NAME:
EMAIL:
SIGNATURE:

Lehman College, CUNY
MAT 456-01: Topics Course: Data Science Spring 2016

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
Total	

1. What will the following code draw:

```
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(0, 2*np.pi, 50)
y = np.sin(x)
y2 = y + 0.1 * np.random.normal(size=x.shape)

fig, ax = plt.subplots()
ax.plot(x, y, 'k--')
ax.plot(x, y2, 'ro')

plt.show()
```

Output:

2. (a) Write a regular regular expression that will match a string that starts with a digit, followed by any two characters:

   (b) Write a regular regular expression that will match any name that begins with a upper case letter and ends in "`a`" or "`y`".

   (c) Write a regular regular expression that will match any string that contains "`x**2`", "`x^2`", or "`x*x`"

3. The New York City Open Data project contains all motor vehicle collisions reported to the New York Police Department. The data can be downloaded as CSV files with the following format:

```
DATE,TIME,BOROUGH,ZIP CODE,LATITUDE,LONGITUDE,LOCATION,ON STREET NAME,CROSS STREET NAME,OFF STREET
02/01/2016,0:09,BRONX,10465,40.8341548,-73.8174815,"(40.8341548, -73.8174815)",BARKLEY AVENUE,DEAN
```

All lines are formatted similarly: they start with the date, then time, the borough, zip code, latitude and longitude, and also include cross streets, types of vehicles involved, number of injuries/fatalities, and possible cause. The first line of the file gives the entries in the order they occur in the rows.

Write a program that takes a file, `bronxCollisions.csv`, and prints out all the dates that crashes occur in the 10468 zip code:

4. The Center for Disease Control (CDC) provides data on the number of occurrences of Lyme Disease. Assuming you have the data stored:

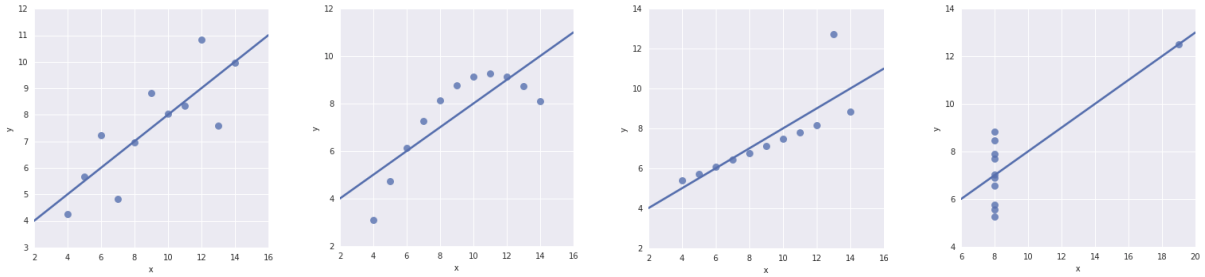
```
years = [2003,2004,2005,2006,2007,2008,2009,2010,2011]
ny = [5399,5100,5565,4460,4165,5741,4134,2385,3118]
nj = [2887,2698,3363,2432,3134,3214,4598,3320,3398]
ct = [1403,1348,1810,1788,3058,2738,2751,1964,2004]
```

Write a program that will plot the state numbers as well as the maximum number of Lyme Disease occurrence (i.e. the maximum of the values for the three states for each year).

5. Most (90%) dark colored cars in the city are black, while the remaining are blue. In a traffic collision involving a dark colored hit-and-run car, a witness claims that the dark colored car was blue. Careful testing has shown that the witness can successfully identify the color of a car 75% of the time.

   Use Bayes Theorem to calculate the probability that the car the witness saw was blue.

6. In 1973, Anscombe gave 4 datasets with similar statistics but that are quite different visually:



That is, for all four, the mean of x is 9 with variance 11, the mean of y is 7.50 with variance of 4.12, correlation is 0.815, and a linear regression line is $y = 3 + x/2$.

