Rachael Kenney

Phone (513) 746-6219

Algorithms for Data Science

Homework 3

Due 7/8/2019

*for all code in hw3.py, data is output to console. Iris.arff is provided in this zip file, a file path to the iris.arff file will be needed to run the code

1) Develop code to analyze the Iris data sets using the test statistics listed in Table 1.
*code for these results can be found in additional file called hw3.py

statistics for full iris dataset

features	min	max	mean	trimmed mean	standard deviation	skewness	kurtosis
sepallength	4.3	7.9	5.387333	5.387838	0.942899	215.051582	443.752752
sepalwidth	2.0	4.4	2.650000	2.643243	0.591580	231.100602	447.882348
petallength	1.0	6.9	3.309333	3.304730	1.815028	69.783535	227.731828
petalwidth	0.1	2.5	0.861333	0.858784	0.832061	150.537532	282.723392

statistics for setosa dataset

features	min	max	mean	trimmed mean	standard deviation	skewness	kurtosis
sepallength	4.3	5.8	4.602	4.612500	0.533835	71.801055	118.964942
sepalwidth	2.3	4.4	3.050	3.060417	0.526972	72.597931	140.527662
petallength	1.0	1.9	1.008	1.002083	0.487278	58.571377	74.507063
petalwidth	0.1	0.6	0.004	0.002083	0.262420	64.528017	95.821326

statistics for versicolor dataset

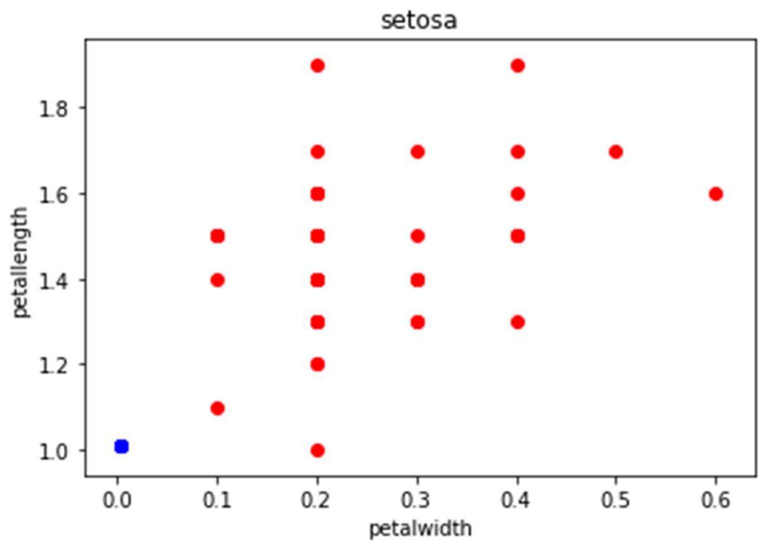
features	min	max	mean	trimmed mean	standard deviation	skewness	kurtosis

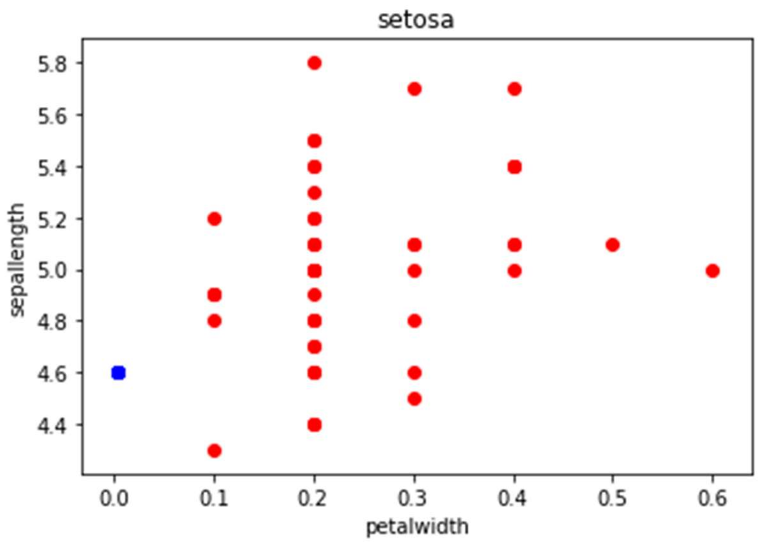
sepallength	4.9	7.0	5.480	5.479167	0.684865	72.478375	126.719986
sepalwidth	2.0	3.4	2.324	2.316667	0.543522	64.548772	91.488206
petallength	3.0	5.1	3.834	3.818750	0.630774	58.704249	96.229957
petalwidth	1.0	1.8	1.008	1.000000	0.373427	65.764177	94.828106

