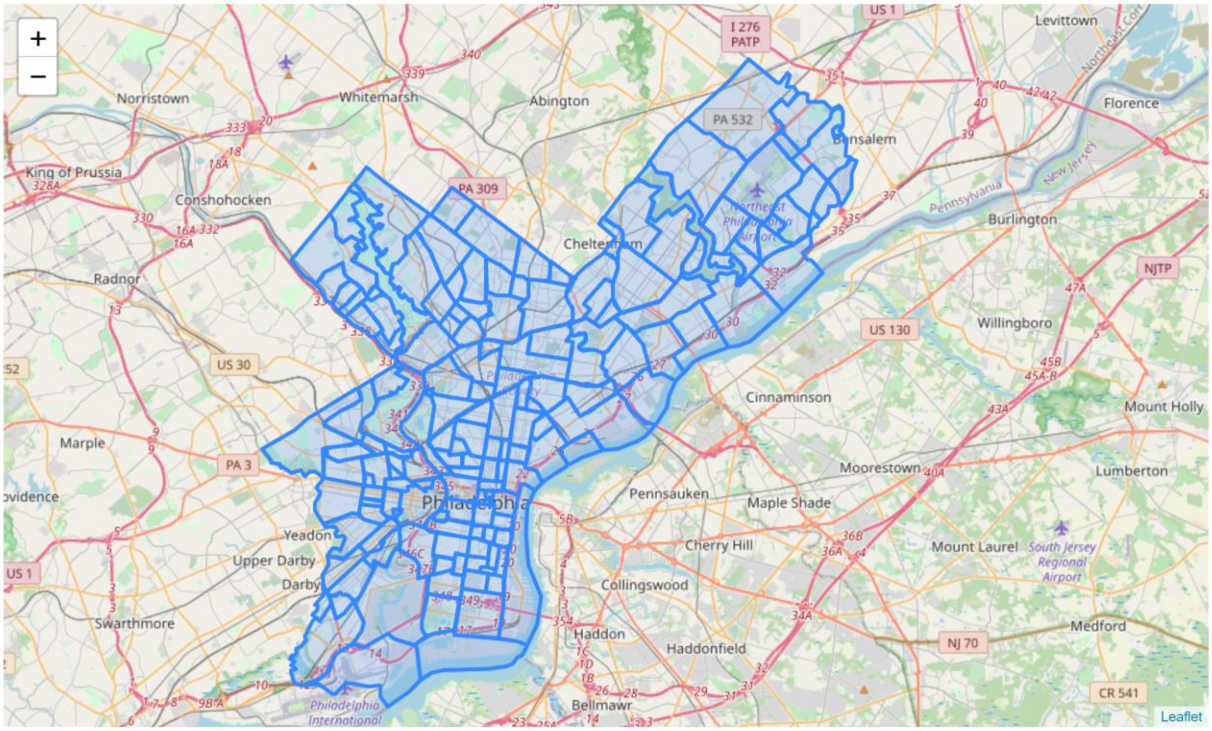
Introduction/Business Problem

For my project I decided to explore whether there is a correlation between neighborhood types derived using ML algorithms and different crime types also broken down by some clusters using K-means. To me this looks to be a very interesting thing to explore: maybe there are some hidden dependencies that we can discover. For example: liquor shops may draw more DUI crimes nearby. In ideal case it can help predict whether neighborhoods having a certain set of attributes (most popular venues in this case), will draw a certain type of crimes more than others.

The map below represents basic geographical segments that we will working with:

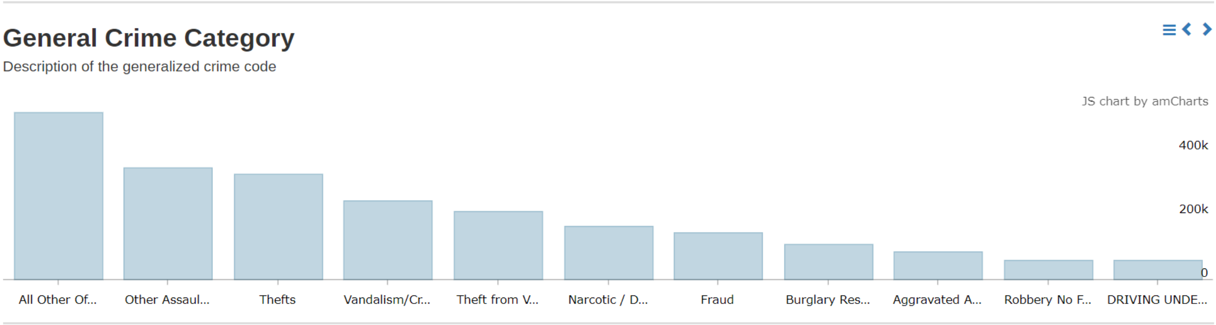


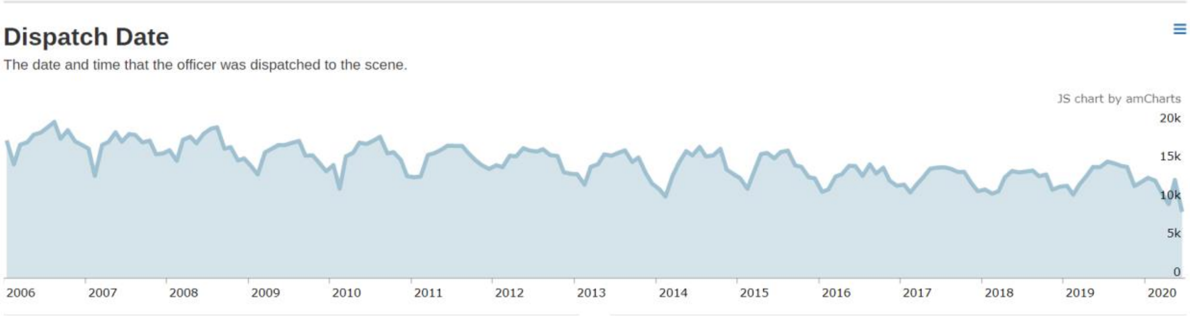
Data Sources:

1. Philadelphia Police Department crimes data set. It includes both violent and nonviolent offences. The original set contained many fields that are not required for the analysis, so only address location, general crime type, datetime and latitude and longitude of that address were left for clustering purposed.
   https://www.opendataphilly.org/dataset/crime-incidents
2. Foursquare API – similar to the course material, I will use it information about neighborhoods
3. Neighborhood shape data in GeoJson format. This will be used to display neighborhoods borders on the map. Latitude and longitude coordinates of each crime record will be compared vs shapes from this source to map each crime into appropriate neighborhood.
   https://github.com/azavea/geo-data/tree/master/Neighborhoods_Philadelphia

Philadelphia police department dataset contains data from 2006 to present, hence the file size is pretty significant.

Here are some visuals to provide a quick snapshot of the data.





FourSquare is a local search place app which provides recommendations of places to go . We use thi service to harvest information about Philadephia neighborhoods.

The API use requires credentials that have been deleted from the workbook published because of privacy/security reasons. Hence it is recommended to load presaved datasets from csv files that this team has saved in the same folder.

The query brought 290 unique categories of venues. Our hypothesis that such a big amount of clustering features will introduce some challenges in interpreting clustering results. So, to avoid it we applied method to produce top 20 venue types for each neighborhood in hope that this will produce "human readable " cluster results.

Methodology:

I used the datasets from above to generate datasets ready to be used for clustering analysis.

DATA CLEANSING:

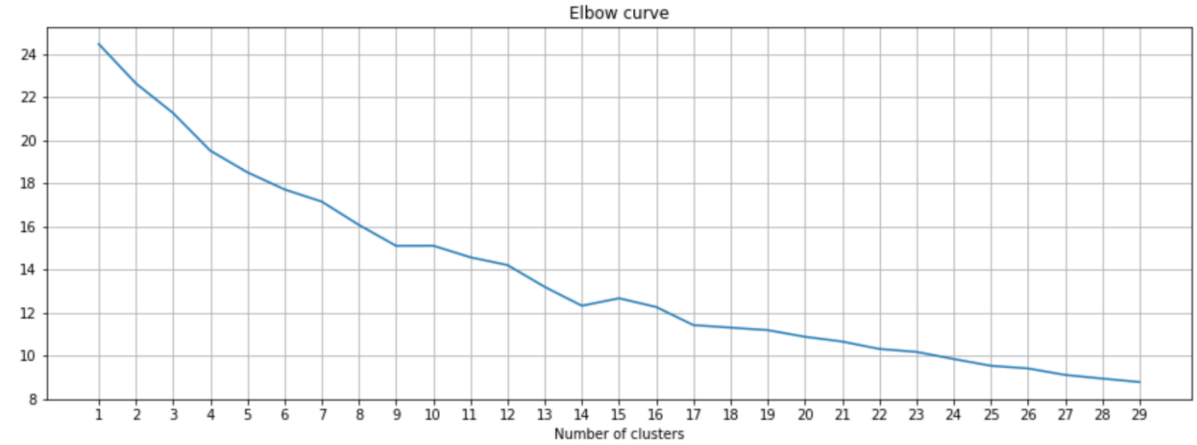
- Neighborhood Crime dataset: Just a regular cleansing with datatypes/ categorical types. The most cumbersome task was to find a proper neighborhood given latitude and longitude of a crime. This team used function borrowed from the link in the code to run through all 3M records of crime data and assign each record to a neighborhood based on these coordinates. It was a very long process and it took more than 8 hours to process all rows on a standard "cognitive labs" python virtual server. Next, crime type was one hot encoded and fed into standard scikit learn k-means clustering algorithm.

- Neighborhood venues dataset. Also, just a regular cleansing like in the course workbook.

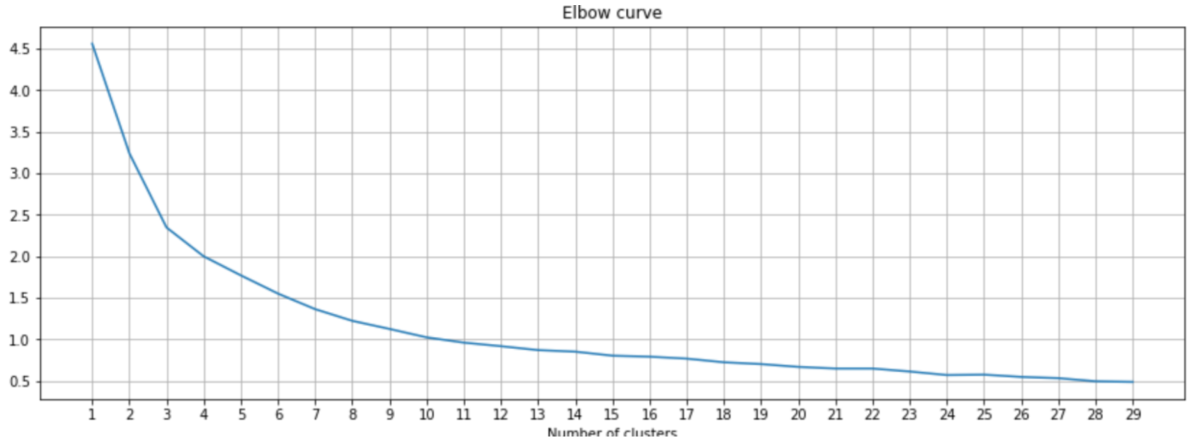
CLUSTERING:

I used K-means clustering method for analyzing for patterns and grouping. To come up with the optimal clusters amount I used Elbow approach for both crime rates and neighborhood analysis. The inertia amount was starting to show less degree of declining with cluster amounts of 10 for both subsets. Hence, I choose 10 as an optimal N for both datasets.

Elbow curve "Venues clusters"

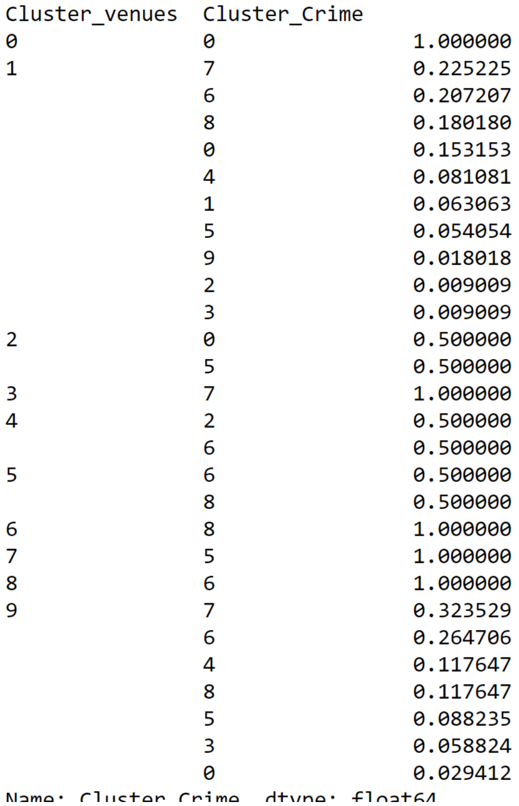


Elbow curve "Crime clusters"



