Analysis of Crime Neighboring Sporting Events

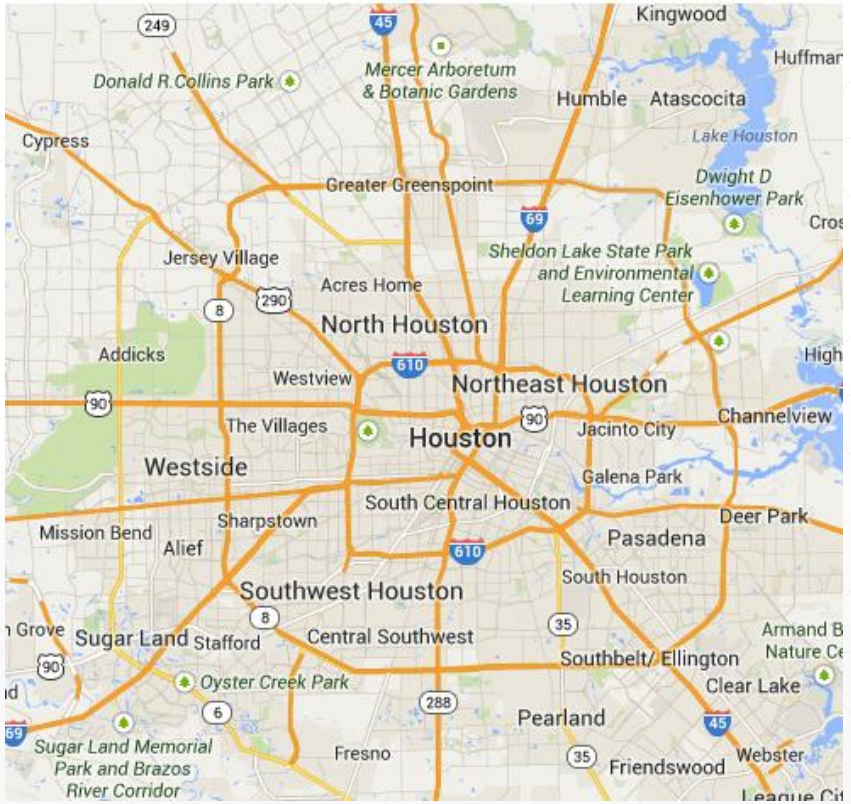
Springboard Data Science Capstone Project

Francisco Salas

https://github.com/franksalas/SpringboardCapstone

Problem

- Houston Texas has around 2.3 million residents
- ~600 square miles
- Four professional sport teams
- Two NCAA Division I-A schools.
- What is the likelihood of a crime around a sports stadium given the event?



Who Might Care?

- Houston Police Department
- City planners
- Sporting event attendees.

Venue : Toyota Center

- NBA
  - Houston Rockets
- Capacity
  - 18,500
- Coordinates
  - 29° 45′ 3″ N, 95° 21′ 44″ W

Venue : Minute Maid Park

- MLB
  - Houston Astros
- Capacity
  - 41,000
- Coordinates
  - 29°45′25″N 95°21′20″W

Venue : BBVA Compass Stadium

- MLS
  - Houston Dynamo
- Capacity
  - 22,000
- Coordinates
  - 29° 45′ 7.92″ N, 95° 21′ 8.64″ W

Venue: NRG Stadium

- NFL
  - Houston Texans
- Capacity
  - 72,000
- Coordinates
  - 29° 41′ 5″ N, 95° 24′ 39″ W

Venue : Rice Stadium

- College Football
  - Rice University
- Capacity
  - 47,000
- Coordinates
  - 29° 42′ 59″ N, 95° 24′ 33″ W

Venue : TDECU Stadium

- College Football
  - University of Houston
- Capacity
  - 40,000
- Coordinates
  - 29° 43′ 19″ N, 95° 20′ 57″ W

Datasets

https://github.com/franksalas/SpringboardCapstone/tree/master/data_wrangling_v2

Uniform Crime Report Program

- Published by the FBI
- Program that compiles official data on crime in the US
- Data is collected by local law enforcement agencies
- Two types
  - Type I
  - Type II

Offense Types

Type I Crimes : On Report	Type II Crimes : Not on Report
Aggravated assault	Simple assault
Rape	Curfew offenses
Murder	Loitering
Robbery	Embezzlement
Burglary	Forgery
Theft	Counterfeiting
Auto Theft	*.. any offense not on Type I*

