Data Analysis of Arlington County Police Incidents

Leland Ball

Final Project Report
June 2020
IST 652

This report covers the data analysis carried out on Police Incidents available via the the Arlington County Open Data initiative. The data and analysis covers the 2015-2020 timeframes, with emphasis on the most recently completed year of data available, 2019.

Data Sources and Preprocessing

Data was sourced via the Arlington County website, which hosts an API and manually downloadable sources of data, which include GIS polygons for mapping areas of interest, and tabular data including a list of police incidents logged for the past several years. These data sources are available for download at https://gisdata-arlgis.opendata.arcgis.com/datasets/zipcode-polygons and https://data.arlingtonva.us/dataviews/225891/police-incident-log/, respectively. Of interest was the zipcode polygons, for later conversion of police incident lat/lon into a zipcode, and also for the purposes of mapping. This set contains about fifty polygons encompassing thirteen zip codes. The police incident data is 63,000 entries containing a loosely-strict categorical description of the incident, lat/lon, a nearby street address at the block level, and various dates and times that the incident started and stopped at.

Very little preprocessing was undertaken, with the exception of the incident categories. There were 1030 unique categorical values for the incident category, with the top 60 categories accounting for 73.6% of all incidents. Applying a manual heuristic category reduction involving lumping together similar terms (e.g. “theft” gets grouped with “Larceny” and “Petit Larceny” is identical to “Larceny – Petit” reduces these top-60 categories to 42 in total- a 30% reduction. After these top categories were identified and consolidated, a document similarity algorithm was used to further consolidate an additional 13.6% of all categories into the top 42 identified categories, and to put the remaining 12.8% of police incidents into the “OTHER” category.

The WMDistance algorithm (Ofir Pele, Michael Werman, Matt Kusner)[1] was used to determine document similarity between each police incident category and the top 42 police incidents. A threshold value was empirically determined under which incidents with similar or identical wording would be combined. This results in categories such as “AGGRAVATED SEXUAL BATTERY” to be subsumed under “ASSAULT BATTERY” while “MALICIOUS UNLAWFUL WOUNDING” (which was not a common occurrence in the data, and so had no very similar category) was binned into the “OTHER” category. An alternative to the WMDistance algorithm may be a cosine similarity measurement using term vectors similar to the TF/IDF algorithms. The WMDistance algorithm was trained on the entire unique corpus of 1030 incident categories.

Other minor data processing included converting each date/time column into a Datetime object, and adding the "EST" timezone to this information.

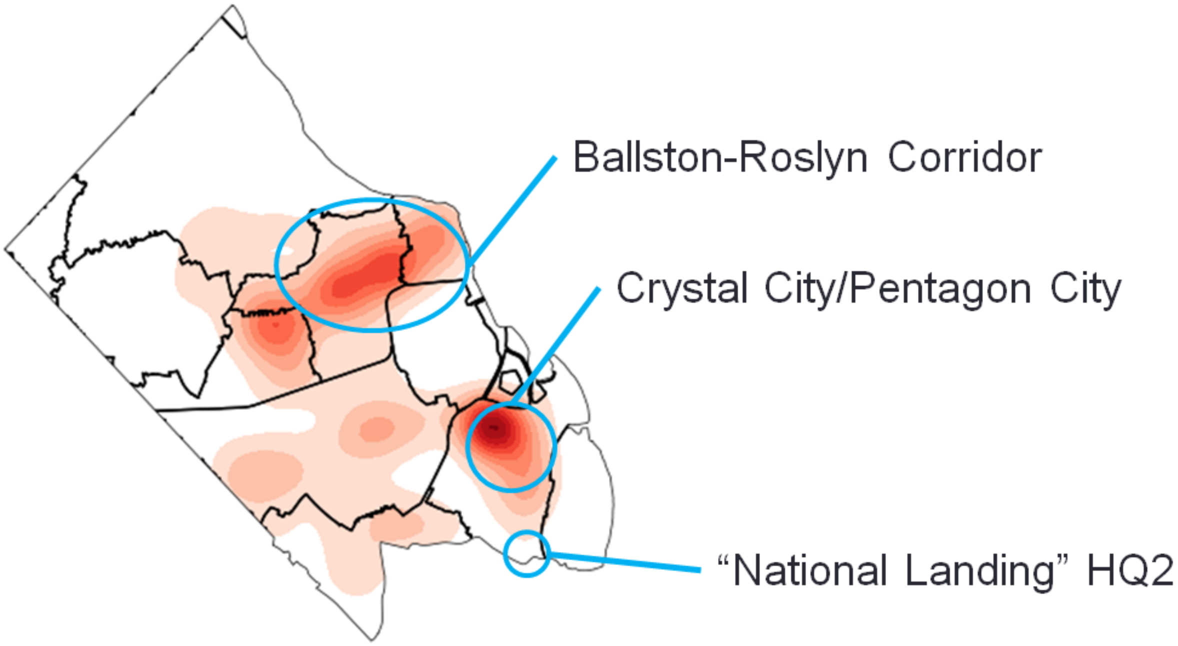
Methods of Analysis and Data Questions

Three major questions of interest included:

- Where are police incidents most common?
- What time are police incidents being recorded?
- Which police incidents are most common?

