CHAPMAN University

Department of Computational and Data Sciences (CADS)
CS501 Introductory Computation for Scientists
Fall 2019
Class Project#3
Matplotlib

Date Given: Nov 6, 2019 Due Date: Dec 14, 2019

==

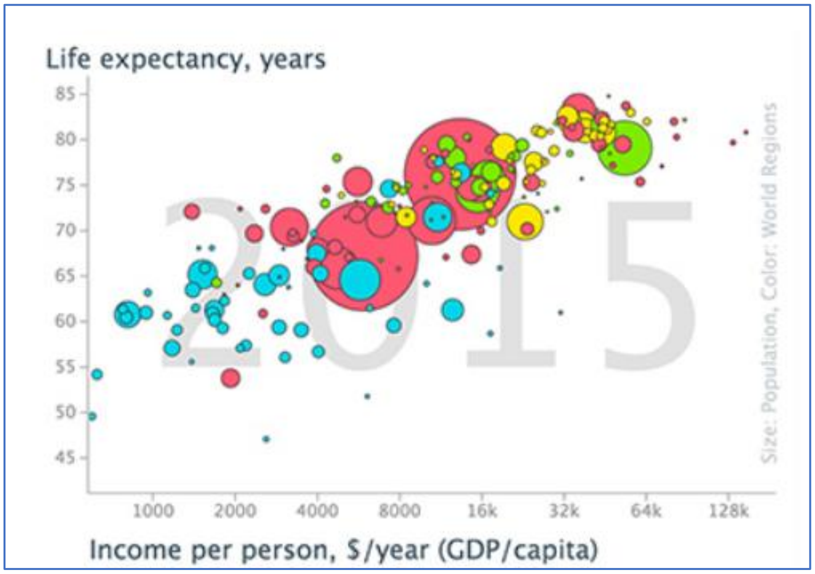
The are 4 programming problems in the Class Project#3. Use Matplotlib + Seaborn Python packages to create graphics.

==

Creating a 3-dimensional plot using Python (Matplotlib)

A “bubble chart” is a type of scatter plot which can depict three dimensions of data through the position (x and y coordinates) and size of the marker. The `plt.scatter` method can produce bubble charts by passing the marker size to its ‘`s′` attribute ($in\ (points)^2$ such that the area of the marker is proportional to the magnitude of the third dimension).

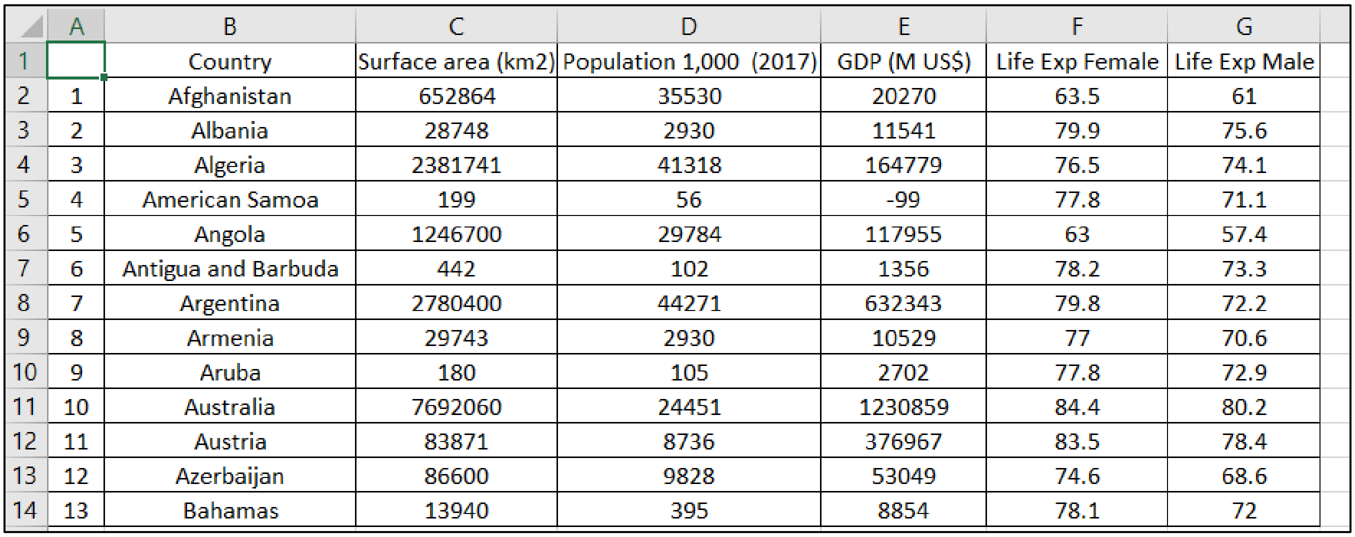
The https://www.gapminder.org/ website displays many 3-dimensional plots using the “bubble chart” technique. The following plot is an example of a “bubble chart” taken from the gapminder.org website.



