The San Francisco Crime Dataset

Authors: Ryan Rosiak and Grant Dawson

The Dataset

Dataset Description:

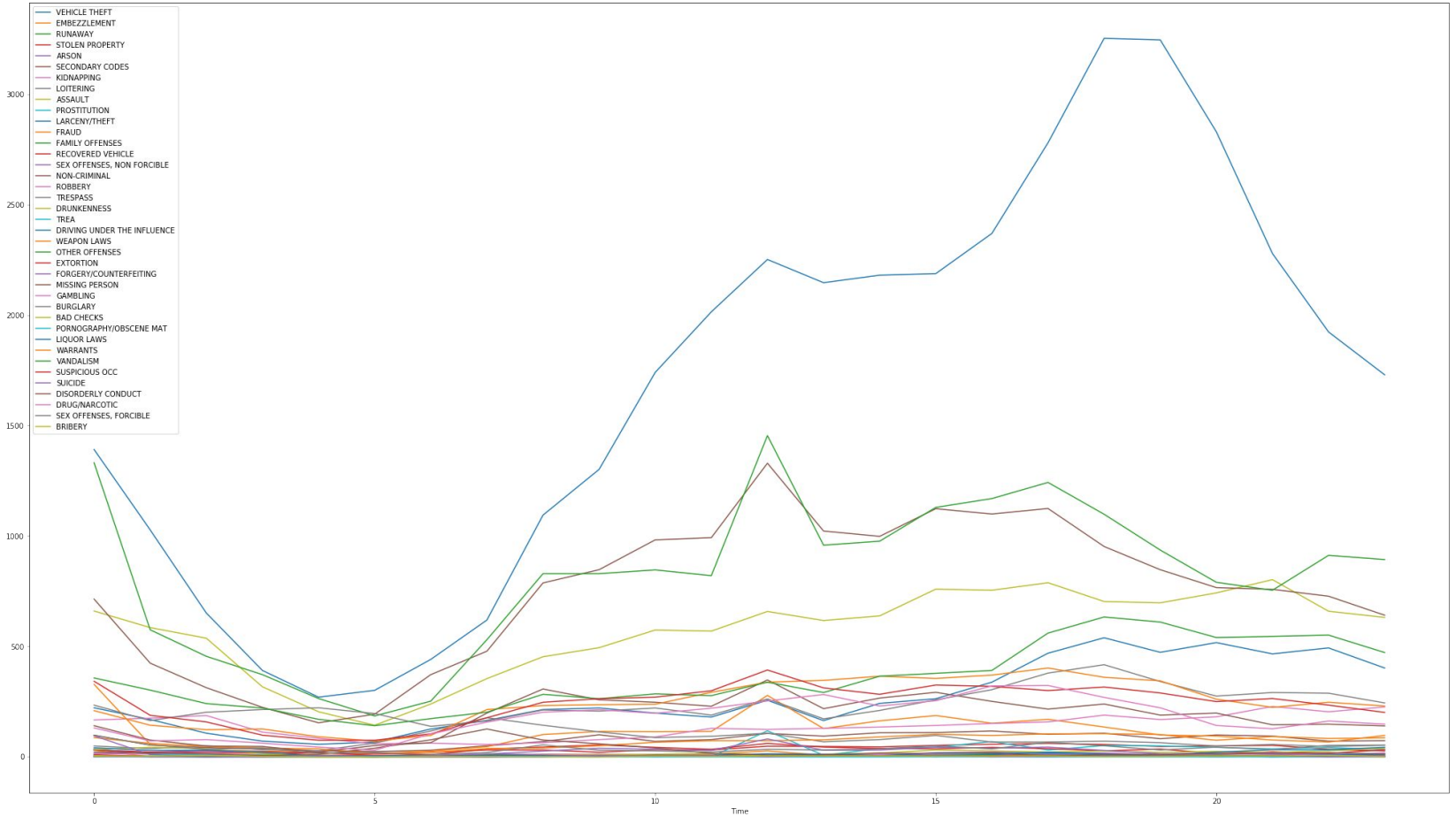
Main Attributes: Incident Number, Category, Description, DayOfWeek, Date, Time, PdDistrict, Resolution, Address, X, Y, Location, PdId