Where are police incidents most common?

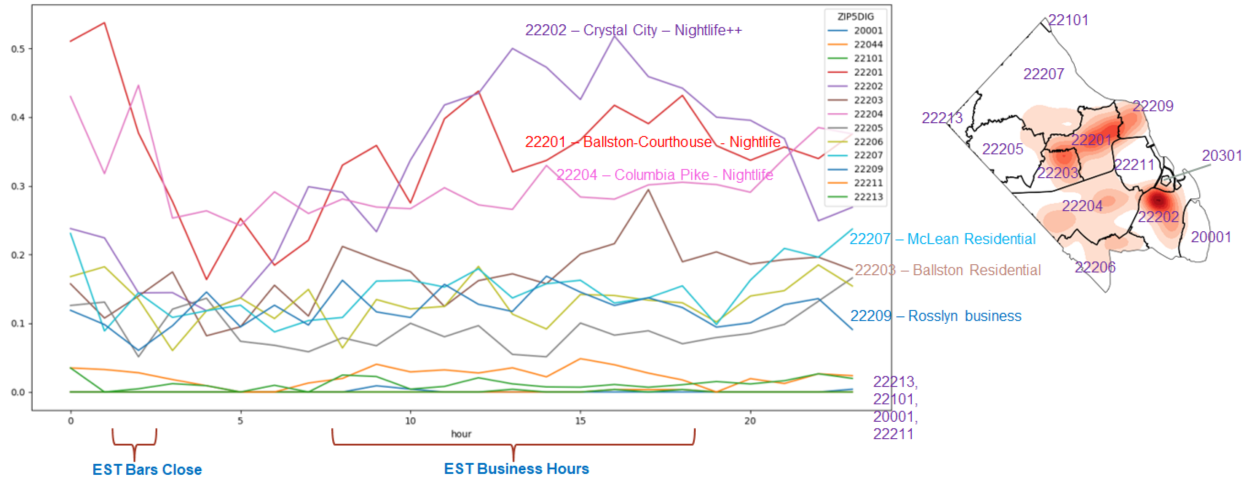
To answer this question, the data was analyzed and displayed in heatmap format using the GeoPandas Python library. This data was then graphed upon the previously mentioned GIS ZipCode information available from Arlington's GIS portal



This required taking the lat/lon of each police incident and projecting it to the same map as the ZipCode GIS polygons

What time are police incidents being recorded?

This question required that the data be binned by hour. After hours were calculated as a separate column for the data, the mean of the sum of all incidents for each hour was calculated and broken out by ZipCode. The time-series was plotted, and then further annotated by hand for clarity.



As can be seen in the plot above, residential areas see much less activity than the more urban and business districts. A distinction can be seen between business districts and the ZipCodes that contain fewer offices and more restaurants/bars (nightlife). This can be seen when looking at Rosslyn (22209) and Ballston-Courthouse (22201) and especially Crystal City (22202). Police incidents are more frequent in these areas, and activity increases later in the day, and into the early morning hours.

Which police incidents are most common?

To address the question of commonality of police incidents, a word-cloud was developed to gain an understanding of just which categories were the most frequent (fig. 1). Then another cloud was created after reducing the number of categories down to as few as possible (fig. 2).