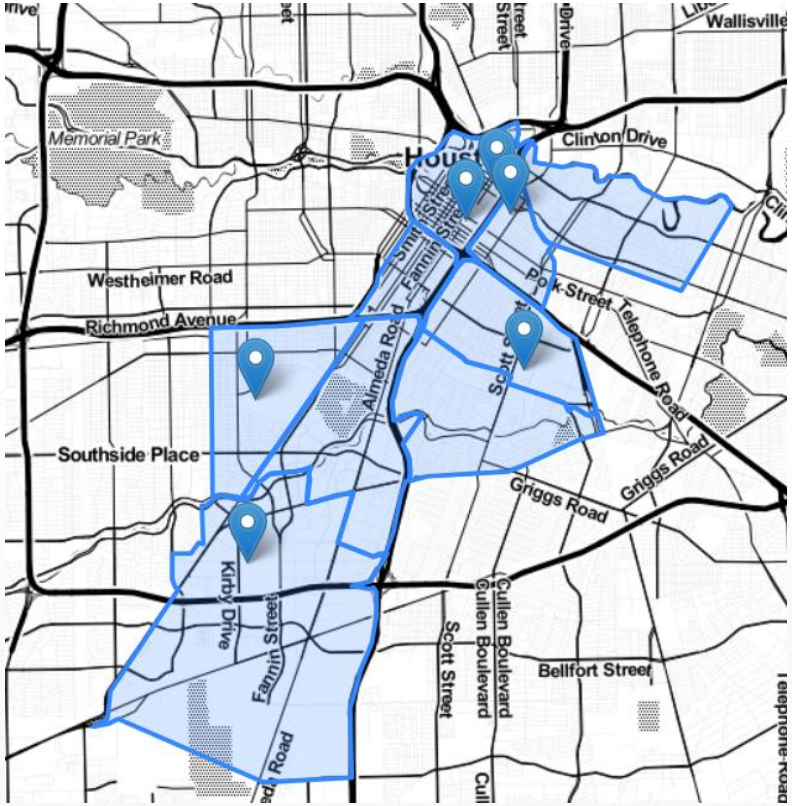
Crime Datasets

- Provided by Houston Police Department
- Monthly files from 2010 -2017
- 96 excel files

Variable	Description
Date	Date of offense, include month/date/year
Hour	Approximate time when an event occurs, value form 0-24
Offense Type	Type I offense
Beat	**The geographic area of the city broken down for patrol and statistical purpose**
Premise	Identify the type of location where crime occurs (apartment complex, parking lot, etc.)
Block Range	The value range of street
Street Name	Name of the street where the offense occurred
Type	Street type, rd, Blvd
Suffix	N, S, E, W
Offenses	Times offense happen within the time frame

Crime Datasets

Police Beats used	
15E40	10H40
10H50	10H60
10H80	10H30
10H10	1A10
10H70	
Crime data was only selected from Police Beats around stadiums	