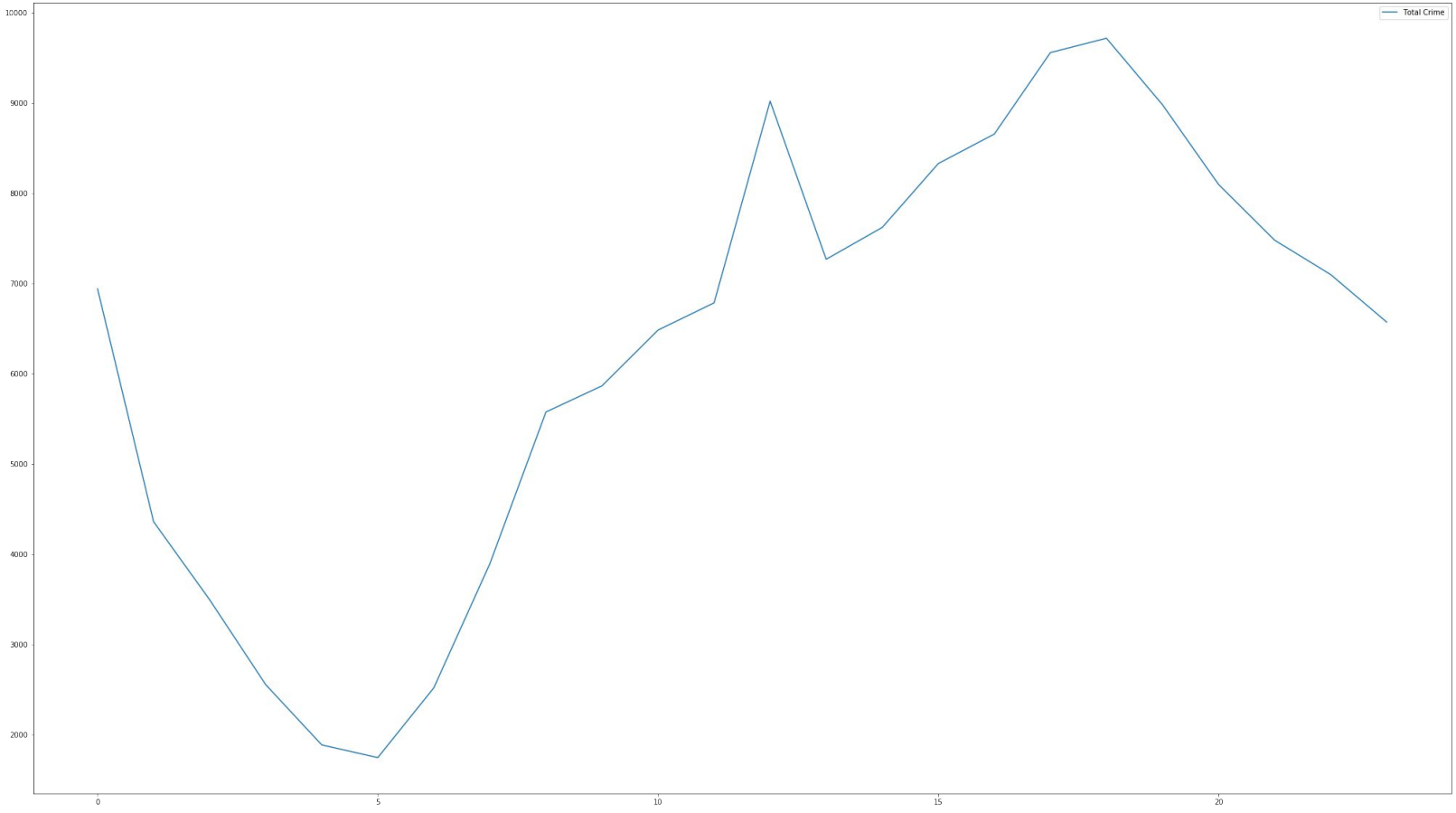
Why did we choose this dataset?

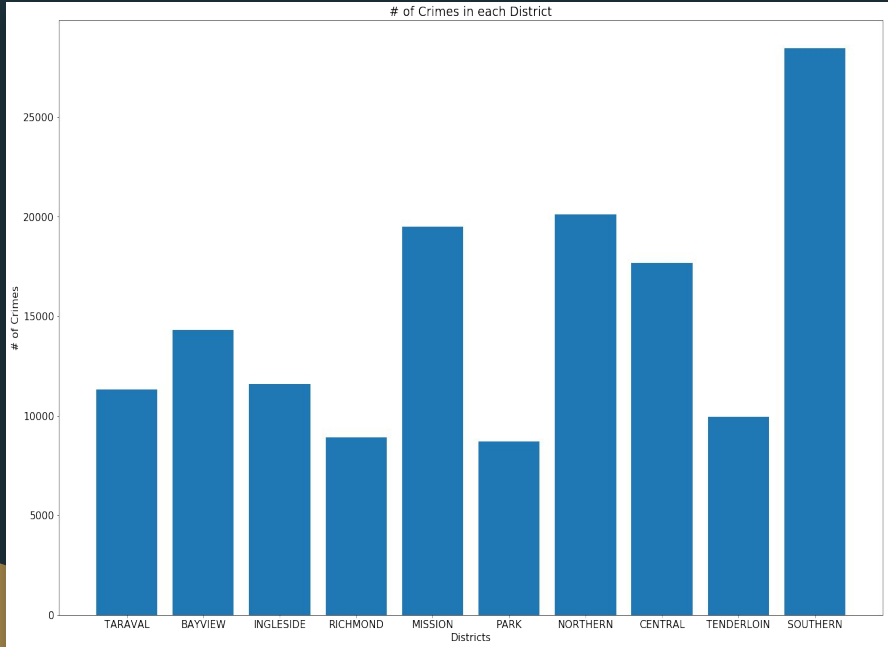
What does this dataset have to offer?

What are some of the obvious correlations we can make?