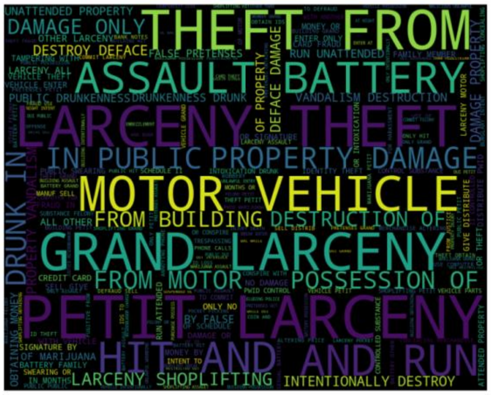


Fig 1

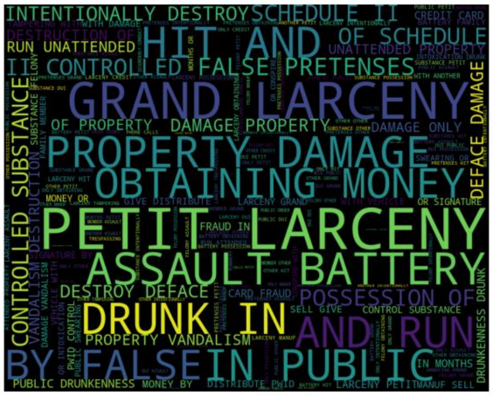
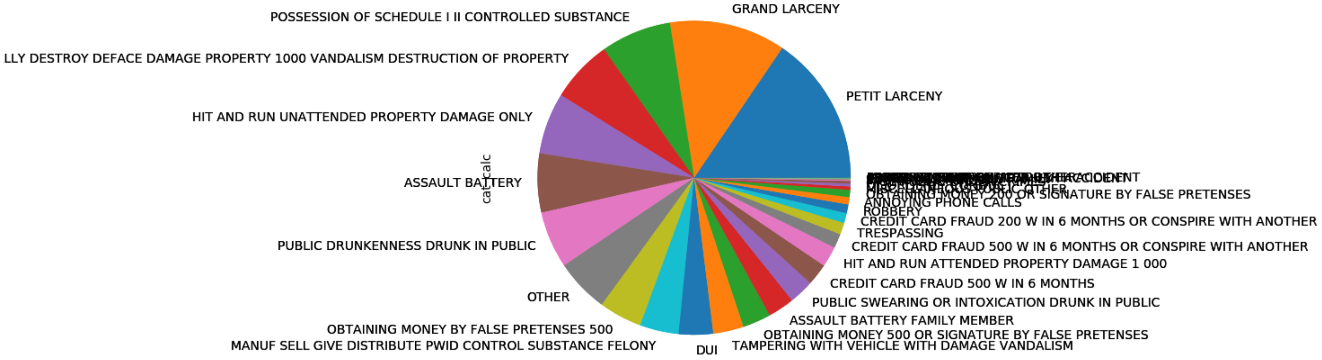
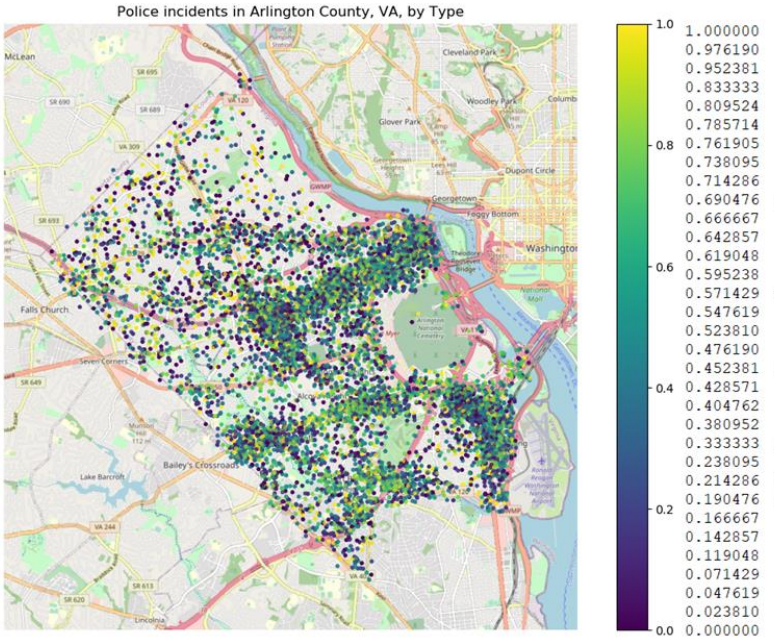


Fig 2

This is where categorization comes in. As explained in the *Data Sources and Preprocessing* section, both manual heuristics and the WMDistance algorithm were used to consolidate the number of categories down to the fewest possible. While the word-cloud was an easy place to start, much work went into re-classifying categories into these top 42 categories. A pie-chart was then made to show how large each category is in relationship to each other, and is displayed below.