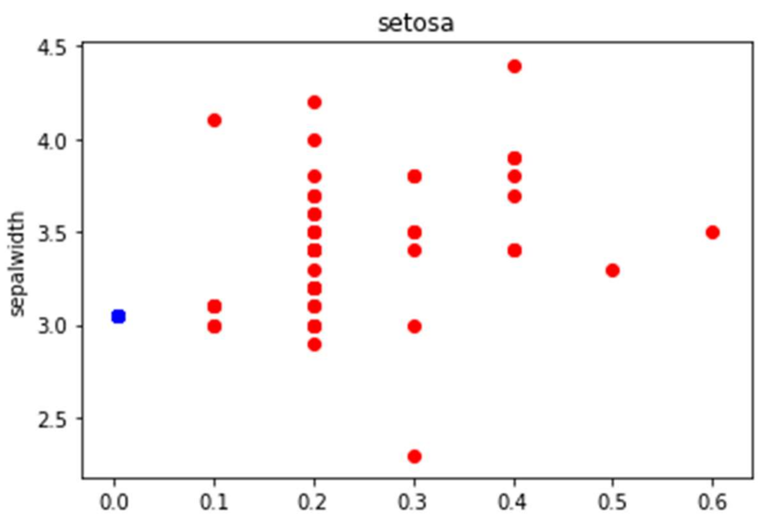
statistics for virginica dataset

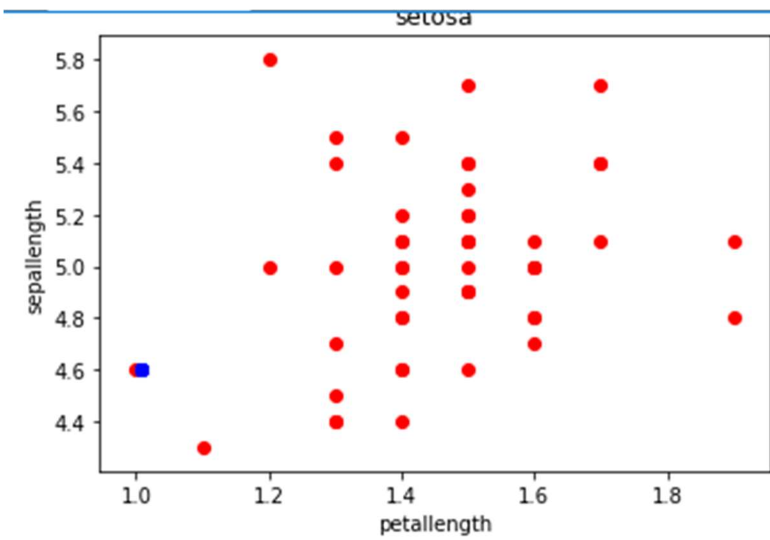
features	min	max	mean	trimmed mean	standard deviation	skewness	kurtosis
sepallength	4.9	7.9	6.086	6.116667	0.805146	72.020620	140.065421
sepalwidth	2.2	3.8	2.586	2.593750	0.502462	74.335825	132.738138
petallength	4.5	6.9	5.100	5.120833	0.709084	81.901140	160.041714
petalwidth	1.4	2.5	1.590	1.593750	0.513829	65.254829	92.409777