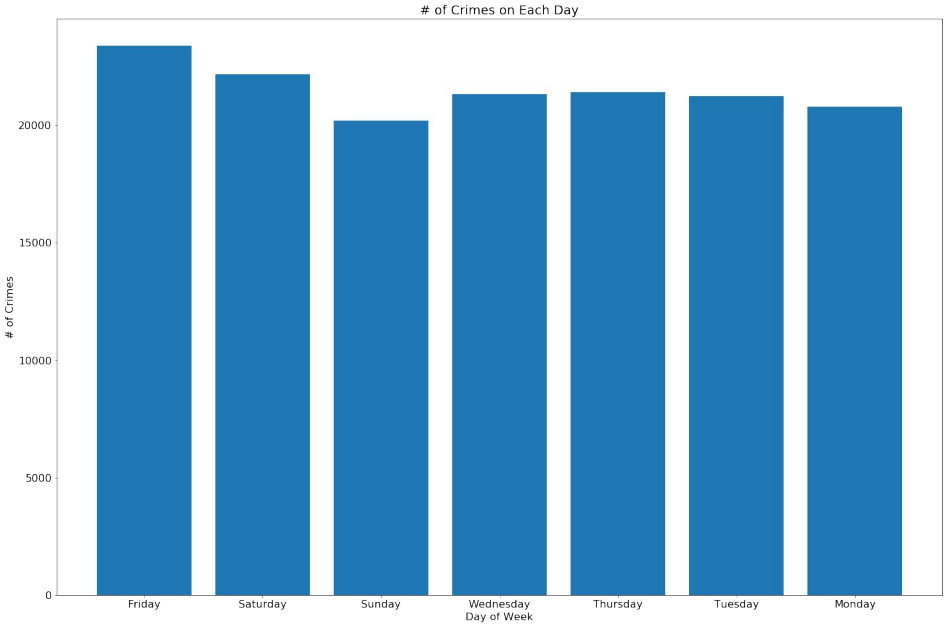
How can what we plan to do translate to the real-world?


VEHICLE THEFT
EMBEZZLEMENT
RUNAWAY
STOLEN PROPERTY
ARSON
SECONDARY CODES
KIDNAPPING
LOITERING
ASSAULT
PROSTITUTION
LARCENY/THEFT
FRAUD
FAMILY OFFENSES
RECOVERED VEHICLE
SEX OFFENSES, NON FORCIBLE
NON-CRIMINAL
ROBBERY
TRESPASS
DRUNKENNESS
TREA
DRIVING UNDER THE INFLUENCE
WEAPON LAWS
OTHER OFFENSES
EXTORTION
FORGERY/COUNTERFEITING
MISSING PERSON
GAMBLING
BURGLARY
BAD CHECKS
PORNOGRAPHY/OBSCENE MAT
LIQUOR LAWS
WARRANTS
VANDALISM
SUSPICIOUS OCC
SUICIDE
DISORDERLY CONDUCT
DRUG/NARCOTIC
SEX OFFENSES, FORCIBLE
BRIBERY
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
Time

