Fatal Police Shootings in the US

Context

Context : Black Lives Matter

2012

Shooting death of African-American teen Trayvon Martin by a police officer

2013

Acquittal of the police officer: **The movement Black Lives Matter was born**

2020

Shooting death of George Floyd : **The movement became international**

Context : Black Lives Matter

- Black Lives Matter (BLM) is a decentralized political and social movement advocating for non-violent civil disobedience in protest against incidents of police brutality and all racially motivated violence against black people

Source : Wikipedia

Data exploration

Data Sources

Washington Post

- Tracking details about each killing
- Gathering this information from:
  - law enforcement websites,
  - local new reports,
  - social media,
  - by monitoring independent databases

The Washington Post

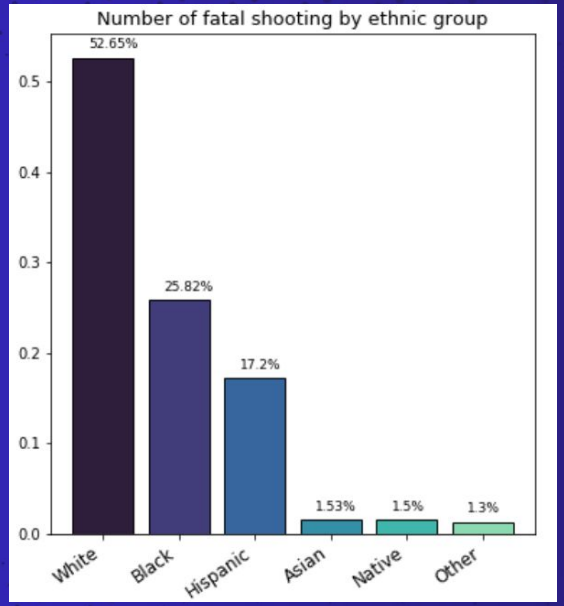
Additional datasets from US census data

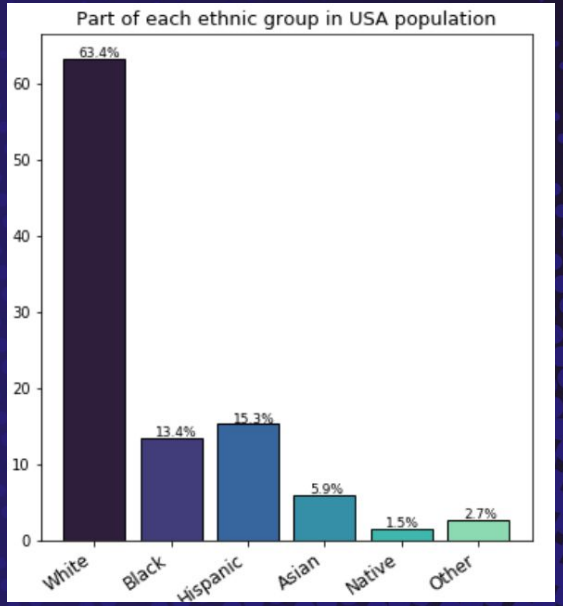
- on poverty rate,
- high school graduation rate,
- median household income,
- racial demographics
- Population by city