Problem#1

Bubble Chart for 3-dimensional data

Read the dataset ‘country_profile_variables’ which contains data about the 212 countries. This dataset was downloaded from UN’s (United Nations) website.

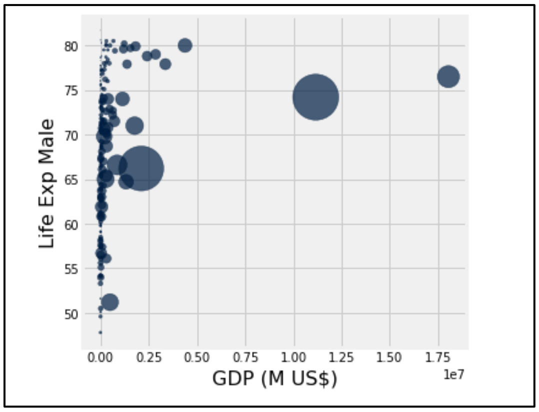


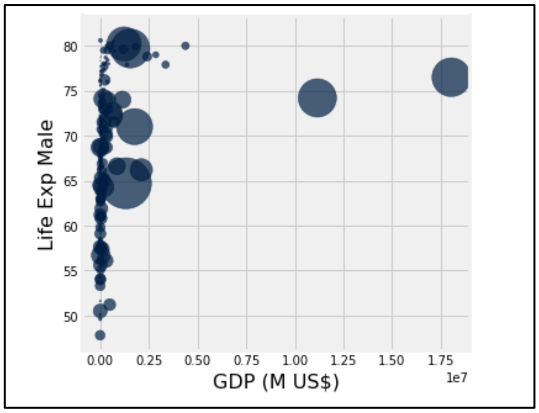
	A	B	C	D	E	F	G
1		Country	Surface area (km2)	Population 1,000 (2017)	GDP (M US$)	Life Exp Female	Life Exp Male
2	1	Afghanistan	652864	35530	20270	63.5	61
3	2	Albania	28748	2930	11541	79.9	75.6
4	3	Algeria	2381741	41318	164779	76.5	74.1
5	4	American Samoa	199	56	-99	77.8	71.1
6	5	Angola	1246700	29784	117955	63	57.4
7	6	Antigua and Barbuda	442	102	1356	78.2	73.3
8	7	Argentina	2780400	44271	632343	79.8	72.2
9	8	Armenia	29743	2930	10529	77	70.6
10	9	Aruba	180	105	2702	77.8	72.9
11	10	Australia	7692060	24451	1230859	84.4	80.2
12	11	Austria	83871	8736	376967	83.5	78.4
13	12	Azerbaijan	86600	9828	53049	74.6	68.6
14	13	Bahamas	13940	395	8854	78.1	72

Create 2 bubble charts between GDP and Life Expectancy (Male or Female) for all the countries in the dataset.

- In the first chart the size of the bubble should be proportional to the population of the country.
- In the second chart the size of the bubble should be proportional to the surface area of the country.