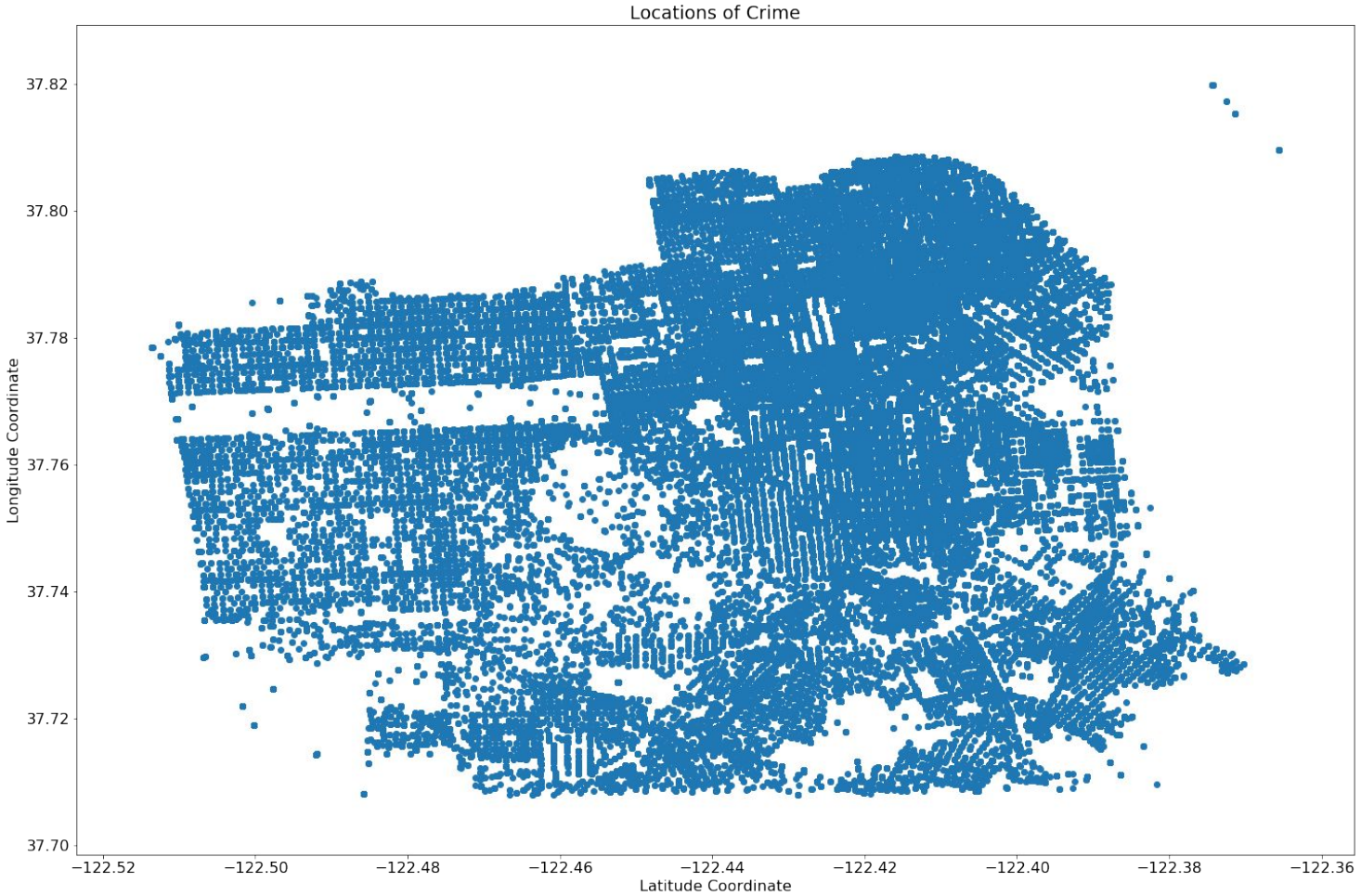

Total Crime
10000
9000
8000
7000
6000
5000
4000
3000
2000
0
5
10
15
20



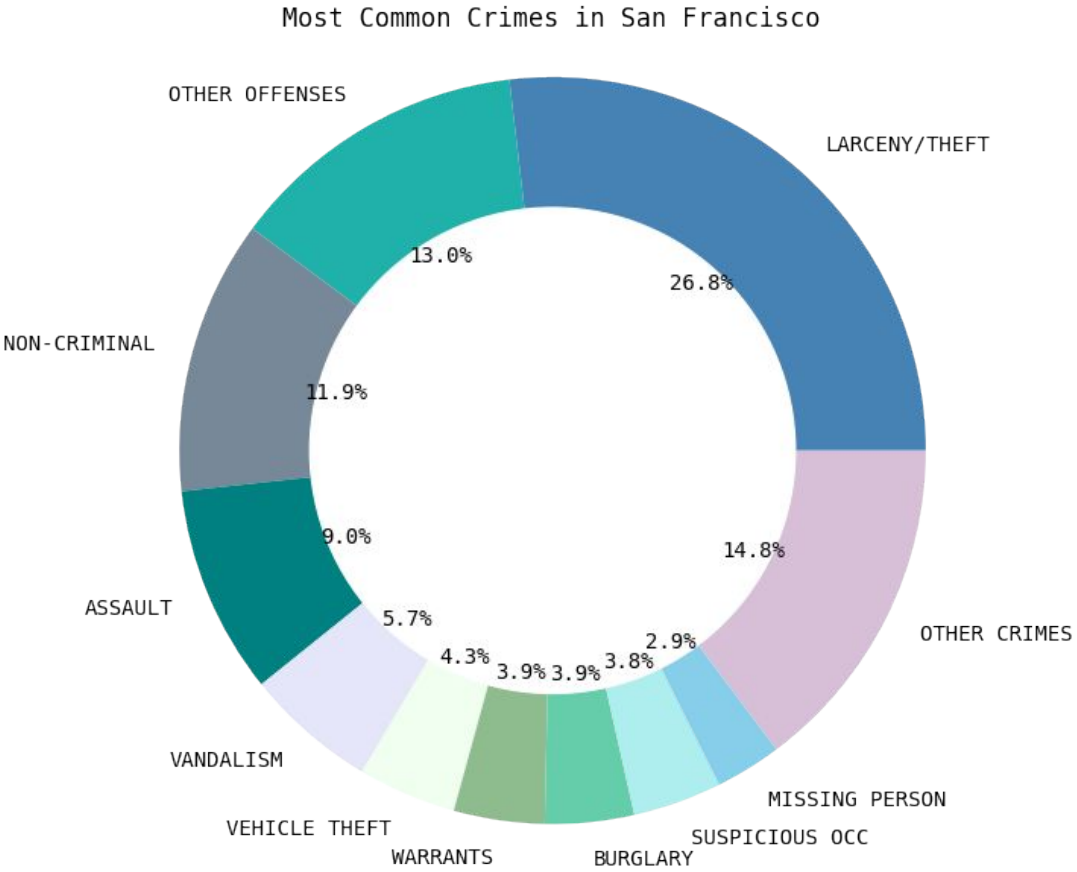
Proportions of Crimes are fairly well dispersed over days of the week. There are no days where crime seems to be super high.