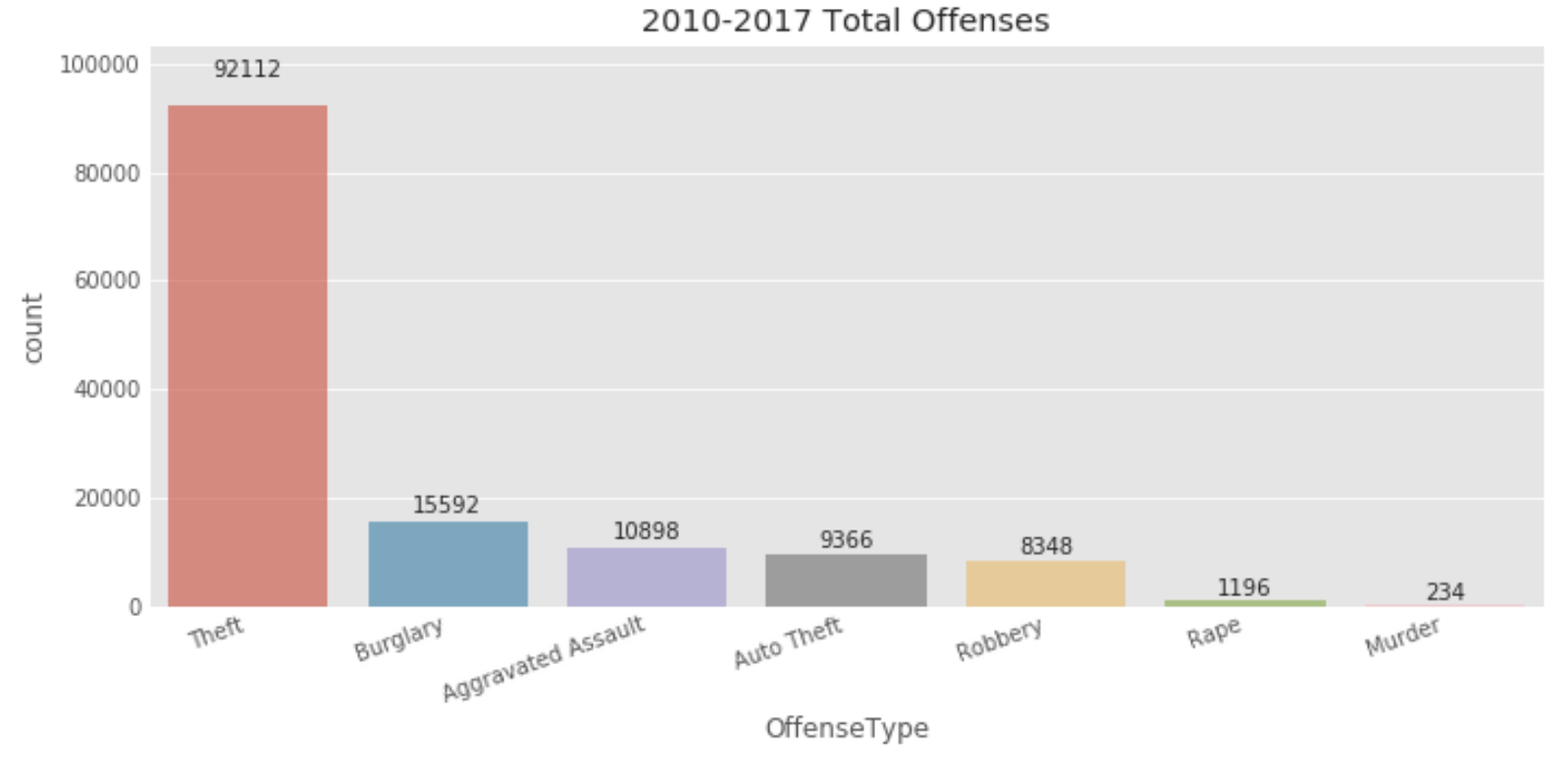
Sports Datasets : Schedule & Scores

Site	Sports Data Extracted
sports-reference.com	• NBA • MLB • CF: UH • CF: RU
github.com/jokecamp/FootballData	MLS
sportsradar.us	NFL

Data Exploration

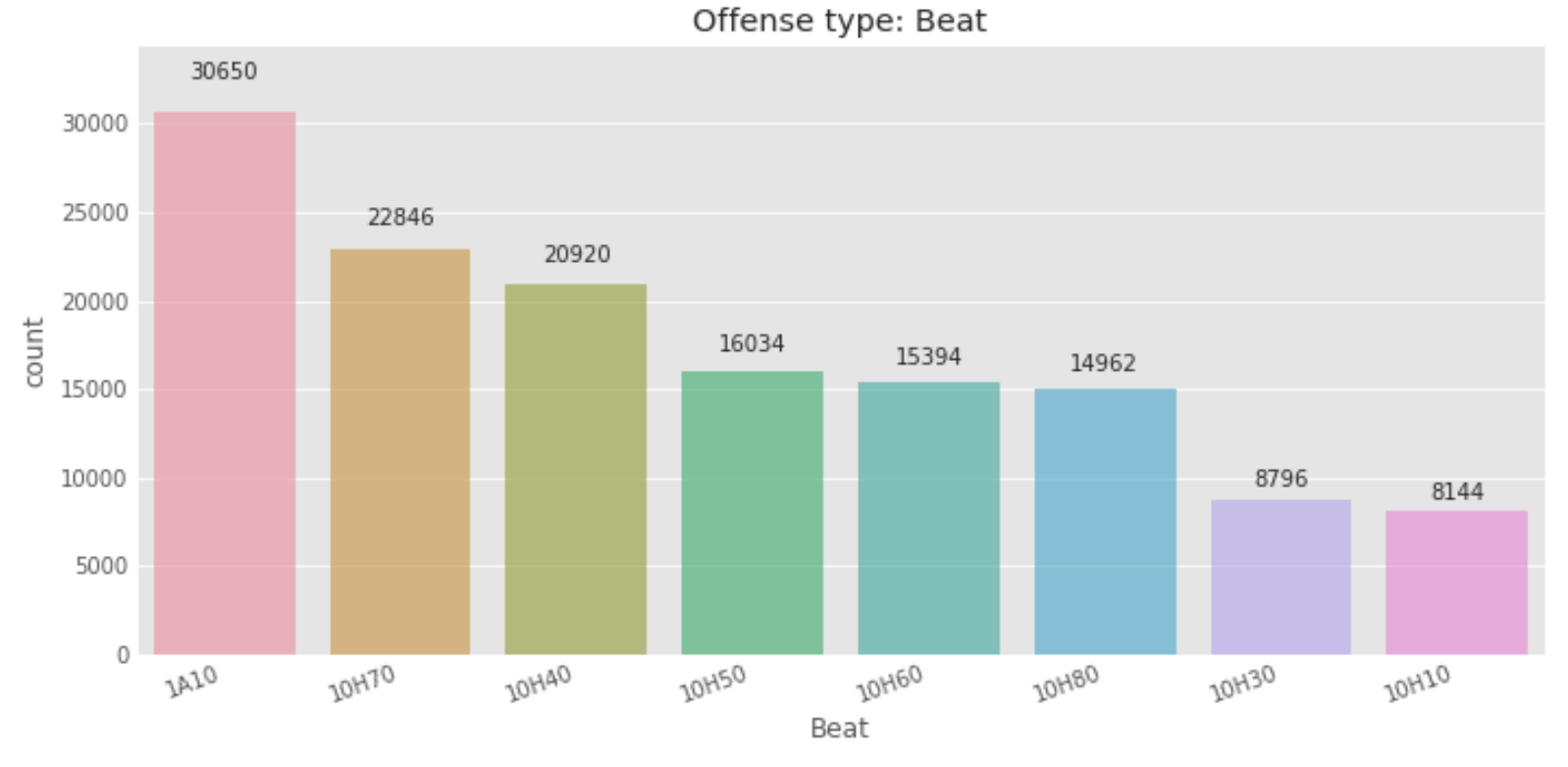
https://github.com/franksalas/SpringboardCapstone/tree/master/data_exploration