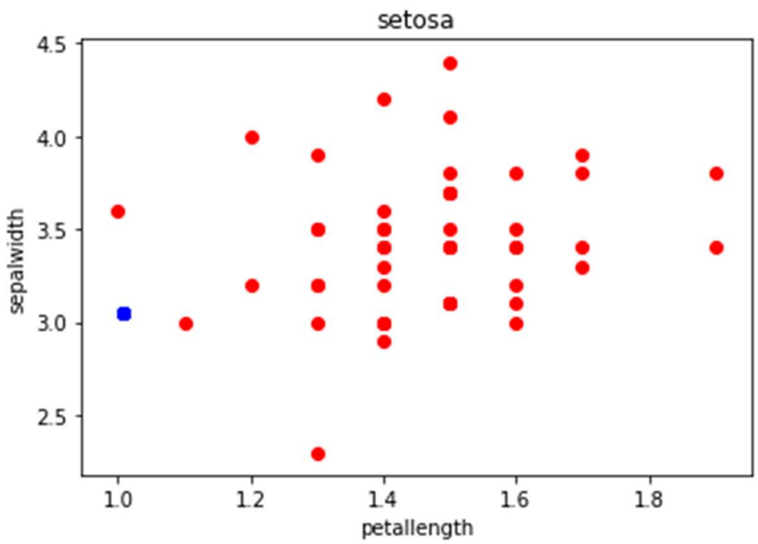
2) Analyze Iris data based on the class of flower type using linear discriminant analysis
   (a) Implement the two class linear discriminant based on the Fisher's Linear Discriminant (FLD) two-class separability (Fisher, 1936) described below. This is also shown in the two class linear discriminant function presented in (Bishop, 2006) Section 4.1.1 Two classes. For this exercise you will want to separate your Iris data into three sets and focus on any two class combination. For example, from the iris data take the first 50 observations for class 1, the next 50 as class 2 and the final 50 as class 3. Using the two class linear discriminant function compare class 1 verses class 2, class 1 verses class 3 and finally compare class 2 versus class 3.
   - code can be found in hw3.py
   (b) For this problem you will want to expand the two class case from part a to a three class case as presented in (Bishop, 2006) from Section 4.1.2 Multiple classes.
   - no code, couldn't attempt
3) Synthetic Data Addition to Iris Dataset (Collaborative)
   - code can be found in hw3.py

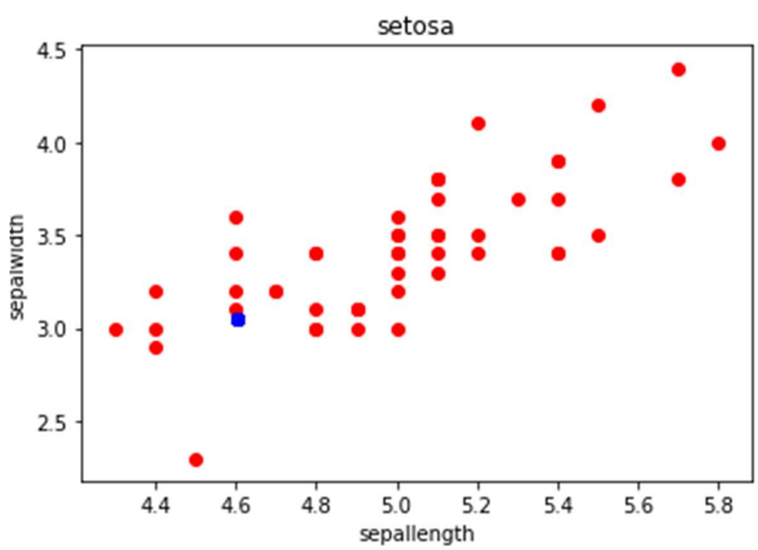

setosa
petallength
petalwidth
1.0
1.2
1.4
1.6
1.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6

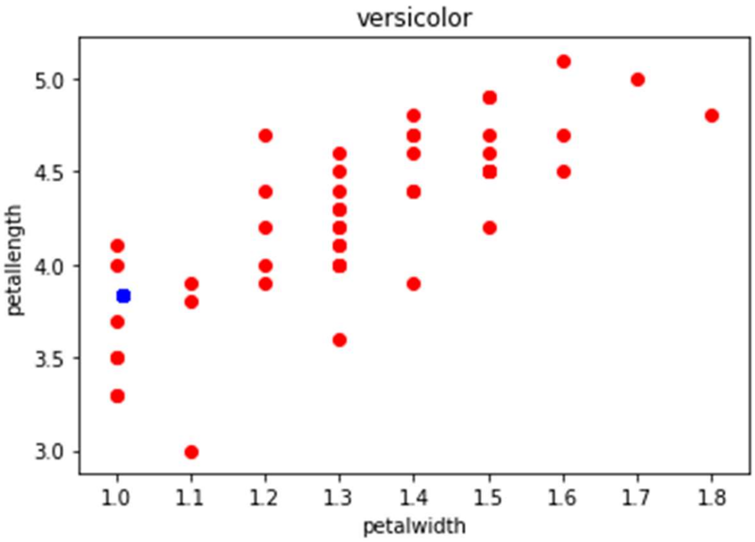

setosa
sepallength
petalwidth
5.8
5.6
5.4
5.2
5.0
4.8
4.6
4.4
0.0
0.1
0.2
0.3
0.4
0.5
0.6