Your plots will look as follows.



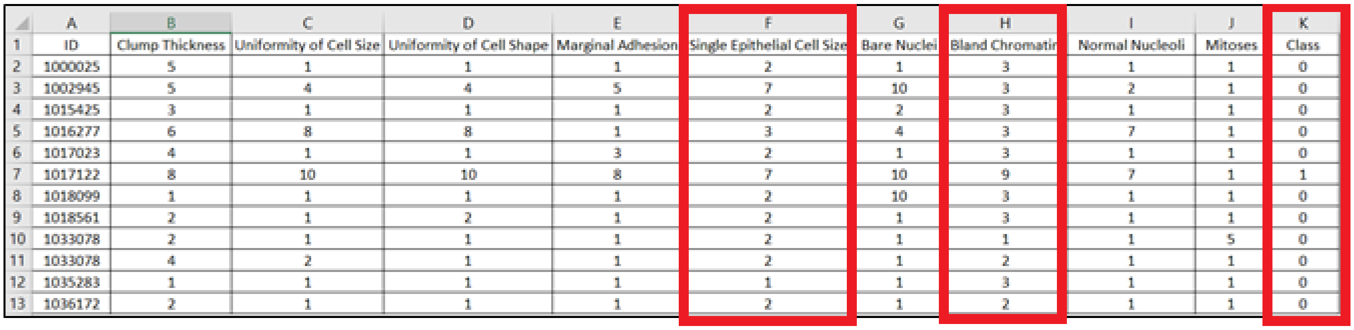


Make both the plots interactive which means when a user clicks on a specific bubble of the plot, it should identify the country’s name.

Problem#2

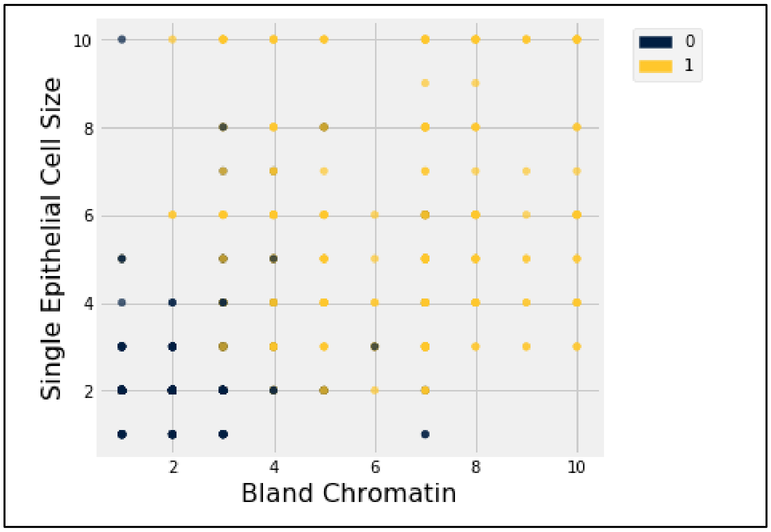
Overlapping Dataset

Read the breast cancer dataset (breast-cancer.csv). This dataset contains medical metrics about 683 patients.