(a) Using these training data sets yields the same linear regression. What does the linear regression estimate the value of y for the following values of x:

x	**y**
2	
16	
18	

(b) Create a data set of 5 points with the same mean of x ($= 9$) but different mean of y ($= 7.5$):

(c) Create a data set of 5 points with the same linear regression line ($y = 3 + x/2$) but different correlation (0.816):

7. The Department of Buildings recently released a CSV file of all registered elevator devices in New York City. The file has the following columns:

```
DV_DEVICE_NUMBER,Device Status,DV_DEVICE_STATUS_DESCRIPTION,BIN,TAX_BLOCK,TAX_LOT,
HOUSE_NUMBER,STREET_NAME,ZIP_CODE,Borough,Device Type,DV_LASTPER_INSP_DATE,
DV_LASTPER_INSP_DISP,DV_APPROVAL_DATE,DV_MANUFACTURER,DV_TRAVEL_DISTANCE,
DV_SPEED_FPM,DV_CAPACITY_LBS,DV_CAR_BUFFER_TYPE,DV_GOVERNOR_TYPE,DV_MACHINE_TYPE,
DV_SAFETY_TYPE,DV_MODE_OPERATION,DV_STATUS_DATE,DV_FLOOR_FROM,DV_FLOOR_TO,,LATITUDE,LONGITUDE
```

A sample line looks like:

```
1D1,W,WK IN PROG,1084781,1480,1,521,EAST   68 STREET,0,Manhattan,Dumbwaiter (D),,,,,,,,,,,,
20090115,,,,40.76445324430363,-73.9541429039059,1D10000,A,ACTIVE,1008273,525,56,494,
WEST BROADWAY,10012,Manhattan,Dumbwaiter (D),20140325,NV,19960524,**,,,,,,,,20140113,,,,
40.727254549594306,-73.99984125121291
```

Write a program that displays on a map the location of all dumbwaiter and freight elevators in Manhattan. Your program should use a different symbol for the different types of device ('Dumbwaiter (D)', 'Freight (F)').

8. Consider the map of middle schools in East Harlem, discussed in class:

(a) Assume you can have two bus drop-offs for the 10 middle schools listed. Where would you put them to minimize walking from the drop-offs to the schools? Explain your answer.

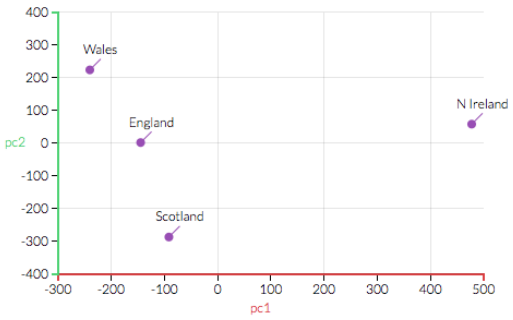
(b) Does your answer change if you use a Euclidean distance (i.e. allowed to 'fly over' buildings) instead of the walking or Manhattan distance (i.e. must stay on streets, cannot cut through buildings). Why or why not? Explain your answer.

(c) What are the best locations of bus stops if you can have 3 (assume Manhattan distance)? Explain your answer.

(d) Explain how you would find the locations in general, given k bus stops to place in a way that minimizes the overall total walking distance (i.e. k-means clustering):

9. In class, we discussed the "Eating in the UK" dataset (from `setosa.io`):

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



(a) The plot on the right is a Principal Components Analysis (PCA) of the 17-dimensional data set. What is PCA? Either the goal or give brief description of the underlying algorithm:

(b) What variation or pattern is highlighted by the first two principal components axis that was not easily seen in the 17-dimensional space? Explain your answer.

(c) Design a program that will read in a data set, perform PCA, and display a plot of the first two PCA axis. List all packages you use. The overall design should be in pseudocode (details of the actual Python, other than which packages you're using where, is not needed).

10. The NYC OpenData project keeps track of the public trash can locations across the city. The file has the following columns

```
BasketType,x,y
```

(a) Given a starting location, `(xStart,yStart)`, and the location of 3 trash cans (that is, the coordinates of each: `(x1,y1),(x2,y2),(x3,y3)`), describe an algorithm to decide which trash can is closest to the starting location:

(b) Extend your algorithm to work for a list of possible trash cans:

(c) Use your algorithm to color in an image of points with the color of the closest trash can. Your program should take a 2D array of points and list of trash cans, it should then randomly pick a color for each trash can, and assign the color of the closest trash can to each (i, j) in the 2D array (i.e. make a Voronoi diagram for the inputted list of trash cans)