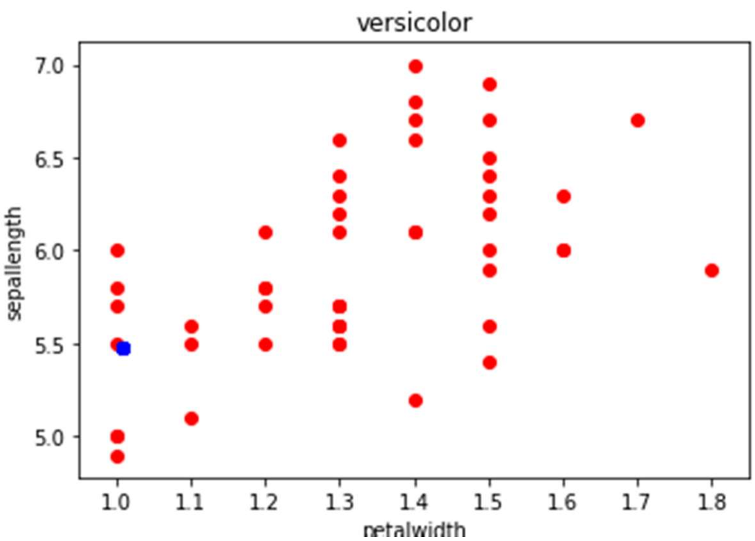

setosa
sepalwidth
4.5
4.0
3.5
3.0
2.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6

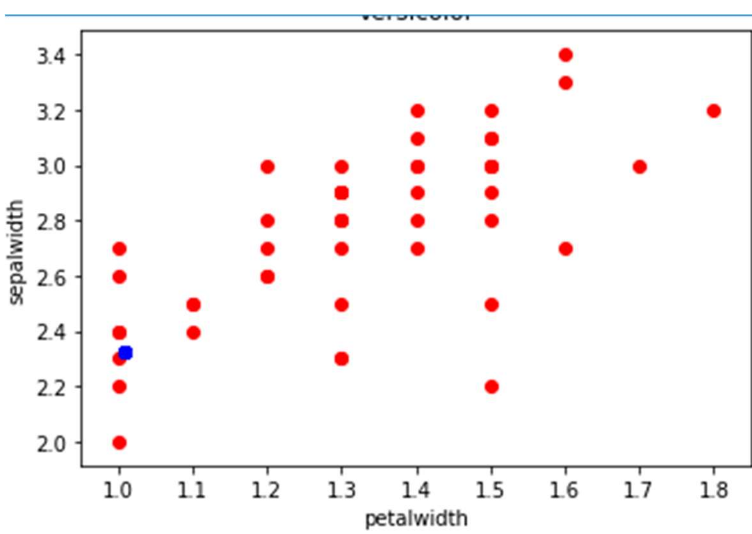

setosa
sepallength
5.8
5.6
5.4
5.2
5.0
4.8
4.6
4.4
1.0
1.2
1.4
1.6
1.8
petallength