United States Census Bureau

Data visualisation by ethnic group

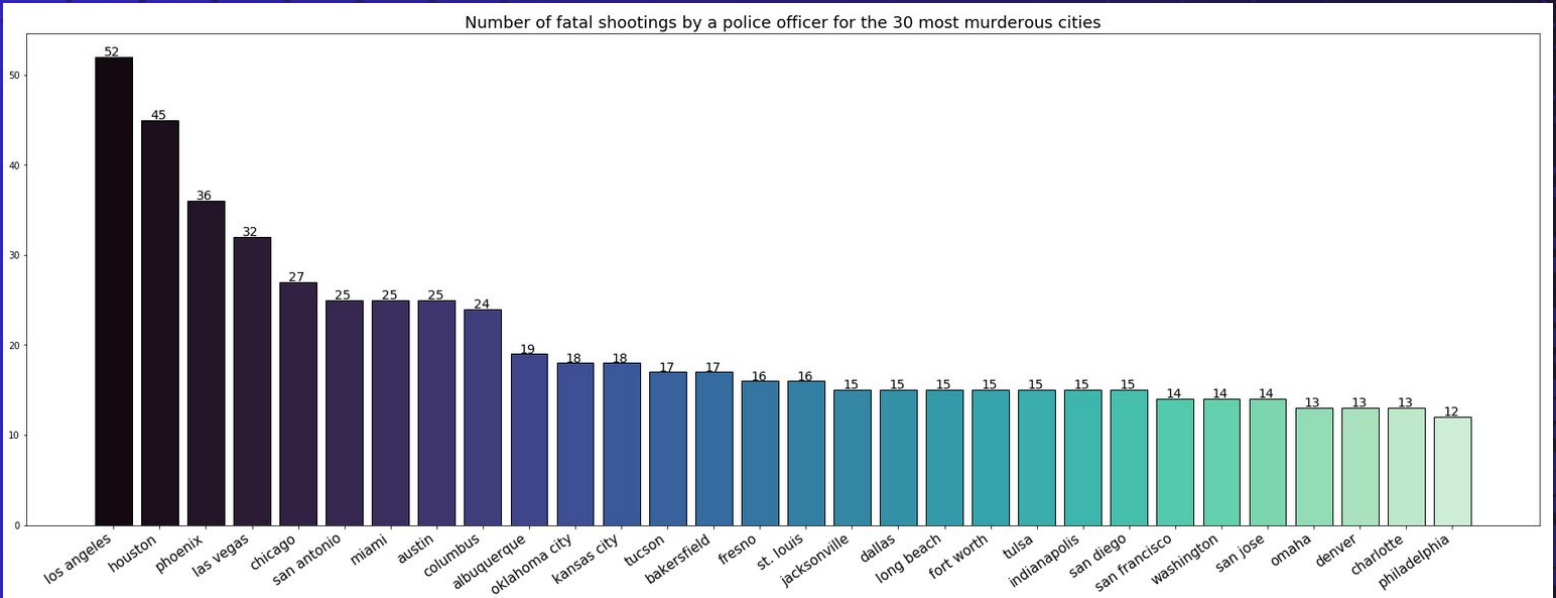
- Which are the part of fatal shootings by ethnic group ?





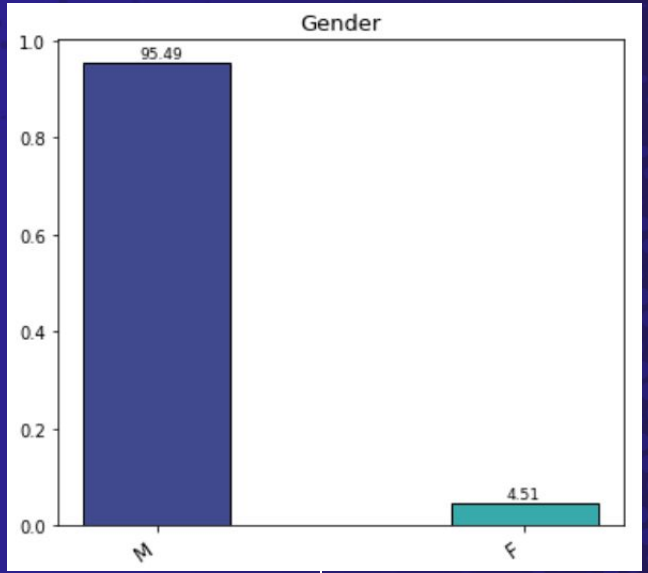
Data visualisation by city

- Which cities counts the highest number of fatal shooting ? *(between 2015 & 2017)*



