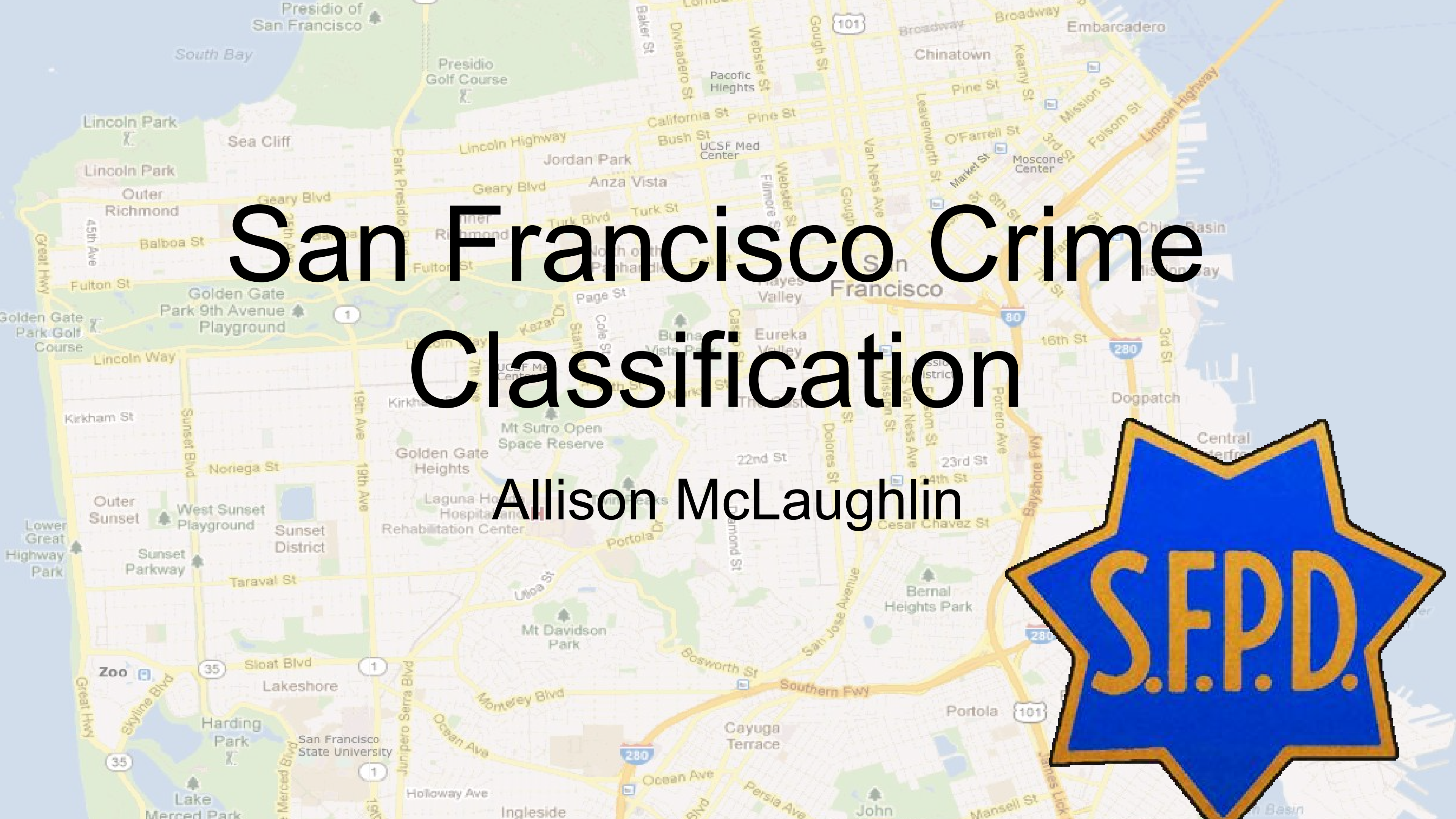
San Francisco Crime Classification

Allison McLaughlin



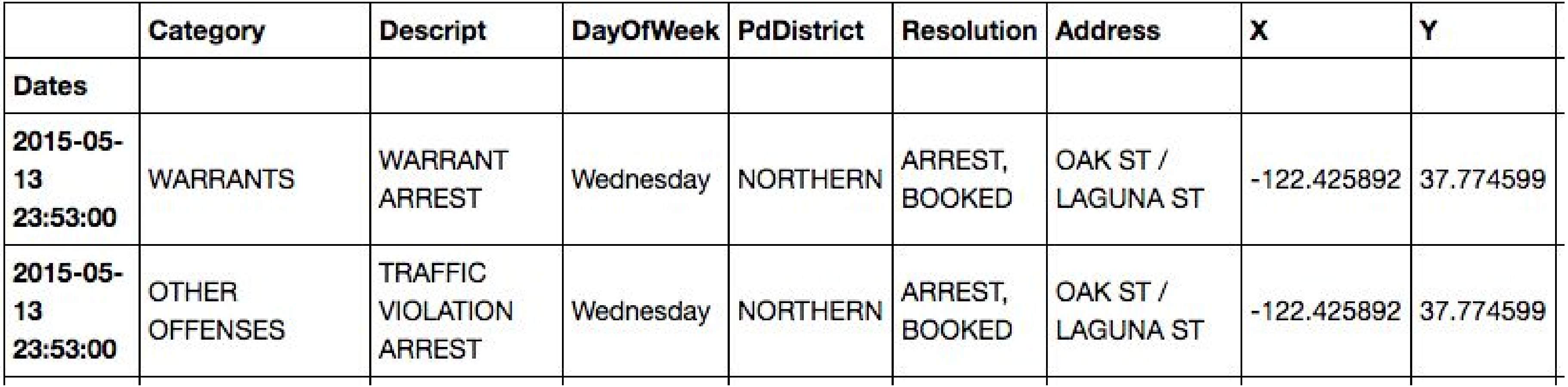
My interest in SF crime data

- Working from home with a broken foot made me the self-declared “best neighborhood watch-person in the city.”

NEIGHBORHOOD
WATCH
PROGRAM IN FORCE