And finally, a rudimentary categorical plot was made on top of a street map of the area. The categorical coloring is approximate, but a key is included as well.



```
1.000000                                               OTHER
0.976190        TAMPERING WITH VEHICLE WITH DAMAGE VANDALISM
0.952381                  PERSONAL INJURY MOTOR VEH ACCIDENT
0.833333    POSSESSION OF SCHEDULE I II CONTROLLED SUBSTANCE
0.809524                  PUBLIC DRUNKENNESS DRUNK IN PUBLIC
0.785714                          SIMPLE ASSAULT NON FAMILY
0.761905              OBTAINING MONEY BY FALSE PRETENSES 500
0.738095                                         TRESPASSING
0.714286                                                 DUI
0.690476  CREDIT CARD FRAUD 200 W IN 6 MONTHS OR CONSPIR...
0.666667     PUBLIC SWEARING OR INTOXICATION DRUNK IN PUBLIC
0.642857                          MISCELLANEOUS PUBLIC OTHER
0.619048                                    PARKING VIOLATION
0.595238                                   SUSPICIOUS PERSON
0.571429  MANUF SELL GIVE DISTRIBUTE PWID CONTROL SUBSTA...
0.547619                 CREDIT CARD FRAUD 500 W IN 6 MONTHS
0.523810                                             ROBBERY
0.476190                                      FOUND PROPERTY
0.452381  CREDIT CARD FRAUD 500 W IN 6 MONTHS OR CONSPIR...
0.428571                                       GRAND LARCENY
0.404762                        ASSAULT BATTERY FAMILY MEMBER
0.380952           HIT AND RUN ATTENDED PROPERTY DAMAGE 1 000
0.333333  OBTAINING MONEY 500 OR SIGNATURE BY FALSE PRET...
0.238095                                ANNOYING PHONE CALLS
0.214286  INTENTIONALLY DESTROY DEFACE DAMAGE PROPERTY 1...
0.190476  OBTAINING MONEY 200 OR SIGNATURE BY FALSE PRET...
0.166667                                  DISORDERLY CONDUCT
0.142857                  PROPERTY DAMAGE MOTOR VEH ACCIDENT
0.119048                                     ASSAULT BATTERY
0.071429          HIT AND RUN UNATTENDED PROPERTY DAMAGE ONLY
0.047619                             MISC PUBLIC ORDER CRIMES
0.023810                          MISCELLANEOUS OFFICER OTHER
0.000000                                        PETIT LARCENY
```

Program Description

The Python program used to develop this analysis makes use of the MongoDB, GeoPandas, MatPlotLib, GeoPlot, WordCloud, and Gensim Python libraries, amongst others. It was created in an Anaconda environment and developed using PyCharm. An ordered description of the program's activities follow:

1. Connect to MongoDB, and download all police incidents (64k documents)
2. Load data into Pandas DataFrame, get top 60 most common incidents, by name

3. Apply text processing to each category including stopword removal, stripping of non-alpha-numeric characters, and uppercasing
4. Load WMDistance model from disk, or recalculate with corpus of categories
5. Compute categorical similarities for each category in the dataframe, and assign new calculated category (of the 42 and “OTHER”)
6. Cast and convert lat/lon and datetime fields
7. Filter for only 2019 data
8. Import ZipCode GIS polygons and convert to lat/lon instead of non-lat/lon encoding
9. GIS-Join incidents and ZipCode polygons, resulting in a single GeoPandas DataFrame and excluding police incidents from outside of Arlington
10. Produce plot of counts of police incidents by hour, across each ZipCode
11. Produce heat map plot of incidents by lat/lon, on top of ZipCode map
12. Produce word cloud graphic of incidents using newly calculated categories
13. Produce pie chart of police incident categories
14. Produce point plot of each incident on web map of Arlington County (no distinct point colors)
15. Produce arbitrarily-hued categorically-colored plot of points on Arlington County web map
16. Print simple string version approximate color-code key for the above map

Conclusions and Results

This analysis of openly available county law-enforcement data shows some obvious truths and some interesting observations. There is distinct police activity in many areas that have more nightlife, with restaurants and bars. This is not necessarily related with population density during the day, as one of the busiest places in Arlington (the Rosslyn neighborhood) sees a relatively stable and shallow amount of police activity. However, areas that are more popular recreationally see an increase in policing activity during the non-morning hours. Looking at the categories of police incidents available, it is no surprise that sobriety issues (one of the top 42 categories of police incidents) may contribute to this trend. Additionally, the various categories of larceny appear to be the biggest category of police activity, accounting for over 25% of activity recorded by Arlington in 2019.

One of the most interesting explorations of this project was the strategy to consolidate loosely-maintained categorical data into larger categories of interest. While some fidelity is lost, it can be seen that both the manual heuristic approach of lumping categories under their parent-category umbrella, and the use of machine learning for document-proximity re-classification can go a ways to giving a rapid understanding into categorical data.

Citations and URLs:

[0] Arlington County Data

https://gisdata-arlgis.opendata.arcgis.com/datasets/zipcode-polygons https://data.arlingtonva.us/dataviews/225891/police-incident-log/

[1] WMDistance Algorithm

Ofir Pele and Michael Werman "A linear time histogram metric for improved SIFT matching"
<http://www.cs.huji.ac.il/~werman/Papers/ECCV2008.pdf>

Ofir Pele and Michael Werman "Fast and robust earth mover's distances"
<https://ieeexplore.ieee.org/document/5459199/>

Matt Kusner et al. "From Word Embeddings To Document Distances"
<http://proceedings.mlr.press/v37/kusnerb15.pdf>