POLICE INCIDENTS: A DIVE INTO DATA

Police Incidents in Arlington County, VA

Leland Ball

Arlington County Open Data

- Arlington County Open Data
  - Freely accessible API
- GIS shapes
  - https://gisdata-arlgis.opendata.arcgis.com/datasets/zipcode-polygons
  - ZIP code breakdowns, as well as census, voter districts, and school zone polygons
- Police Incident Logs
  - https://data.arlingtonva.us/dataviews/225891/police-incident-log/
  - Includes category of incident, as well as lat/lon

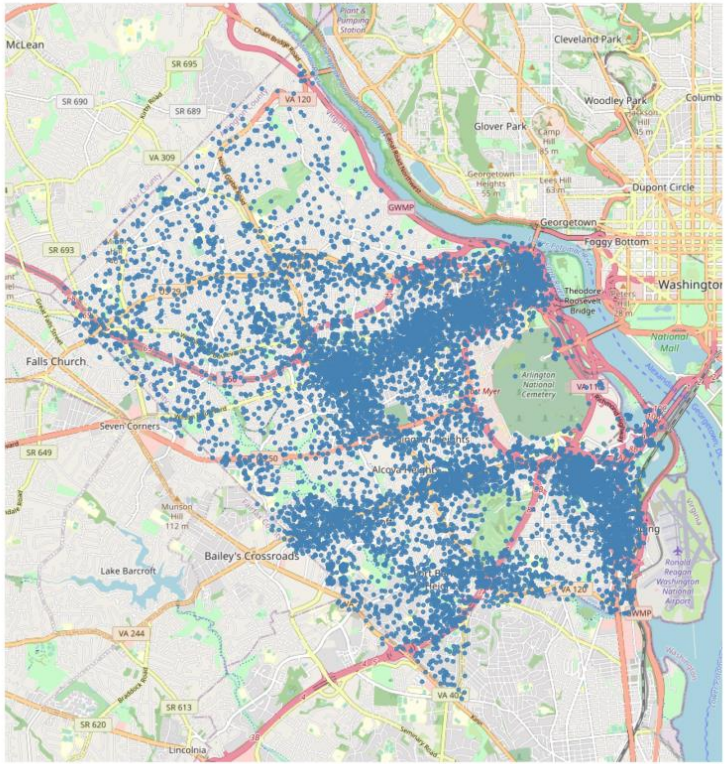
