Capstone Project:

San Francisco Crime Classification

ANUSHREE SRINIVAS

MENTOR : SRDJAN SANTIC

Problem to be solved and Motivation

- From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz.
- Although today, the city is known more for its tech scene than its criminal past, there is no scarcity of crime in the city by the bay.
- From Sunset to SOMA, and Marina to Excelsior, the dataset I worked on provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods.
- The objective of this project is to predict the category of crime that occurred at a given time and location.

Client

- City and County of San Francisco is trying to make the city and the neighborhoods more safe by leveraging the past crime data over the years.
- The City authorities and the San Francisco Police Department will be the client for whom the conducted data analysis as part of the capstone project will be beneficial.
- This analysis can also be helpful for residents, policymakers and the government to create awareness and take the necessary measures.

Feature Engineering

- **Dates**:
  - day(day of month as **Day**),
  - hour
  - minute
  - Dayofweek (**day_num**)
  - weekofyear
  - Dayofyear
  - year
  - Month
- PdDistrict(**District_num**)
- **Season:**
  - Fall
  - Winter
  - Spring
  - Summer
- **Address**:
  - Intersection / Residential
  - Latitude(**X**)
  - Longitude(**Y**)
- **Weekend:**
  - Friday
  - Saturday
  - Sunday

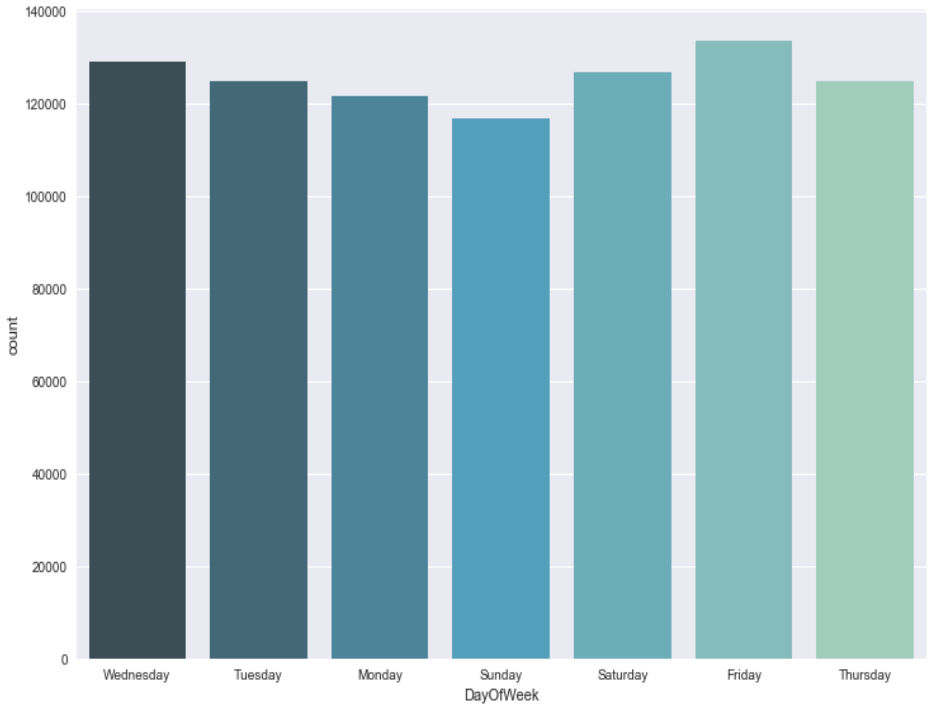
Exploratory Data Analysis

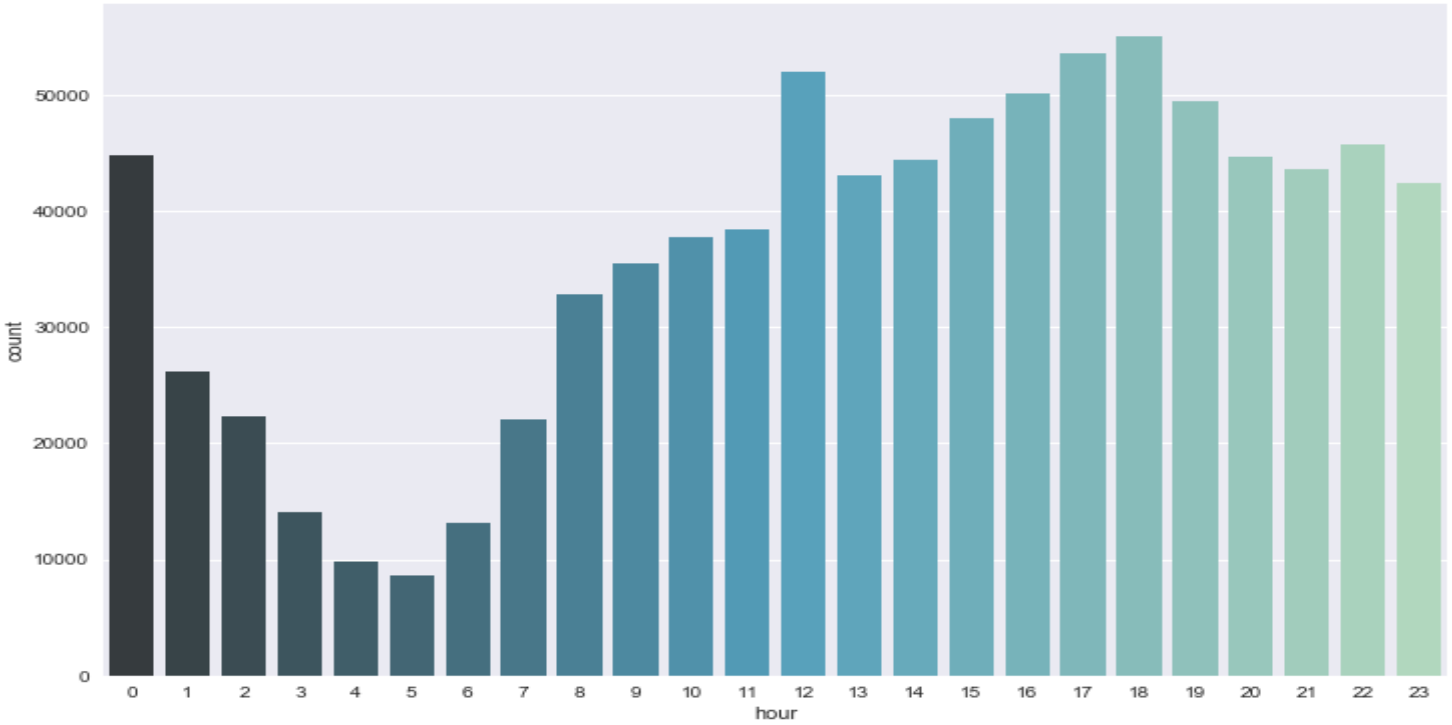
- The number of crimes reached an all-time high in the span of 12 years in 2013
- The number of crimes was high at 2003 after which it was decreasing until 2008 where it increased, after which there was a decrease until 2012-2013.
- A spike in crime rate can be noticed for every 5 years.
- The crime rate is maximum on Friday and Saturday and least on Sundays.
- The months of May and October recorded a high number of crimes.
- Maximum crime activity is around 12pm and in the 5-7pm window the maximum being at 6pm.

Exploratory Data Analysis

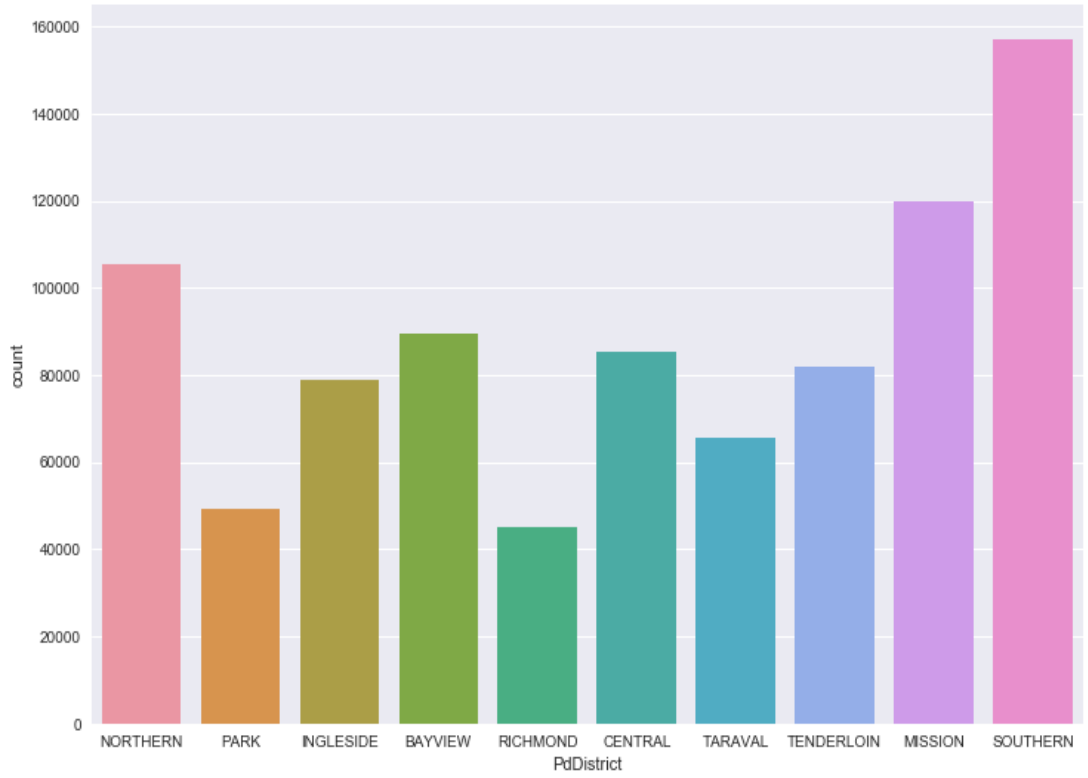
- The Southern region followed by the mission region recorded the maximum crime rate.
- The top 5 categories of crime being Larceny/Theft, Assault, Vandalism, Drug/Narcotic, Vehicle/Theft.
- The burglary rate is maximum on Fridays. On checking the hourly distribution, the conclusion is5pm on Friday is most prone to burglaries.
- Although the overall crime recorded is the least on Wednesday the number of crimes due to drug or narcotic is most on Wednesday and specifically at 2pm.
- Larceny was most recorded on Saturday followed by Friday.
- The maximum crime activity recorded due to vandalism is on Saturday at 10pm.

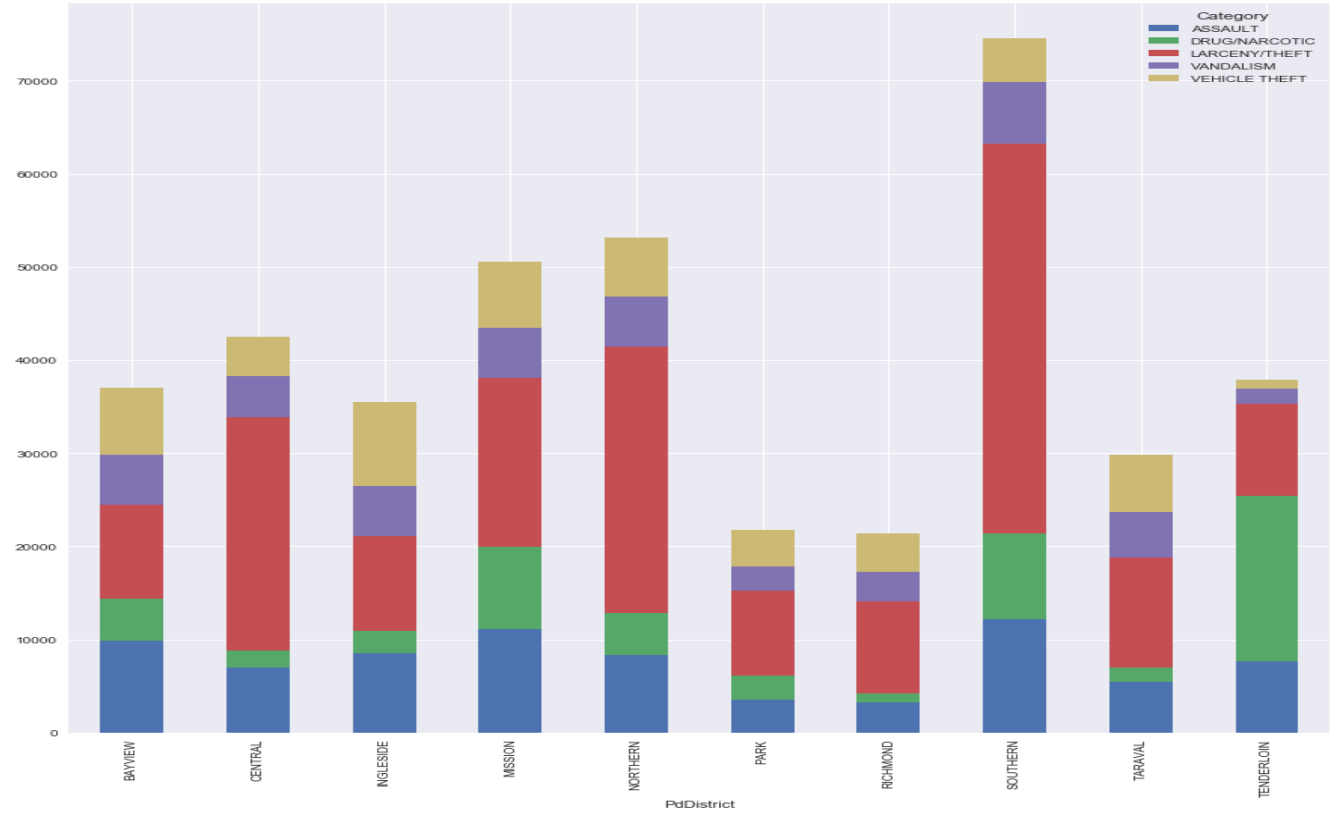
Visualizing the Crimes in San Francisco





Visualizing the Crimes in San Francisco





Algorithms and Results

Models	Log loss score on Kaggle
Random Forest Classifier with all features	2.41
Random Forest Classifier with fewer features(excluding weekends and seasons)	**2.38**
AdaBoost Classifier	3.51
AdaBoost Classifier with best parameters using GridSearch Cross Validation	2.83

Recommendations for the Client

- Policymakers can be put forth local policies to ensure the safety of the districts and neighborhoods.
- For law enforcement agencies that are considering adopting predictive policing tools, situational awareness can be increased.
- The police force can be distributed according to the day and type of crime in a particular neighborhood thereby reducing the extent of the damage.
- Advertisement and awareness campaigns can be conducted in different areas for the area specific crime to involve the citizens and work with them.
- Knowing the category of the crime, we can better equip the police force, district wise for better response time.

Future Research

- **Try non-linear models:** The models that were used in here were all linear models. Non-linear models could be implemented to see if better results can be achieved.

- **New features:** New features could be created to help us generalize better on the test dataset thereby achieving better results.