The southern district does not look like the place you would want to live.



Can you guess where the South District is?


Most Common Crimes in San Francisco
OTHER OFFENSES
LARCENY/THEFT
13.0%
26.8%
NON-CRIMINAL
11.9%
9.0%
14.8%
ASSAULT
5.7%
OTHER CRIMES
2.9%
4.3%
3.9%
3.9%
3.8%
VANDALISM
MISSING PERSON
VEHICLE THEFT
SUSPICIOUS OCC
WARRANTS
BURGLARY

ML Algorithm 1: K-Nearest Neighbors

K-Nearest Neighbors Theory

Hypothesis: By using the X and Y coordinates from the dataset, we can use K-Nearest Neighbors to accurately predict what crime was committed at a specific geographical location.

Why only use these attributes?

- Look back at scatter plot - creates an outline of San Francisco
- Theoretically makes sense
- Certain districts may be home to certain crimes (poor vs. rich)
- Classifying numerical points (match made in heaven! KNN <3)

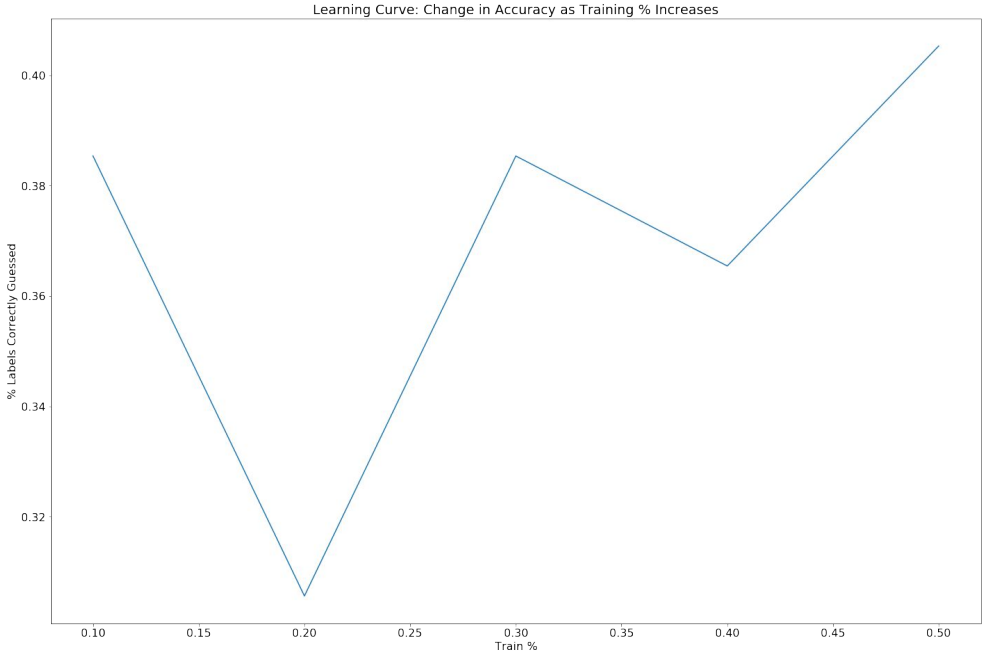
How we trained

We wanted to pick a baseline training value so that we could efficiently analyze multiple ‘k’ values and see how well KNN could pick up on what crimes were where.

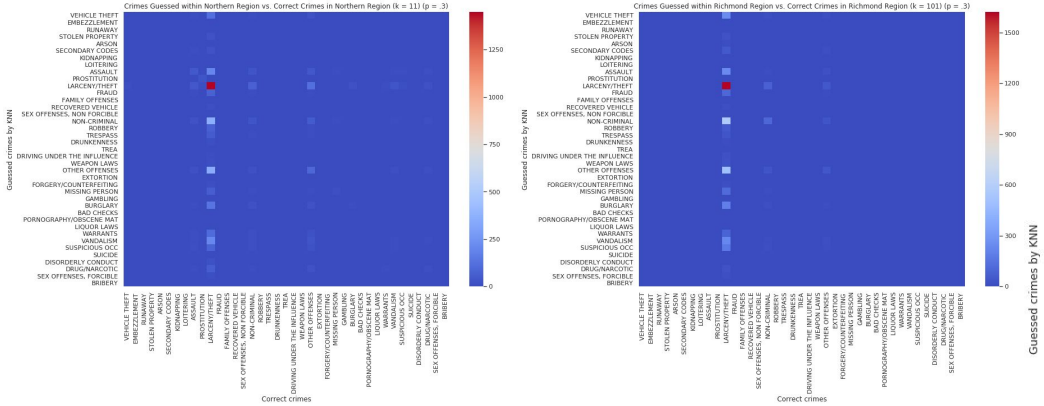
We also decided to separate by districts, to hopefully cancel out a level of ambiguity in location.