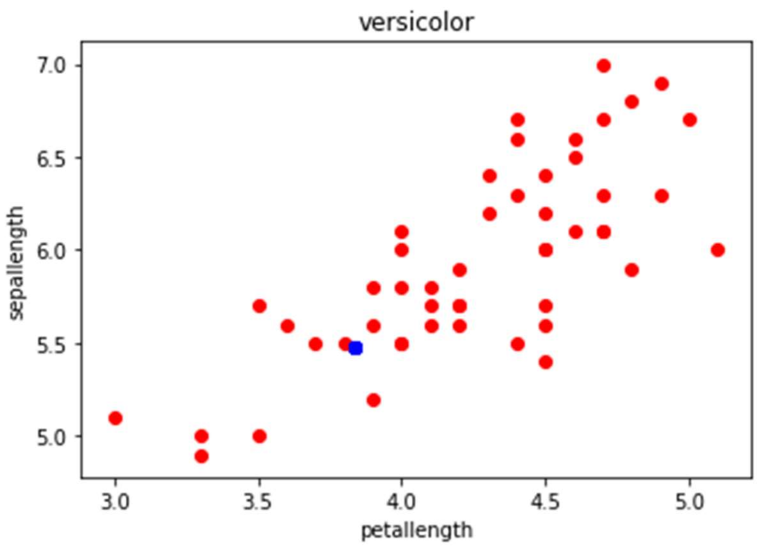

setosa
sepalwidth
4.5
4.0
3.5
3.0
2.5
1.0
1.2
1.4
1.6
1.8
petallength

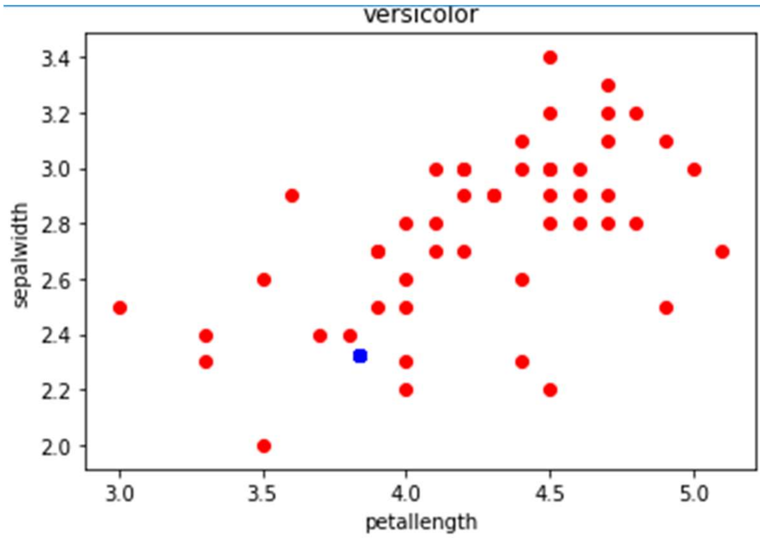

setosa
sepalwidth
4.5
4.0
3.5
3.0
2.5
4.4
4.6
4.8
5.0
5.2
5.4
5.6
5.8
sepallength

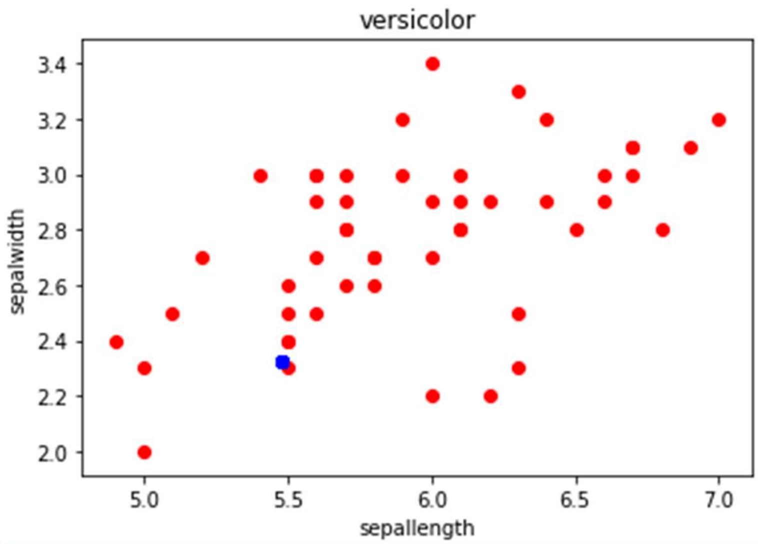

versicolor
5.0
4.5
4.0
3.5
3.0
petallength
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
petalwidth