Questions

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

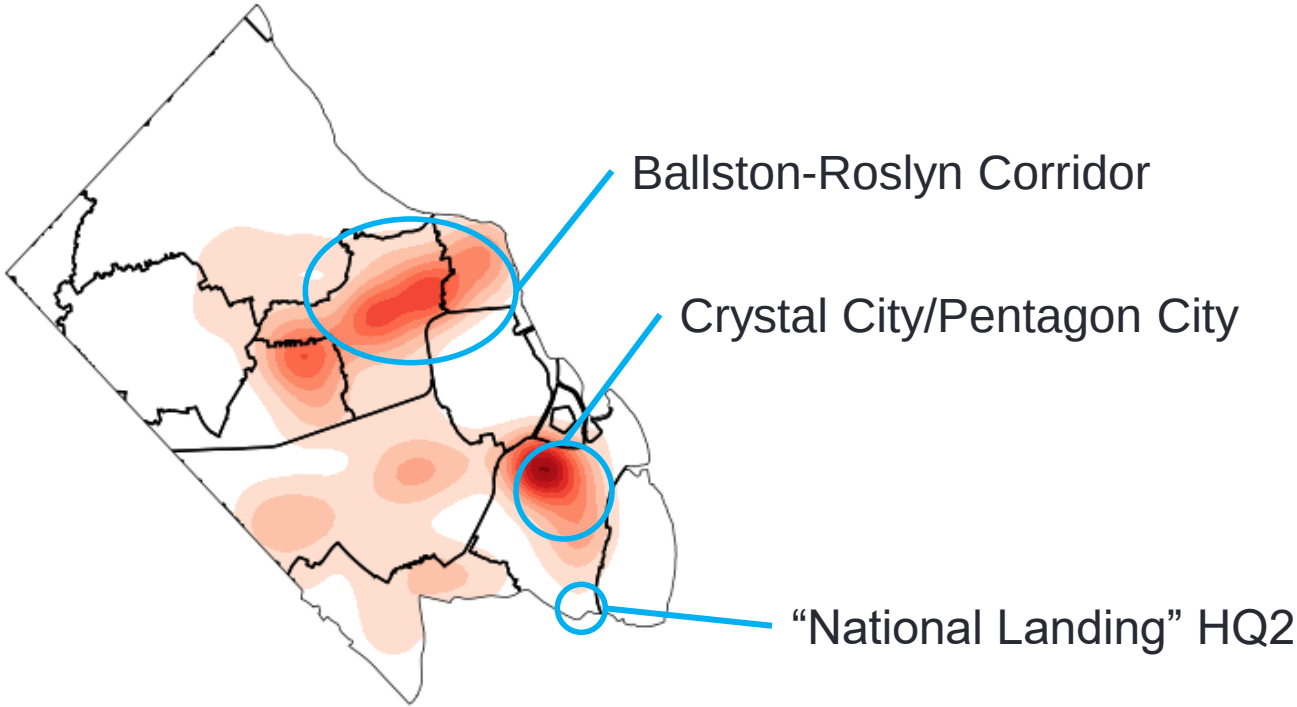
Data Quality

- High quality data from the Arlington County Open Data API
- Categories of police incidents were heavily abridged
- Times were adjusted to reflect the EST timezone
- Data was loaded into Pandas and GeoPandas DataFrames for analysis
  - This allowed for a geo-lookup from Lat/Lon to ZIP code

Where are incidents most common?

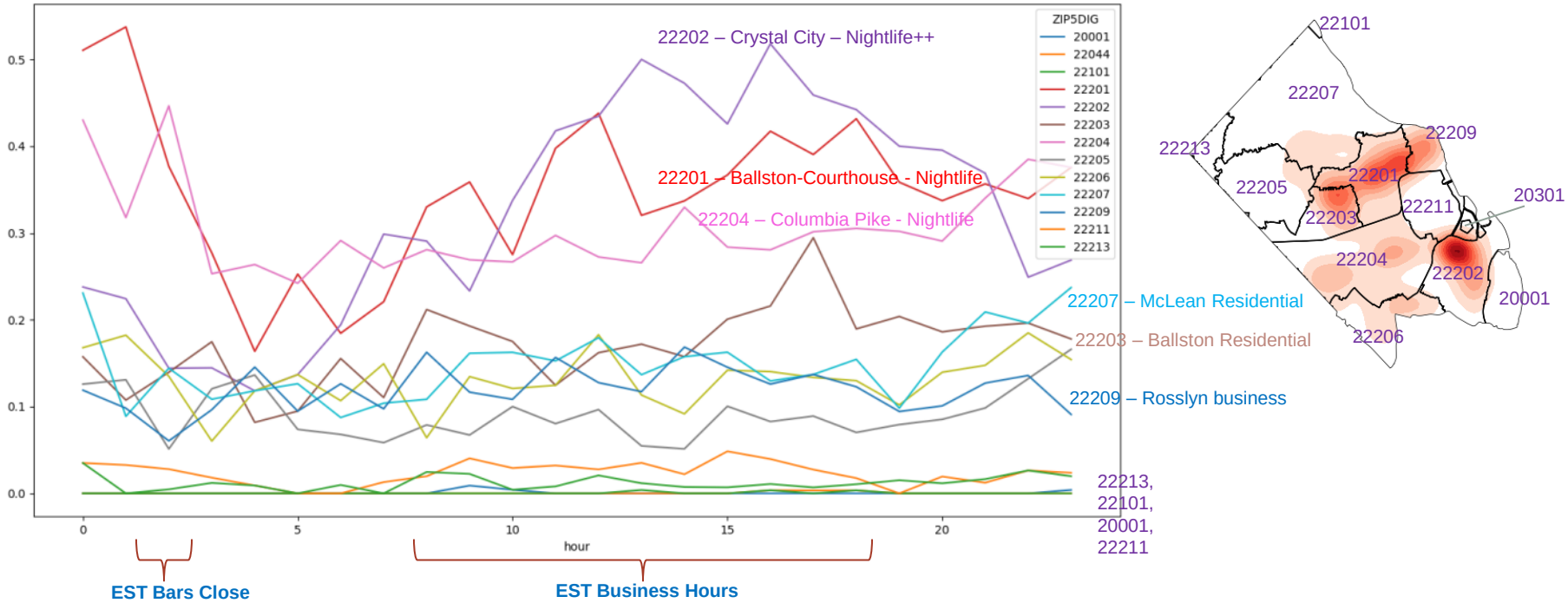


Where are incidents most common?