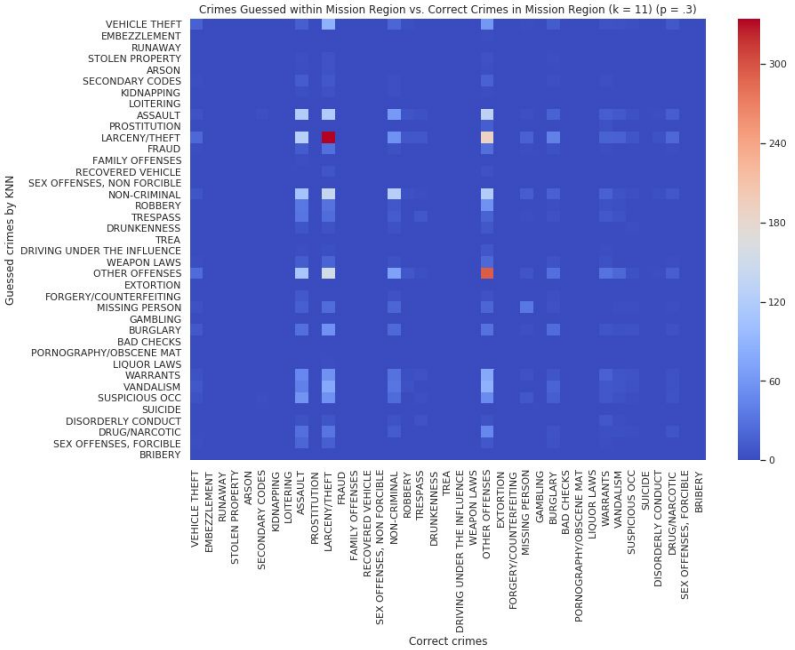
We also needed to be able to run tests in a reasonable amount of time.

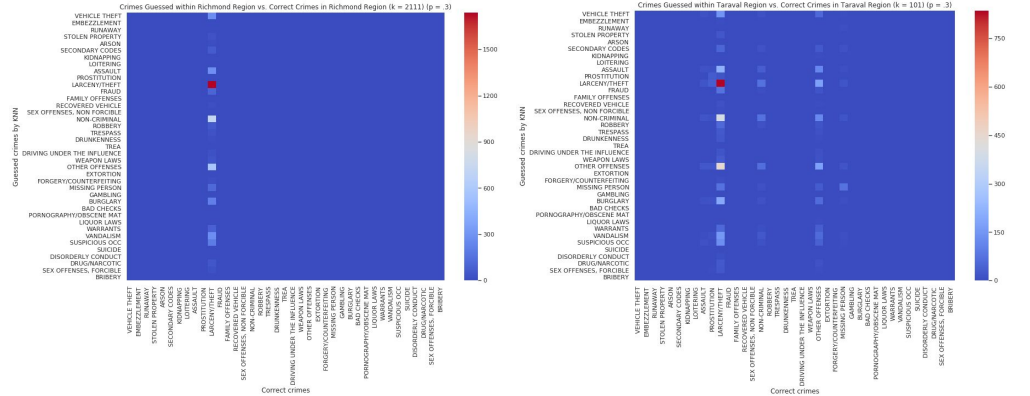
- A test on the full sample of data was done analyzing the learning of KNN over various amounts of training given



Time to show some results... HEAT MAPS!



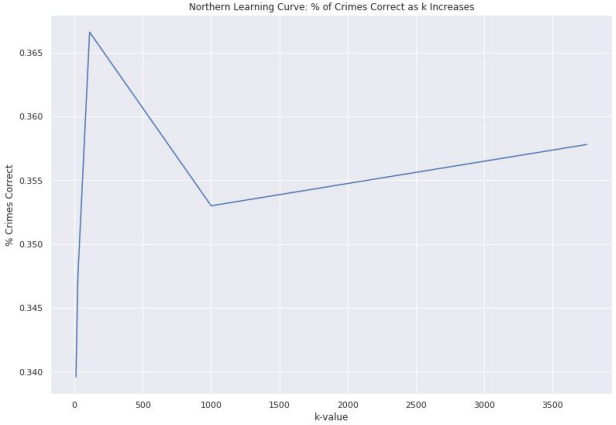


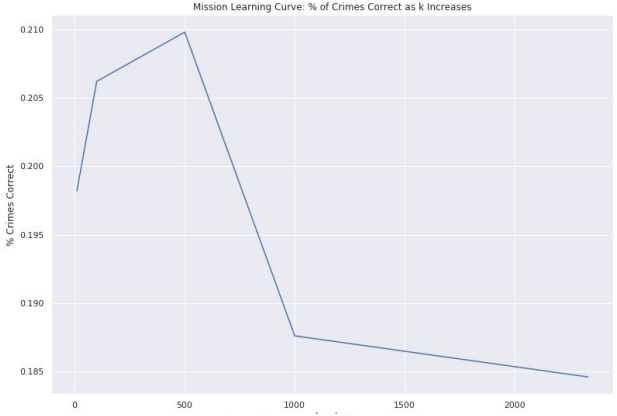


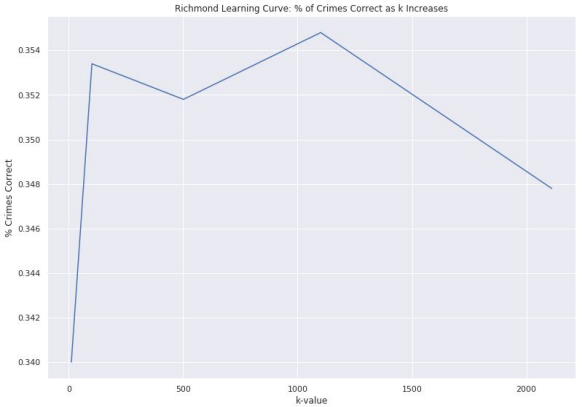
Over 50 heat maps were generated, here are just a few cases.