We report all Suspicious Persons & Activities to our Law Enforcement Agency

The Data

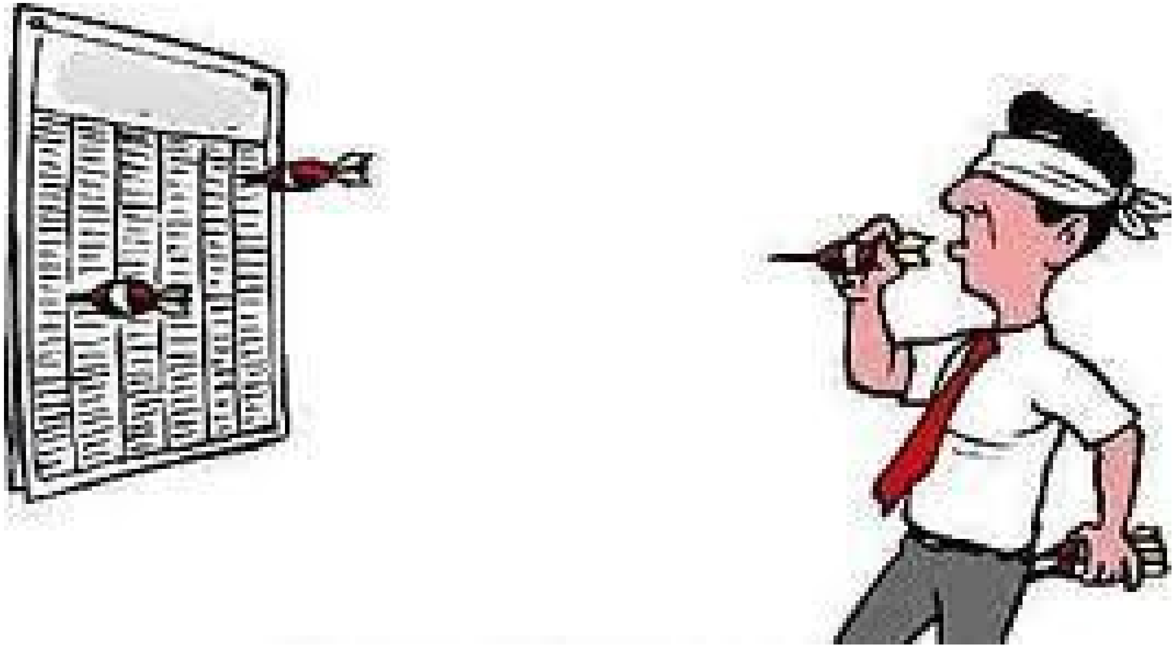
- SFPD incident reports from January 1, 2003 to May 13, 2015.
- Available at data.sfgov.org open data website as a CSV and is currently part of a kaggle competition.
- Goal is to predict the category of crime based on time and location.