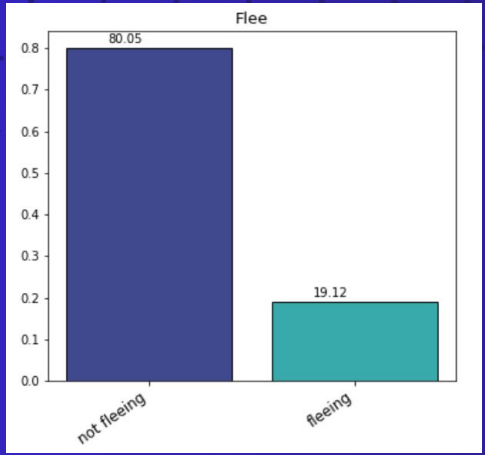
Data visualisation

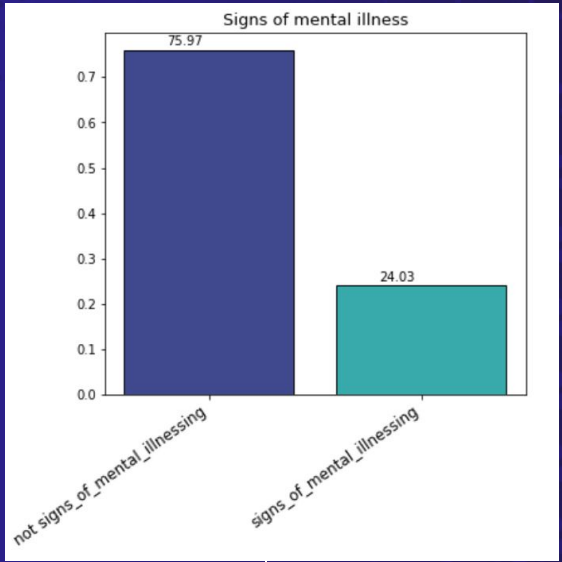
- What's the part of men and women?

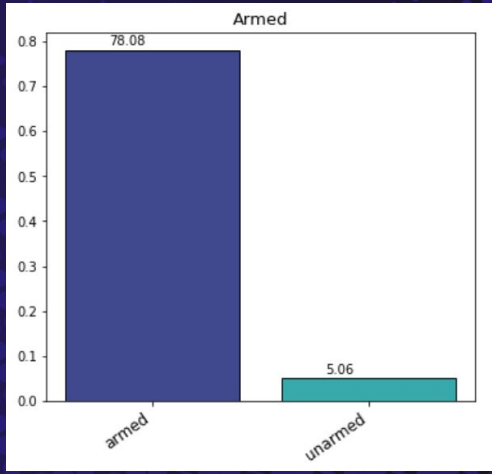


Data visualisation

- What condition was the victim in?

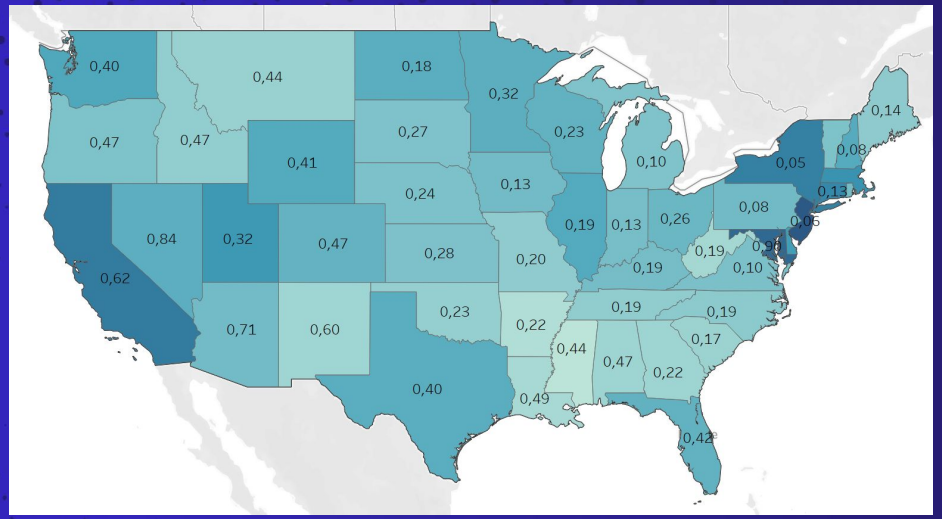


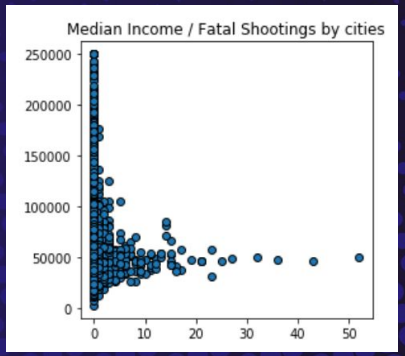




Data visualisation : median income

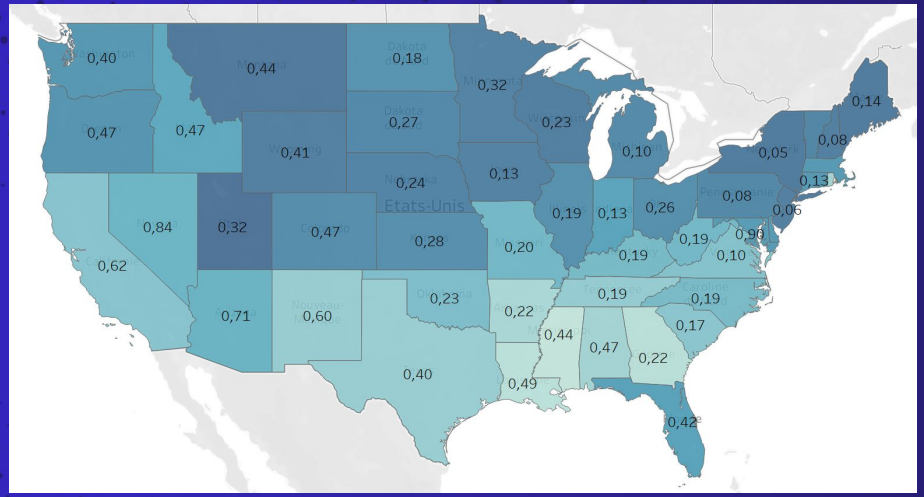
- Median Income & number of fatal shootings by state for 10,000 inhabitants

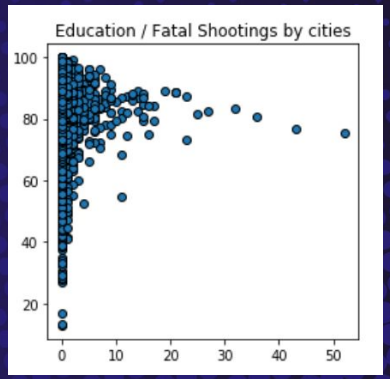


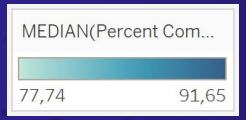


Data visualisation : education

- Percent completed high school & number of fatal shootings by state for 10,000 inhabitants



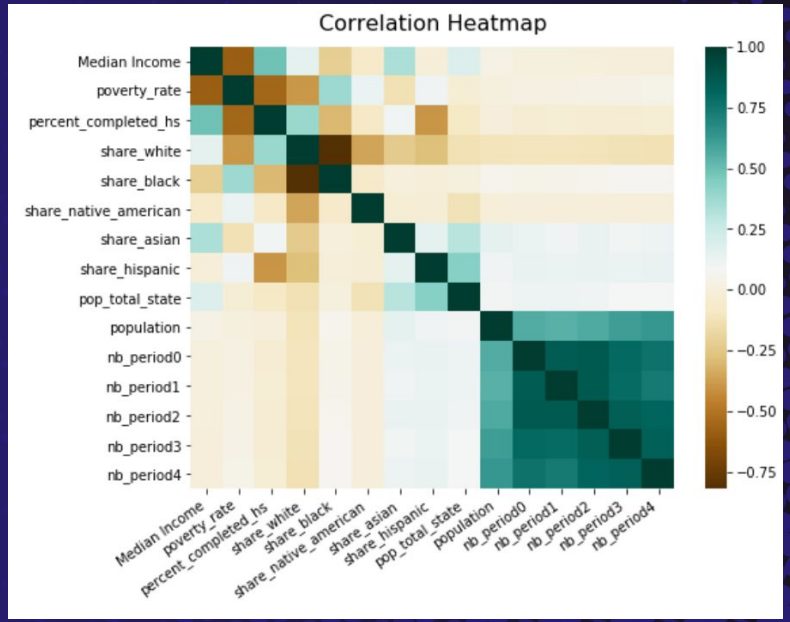




Data visualisation : Correlation

- Highly correlated columns

```
[['share_white', 'share_black', -0.82],
 ['nb_period0', 'nb_period1', 0.85],
 ['nb_period0', 'nb_period2', 0.87],
 ['nb_period1', 'nb_period2', 0.86],
 ['nb_period0', 'nb_period3', 0.81],
 ['nb_period1', 'nb_period3', 0.79],
 ['nb_period2', 'nb_period3', 0.85],
 ['nb_period0', 'nb_period4', 0.77],
 ['nb_period1', 'nb_period4', 0.74],
 ['nb_period2', 'nb_period4', 0.82],
 ['nb_period3', 'nb_period4', 0.84]]
```



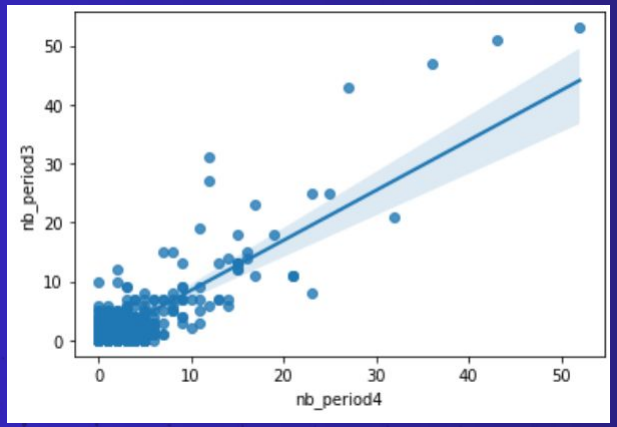
Machine Learning

Structure of the dataset

- Target : predict the number of fatal shootings by police by city between 2015 & 2017
- Features :
  - Median Income,
  - poverty_rate,
  - percent_completed_hs, *(education)*
  - share_white,
  - share_black,
  - share_native_american,
  - share_asian,
  - share_hispanic,
  - pop_total_state,
  - population,
  - nb_period0, (*number of fatal shootings between 2003 & 2005)*
  - nb_period1, (*number of fatal shootings between 2006 & 2008)*
  - nb_period2, (*number of fatal shootings between 2009 & 2011)*
  - nb_period3, (*number of fatal shootings between 2012 & 2014)*

Linear Regression

- Target : predict the number of fatal shootings by police by city between 2015 & 2017
- The number of fatal shootings between 2015 & 2017 is highly correlated to the other periods, and more especially between 2012 & 2014.



Dep. Variable:	y	**R-squared:**	0.707
Model:	OLS	**Adj. R-squared:**	0.707
Method:	Least Squares	**F-statistic:**	4.737e+04
Date:	Tue, 10 Nov 2020	**Prob (F-statistic):**	0.00
Time:	13:58:27	**Log-Likelihood:**	-16617.
No. Observations:	19639	**AIC:**	3.324e+04
Df Residuals:	19637	**BIC:**	3.325e+04
Df Model:	1		
Covariance Type:	nonrobust		

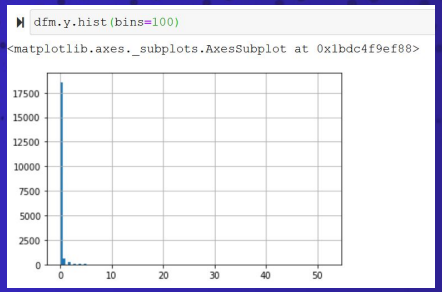
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0379	0.004	9.364	0.000	0.030	0.046
nb_period3	0.8336	0.004	217.648	0.000	0.826	0.841

Omnibus:	22637.964	**Durbin-Watson:**	1.856
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	20710837.613
Skew:	5.231	**Prob(JB):**	0.00
Kurtosis:	161.746	**Cond. No.**	1.13

Distribution of the target

- Distribution with the value '0'

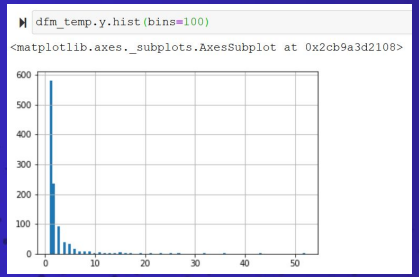
```
dfm.y.hist(bins=100)
<matplotlib.axes._subplots.AxesSubplot at 0x1bdc4f9ef88>
```

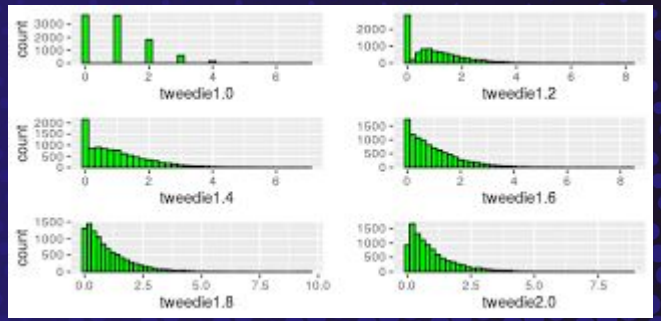


The distribution is not normal, right-skewed with a huge peak at '0'

- Distribution without the value '0'

```
dfm_temp.y.hist(bins=100)
<matplotlib.axes._subplots.AxesSubplot at 0x2cb9a3d2108>
```





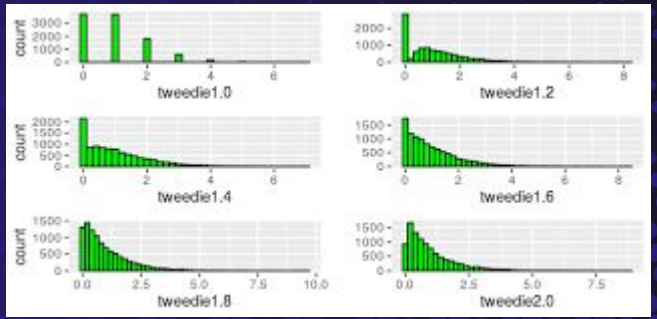
Tweedie distribution

What's Tweedie Regression?

- Tweedie distribution is :
  - a special case of exponential dispersion models
  - with a power parameter p, which determines the underlying target distribution according to the following table:

Power	Distribution
0	Normal
1	Poisson
(1,2)	Compound Poisson Gamma
2	Gamma
3	Inverse Gaussian

  - This model can be applied where you find a mixture of zeros and non-negative continuous data points.



Tweedie distribution

Tweedie Regression : comparison results

Results with Tweedie Regression

subset metric	train	test
D² explained	0.4966	-0.2237
mean Tweedie dev p=1.3000	2.4774	3.4891
mean Tweedie dev p=1.5000	1.6477	2.6495
mean Tweedie dev p=1.8000	0.9726	2.0093
mean Tweedie dev p=1.9000	0.8383	1.9063
mean Tweedie dev p=2.0000	0.7348	1.8454
mean Tweedie dev p=2.1000	0.6575	1.8228
mean Tweedie dev p=2.2000	0.6043	1.8378
mean Tweedie dev p=2.2200	0.5966	1.8454
mean Tweedie dev p=2.5000	0.6455	2.1810
mean Tweedie dev p=3.0000	7.3867	10.4233
mean abs. error	5.1705	4.9667
mean squared error	73.0003	60.0414

Results with Feature Selection (SFS)

subset metric	train	test
D² explained	0.1833	0.1052
mean Tweedie dev p=1.3000	4.6902	4.5944
mean Tweedie dev p=1.5000	3.0285	3.0203
mean Tweedie dev p=1.8000	1.6448	1.6760
mean Tweedie dev p=1.9000	1.3599	1.3940
mean Tweedie dev p=2.0000	1.1334	1.1682
mean Tweedie dev p=2.3000	0.7160	0.7497
mean Tweedie dev p=2.3100	0.7082	0.7420
mean Tweedie dev p=2.3200	0.7009	0.7347
mean Tweedie dev p=2.5000	0.6641	0.7055
mean Tweedie dev p=3.0000	7.1544	7.8260
mean abs. error	7.5226	7.0451
mean squared error	115142.3160	113.7477

Results with Dummy Regression

subset metric	train	test
D² explained	-0.5407	-0.6426
mean abs. error	8.2781	7.5482
mean squared error	195.2634	145.6299

- Dummy regression is a simple baseline to compare with other (real) regressors.
- Different strategies, but here :
  → strategy='mean' : always predicts the mean of the training set

Conclusion

- The best method to predict the number of fatal shootings by a police officer is **Linear Regression**.
- It seems that poverty, size of cities, share of ethnic groups or education have no impact on the number of fatal police shootings.
- **The prevention should be done in the city already affected in the past.**

Thanks!

https://github.com/AuroreSupiot/Fatal_Police_Shootings_USA