .csv file



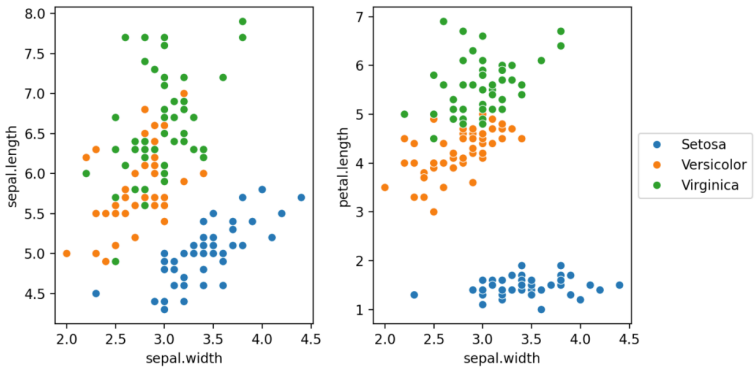
3. Navigate to your .csv file, select it, and click Open.
4. Select which column you would like to use as the group indicator from the dropdown in step 2, and which groups you would like to explore. You can visualize data for all groups, or for just a single group. In this case, I selected all iris varieties.



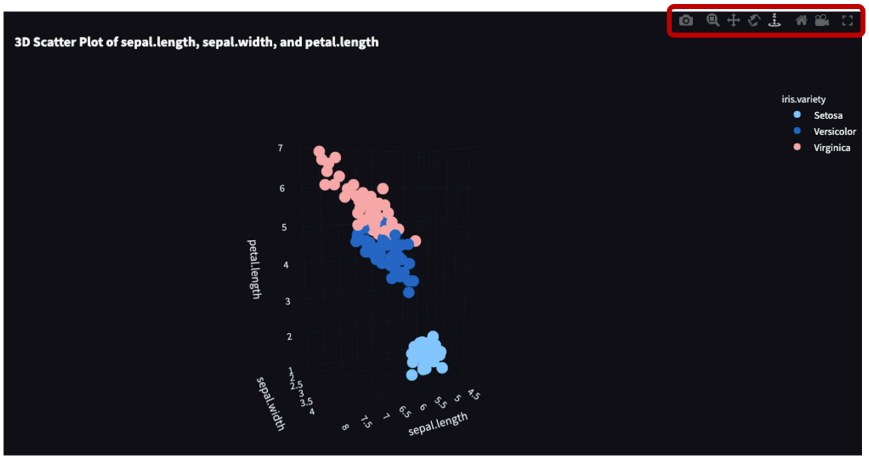
5. Select the variables that you would like to visually explore. Three types of plots will appear with the selected variables. The scatterplots will be colored by the groups selected above.
    a. *Histogram*
    b. *2D scatterplot of the first and second variable, and a separate 2D scatterplot of the second and third variable*
    c. *3D scatterplot of all three variables*. **Note:** the 3D scatterplot is fully interactive. You can enlarge the plot, take screenshots, change controls, etc. with the icons in the top right corner of the plot (controls are outlined in red in screenshot below). The controls are the same as found here: https://plotly.com/python/3d-camera-controls/



Histograms



2D scatterplots

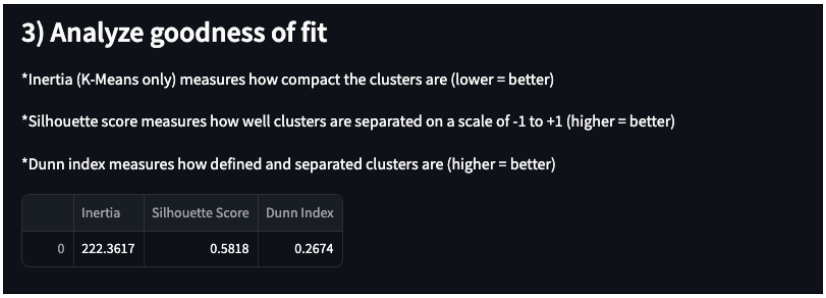


3D scatterplot

6. The next part of the dashboard allows you to explore a couple of clustering models. It currently supports K-means and DBSCAN.
7. First, select the features you'd like to use to cluster your data.

8. Select the model you'd like to use to cluster your data. Parameters depend on the chosen algorithm. For this case, I went K-means and 2 total clusters.
    a. *K-means: number of clusters*
    b. *DBSCAN: Epsilon and Min. Samples*

9. Analyze the goodness of fit of your clusters. Descriptions of the metrics used to evaluate the goodness of fit of the clusters are displayed on the dashboard for easy reference.
    a. *DBSCAN does not support the Inertia metric as a centroid of the cluster is not computed.*



10. Lastly, select the variables you'd like to visualize in a 3D scatterplot. This scatterplot will be colored by the clusters identified by the clustering algorithm. As you can see, we have two distinct clusters using the features of the iris varieties chosen above.