Final Project

Laura Godleski
Lindsay MacDonald
Natalie VanDyke

Overview

WHO IS OUR TEAM?

Lindsay MacDonald

Natalie VanDyke

Laura Godleski

WHAT IS THE TOPIC?

Influence of demographics on the homicide solve rate in California.

WHY THIS TOPIC?

Data Sources

U.S. Homicide Reports, 1980-2014

https://www.kaggle.com/jyzaguirre/us-homicide-reports

US Census Demographic Data

https://www.kaggle.com/muonneutrino/us-census-demographic-data/data?select=acs2015_county_data.csv

Questions to Answer

- Is there a correlation between the personal attributes of the victim and the solve rate?
- Is there a correlation between county demographics where the murder occurred and the solve rate?
- Can we develop a machine learning model that predicts whether or not a crime would be solved given hypothetical sets of circumstances?

Data Exploration: Preprocessing

Kaggle Dataset #1 (US Homicide Reports, 1980-2014)

- Filtered ‘State’ column for California
- Renamed ‘City’ column as ‘County’ in order to merge with with the second dataset
- Dropped irrelevant columns
- Checked for null values (0)

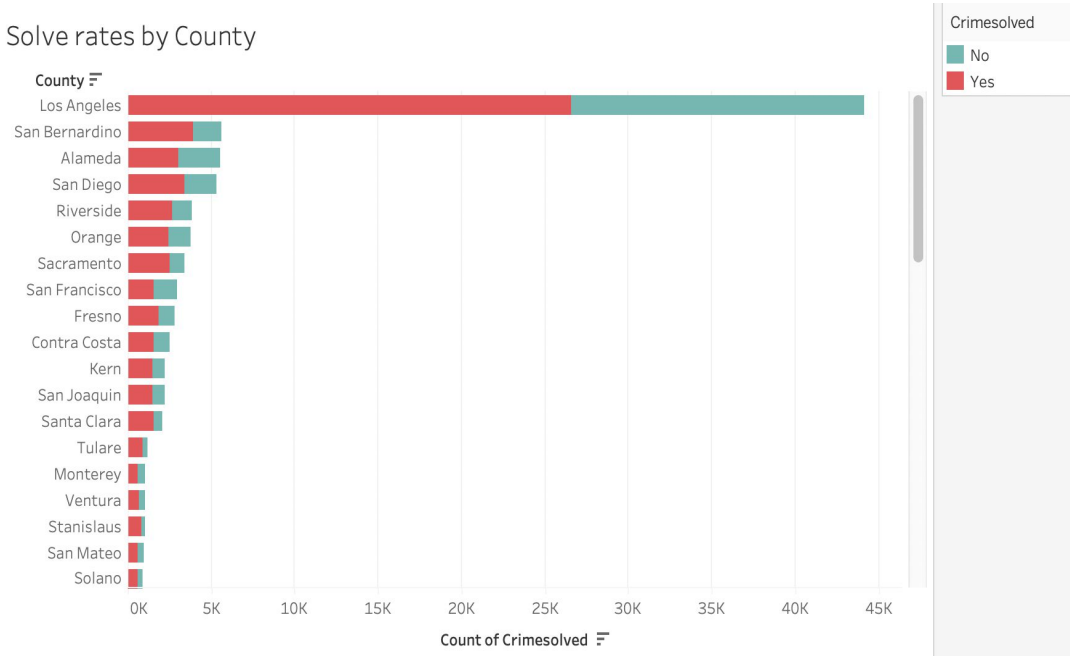
Kaggle Dataset #2 (US Census Demographic Data)

- Filtered ‘State’ column for California
- Dropped irrelevant columns
- Checked for null values (0)

Datasets were imported into Postgres and joined; merged dataset was then uploaded to AWS S3 bucket

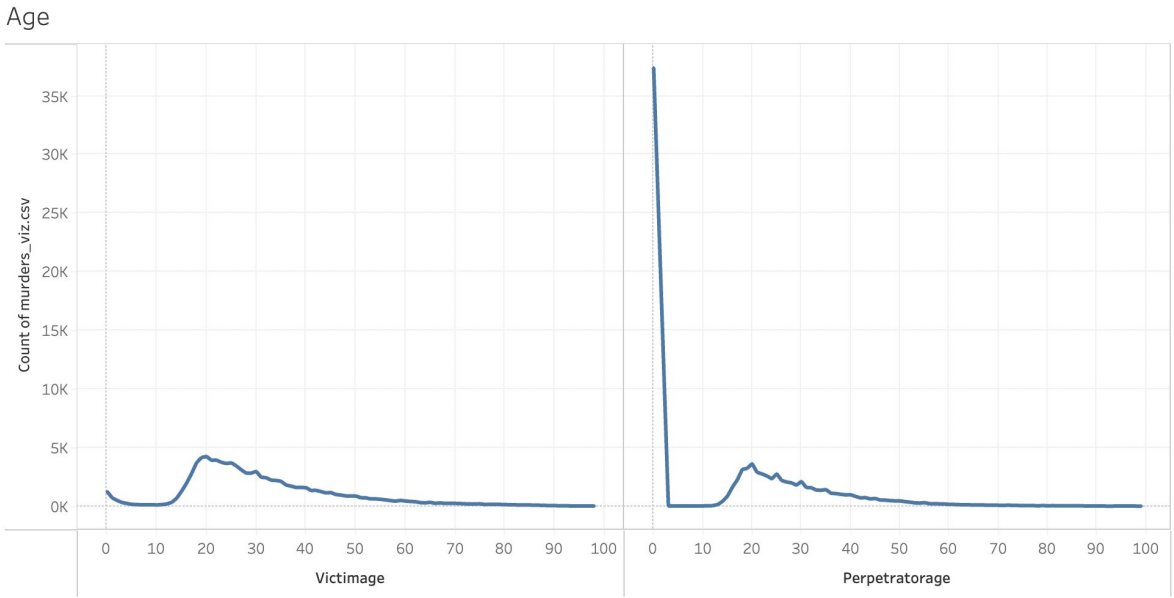
Data Exploration: Initial Findings

Due to the population disparity among counties in CA, a vast majority of the data was for Los Angeles county

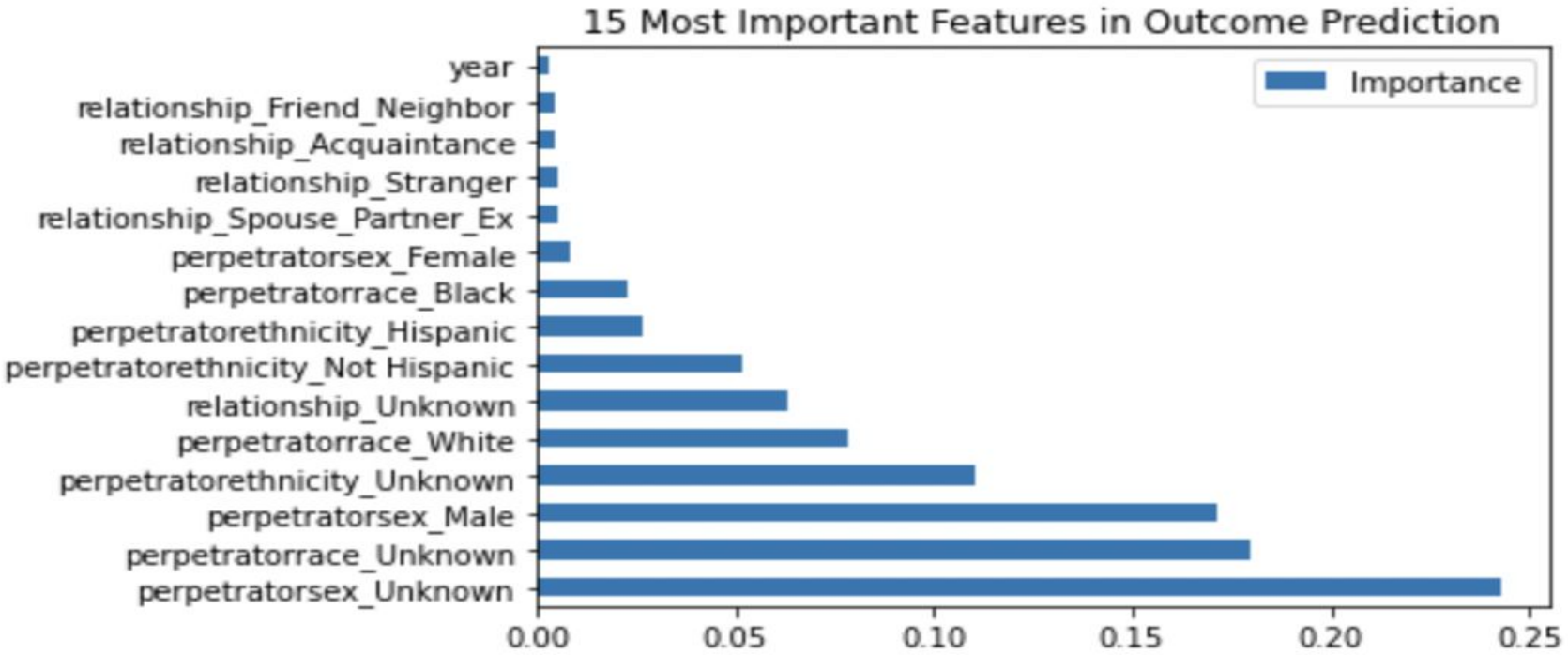


Data Exploration: Initial Findings

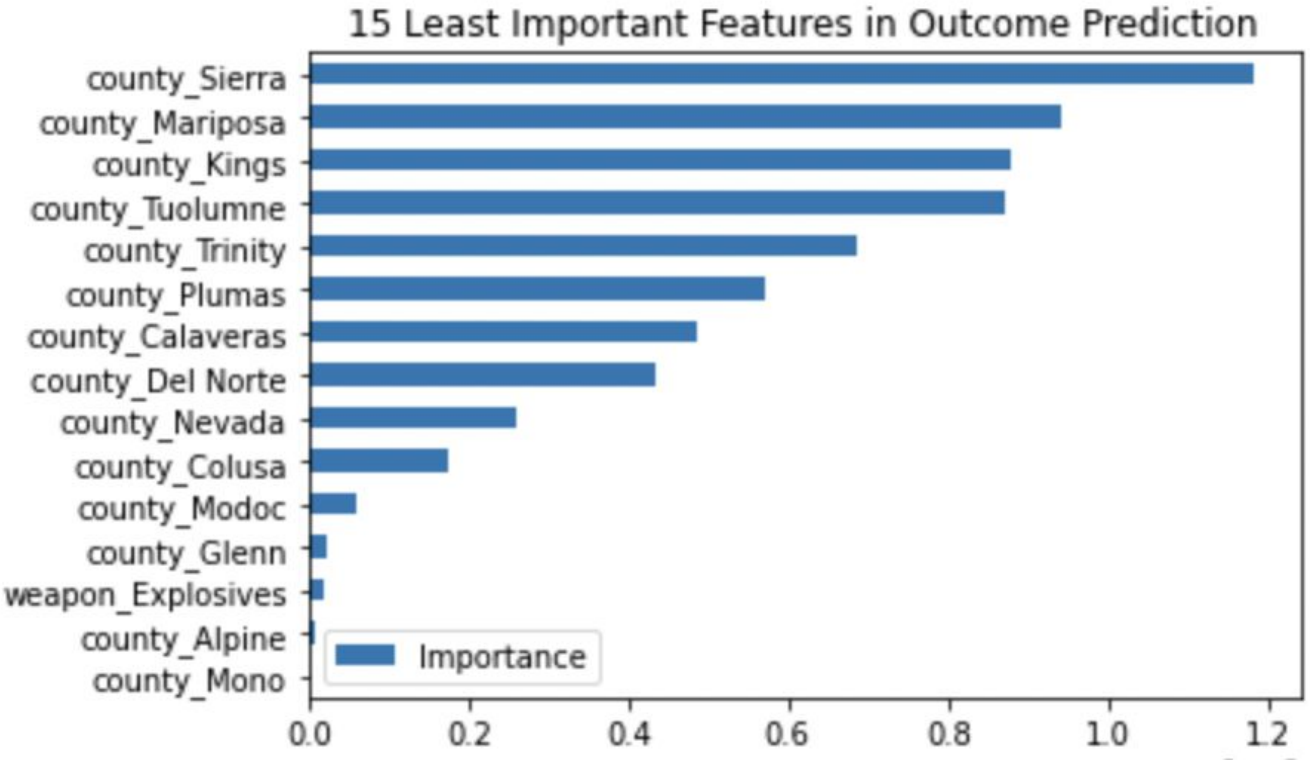
There is an unusual amount of records capturing an age of 0 for perpetrators which suggests these individuals were largely unidentified and therefore their age was unknown



Analysis


15 Most Important Features in Outcome Prediction
Importance
year
relationship_Friend_Neighbor
relationship_Acquaintance
relationship_Stranger
relationship_Spouse_Partner_Ex
perpetratorsex_Female
perpetratorrace_Black
perpetratorethnicity_Hispanic
perpetratorethnicity_Not Hispanic
relationship_Unknown
perpetratorrace_White
perpetratorethnicity_Unknown
perpetratorsex_Male
perpetratorrace_Unknown
perpetratorsex_Unknown
0.00
0.05
0.10
0.15
0.20
0.25

Analysis


15 Least Important Features in Outcome Prediction
county_Sierra
county_Mariposa
county_Kings
county_Tuolumne
county_Trinity
county_Plumas
county_Calaveras
county_Del Norte
county_Nevada
county_Colusa
county_Modoc
county_Glenn
weapon_Explosives
county_Alpine
county_Mono
Importance
0.0
0.2
0.4
0.6
0.8
1.0
1.2