Total offenses 2010-2017



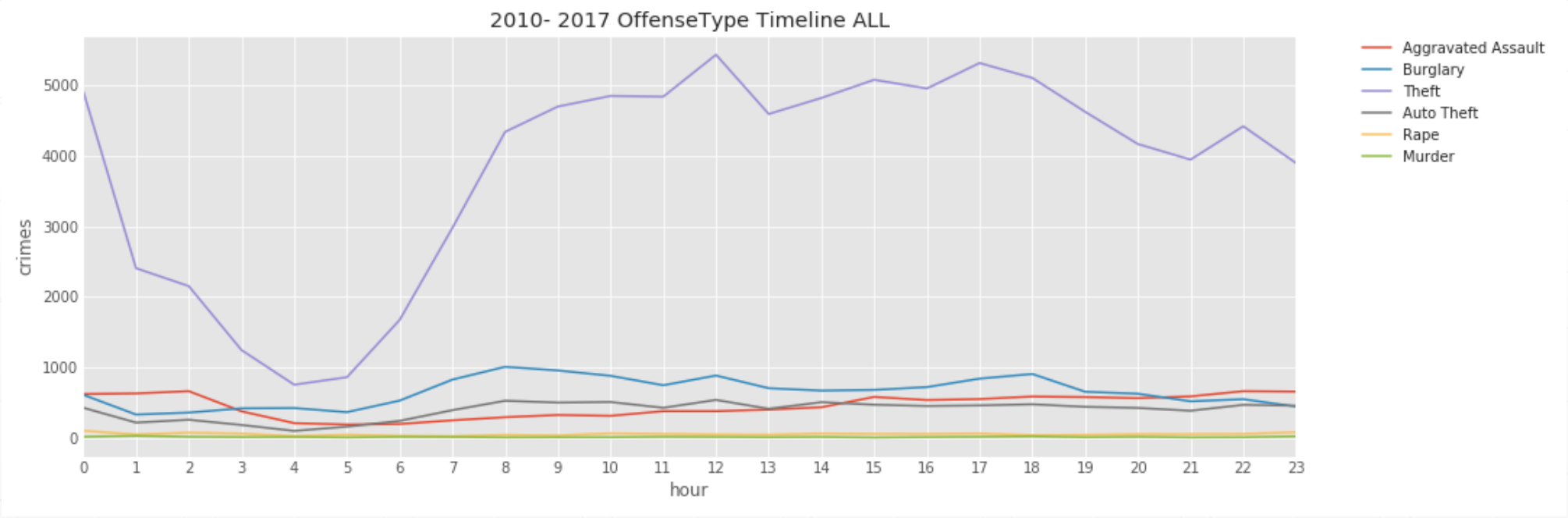
*Selected Police Beats

Sum of Crimes by Police Beat



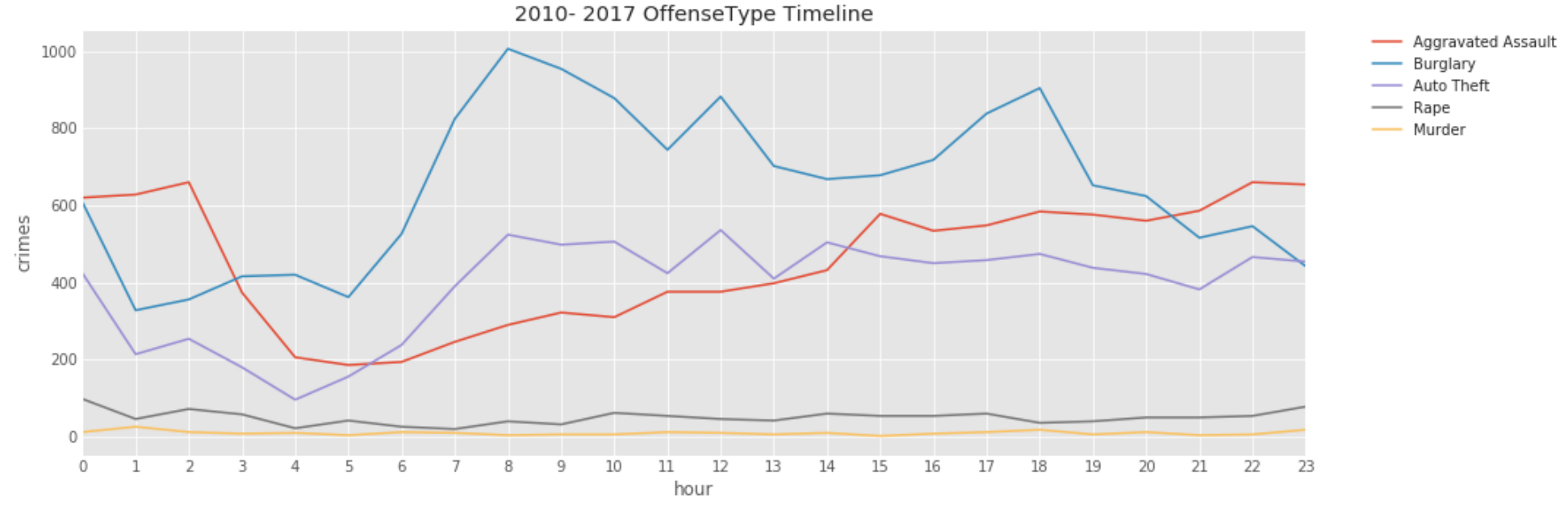
*Selected Police Beats

Timeline Offense Type



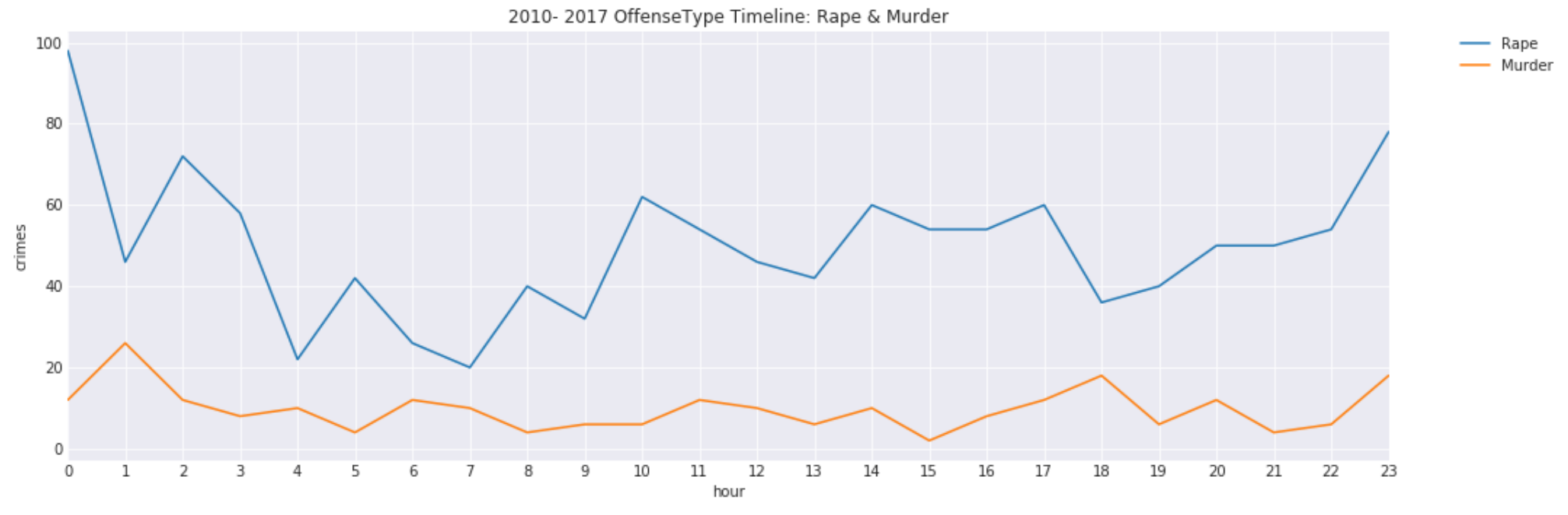
*Selected Police Beats

