Name:
Email:
Signature:

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
Total	

Lehman College, CUNY
CMP 464-C401: Topics Course: Data Science Spring 2016

You may have a 2-sided 8.5”×11” page of notes.

1. What will the following code draw:

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

t = np.arange(0., 1., 0.1)

plt.plot(t, t, ’r--’)
plt.plot( t, t**2, ’bs’)
plt.plot( t, t**3, ’g^’)
plt.show()
```

Output:

2. For each of the regular expressions, give a string that will matches it:

   (a) Write a regular regular expression that will match any string that starts and ends with a digit with any number of any type of characters in between:

   (b) Write a regular expressions that will match any string that contains two 5-letter words in a row (a consecutive string of 10 alphabetic characters):

   (c) Write a regular expressions that will match any string that contains the string “`def`”, followed by a function name (a consecutive string of alphanumeric characters), and parenthesis surrounding 0 or more characters.

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 times 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:

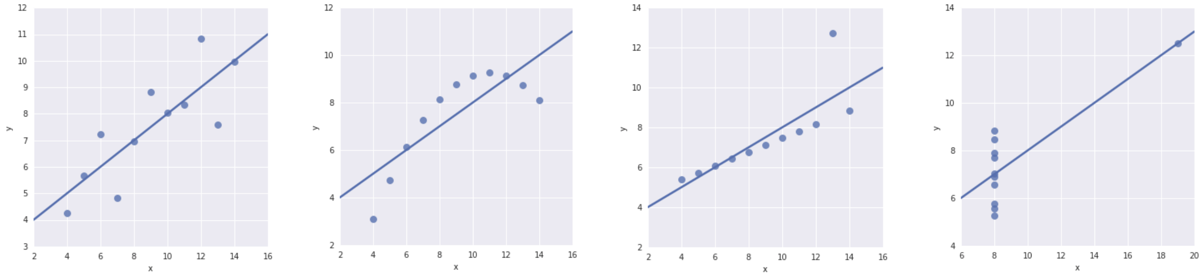
   ```
   years = [2003,2004,2005,2006,2007,2008,2009,2010,2011]
   ny = [5399,5100,5565,4460,4165,5741,4134,2385,3118]
   nj = [2887,2698,3363,2432,3134,3214,4598,3320,3398]
   ct = [1403,1348,1810,1788,3058,2738,2751,1964,2004]
   ```

   Write a program that will print the maximum number of disease occurrences each state and the correlation between each state's occurrences, that is, $\rho(ny, nj)$, $\rho(ny, ct)$, and $\rho(ct, nj)$. You may compute the correlation directly or use the `correlation(x,y)` function from the book's `statistics` module. If you choose the latter, include appropriate `import` statements in your answer.

5. Most (90%) email messages are spam, while the remaining are real ("ham") messages. Your spam filter can identify spam messages about 75% of the time. You just received a message that the filter has marked as spam.

   Use Bayes Theorem to calculate the probability that the message is truly spam.

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:

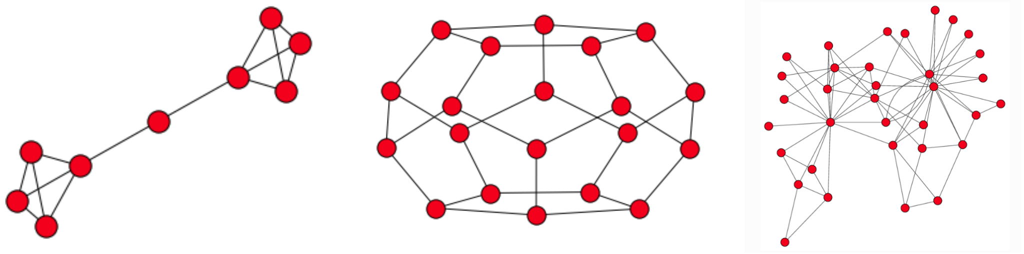
```
DV_DEVICE_NUMBER,Device Status,DV_DEVICE_STATUS_DESCRIPTION,BIN,TAX_BLOCK,TAX_LOT,
HOUSE_NUMBER,STREET_NAME,ZIP_CODE,Borough,Device Type,DV_LASTPER_INSP_DATE,
DV_LASTPER_INSP_DISP,DV_APPROVAL_DATE,DV_MANUFACTURER,DV_TRAVEL_DISTANCE,
DV_SPEED_FPM,DV_CAPACITY_LBS,DV_CAR_BUFFER_TYPE,DV_GOVERNOR_TYPE,DV_MACHINE_TYPE,
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. In network analysis, the number of connections a node (the "degree" of the node) and redundancy of paths are important concepts for assessing centrality and importance of nodes.



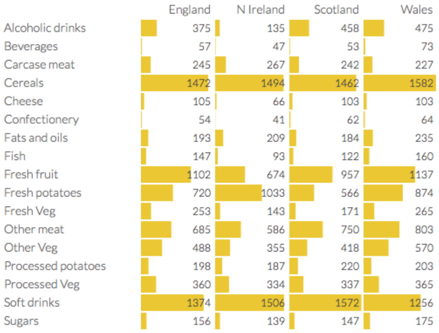
(a) What is the highest degree node in each graph above? Explain your answer.

(b) Write pseudocode that takes a graph and returns the k vertices with the highest degree:

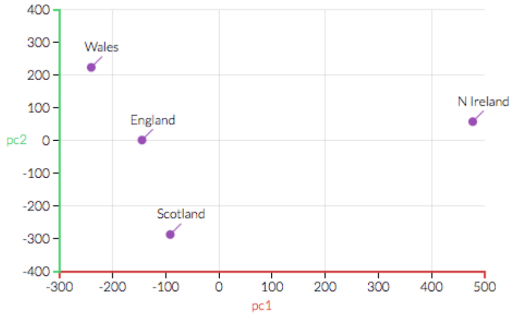
(c) Can you disconnect any of the graphs above by removing a single node? That is, if a single node on the network was removed, could you still communicate between the remaining nodes on the remaining edges?

(d) Write pseudocode that takes a graph and a node and returns true if the graph remains connected when the node is removed and returns false otherwise.

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. In class, we built simple neural networks to capture and combined them to recognize complex patterns. We used a perceptron from the textbook for simple modeling:

```
def step_function(x):
    return 1 if x >= 0 else 0
def perceptron_output(weights, bias, x):
    """returns 1 if the perceptron ’fires’, 0 if not"""
    return step_function(dot(weights, x) + bias)
```

(a) What weights and bias are needed to take two inputs and return 1 if both inputs are 1 and 0 otherwise (i.e. an AND gate):

(b) What weights and bias are needed to take one input and return 0 if the input is 1 and 0 otherwise (i.e. a NOT gate):

(c) We also trained neural networks to determine number captchas, where each number was represented by a 5 × 5 grid of values. For example, the number three is:

```
11111
....1
11111
....1
11111
```

It was able to recognize the following as the number “3” about 90% of the time. How could you improve that recognition rate of numbers that are not quite drawn the standard way?

```
.@@@.
...@@
..@@.
...@@
.@@@.
```