NAME:
EMAIL:
SIGNATURE:

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
Total	

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

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

1. What will the following code draw:

```
n = 10
X = np.arange(n)
Y1 = X/2.0
Y2 = X/4.0

plt.bar(X, +Y1, facecolor='blue')
plt.bar(X, -Y2, facecolor='red')

for x, y in zip(X, Y1):
    plt.text(x + 0.4, y + 0.05, '%.2f' % y, \
             ha='center', va='bottom')

plt.ylim(-5, +5)
```

Output:

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

   (a) `(\d){3}\w`

   (b) `[aA]+[bB]+[cC]+`

   (c) `Bro((nx)|(oklyn)|(ther))`

   (d) `\d{3}-\d{3}-\d{4}`

   (e) `\w+@([\w\.-]+)`

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 locations 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]
```

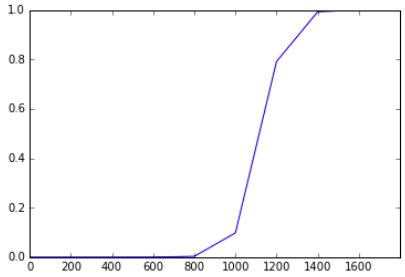
Write a program that will plot the percent increase in Lyme Disease occurrence with respect to the first year in the list (note some numbers plotted could be negative, since the disease occurrence has both decreased and increased from the initial observations).

5. You are responsible for testing two different front pages for a website. The first option has aquamarine colored buttons ('Option A'), and the second option has black colored buttons ('Option B'). If 200 out of 1,000 viewers click through on Option A and 180 out of 1,000 viewers click through on Option B.

   What is the A/B test statistic for Option A and Option B (i.e. the difference of the estimated means divided by the square root of the sum of the variances)?

6. You culled 20,000 admissions records to determine if SAT score could be used to predict admission. Using this training data, you fit the following logistic function:

$$f(x) = \frac{1}{1 + e^{200 - x/50}}$$



(a) With what probability would you expect that a student with a 500 on the SAT is admitted? Explain your answer.

(b) With what probability would you expect that a student with a 1100 on the SAT is admitted? Explain your answer.

(c) With what probability would you expect that a student with a 1500 on the SAT is admitted? Explain your answer.

(d) Say you sampled data from a single high school and found that average SAT of score of admitted students was 400 with a standard deviation of 53. Does that fit with your prediction function above? Why or why not?

7. You are helping a friend find a new apartment. After going through 100 listings with them, you created a training data set of tuples of important features `(price, size, closestStop, crimeRate, like)` where:

   - `price` is the monthly rent, in dollars
   - `size` is the square footage of the apartment
   - `closestStop` is the walking distance to the nearest subway entrance
   - `crimeRate` is the number of felonies per 1000 residents over the last year
   - `like` is 0 if your friend did not like the apartment and 1 if they did.

   There are 50 new listings today for apartments. Your friend would like to know which are the 5 apartments that they would like most.

   (a) Explain how the k-Nearest Neighbor algorithm would work for this data:

   (b) Write a function that takes the training data, the value k, and a new apartment tuple and returns true if the majority of the k closest neighbors are liked and false if not.

8. You are allowed 5 colors to print an image. Write a program that uses clustering to choose the 5 colors that would best capture the image.

   - displays the original image to the screen,
   - uses clustering to choose the best 5 colors, and
   - displays the image recolored with just those 5 colors.

   You may use any method to cluster to the colors.

9. Design a program that will allow the user to visualize actual distance and transit time distance between cities.

   Inputs:

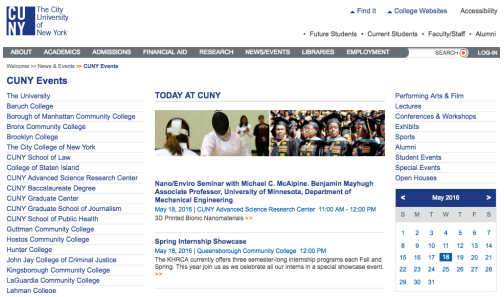
   - A dictionary of city names with values tuples (x,y) of their GIS coordinates.
   - A distance matrix of transit times between cities

   Output:

   - A map with the original cities (you may use any Python map drawing program– i.e. basemap, bokeh's mapping functions, etc.)
   - A Multidimensional Scaling (MDS) plot of the cities under the transit distances.

   Write the pseudocode and include a list of all packages you would use in the final design.

10. Design a program that scrapes the `events.cuny.edu` page and prints out the date and title of all events listed. The page and raw HTML look like:




```
                <div id="listing" class="listing">


                        <div class="block-txt">

            <div class="event-txt">
                        <!-- insert  getDetailHRef(eventId, CollegeId),  -->

                            <h3><a href="eventDetail.asp?EventId=76556">Nano/Enviro Seminar with Michael C. McAlpine.
Associate Professor, University of Minnesota, Department of Mechanical Engineering</a></h3>
                  <h4><a href="eventDetail.asp?EventId=76556">
               May 18, 2016

               | CUNY Advanced Science Research Center
                         11:00 AM - 12:00 PM
               </a></h4>
                 <p class="short-description">
                            3D Printed Bionic Nanomaterials
                  <a href="eventDetail.asp?EventId=76556" class="suffix">&gt;&gt;</a>
                 </p>
```