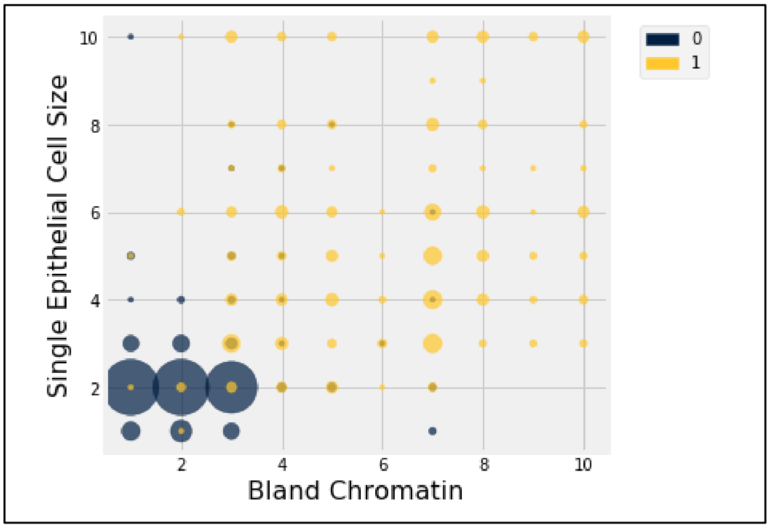
	A	B	C	D	E	F	G	H	I	J	K
1	ID	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatir	Normal Nucleoli	Mitoses	Class
2	1000025	5	1	1	1	2	1	3	1	1	0
3	1002945	5	4	4	5	7	10	3	2	1	0
4	1015425	3	1	1	1	2	2	3	1	1	0
5	1016277	6	8	8	1	3	4	3	7	1	0
6	1017023	4	1	1	3	2	1	3	1	1	0
7	1017122	8	10	10	8	7	10	9	7	1	1
8	1018099	1	1	1	1	2	10	3	1	1	0
9	1018561	2	1	2	1	2	1	3	1	1	0
10	1033078	2	1	1	1	2	1	1	1	5	0
11	1033078	4	2	1	1	2	1	2	1	1	0
12	1035283	1	1	1	1	1	1	3	1	1	0
13	1036172	2	1	1	1	2	1	2	1	1	0

First create a scatter plot between “Bland Chromatin” (column H) and “Single Epithelial Cell Size” (column F) using ‘class’ (column K) to classify by color whether a patient has cancer or not. The class value of ‘1’ indicates cancer. Your plot will look as follows.

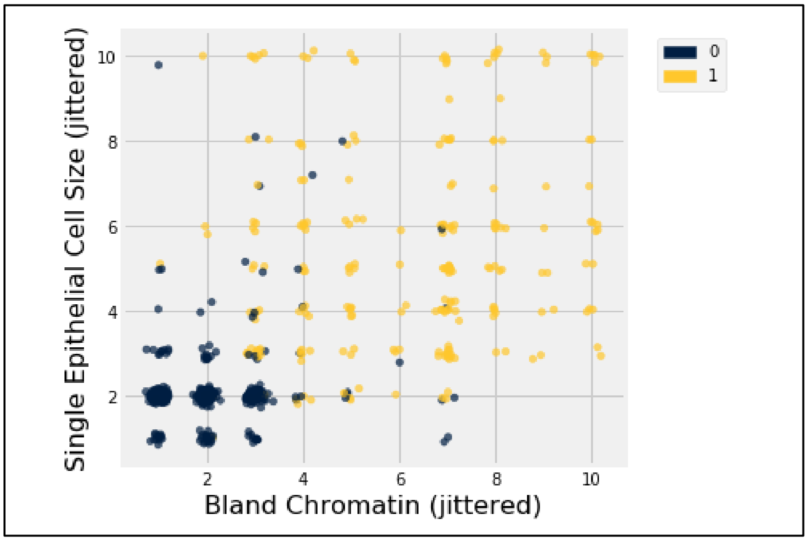


There are 683 patients and the plot above displays less than 683 points. The problem will this scatter plot is that points are **overlapping**. Therefore, we cannot tell how many patients are represented by a single point.

Create a bubble scatter plot (like Problem#1) where the size of the marker is proportional to the number of points at that location. Your plot should look as follows.