Timeline Offense Type *(theft removed)*



*Selected Police Beats

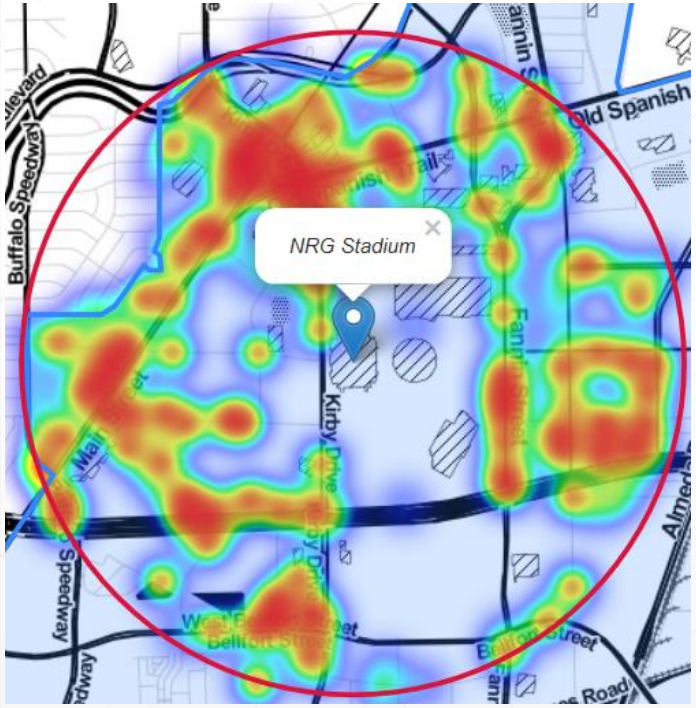
Timeline : Rape & Murder



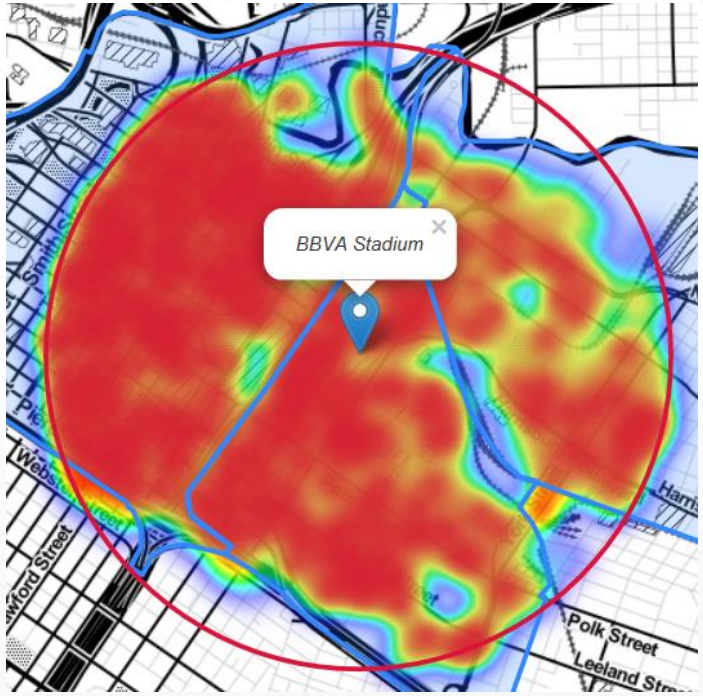
**Selected Police Beats*

Crime Heatmap

NRG Stadium



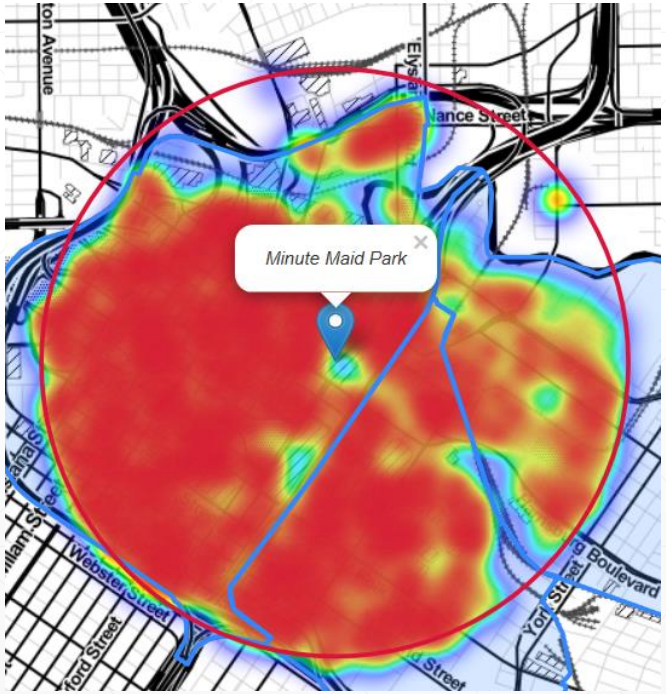
BBVA Stadium



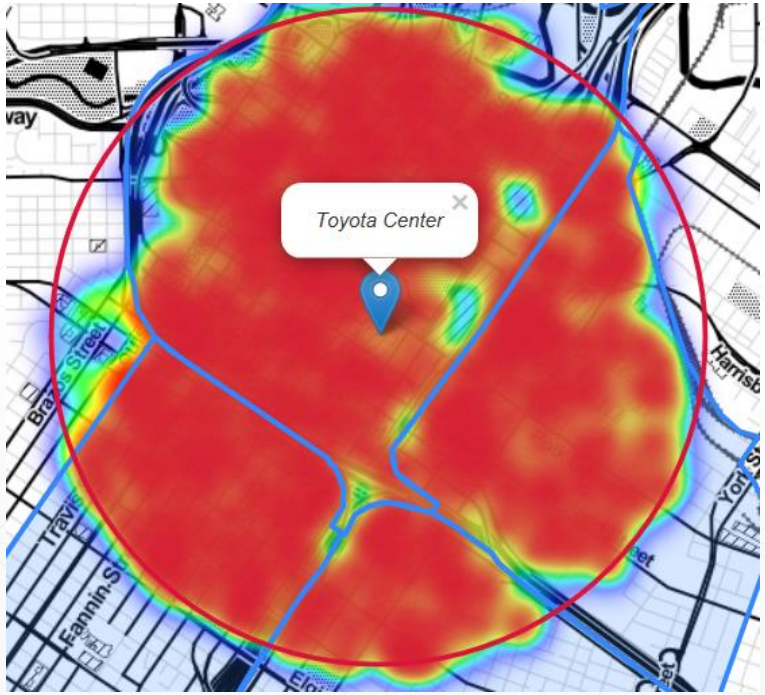
1 mile radius around stadium

Crime Heatmap

Minute Maid Park



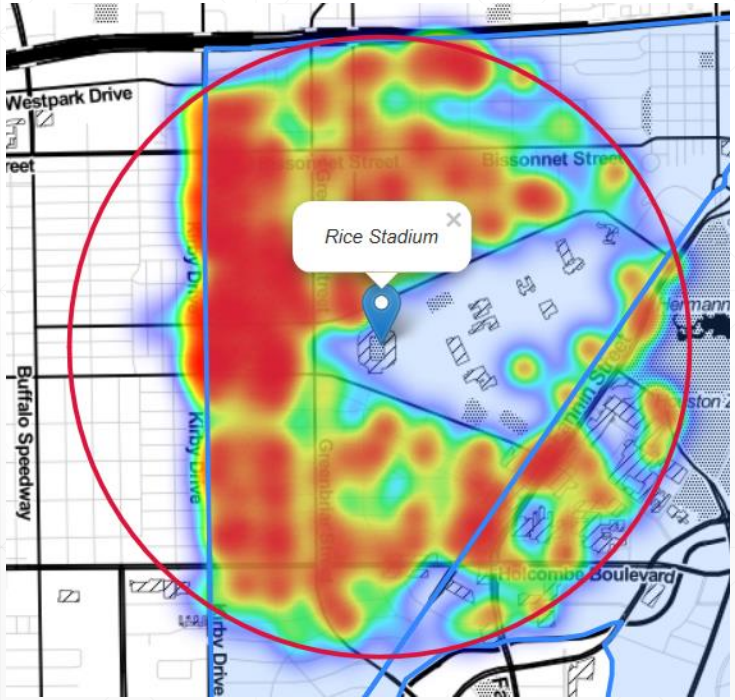
Toyota Center



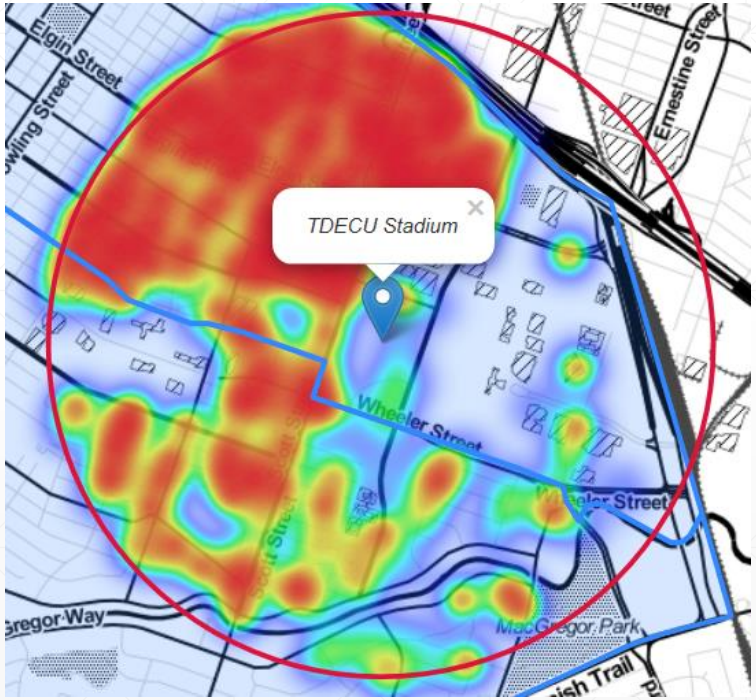
1 mile radius around stadium

Crime Heatmap

Rice Stadium



TDECU Stadium



1 mile radius around stadium

Modeling

https://github.com/franksalas/SpringboardCapstone/tree/master/final

Supervised Learning

- **Linear Regression:**
  - Ordinary least squares Linear Regression.
- **Ridge:**
  - Linear least squares with l2 regularization.
- **Lasso:**
  - Linear Model trained with L1 prior as regularized.

