PYTHOMCRUISE

STEVEN SPIELBERG

MAJORITY REPORT

"Two Thumbs Way Up! A Masterpiece."

— Ebert & Roeper

Introduction

We wanted to look into, if a machine can look at a dataset regarding a particular homicide case and provide information on what profile the police/FBI/criminal investigators should be looking for. This could be different variables such as the perpetrators race, age and sex.

Selection of dataset

The dataset that has been used in this paper has been obtained through Kaggle.com and contains over half a million different cases of manslaughters/homicides in the United States of America, with negligence as a differentiator. The dataset has a span of 34 years from 1980 to 2014.

Cleaning of data

When going through the data, we get lots of information, some of which is valuable and some of which is not needed.

Due to this we cleaned the dataset by filtering out all the columns, which we are not interested in. These are:
Agency Code, Agency Name, Incident, Victim Ethnicity, Perpetrator Ethnicity and Record Source. Ethnicity would normally be fine to include, but since this only ranges from hispanic/non-hispanic, it is deemed irrelevant.

These are unnecessary for us to include in the dataset, since the Agency information is irrelevant when the objective is to identify a perpetrator based on information of the crime. Filtering the above out of the dataset is not enough, since a deep dive into the data reveals other implications such as missing data in the columns about Perpetrator Count and Victim Count. These should at all times be at least 1. But in a majority of the data, this has been left out. Therefore we simply remove these columns from the dataset instead of filtering out all cases, where these are not filled in.

Besides filtering columns out, we also cleaned out entries with faulty data. We wanted to use the Perpetrator Age, but in some cases, this was set at 0. This is a mistake, since the Perpetrator needs an age. We choose to restrict this to be at the lowest convictable age in the US, at 6 years of age. This accounted for some 2.000 entries that were removed.

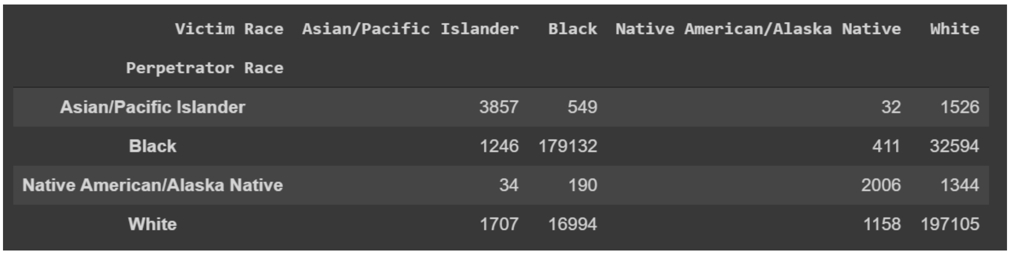
Data exploration

In this section we do some basic statistics on our dataset, to debunk or confirm our hypothesis.

The Mindhunter Hypothesis

In season 2 of the series, Mindhunter, there is a lot of profiling done by the protagonist, Holden Ford. He specifically mentions, that murders rarely happens outside of one's own race.

We tested this by doing a crosstab between the race of the victim and the race of the perpetrator (Table 1). This showed a significant correlation between the two.



Victim Race	Asian/Pacific Islander	Black	Native American/Alaska Native	White
Perpetrator Race				
Asian/Pacific Islander	3857	549	32	1526
Black	1246	179132	411	32594
Native American/Alaska Native	34	190	2006	1344
White	1707	16994	1158	197105

(Table 1 - Crosstab)

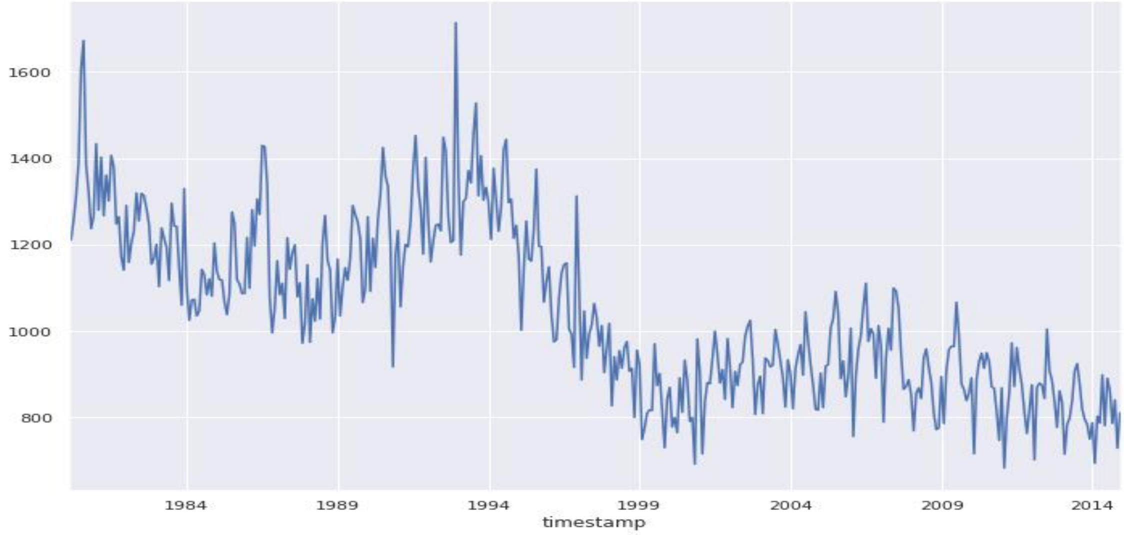
To summarize the findings:

White Victim / White Perpetrator	= 90.85%
Black Victim / Black Perpetrator	= 83.95%
Asian/Pacific Islander Victim / Asian/Pacific Islander Perpetrator	= 64.67%
Native Victim / Native Perpetrator	= 56.13%

It should be noted, that the percentages for the Asian group and the Native group can change significantly if there had been more cases of murders where the victims and perpetrators were of Asian or Native race.

Crime Development from 1980 to 2014

Looking at homicide statistics from 1980 to 2014 (Figure 1), we can see a distinct drop in homicides throughout the years.



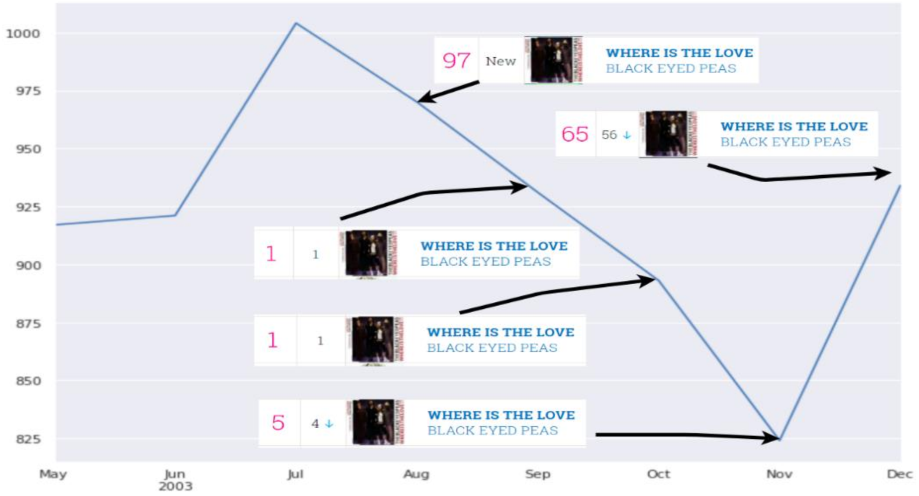
(Figure 1 - *Homicides from 1980 to 2014*)

Fun with statistics

In this part of the assignment, we look into a specific case and its correlation with the homicide statistics at the given moment.

2003 - "Where is the love"

The case we wanted to look into, was the correlation between the hit-song "Where is the love", made by The Black Eyed Peas. This song was released late June 2003, starting it's incline on the american hit-charts in August 2003. In the graph below, we have displayed the homicide statistics with the corresponding popularity of the song (Figure 2).



(Figure 2 - Where is the love?")

Looking at the graph, the findings would suggest that there seems to be a connection between how well "Where is the love" performed on american charts and the amount of homicides in the US.

Although there could be a *correlation*, it is highly unlikely that there is a *causation*. This simply shows, that data can be manipulated, and presented in such a way, that your point can be proven even though the statistical or scientific method for reaching your conclusion is faulty.

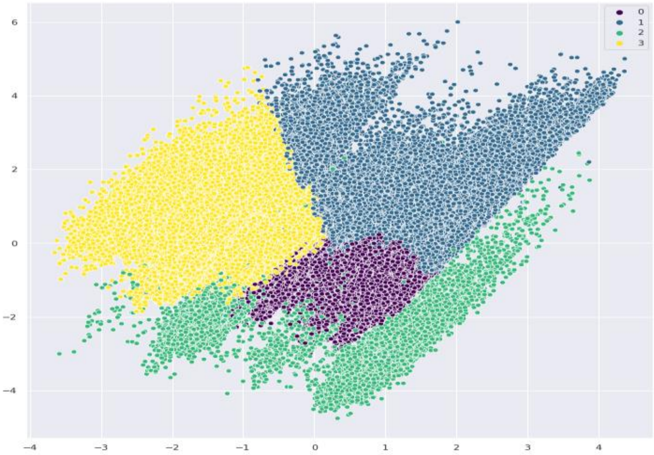
Unsupervised Machine learning

The unsupervised part of the Machine Learning, is the part, where the machine does not solely rely on input from us to calculate some sort of statistic. An example of this is clustering. We don't have to feed the machine information about what should be in the individual clusters, but it does so on its own instead.

In this part of the assignment we used a PCA and made the machine cluster, based on the variables in the dataset. But to do this, we had to convert the information in the dataset from being mostly strings/objects into being integers that can be used in calculations. An example of this can be how Perpetrator and Victim Sex (Male or Female) are converted into numeric values, where Female = 0, and Male = 1.

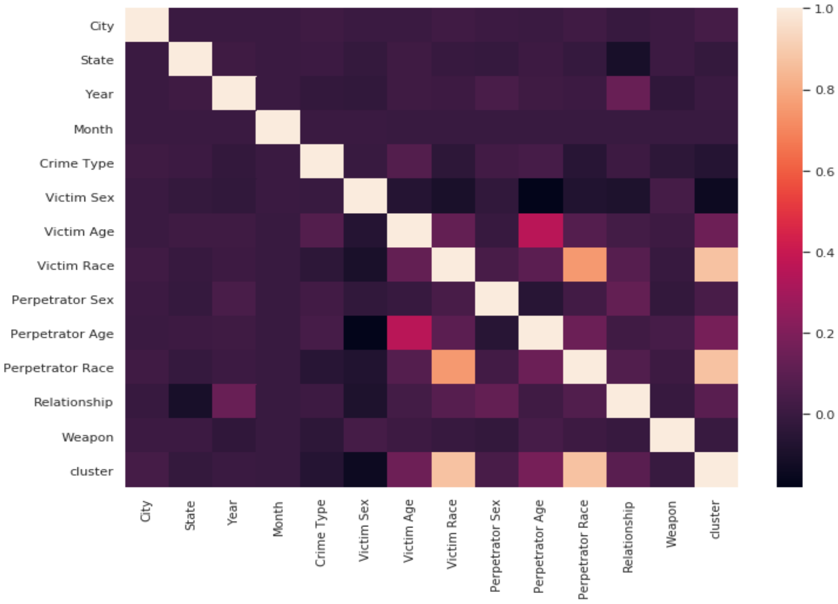
We also look at the "evr" with 8 components. The "evr" showed that with 8 components the machine could calculate approximately 73% of the variance in data.

Next step was clustering the data and inspecting the clusters. We made our clusters based on the results from using the elbow method. This resulted in using 4 clusters, which looked like this (Figure 3):



(Figure 3 - Cluster data)

The cluster was divided into 4 mostly compact areas, with cluster 1 and 3 being the largest of the four. To determine what the machine clustered by, and the other correlations in the dataset, we created a heatmap.



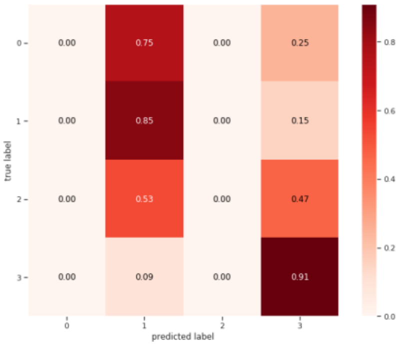
(Figure 4 - Heatmap)

The heatmap showed a couple of interesting things. The first being the parameters which the machine clustered by. This revealed that the clusters were dependent on the Victims and the Perpetrators race. It also revealed a correlation between the two, meaning that it can be expected, that a perpetrator will most likely be of the same race as the victim. This supports our hypothesis from Mindhunter, where murders rarely occurs outside one's own race.

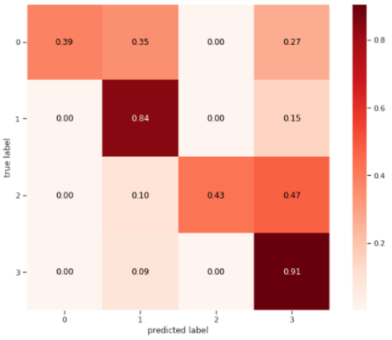
Supervised Machine learning

This part of the assignment focuses on the supervised machine learning. This includes selecting algorithms and training the machine on our datasets. We used a split off 25/75 (test/training). We selected 3 algorithms to show the differences between them, and also highlight which of the algorithms are the best at processing multiple variables. We chose Logistic Regression, Decision Tree and XGBoost as our algorithms. The models were tasked with determining the perpetrator race.

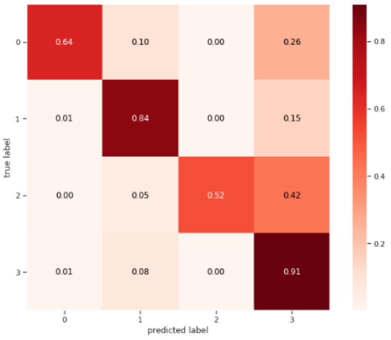
The three algorithms were tested on both the training/test set but also tested on the full set, just to see how they handled a bigger dataset. We've included our confusion matrix' for each of the algorithms and explained the results below.



Logistic Regression



Decision Tree



XGBoost

Due to the nature of the "Logistic Regression", which investigates the correlation between two factors, it does not show any valuable information as we investigate race which in our dataframe is divided into four categories. However the "Decision Tree" and "XGBoost" methods shows that our algorithm can predict the race of a perpetrator to a certain degree of accuracy. For example the Decision Tree can predict that a perpetrator is white with 91% certainty.

We also tested the results from our heatmap, where we noticed a correlation between victim and perpetrator race. This is also the case when investigating the results form the models. This showed a 57% dependency on knowing victim race. when determining perpetrator race.

Results:

We can conclude that our hypothesis, based on Mindhunter, is true - the perpetrators race is in most cases the same as the victims race. This hypothesis is only proven true for the white and black perpetrators as there is insufficient data compared to data on black and white perpetrators to conclude on the asian and native american perpetrators. Furthermore the usability of data highly depends on the methods used, and how data is being shown to stakeholders. In our data exploration we showed that there is a correlation between homicides and unrelated events, like music releases. In our unsupervised machine learning part we visualized a heat map to support our hypothesis. In the supervised machine learning part we used three algorithms to predict the race on the perpetrator based on different parameters. We can conclude that the logistic regression algorithm is less useful when dealing with non-binary outcomes.

For future assignments, this work could be extended to calculating more parameters by using multiple labels to not only calculate perpetrator race, but also age and sex. This could also be used on the unsolved crimes to give an estimate on what the perpetrator could look like in those cases.