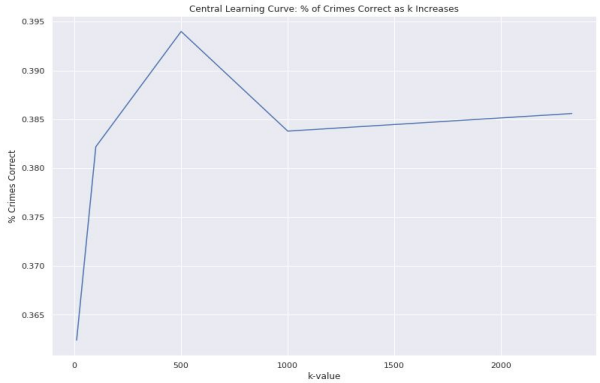
What did we learn?

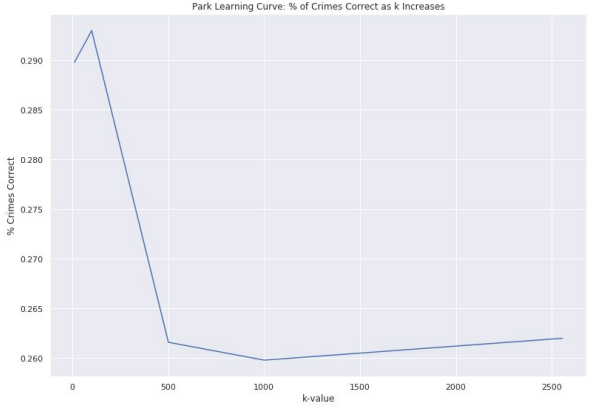
Once again, a few of many!

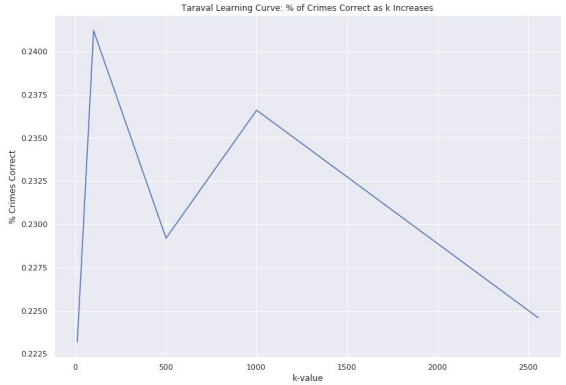












Conclusions

- Results are generally on the unreliable side (many learning curves expressed inconsistent learning as k changed)
- Heat maps showed a lot of promising data
- Got lucky because of the overwhelming amount of certain crimes
- KNN needs more time to work (time for training and time for classifying)
- In theory, location based predictions made a lot of sense

Possible Issues:

- There may not be enough data given (ambiguity of X and Y coordinates
- Not enough training. Where is the sweet spot????
- Shear size of dataset and speed of KNN
- Too many points ‘close to’ one another… crimes in the same location?
- Needed to predict more points

Where to go from here?

- Speed up KNN…. parallelize????
- Increase the training
- Increase the k even more
- Use a larger batch of prediction points
- Subset various districts into sub districts
- Add more attributes (tackle the ambiguity)

ML Algorithm 2: Neural Network

Neural Network Theory

Hypothesis: By using relevant data from the dataset, we can use a Neural Network to accurately and reliably predict what crime was committed.

What and why do we use these attributes?

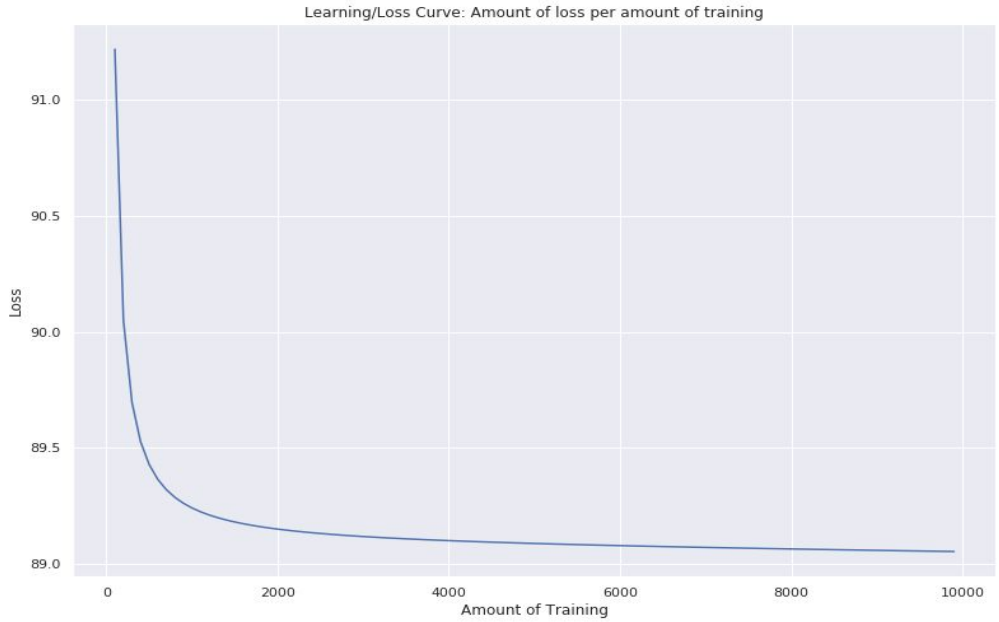
- Time/Date
- Day of Week
- Geological location
- PD District