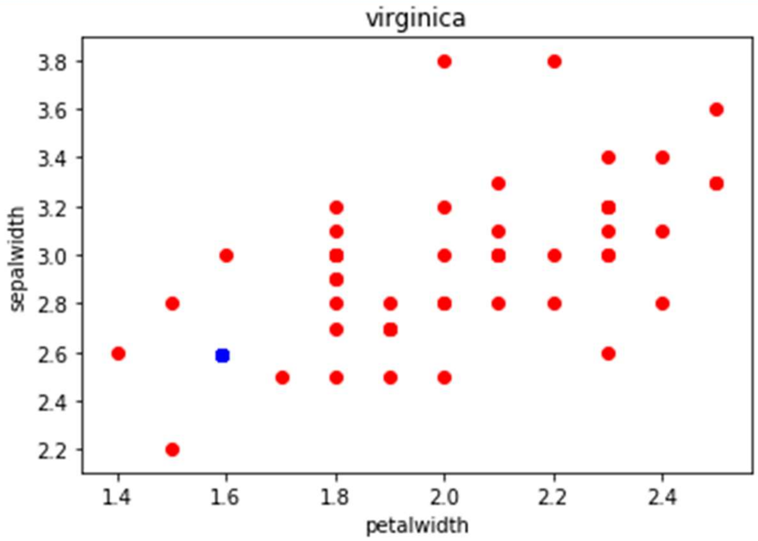

versicolor
7.0
6.5
6.0
5.5
5.0
sepallength
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
petalwidth

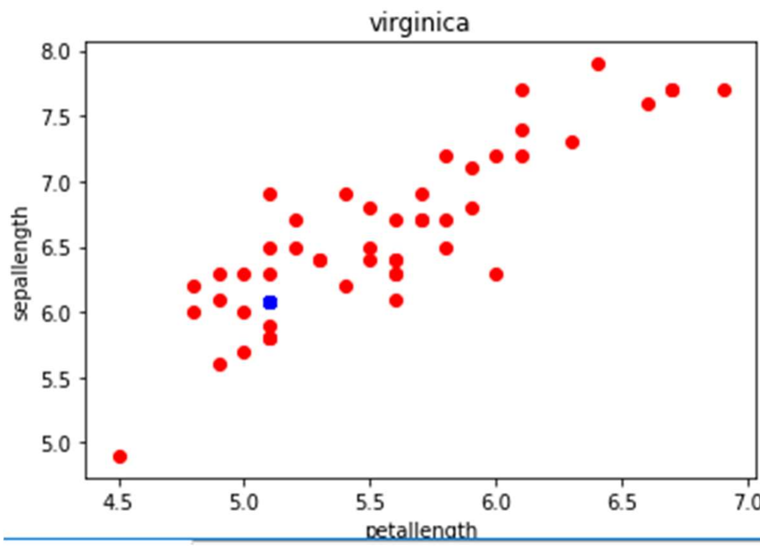

versicolor
sepalwidth
petalwidth
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8

