

NAME:
EMAIL:
SIGNATURE:

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	
2	
3	
4	
5	
6	
7	
8	
9	
10	
Total	

1. What will the following code draw:

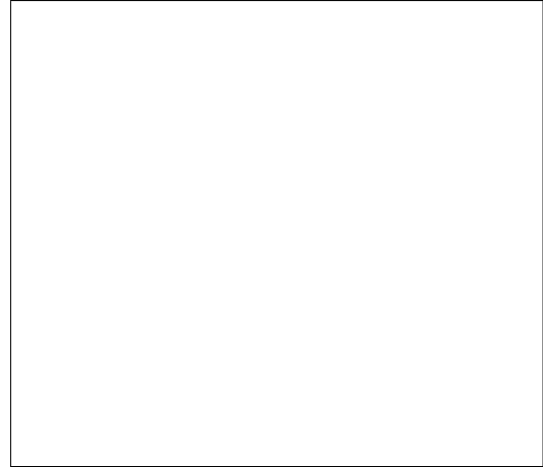
```
n = 10
X = np.arange(n)
Y1 = X/2.0
Y2 = X/4.0

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]
```

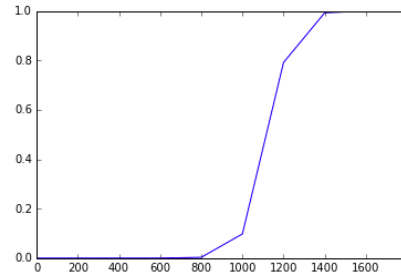
```
ct = [1403,1348,1810,1788,3058,2738,2751,1964,2004]
```

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:

The screenshot shows the CUNY Events page. On the left, there is a navigation menu with links to various CUNY entities. The main content area is titled "TODAY AT CUNY" and features a large image of a group of people. Below the image, there is a list of events, including a "NanoEnviro Seminar with Michael C. McAlpine, Benjamin Mayhugh Associate Professor, University of Minnesota, Department of Mechanical Engineering" on May 18, 2016. A calendar for May 2016 is also visible, showing the date May 18 highlighted.

```

221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```