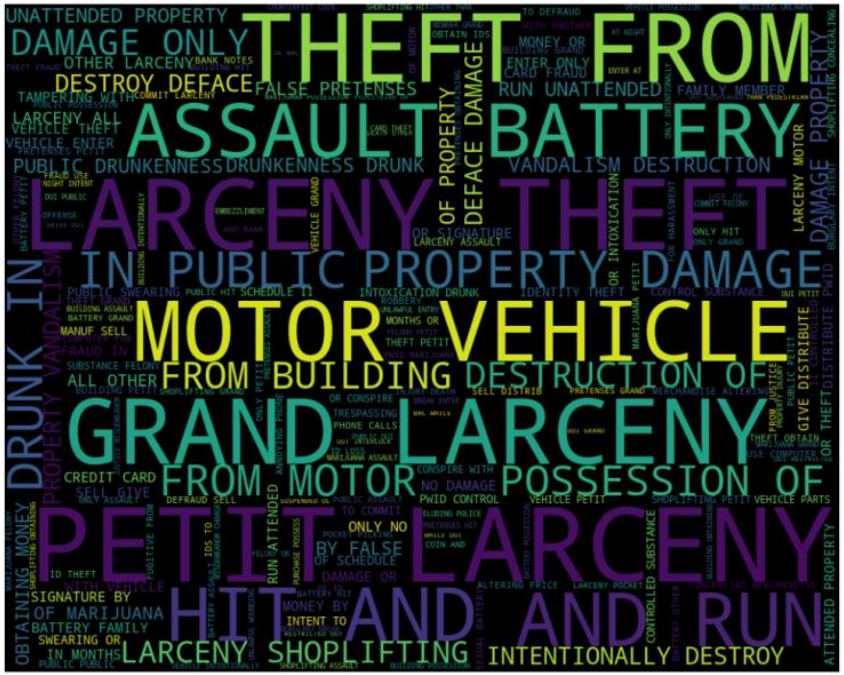
What time are incidents being recorded?

- Mean incident count over 2019, by Zip code,



Which incidents are most common?

- Unprocessed Word-Cloud of Police Incident Categories



Which incidents are most common?

- Post-Category Processed Word-Cloud

Categorizing Binning Incident Types

- 63,724 incidents in the data set (2015-2020)
- 1030 unique incident categories
- Top 60 categories account for 73.6% of all incidents
- Applying heuristic category reduction results in a 30.0% reduction in total category number
  - 42 categories now account for 73.6% of all incidents
- Document Similarity bins additional 13.6% into those 42
  - This leaves 12.8% in new “OTHER” category

Categorizing - Heuristics

```
manual_cat_combos = OrderedDict({
    'PETIT,LARCENY': 'PETIT LARCENY',
    'GRAND,LARCENY': 'GRAND LARCENY',
    'LARCENY': 'LARCENY',
    'LARC': 'LARCENY',
    'THEFT': 'LARCENY',
    'MARIJUANA': 'MARIJUANA',
    'DUI': 'DUI',
    'DRIVING,UNDER,INFLUENCE': 'DUI'
})
```

Any “Larceny” or “Larc” gets binned as “Larceny”, but if the incident has “Petit” or “Grand” then it is binned as that specific type of larceny