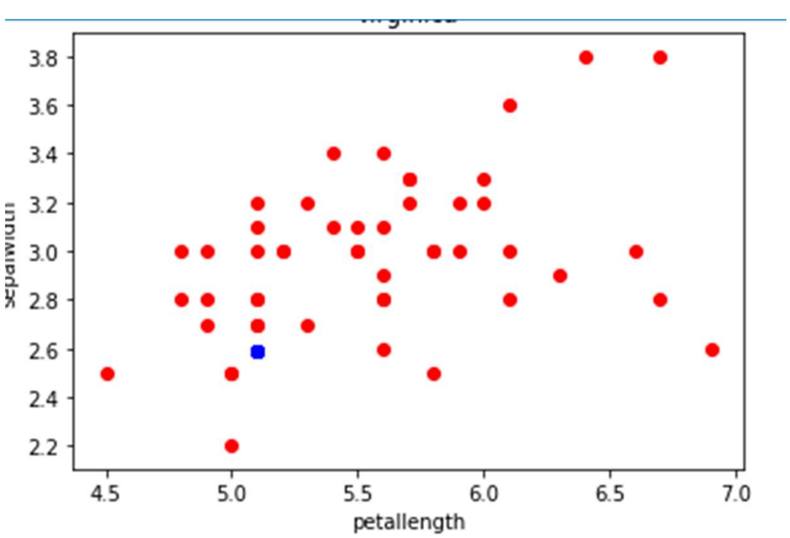

versicolor
sepallength
petallength
7.0
6.5
6.0
5.5
5.0
3.0
3.5
4.0
4.5
5.0

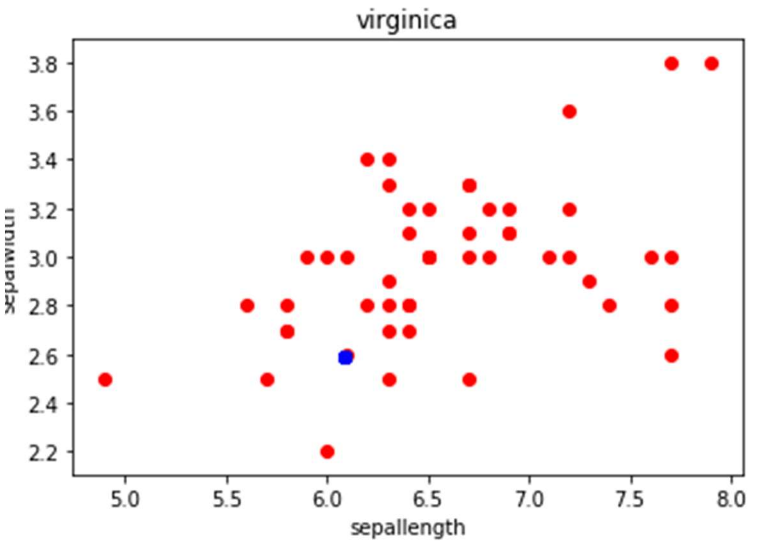

versicolor
sepalwidth
petallength
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
3.0
3.5
4.0
4.5
5.0

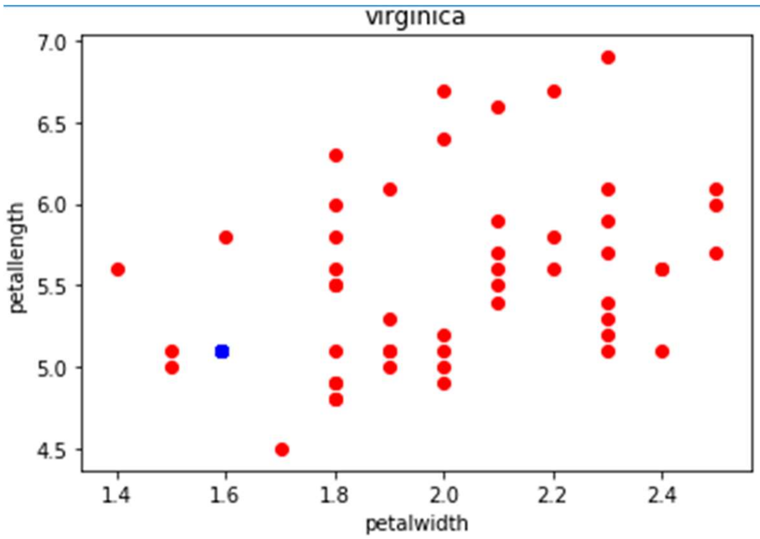

versicolor
sepalwidth
sepallength
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
5.0
5.5
6.0
6.5
7.0

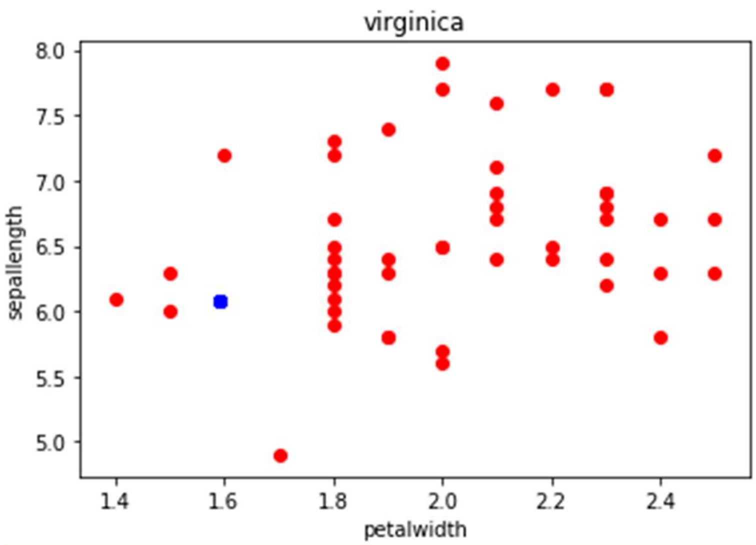

virginica
sepalwidth
petalwidth
3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
2.2
1.4
1.6
1.8
2.0
2.2
2.4


virginica
sepallength
petallength
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
5.0
5.5
6.0
6.5
7.0