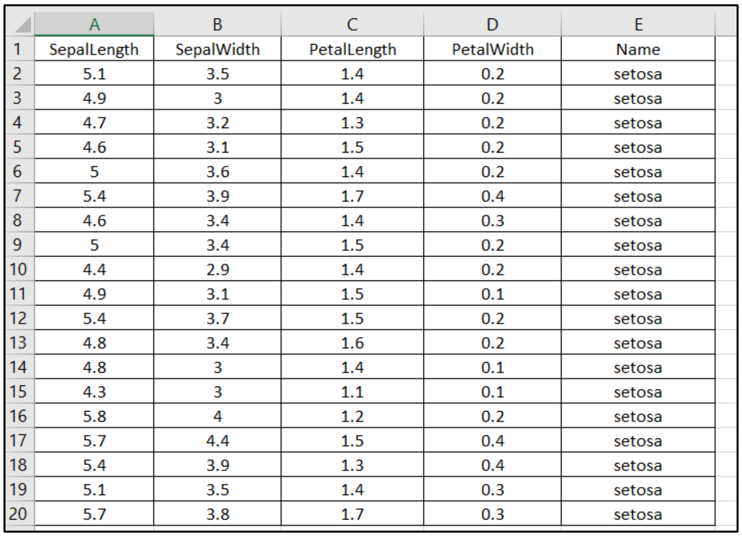
Another way to handle this situation is to add random noise (jitter) to each data points. Therefore, when the plot is created the marker is shifted slightly from the exact position. The ‘jitter’ plot looks as follows.



Create “bubble” and “jitter” scatter plots between “Bland Chromatin” (column H) and “Single Epithelial Cell Size” (column F) using ‘class’ (column K) to classify by color whether a patient has cancer or not.

Problem#3

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species.

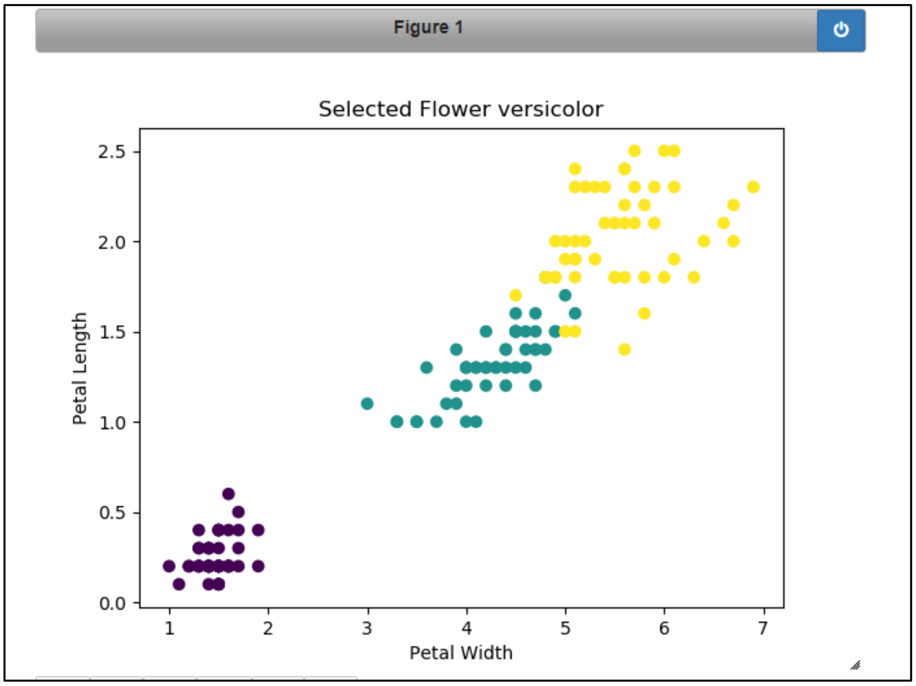


	A	B	C	D	E
1	SepalLength	SepalWidth	PetalLength	PetalWidth	Name
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa

Create an interactive plot (using Matplotlib) using Iris data set.

- X value: Petal Length
- Y value: Petal Width

Create a scatter plot. When a user clicks on any point on the graph, system should display the Iris flower type. Your plot should look as follows.



Problem#4

Generate 2 sets of univariate data. Create a 2-dimensional KDE Joint plot using Seaborn.

Univariate data set#1: 1000 normally distributed data with $\mu = 50, \sigma = 10$

Univariate data set#2: 1000 normally distributed data with $\mu = 75, \sigma = 50$

Also create ‘Hex’ and ‘KDE’ Joint plots.