Neural Network Variations

We made 5 iterations of the neural network algorithm each with different hidden layer sizes and widths.

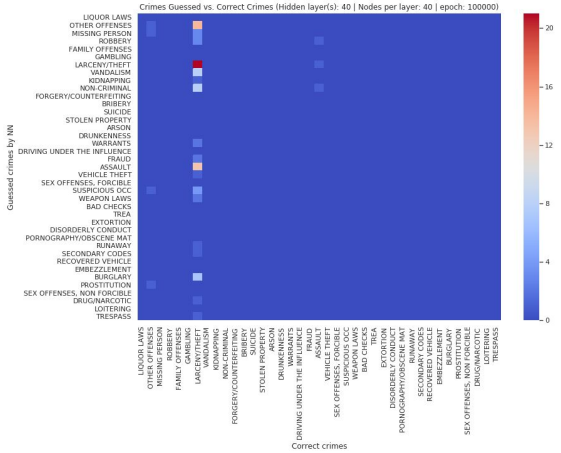
There is so much more to do and test!

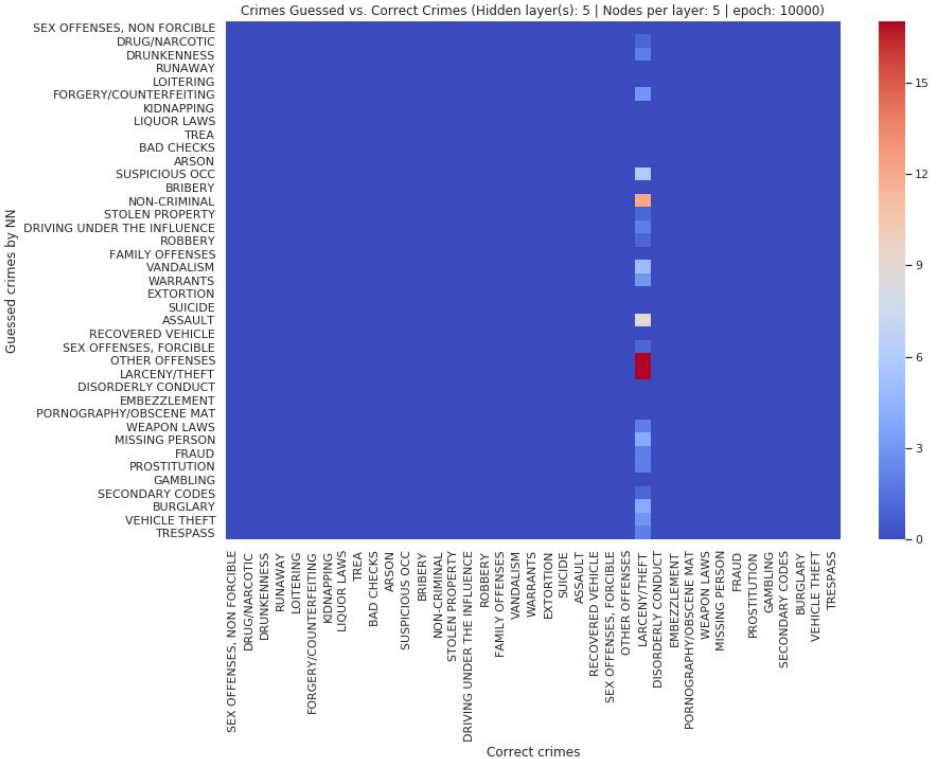
- A sample of a lose over number of iterations of training data being trained on.



I Guess This is Right?

- Unfortunate data
  - Shows a worse case
- Correlations between guesses
- Why is it so bad? Is it bad?





Conclusions

- Results are unreliable (for now)
- Guesses wrong guesses made sense compare too
- NN takes a long time and you need patients
- In theory knowing all the facts about a crime would make sense to make a good guess.

Possible Issues:

- Sweet spots for all settings
- Not enough TIME!!
- The iterations of NN

Improvements:

- Pause button for a neural network?
- Change settings itself
- Jupyter notebooks

Any Questions?