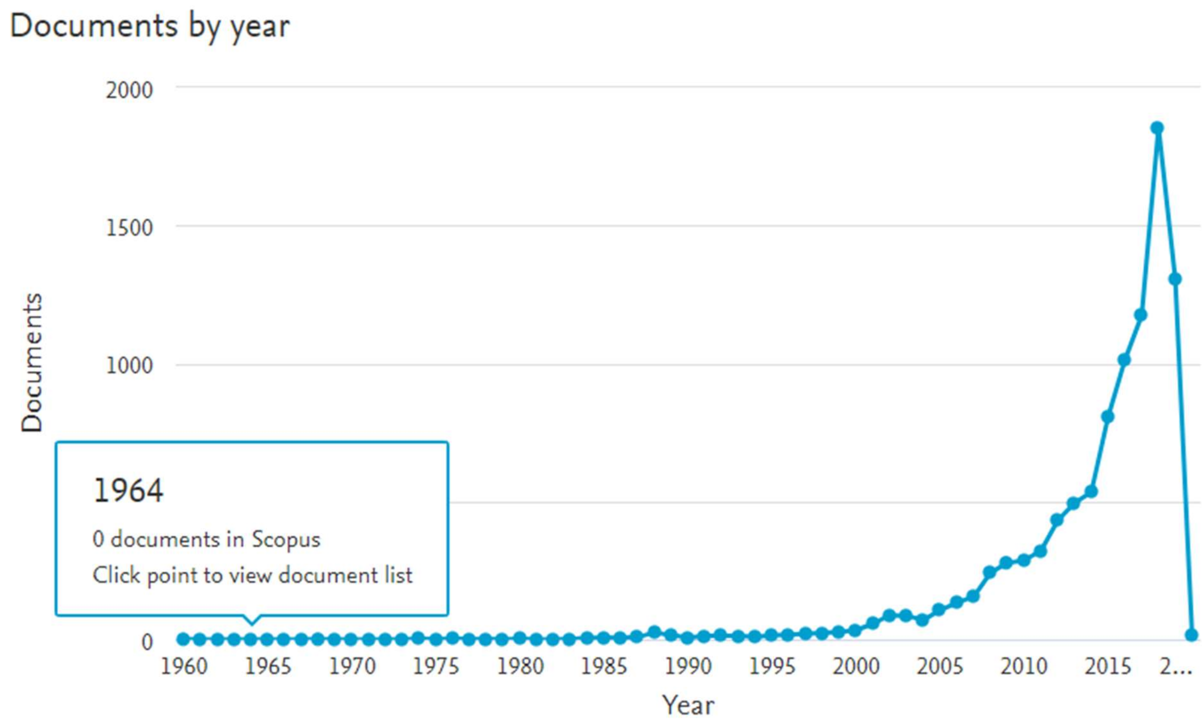

virginica
sepalwidth
petallength


virginica
sepalwidth
sepallength





4) Scopus Data Science and Machine Learning Document Analysis (Collaborative)
   (a) Go to the Scopus website and search for data science and machine learning related documents. Plot the distribution of the number of documents by year from at least the last 10 years. What is the story that the plot tells you?



- Looking at this graph, it is clear that machine learning and data science have only recently gained popularity in the past five years. Before 2015, these topics were steadily being written about a bit more, but around 2015 the number of papers written about these topics jumped up significantly with about 150% growth from 500 to almost 2000. If searching other related fields, you can see similar jumps in the early 2000s for "robots" or "artificial intelligence", or even "Bayesian statistics".

(b) Limit the search to 2016 and 2017. List the possible data fields/columns you may need to export in order to answer the question of author and/or institution collaborations in this scientific area during this timeframe.

- The fields that would be most useful in an analysis about author and institution collaborations would be author, affiliation, and funding sponsor. Along with funding sponsor, country might be a valuable field to include. There are many "National Instituted of…" in funding sponsor. So in order to remove potential duplicates, it would help to include the country that each goes with.

(c) Within the possible fields you suggest to export, which fields need data cleansing and why, in order to provide robust input for performing portfolio analysis?

- The author field has consistent formatting with Last Name, First Initial. So this field is probably fine to leave as is, but it might be noteworthy that some of the last names with two names are separated by a space and some are separated by a hyphen. Affiliation and funding sponsor both have some data quality issues. There were a few affiliations that used acronyms. In both affiliation and funding sponsor, there were some inconsistent uses of commas to denote a more detailed name of an item. For example, in affiliation there was: University of Michigan, Ann Arbor. Also, in funding sponsor, there was: City, University of London. In addition to this, some duplicate naming in funding sponsor across countries, as discussed above, may need to be cleaned up as well.