Your plots should look as follows.

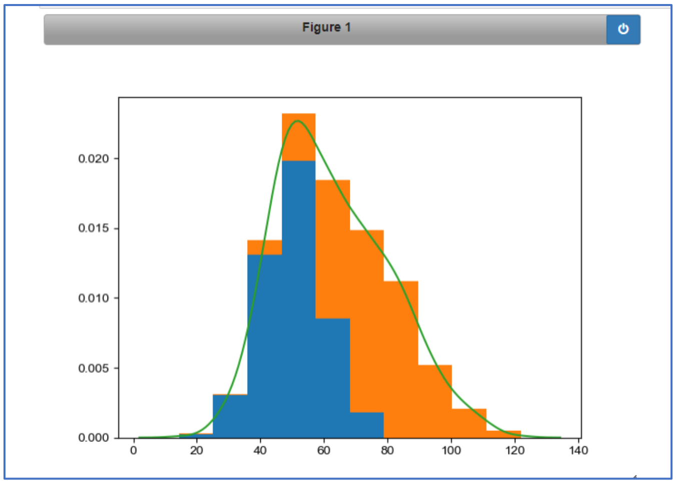


Figure 1
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
140

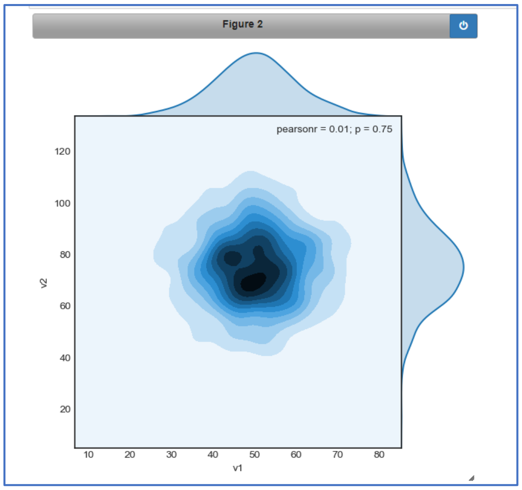


Figure 2
pearsonr = 0.01; p = 0.75
120
100
80
60
40
20
v2
10
20
30
40
50
60
70
80
v1

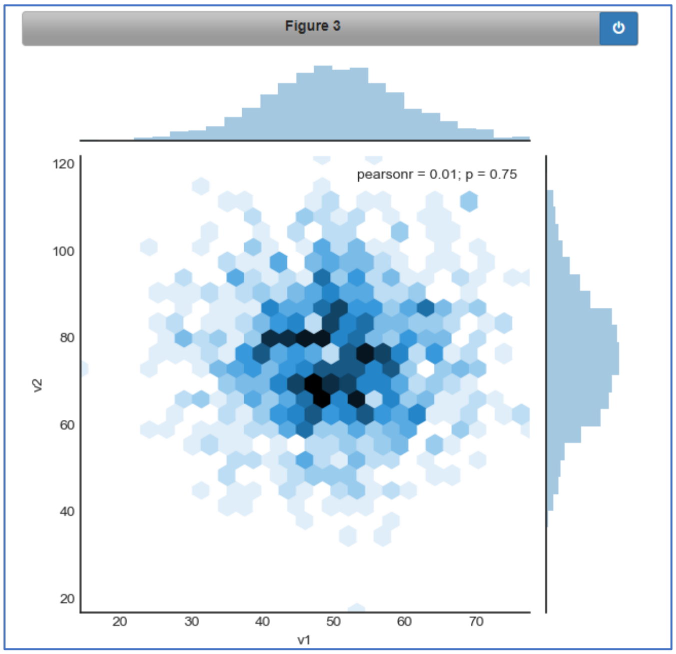


Figure 3
pearsonr = 0.01; p = 0.75
120
100
80
60
40
20
v2
20
30
40
50
60
70
v1