

Process book

Group 23

Jonas Kouwenhoven

Martijn Maiwald

Folkert Stijnman

Casper Wortmann

Juni 2018




## Update 6 juni 2018

- Een aantal lijnen van de DataFrame wilden niet laden doordat een `|` ineens `||` werd
- Sommige lines gaven problemen tijdens het preprocessen doordat er quotes in quotes stonden “...;tekst’....”
- Regels kunnen isoleren en vervolgens alle regels met 1 kleine aanpassingen kunnen lezen en schrijven
- In de *GitHub* opstart gids werd vertelt dat we een private *repository* moesten creëren, dit koste geld en na overleg met Nick zijn we overgestapt op een *public repository*
- Jonas maakte de overstap van CS50 naar normaal programmeren, programma’s runnen verliep nu veel sneller.

## