Debbie Yu | General Assembly | DAT7 | August 10, 2015

PREDICTING CRIME IN DC

PROBLEM

While crime rates in DC have steadily decreased over the past 20 years, DC still struggles with relatively high crime rates like most major US cities.

Washington Post article from June 25, 2015:


Public Safety

Homicides up 20 percent in D.C. this year, with nearly 30 killed since May 1

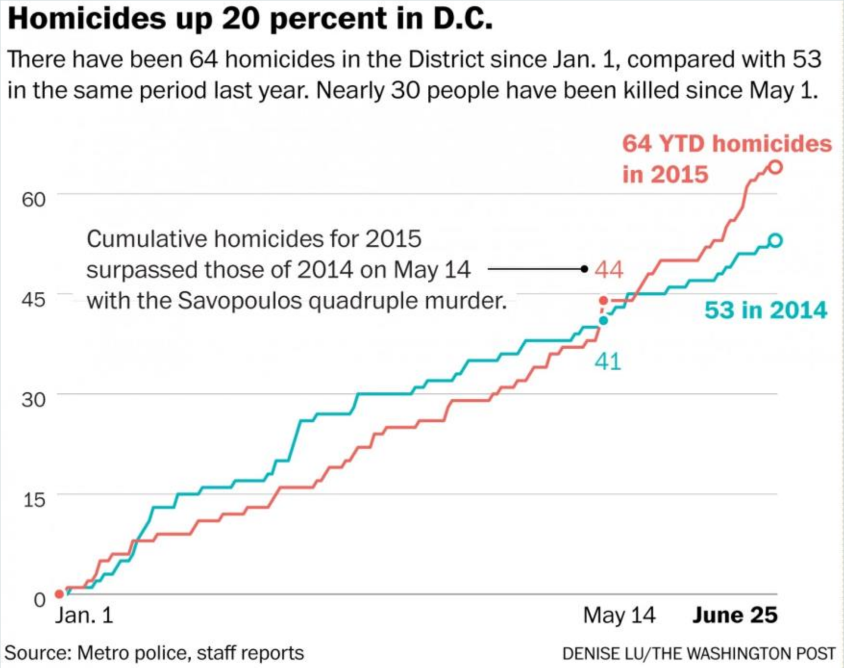
Police investigate at the fire-damaged home in Northwest Washington where 46-year-old Savvas Savopoulos, his 47-year-old wife, Amy Savopoulos, the couple's 10-year-old son, Philip, and housekeeper Veralicia Figueroa were found dead. (Jacquelyn Martin/AP)

By Peter Hermann and Keith L. Alexander June 25

Most Read Local

1 Hard feelings over school boundaries prompt an alleged assault in Virginia

2 How this stressed-out working mom calmed her chaotic life



SOURCE: http://www.washingtonpost.com/local/crime/dc-homicides-jump-20-percent-amid-a-surge-in-may-june/2015/06/25/2effd93c-1b37-11e5-93b7-5eddc056ad8a_story.html

QUESTION

- Given a crime committed, can I predict whether or not it is a violent crime or nonviolent crime?
  - Particularly, with a focus on location in DC (e.g. neighborhood), and time of year (e.g g season and months)
  - Violent crime includes: homicide, assault, robbery, sexual abuse

- Data sources:
  - DC Crime data for 2014
  - DC Neighborhood data
  - DC Weather for 2014

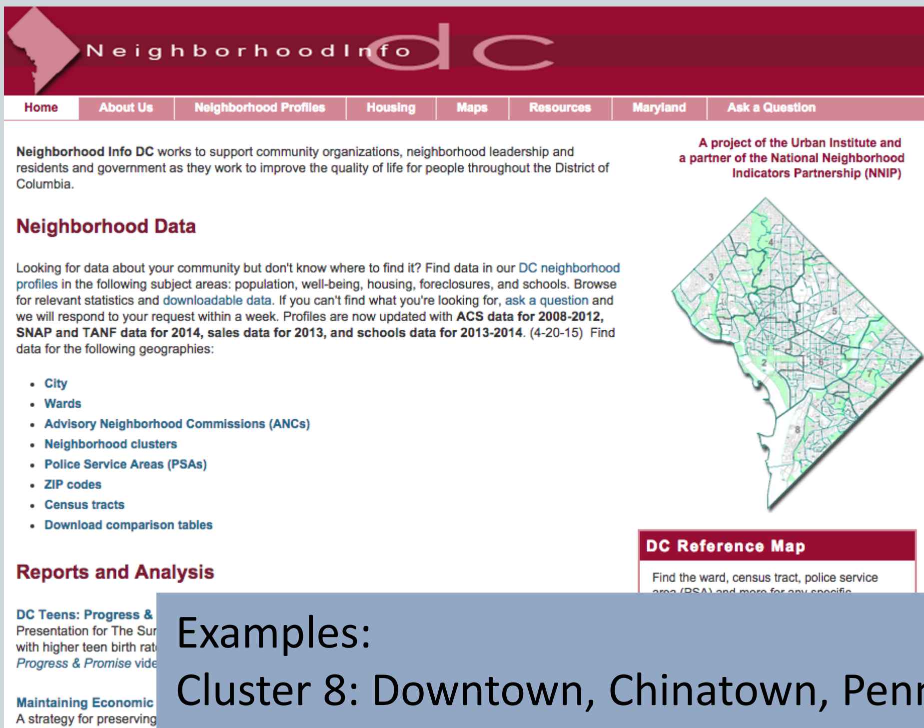
DATA SOURCES

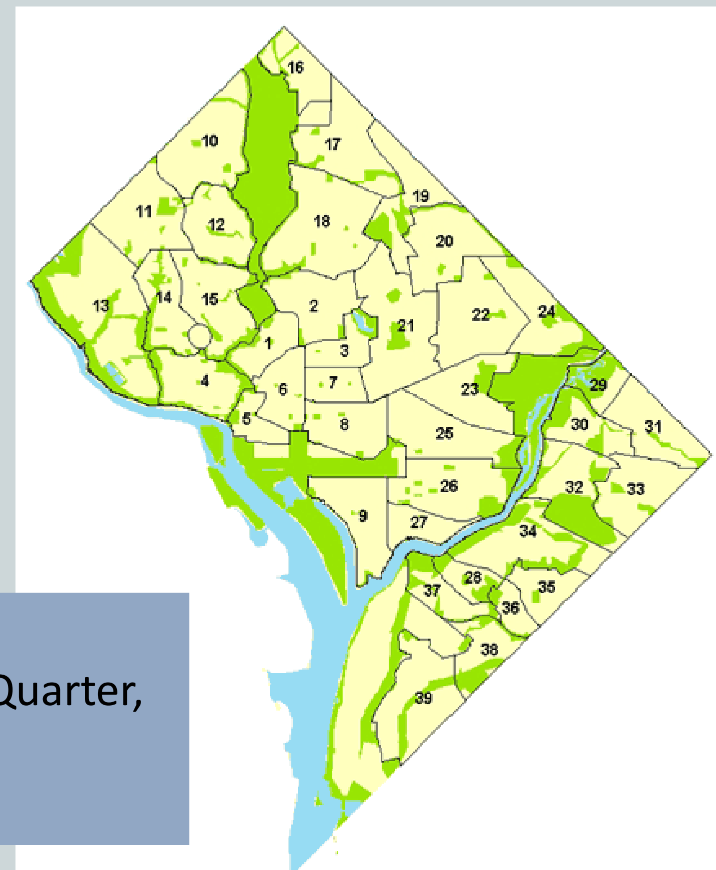
- 2014 DC Crime data
  - Source: DC Government Open Data portal
  - Shape: (38388, 21)
  - Rows: Crime incidents for all of 2014
  - 21 Columns:

1. ccn
2. reportdatetime
3. policeshift
4. offense
5. method
6. lastmodifieddate
7. blocksiteaddress
8. blockxcoord
9. blockycoord
10. ward
11. anc
12. district
13. psa
14. neighborhoodcluster
15. businessimprovementdistrict
16. block_group
17. census_tract
18. voting_precinct
19. start_date
20. end_date
21. esri_oid

DATA SOURCES

- Neighborhood Data
  - Data is organized by “neighborhood cluster” (groupings of 3-5 neighborhoods throughout DC)





Examples:
Cluster 8: Downtown, Chinatown, Penn Quarter, Mt. Vernon Sq, North Capitol
Cluster 26: Capitol Hill, Lincoln Park

DATA SOURCES

- Neighborhood Data Web Scraper

```python
DC Neighborhood Cluster Web-Scraping
from https://neighborhoodinfodc.org
'''
import pandas as pd
from bs4 import BeautifulSoup
import requests

cluster_id=range(1,40)
#population tab
Pop2010=[]
Bl2010=[]
Wh2010=[]
Hs2010=[]
As2010=[]
#well-being tab
poverty08_12=[]
unemployment08_12=[]
employed08_12=[]
nohsdiploma08_12=[]
avgfamincome08_12=[]
foodstamps2014=[]
tanf2014=[]
#chousing tab
medianhomeprice2013=[]

for num in cluster_id:
    r = requests.get('http://neighborhoodinfodc.org/nclusters/Nbr_prof_clus' + str(num) + '.html')
    b = BeautifulSoup(r.text)
    Pop2010.append(b('table')[2].find_all('tr')[6].find_all('td')[1].text)
    Bl2010.append(b('table')[2].find_all('tr')[29].find_all('td')[1].text)
    Wh2010.append(b('table')[2].find_all('tr')[32].find_all('td')[1].text)
    Hs2010.append(b('table')[2].find_all('tr')[35].find_all('td')[1].text)
    As2010.append(b('table')[2].find_all('tr')[38].find_all('td')[1].text)

    r = requests.get('http://neighborhoodinfodc.org/nclusters/Nbr_prof_clusb' + str(num) + '.html')
    b = BeautifulSoup(r.text)
    poverty08_12.append(b('table')[2].find_all('tr')[6].find_all('td')[1].text)
    unemployment08_12.append(b('table')[2].find_all('tr')[17].find_all('td')[1].text)
    employed08_12.append(b('table')[2].find_all('tr')[21].find_all('td')[1].text)
    nohsdiploma08_12.append(b('table')[2].find_all('tr')[26].find_all('td')[1].text)
    avgfamincome08_12.append(b('table')[2].find_all('tr')[36].find_all('td')[1].text)
    foodstamps2014.append(b('table')[2].find_all('tr')[55].find_all('td')[1].text)
    tanf2014.append(b('table')[2].find_all('tr')[71].find_all('td')[1].text)
```

Features Scraped for each neighborhood cluster:

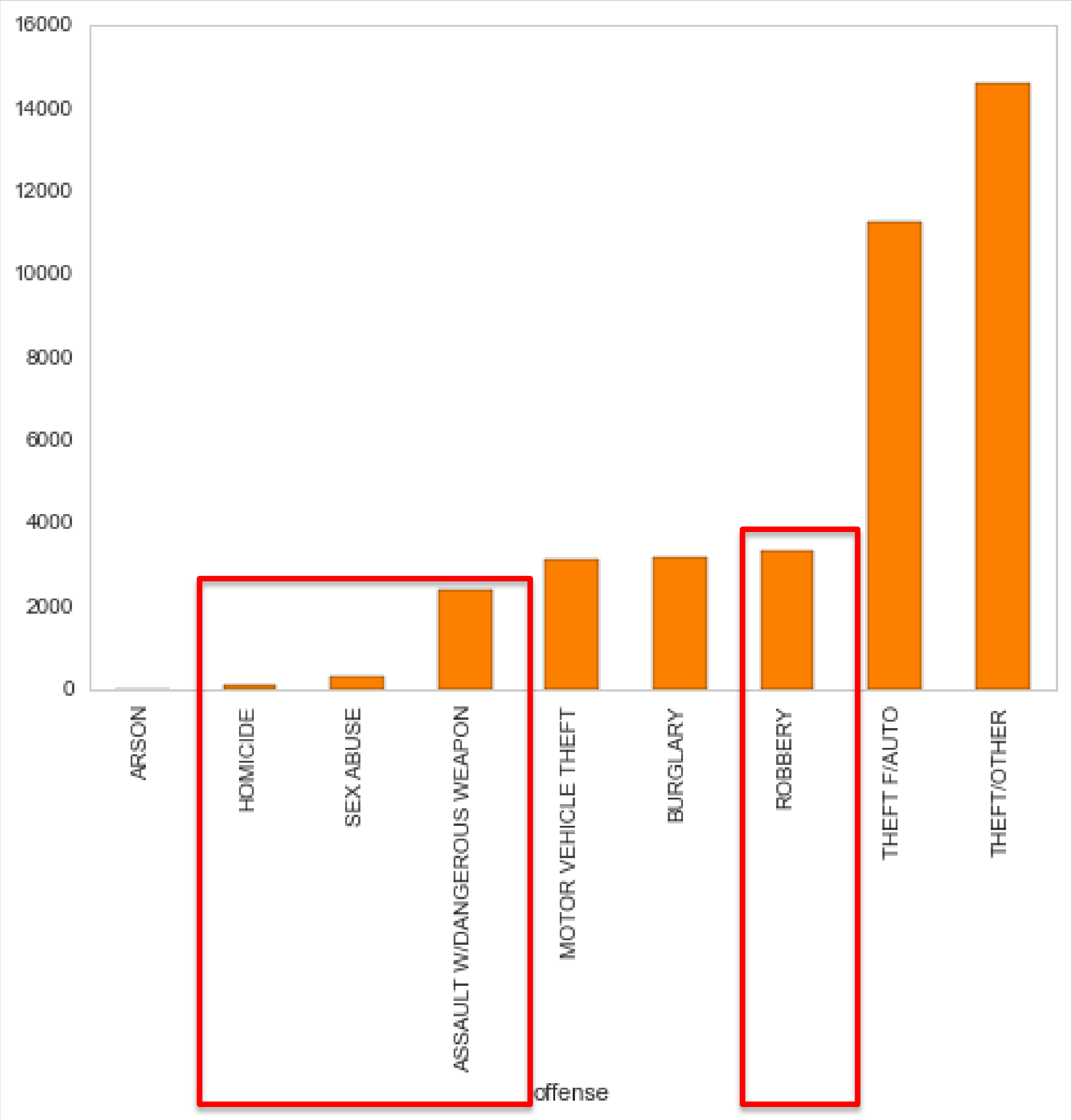
1. population 2010
2. %African American
3. %white
4. %hispanic
5. %asian
6. poverty_rate 2008-2012
7. unemployment rate 2008-2012
8. employment rate 2008-2012
9. no high school diploma 2008-2012
10. average family income 2008-2012
11. % receiving food stamps 2014
12. % receiving TANF 2014

DATA SOURCES

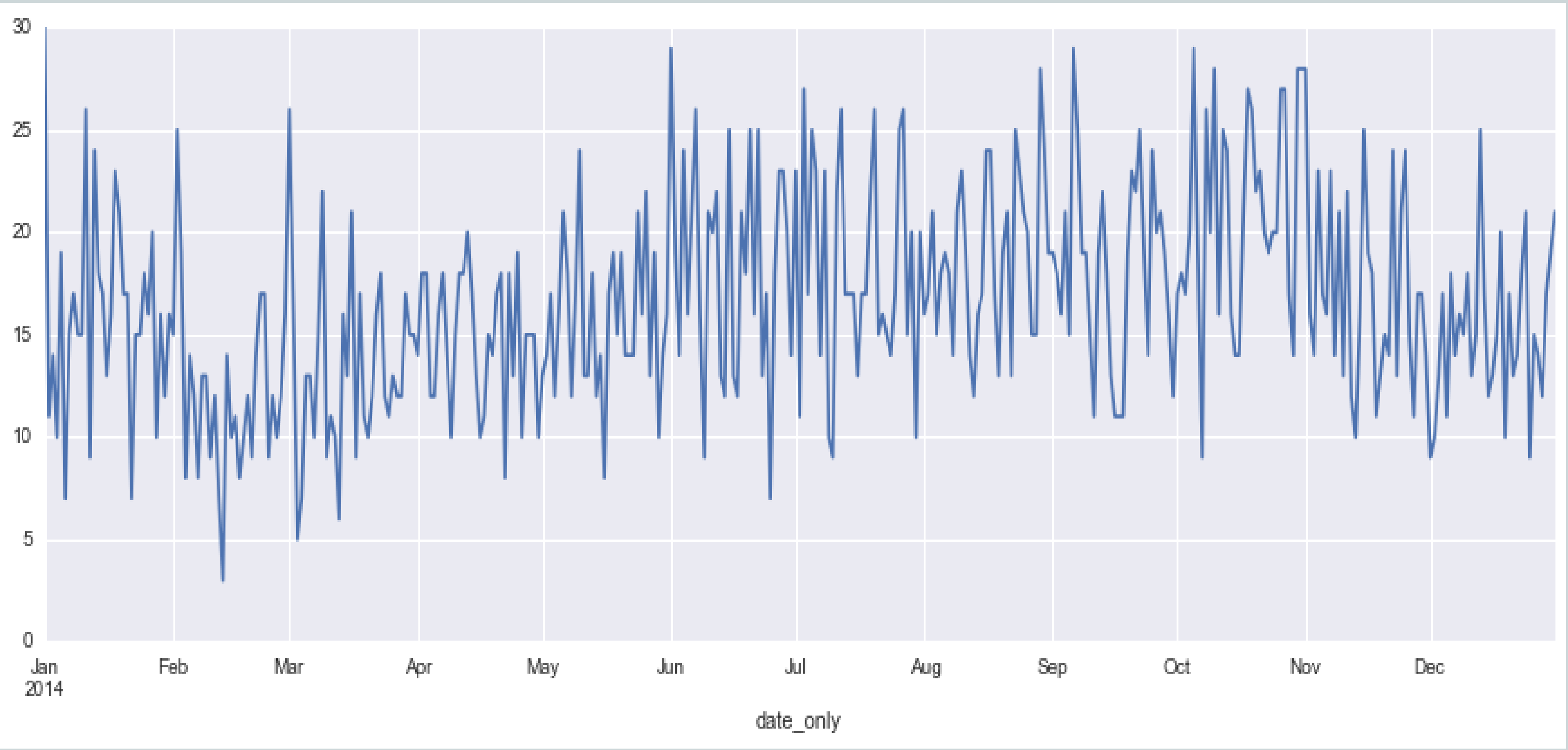
- DC Weather for 2014 from NOAA
- Retrieved csv file of daily temperature recordings for DC weather stations in 2014
  - Reagan National Airport weather station
- Features:
  1. Date
  2. Max Temp
  3. Min Temp
  4. Snow
  5. Rain

DISTRIBUTION OF CRIME TYPES IN DC

- Class imbalance between violent and nonviolent crimes
- Data was downsampled to even the distribution of violent vs. nonviolent crimes
- Dataframe size downsampled to ~12000 rows

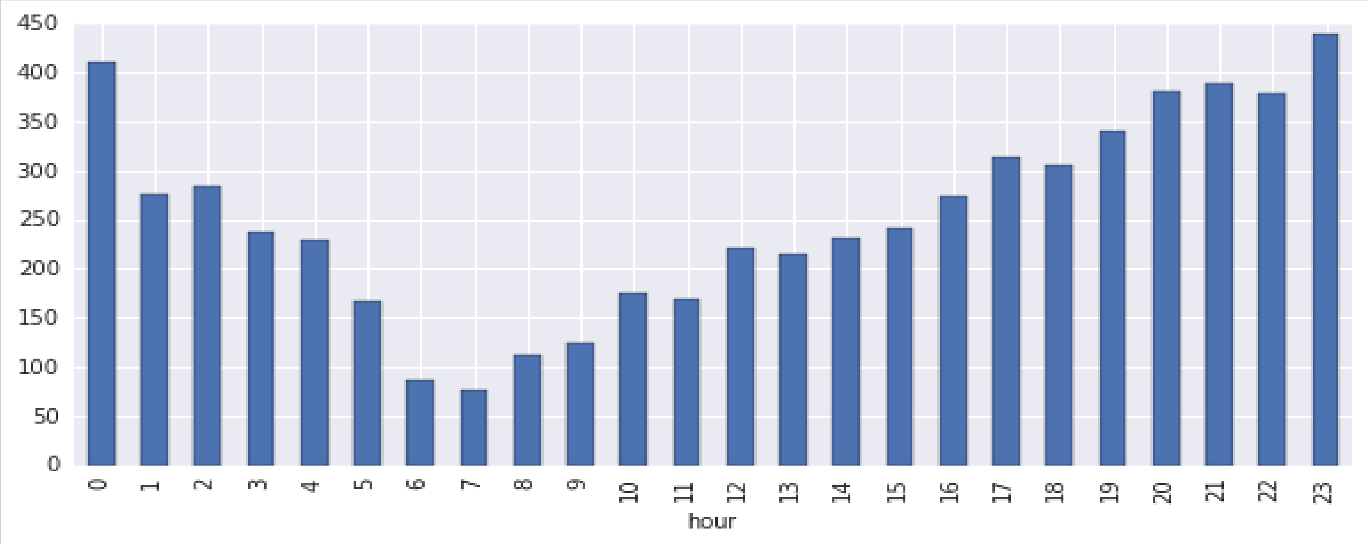


2014 DAILY VIOLENT CRIME COUNT

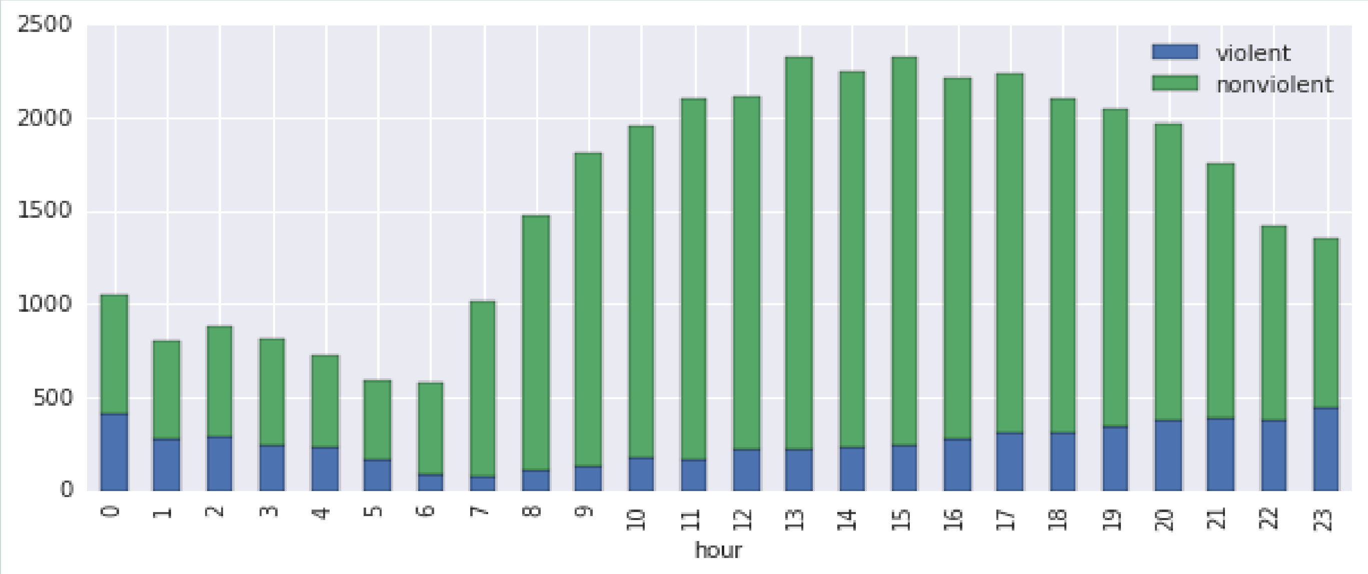


VIOLENT & NONVIOLENT CRIMES BY HR

VIOLENT CRIMES BY HOUR



VIOLENT & NONVIOLENT CRIMES BY HOUR



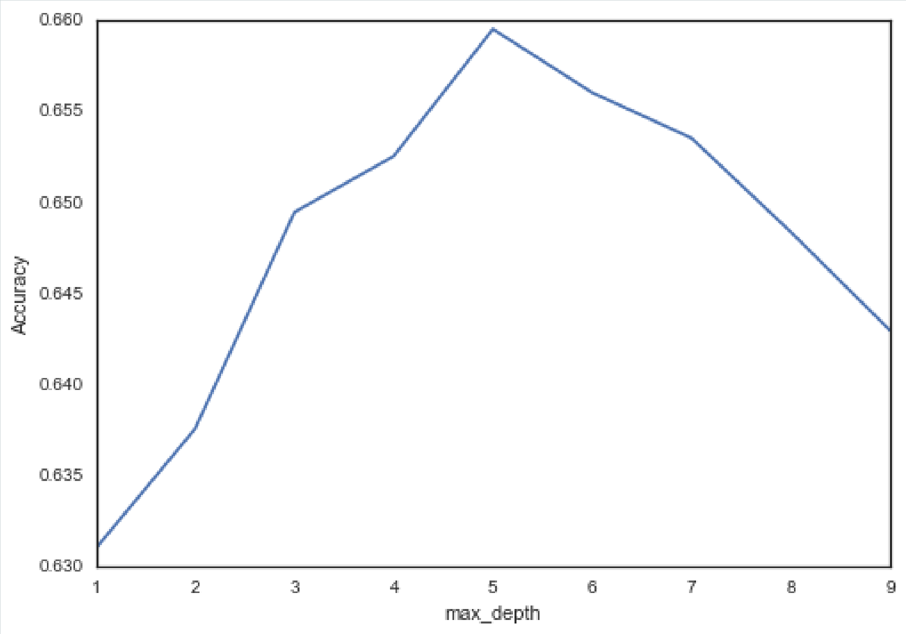
MODELS USED

- Decision Trees
- Random Forests
- Logistic Regression

- Use DTs to determine most important features; use those features for logistic regression model
- Null hypothesis: 0.501

MODEL #1: DECISION TREES

- Decision Tree Classifier – looped through max depth ranges using 11 possible features
  - Day/Night
  - Median home prices
  - Census tract
  - Poverty rate
  - Rain
  - Snow
  - People employed over 16
  - Weekend
  - Month
  - Avgfamily income
  - Unemployment rate
  - % pop no hs diploma



- I had a lot of features that are highly correlated (e.g. poverty rate and % of population receiving food stamps); I only picked subsets of these features and then wanted to see how DTs handled the relationship between different variables

MODEL #1: DECISION TREES

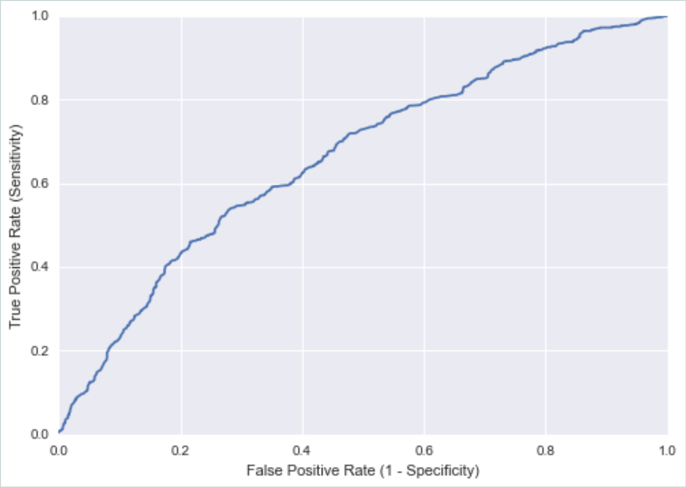
- Decision Tree with max depth 5
- Observations:
  - Tree split in places I would have expected (e.g. time of day and day of the week)
  - Top 5 features:
    - Avg family income
    - Time of day (day vs night)
    - Census Tract
    - Weekend
    - Unemployment rate
- Pitfalls:
  - Since DTs make locally optimal splits, it could still leave in features that are globally correlated (e.g. avg family income and unemployment rate)
  - Gini scores are fairly high (above 0.45) for many of the nodes; thus nodes are less “pure”
  - Consistent with the best possible accuracy score just under 0.66

MODEL #2: RANDOM FORESTS

- Using the same features used for DTs, my next model used was Random Forests
  - # of estimators: 150
  - Max features: 5
- Out of bag error: 0.621, better than the null but not by much
- Different ordering of feature importance!
  - Month
  - Census Tract
  - Precipitation
  - Avg family income
  - Day vs night
- OOB is lower than accuracy score for DTs
- ‘Month’ was not chosen in DTs

MODEL #3a: LOGISTIC REGRESSION

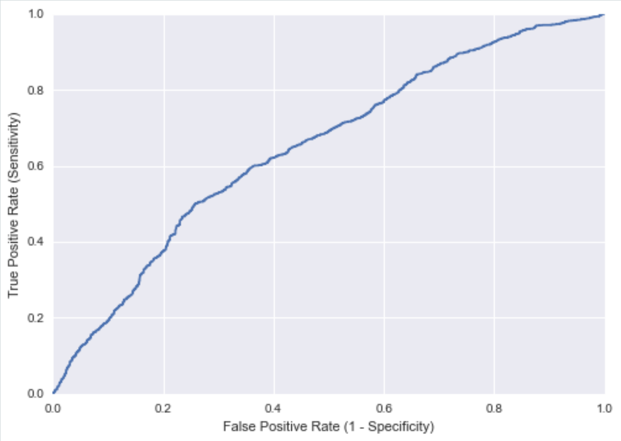
- First model using top features determined by DTs
  - Avg family income
  - Time of day (day vs night)
  - Census Tract
  - Weekend
  - Unemployment rate
- Train, test, split accuracy score: 0.6115
- Cross val score (mean): 0.6230
- AUC: 0.6617, ROC:



	Predicted No	Predicted Yes
Actual No	1058	455
Actual Yes	689	836
Sensitivity: 836/(689+836) = 0.54		
Specificity: 1058/(1058+455) = 0.69		

MODEL #3b: LOGISTIC REGRESSION

- Second model using top features determined by random forests
  - Month
  - Time of day (day vs night)
  - Census Tract
  - Avg family income
  - Rain
- Train, test, split accuracy score: 0.6119
- Cross val score (mean): 0.6167
- AUC: 0.6488, ROC:



	Predicted No	**Predicted Yes**
Actual No	950	582
Actual Yes	599	907
Sensitivity: 907/(599+907) = 0.60		
Specificity: 950/(950+582) = 0.62		

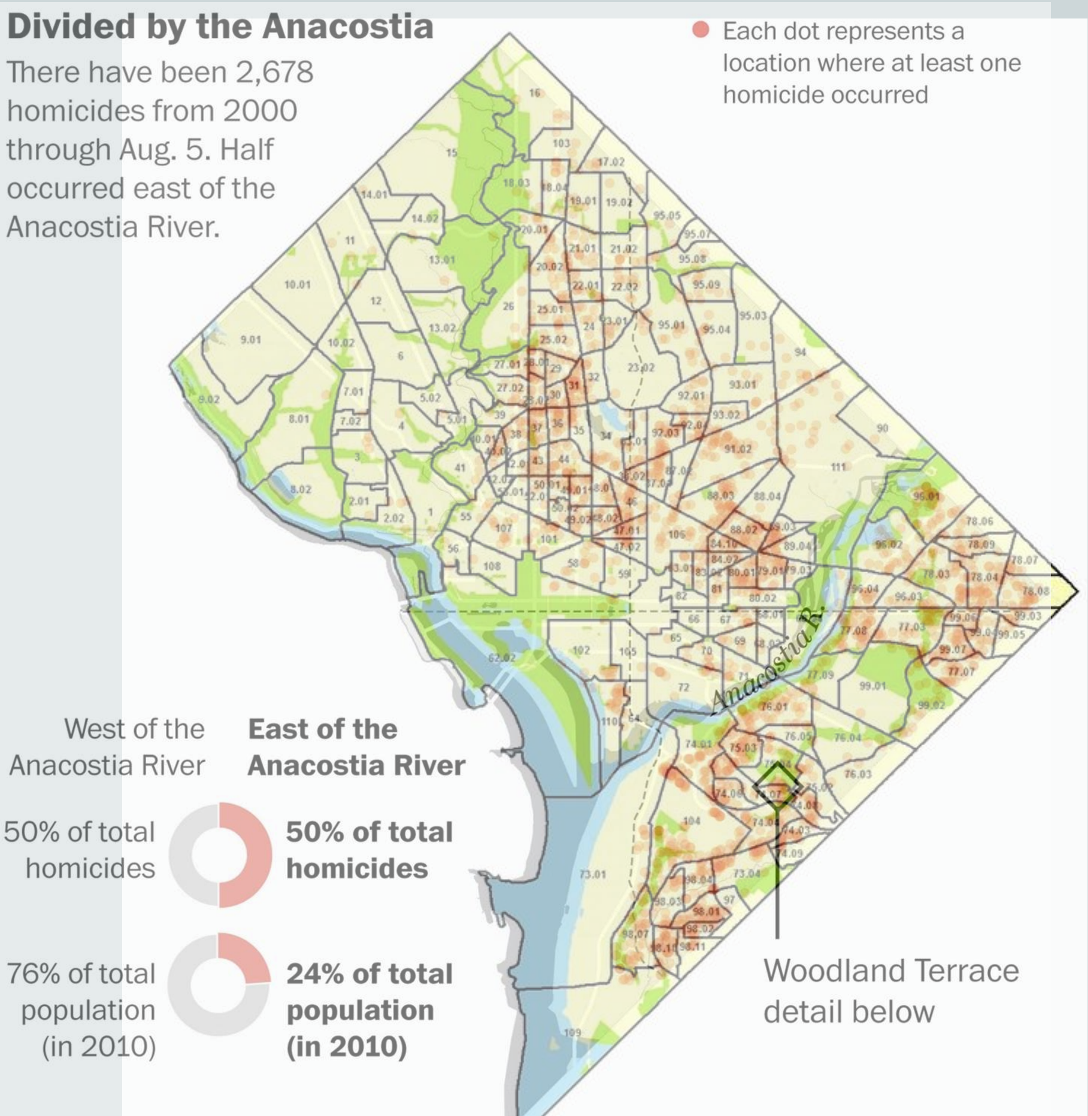
TAKEAWAYS/NEXT STEPS

- Grand lesson: more opportunities for better FEATURE COLLECTION AND ENGINEERING!
  - Lots of features…not a whole lot of which were useful
  - Gathered too many features that were correlated
  - Gathered features that were possibly at the wrong ‘scale’, geographically (e.g. neighborhood cluster vs. census tract)
  - Probably could have explored the features I had a bit more

- Next steps:
  - Continue to explore the features I have
    - Use dummy variables for neighborhood clusters or census tract
  - Explore geopandas/how to use geospatial data better

NOTE ON 'CENSUS_TRACT'

- Census tract is currently represented in the data as a number (e.g. 1 through 111)
- There is no ordinal value to the numbering, but the numbers correspond generally to the geographic quadrants in DC and homicide rates
- Census tracts numbered above 75 contain all of the areas in DC that have the highest homicide rates
- Probably should have created dummy variables with this feature



Homicide graphic source:http://www.washingtonpost.com/local/crime/a-deadly-dc-community-that-has-yet-to-improve/2015/08/06/7abf3f60-3c94-11e5-b3ac-8a79bc44e5e2_graphic.html