Dataset

- Target
  - *crime_total*
- Predictors
  - *~ crime_total*

crime_total	year	dist_stadium_median	hour_mode	Weekday_Monday	…

Supervised Learning : Plugin Function

```
def modelfit(alg, X,y):
    '''target = y, predictors = X, alg = algorithm used
    '''
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state=42)
    #Fit the algorithm on the data
    alg.fit(X_train,y_train)
    #Predict training set:
    train_predictions = alg.predict(X_train)
    #Perform cross-validation:
    cv_score = cross_val_score(alg, X, y, cv=10, scoring='neg_mean_squared_error')
    cv_score = np.sqrt(np.abs(cv_score))

    #Print model report:
    print ("\nModel Report")
    print( "RMSE : %.4g % np.sqrt(metrics.mean_squared_error(y_train, train_predictions)))
    print ("CV Score : Mean  %.4g | Std  %.4g | Min  %.4g | Max  %.4g" % \
           (np.mean(cv_score),np.std(cv_score),np.min(cv_score),np.max(cv_score)))
```

Function

Algorithms used

```
alg1 = LinearRegression(normalize=True)
alg2 = Ridge(alpha=0.1,normalize=True)
alg3 = Lasso(alpha=0.1,normalize=True)
```

Results

https://github.com/franksalas/SpringboardCapstone/tree/master/final

TDECU Stadium

	Results
Linear	Model Report RMSE : 1.077 CV Score : Mean 5.234e+13 | Std 1.317e+14 | Min 3.493 | Max 4.45e+14
Ridge	Model Report RMSE : 1.704 CV Score : Mean 4.767 | Std 3.785 | Min 1.255 | Max 13.5
Lasso	Model Report RMSE : 2.782 CV Score : Mean 4.701 | Std 4.044 | Min 1.393 | Max 12.09

Rice Stadium

	Results
Linear	Model Report RMSE : 0.0716 CV Score : Mean 2.004e+14 | Std 4.137e+14 | Min 1.32 | Max 1.32e+15
Ridge	Model Report RMSE : 0.6496 CV Score : Mean 1.595 | Std 0.7066 | Min 0.2395 | Max 2.417
Lasso	Model Report RMSE : 1.357 CV Score : Mean 1.283 | Std 0.4084 | Min 0.7061 | Max 2.274

NRG Stadium

	Results
Linear	Model Report RMSE : 4.733 CV Score : Mean 3.006e+13 | Std 9.019e+13 | Min 5.866 | Max 3.006e+14
Ridge	Model Report RMSE : 5.195 CV Score : Mean 8.908 | Std 3.495 | Min 4.1 | Max 15.28
Lasso	Model Report RMSE : 6.08 CV Score : Mean 7.667 | Std 3.242 | Min 3.67 | Max 12.71

Toyota Center

Results	
Linear	Model Report RMSE : 3.04 CV Score : Mean 2.718e+06 | Std 8.084e+06 | Min 2.718 | Max 2.697e+07
Ridge	Model Report RMSE : 3.079 CV Score : Mean 3.47 | Std 0.7354 | Min 2.501 | Max 5.116
Lasso	Model Report RMSE : 3.907 CV Score : Mean 3.689 | Std 0.8254 | Min 2.687 | Max 5.62

Minute Maid Park

	Results
Linear	Model Report RMSE : 2.811 CV Score : Mean 6.38e+13 | Std 1.081e+14 | Min 2.994 | Max 3.154e+14
Ridge	Model Report RMSE : 2.818 CV Score : Mean 3.085 | Std 0.3041 | Min 2.538 | Max 3.513
Lasso	Model Report RMSE : 3.231 CV Score : Mean 3.249 | Std 0.3112 | Min 2.772 | Max 3.79

BBVA Compass Stadium

	Results
Linear	Model Report RMSE : 2.478 CV Score : Mean 2.521e+13 | Std 4.274e+13 | Min 2.7 | Max 1.355e+14
Ridge	Model Report RMSE : 2.554 CV Score : Mean 3.653 | Std 0.9231 | Min 2.467 | Max 5.433
Lasso	Model Report RMSE : 3.515 CV Score : Mean 3.602 | Std 0.8621 | Min 2.05 | Max 5.239

Thank you