Categorizing - WMDistance

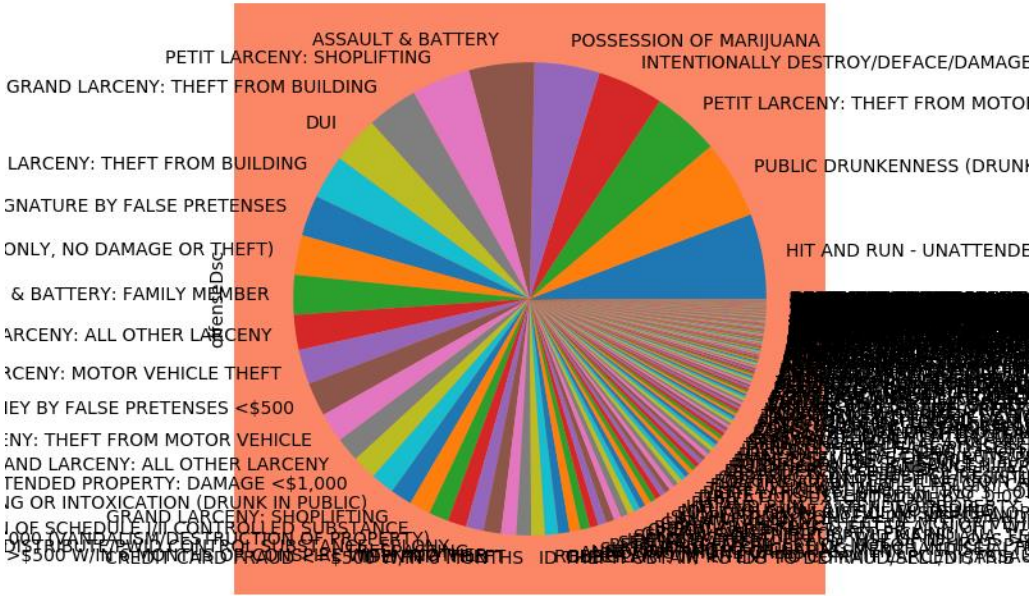
- WMDistance is a measure of document similarity
  - Requires training on a corpus (trained on 1030 documents)
- Remaining categories get binned by document similarity
- Could probably have used cosine similarity as a lighter way to measure similarity

- **ASSAULT BATTERY** is binned as **ASSAULT BATTERY**
  - Identical document results in 0 score
- **AGGRAVATED SEXUAL BATTERY** is binned as **ASSAULT BATTERY**
  - Similar document results in ~0.03 score or less
- **MALICIOUS UNLAWFUL WOUNDING** is binned as **OTHER**
  - Documents dissimilar enough from the “top N” most common terms, with scores above 0.03 are binned as “OTHER”