	Category	**Descript**	**DayOfWeek**	**PdDistrict**	**Resolution**	**Address**	**X**	**Y**
Dates								
2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599

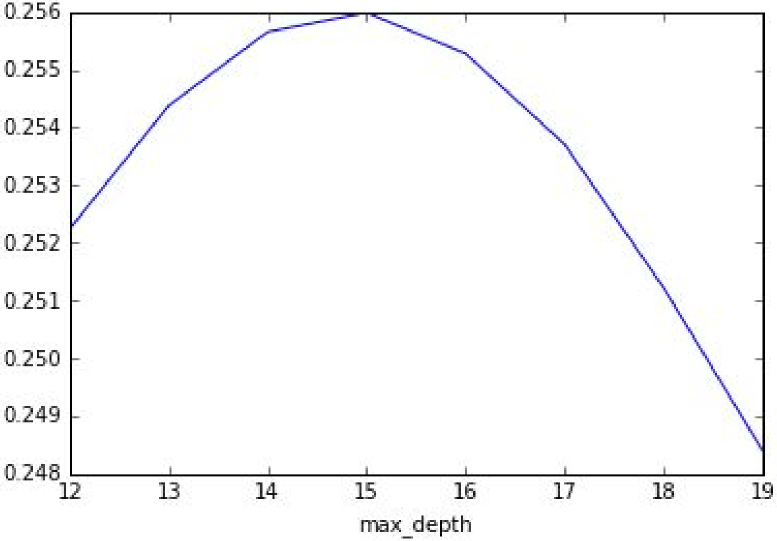
Initial Results

Not good…

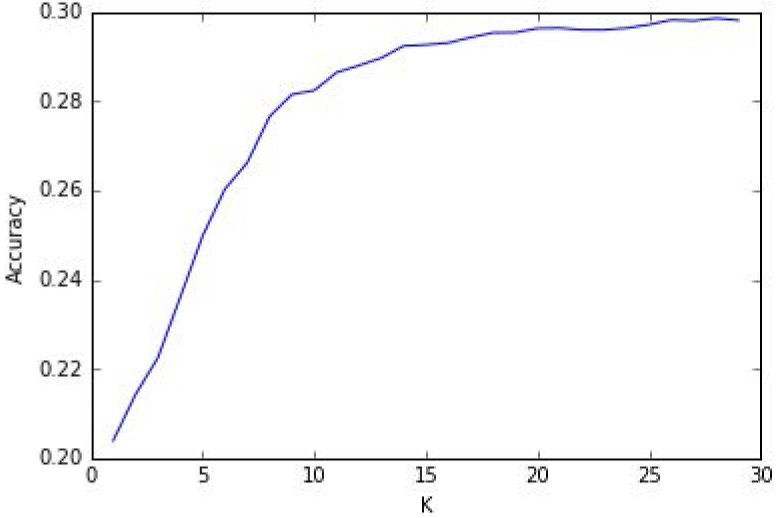


Models were unsuccessful on the entire dataset

- Accuracy when classifying with a decision tree:



- Accuracy using KNN:



.199

Null Accuracy

Why were the results so low?

- There are 39 crime categories with 917 unique subcategories or descriptions.
- There is overlap and inconsistency of classification between categories.
- Each category contains an ENORMOUS range of crimes.
- Location and time alone are not good predictors of human behavior.

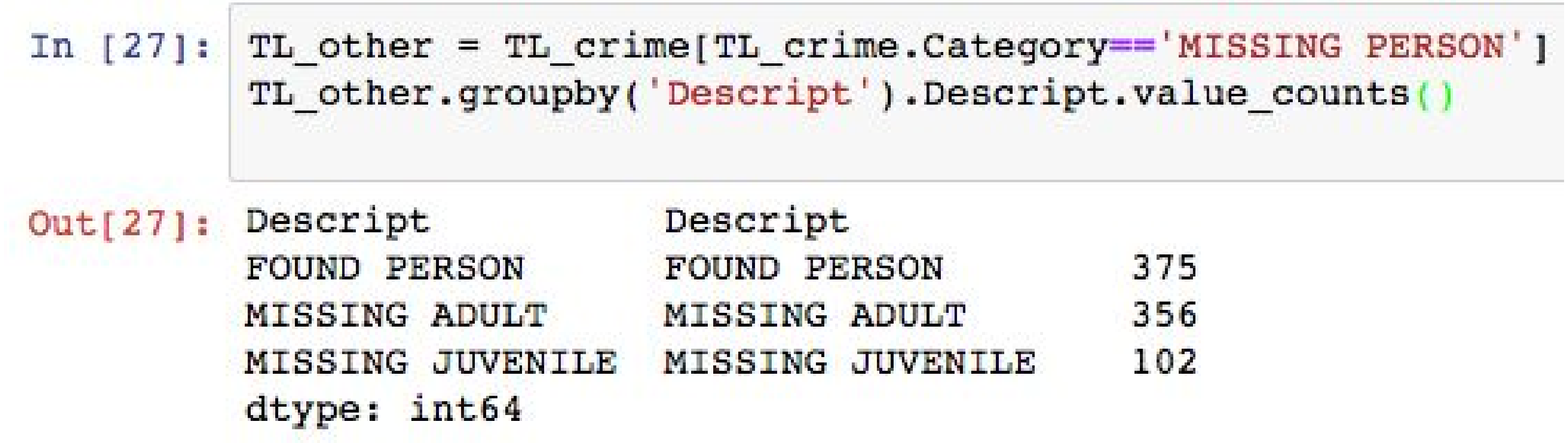
Category = Assault

- Threatening Life
- Threatening Phone Calls
- Stalking
- Inflict Injury on cohabitee
- Aggravated Assault
- Attempted Mayhem with Bodily Force
- Child Abuse
- Hate Crimes

Category = Larceny/Theft

- Attempted theft of a phone booth
- Grand theft of a phone booth
- Grand theft by prostitute
- Trade secrets, theft, or unauthorized copying

Category = Missing Person

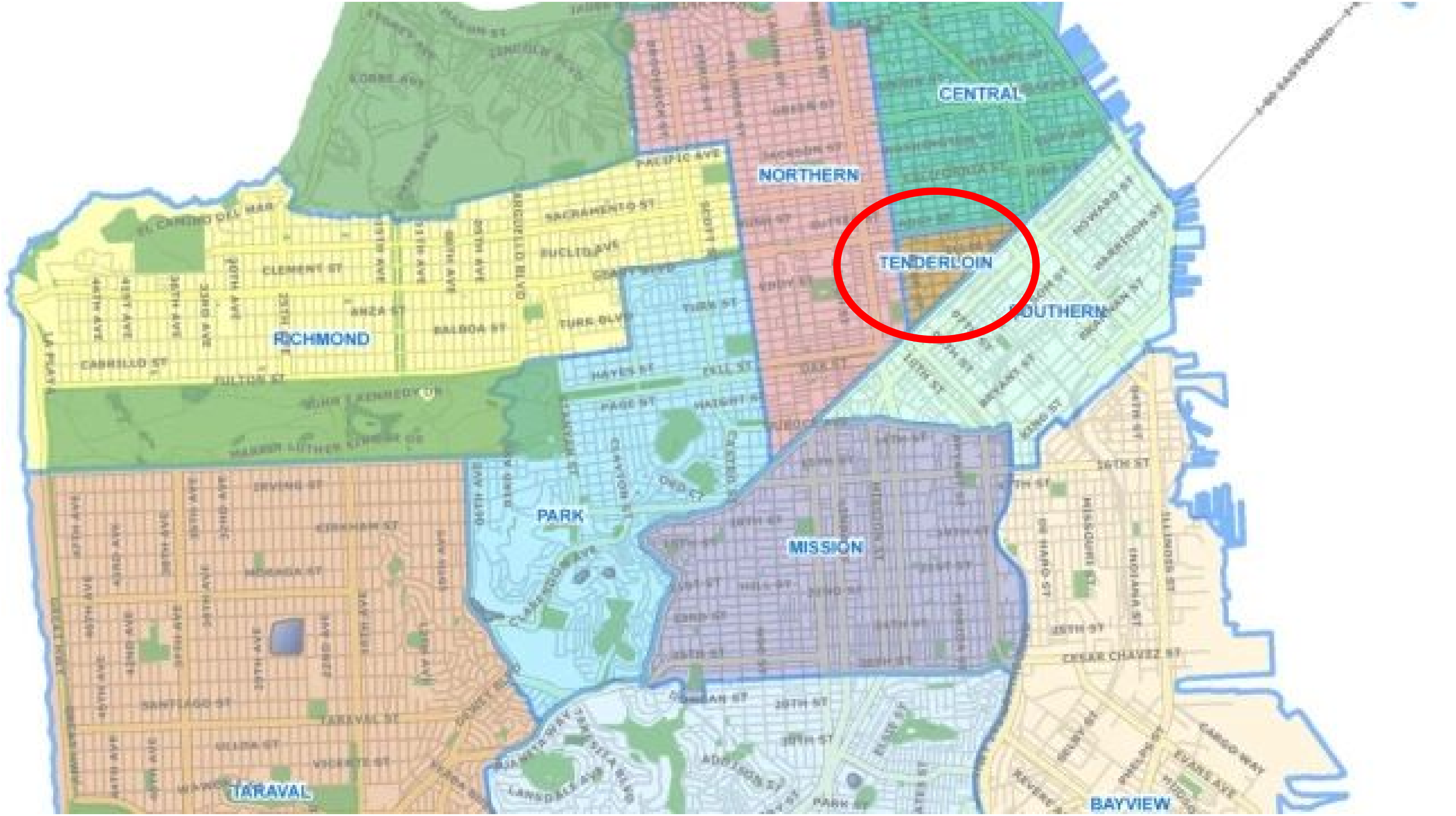


```
In [27]: TL_other = TL_crime[TL_crime.Category=='MISSING PERSON']
         TL_other.groupby('Descript').Descript.value_counts()
```

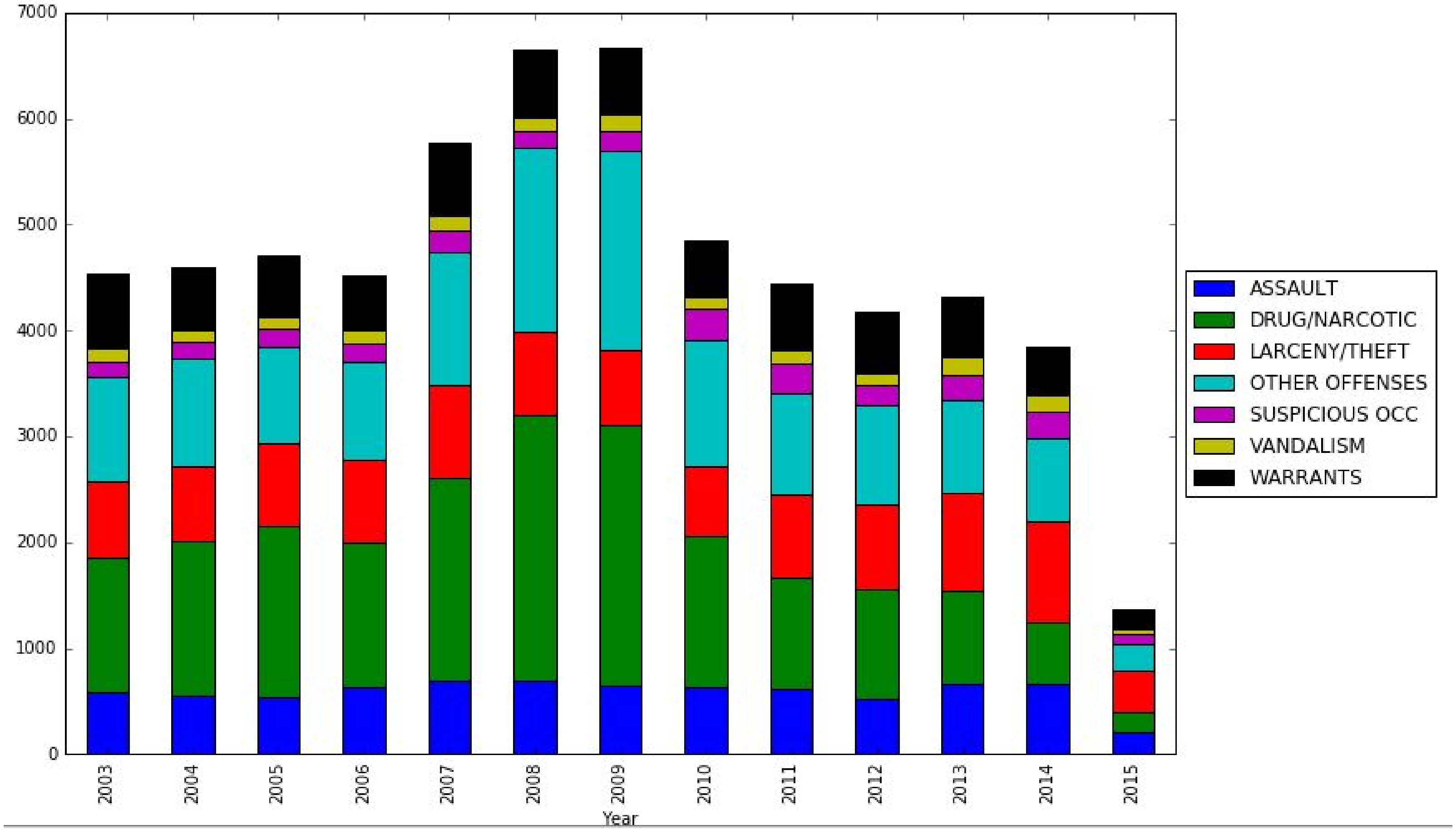
```
Out[27]: Descript          Descript
         FOUND PERSON      FOUND PERSON        375
         MISSING ADULT     MISSING ADULT       356
         MISSING JUVENILE  MISSING JUVENILE    102
         dtype: int64
```

These categories make predictions across the whole city pretty tough.

What about one police district?



Crime in the TL



Some Domain Knowledge

If I hit ***Turk*** Street, then ***Hyde*** and ***Leavenworth*** get neglected. I have to be very surgical in how I do enforcement.

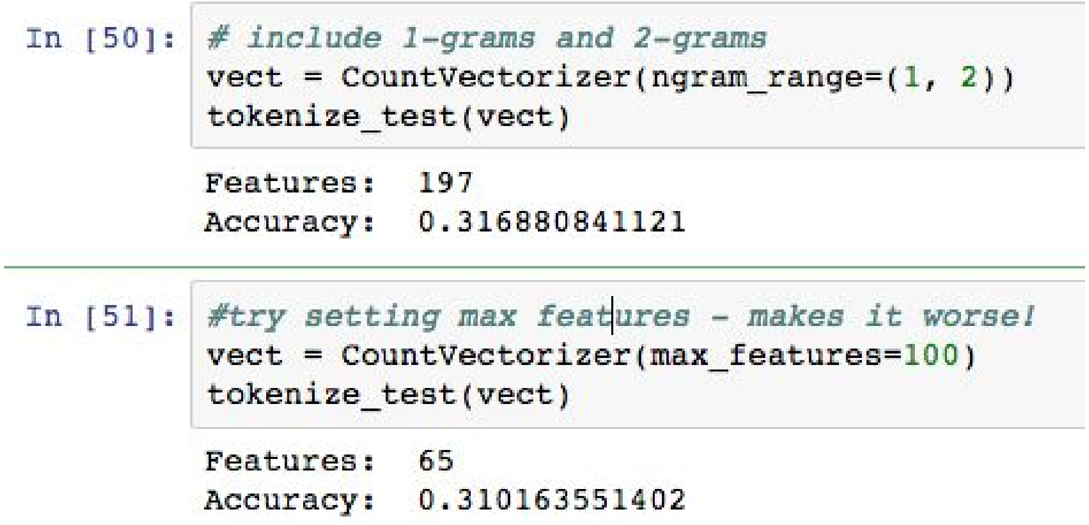
--Captain Jason Cherniss, SFPD

Arresting drug dealers is not the most effective solution to the Tenderloin’s problems. They’re going to go away and come back. Most of the drug dealers in the Tenderloin come in from the outside and that the neighborhood’s ***proximity to two BART stations***, Powell Street and Civic Center, provides them with something that all Bay Area residents long for: an easy commute to work.

--Captain Jason Cherniss, SFPD

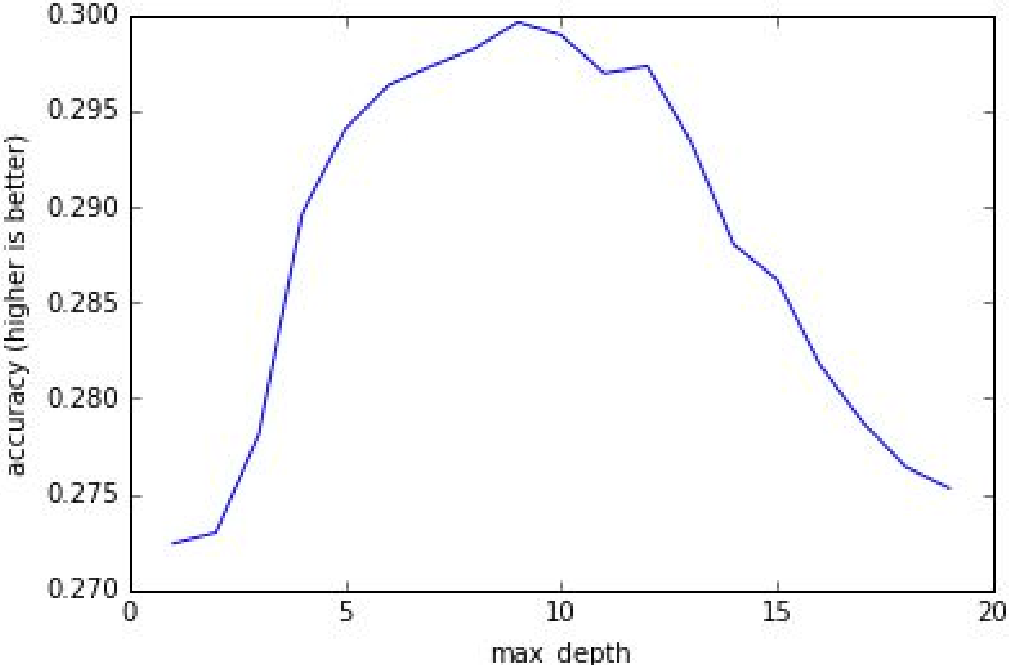
Countvectorizer with street addresses in the TL

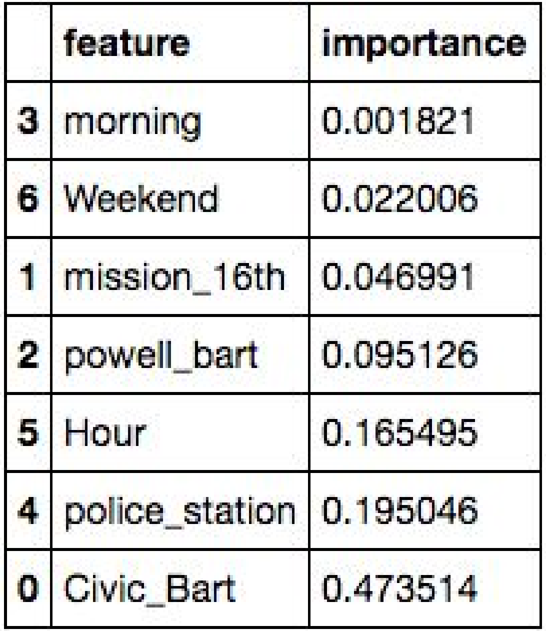
- Removed stop words like St, ave, ct, way to get just the street name.
- Performed about as well as other models using latitude and longitude.



```
In [50]: # include 1-grams and 2-grams
         vect = CountVectorizer(ngram_range=(1, 2))
         tokenize_test(vect)

         Features:  197
         Accuracy:  0.316880841121

In [51]: #try setting max features - makes it worse!
         vect = CountVectorizer(max_features=100)
         tokenize_test(vect)

         Features:  65
         Accuracy:  0.310163551402
```

Decision Trees Based on Bart and Police Station Locations





	feature	importance
3	morning	0.001821
6	Weekend	0.022006
1	mission_16th	0.046991
2	powell_bart	0.095126
5	Hour	0.165495
4	police_station	0.195046
0	Civic_Bart	0.473514

.323

Null Accuracy

Conclusion

- In the TL, the location of Civic Center Bart and police station proved to be strong predictors.
- Adding additional columns for places like police stations and Bart stations improved accuracy scores more than using different models.
- Crime is tied to relative locations (ie near bars, parks, transit) rather than absolute locations.
- Next steps include adding additional location features such as bars/liquor stores from data released about liquor licenses.