Results:

The final dataset consists of each neighborhood clustered by venues and crimes. Also, each line includes top ten venues and top ten crimes. During EDA, few neighborhood clusters showed clear relationship with definite crime clusters, while majority of them had at least 2-3 crime clusters without and dominant ones. Below is the grouping by venue and then crime cluster. Numbers represent share of each crime cluster in venues cluster.



```
Cluster_venues  Cluster_Crime
0               0                1.000000
1               7                0.225225
                6                0.207207
                8                0.180180
                0                0.153153
                4                0.081081
                1                0.063063
                5                0.054054
                9                0.018018
                2                0.009009
                3                0.009009
2               0                0.500000
                5                0.500000
3               7                1.000000
4               2                0.500000
                6                0.500000
5               6                0.500000
                8                0.500000
6               8                1.000000
7               5                1.000000
8               6                1.000000
9               7                0.323529
                6                0.264706
                4                0.117647
                8                0.117647
                5                0.088235
                3                0.058824
                0                0.029412
Name: Cluster_Crime, dtype: float64
```

For example, neighborhoods clusters under 0,3 and 6 cluster only have crimes clustered 0,7 and 8 respectively.

Some neighborhoods (cluster 2,4,5) showed predominant crime type (crime clusters 5,6), while roughly half of neighborhood clusters had no clear dominant clustering in crime.

Let’s try to determine if there are any similarities in clusters and whether we can spot any trends. Again, we will focus only on Neighborhood clusters that only have 1 or 2 predominant crime types (0,3,6) in it, ignoring neighborhoods with all sort of crimes.

Crime cluster 0 is a cluster with predominant all other offences and thefts from vehicles. Then Thefts, Vandalism/Fraud are trailing behind. While all of this is not really great, still it is much more “safer” than Cluster crime 7 for example, where “All other assaults” are taking 1 and the second most frequently happening place.

Let’s see if we draw any conclusions: Neighborhood cluster 0 which draws so many “light” crimes is a cluster where many service establishments are located: Home Service/Vintage Store/Barbershop/Zoo Exhibit – all nonfood places with some exceptions. Accordingly, these spots with mass gatherings, though not as social as eating/drinking draw many “light crimes”. On the contrary, Neighborhood cluster 6 is more of “eatery” style borough with Thai/Fast Food/Filipino restaurants. It looks like that these (not necessary these specific types of restaurants, but more like general places where people can get food and drinks) – tend to attract more violent type of crimes.

Discussion:

Clearly there is not enough information gained by using this data and trying to see how clusters obtained for both crime and venues dataset relate to each other. We can observe that small bars and eateries can relate to more serious crimes vs neighborhoods that have more like public gathering venues, where predominantly other offences/thefts occur mostly. Still we identified only few clusters out of 10 where this relationship has a clear effect. For the rest – we can’t say with a great certainty what type of crime will prevail in that specific neighborhood and this is clearly something to work on in future developments.

Also, there can be many other steps to improve model accuracy (specifically on the crimes side, as there are too many “Other” types of crimes). Though, 290 features of different venues types also introduced some noise. For future work we would recommend go across each venue type and aggregate that into higher level categories.

Conclusion:

In this project I applied the approach discussed during the class to analyze city neighborhoods.   I decided to work on city of Philadelphia neighborhoods. Additionally, as I was interested in seeing whether there is a relationship between clusters of venues types and crime types, I applied similar K-means clustering technique on Philadelphia police dataset and came up with neighborhoods split by crime cluster.

In the end out of 10 clusters I was able to relate 3 definite venue clusters with specific crime rates clusters (see results section). The other 7 clusters had some mixed results without any predominant crime cluster.