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>

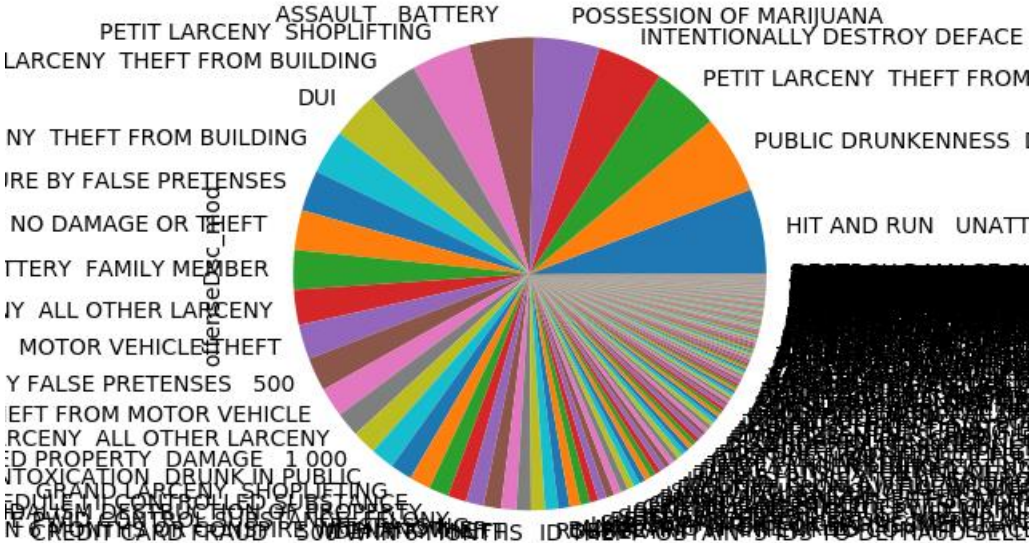
Categorizing

- Default 1030 categories in a pie chart



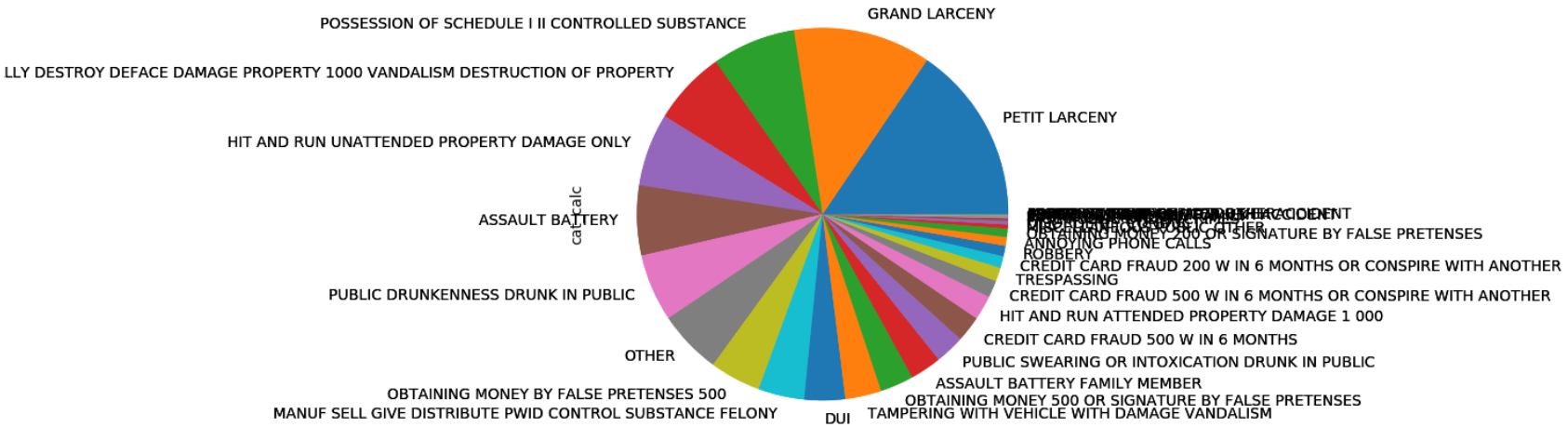
Categorizing

- After String Manipulations (no drastic changes)

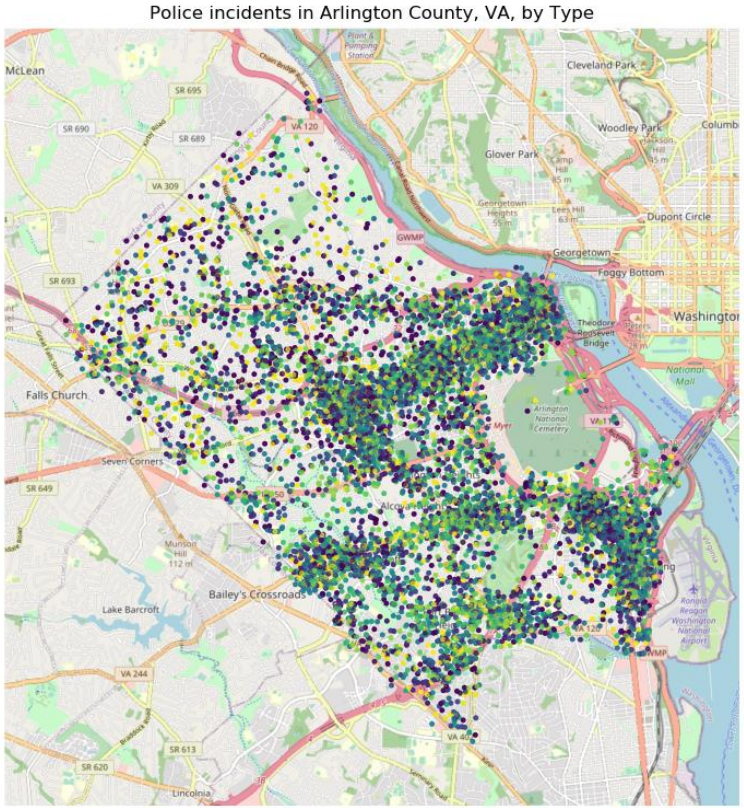


Categorizing

- Post Top-60, Heuristics, and WMDistance



Categorizing - Plot



Interesting Finds?

- Unusual/Incorrect Incidents:

  VERY OUT OF TOWN
  GRAND SHOPLIFTING
  TRAFFIC:WINDSHIELD OBSTR:-SIGN-POSTER-DECAL
  SUNSHIELD WINDOWS-NO MEDICAL NEED>=70%WINDSHIELD

Thanks!