**Microsoft:** DAT210x Programming with Python for Data Science

Bookmarks

- Start Here
- 1. The Big Picture
- 2. Data And Features
- 3. Exploring Data
- 4. Transforming Data
- **5. Data Modeling**

**Lecture: Clustering**
Quiz

**Lab: Clustering**
Lab

**Lecture: Splitting Data**
Quiz

**Lecture: K-Nearest Neighbors**
Quiz

**Lab: K-Nearest Neighbors**

5. Data Modeling > Lab: Clustering > Assignment 1

Bookmark

## Welcome to Module 5's Clustering Labs!

In order to complete the Clustering labs in this module, please make sure you download and unarchive this .zip file with all the datasets and files necessary.

## Lab Assignment 1

Many U.S. cities, the U.S. federal government, and even other cities and governments abroad have started subscribing to an Open Data policy, because some data should be transparent and available to everyone to use and republish freely, without restrictions from copyright, patents, or other mechanisms of control. After reading their terms of use, in this lab you'll be exploring the City of Chicago's *Crime* data set, which is part of their Open Data initiative.

1. Start by navigating over to the City of Chicago's Crimes dataset exploration page. It lists crimes from 2001 to the present, but you'll only be targeting **Gambling**. The city's website itself has hundreds of other datasets you can browse and do machine learning on.
2. Open up the /Module5/**assignment1.py** starter code, and follow the directions to acquire the dataset and properly set it up.
3. Fill out the doKMeans method to find and plot **seven clusters** and print out their centroids. These could be places a police officer investigates to check for on-going illegal activities.
4. Re-run your assignment a few times over, looking at your printed and plotted results. Then answer the following questions.

Lab

**Lecture: Regression**
Quiz

**Lab: Regression**
Lab

**Dive Deeper**

- 6. Data Modeling II
- 7. Evaluating Data
- Course Wrap-up

*Note: If Pandas complains about your data, you can use dropna() on any row that has nans in it.*

## Lab Questions

(2/2 points)

You'll notice that the cluster assignments are pretty accurate. Most of them should be spot-on, dead-center. Only one cluster might have been assigned to outliers. Given the results, answer the following questions to the best of your ability:

Did your centroid locations change after you limited the date range to +2011?

Only slightly... ▾ ✓ **Answer:** Only slightly...

What about during successive runs of your assignment? Any centroid location changes happened there?

○ All clusters have moved, and the cluster arrangement isn't anything like it was before

● All clusters have moved but only slightly, and the centroid arrangement still has the same shape for the most part ✓

○ The clusters did not really move at all, or if they did, it wasn't noticeable

○ The cluster centroids are identical according to the print statement output

EXPLANATION

There isn't a great deal of data here; but even with just what you have, it would be statistically very difficult for you to get the same cluster centroids between successive runs. And of course if you filter your data and change its samples, that will alter the K-Means centroid arrangement.

*You have used 3 of 3 submissions*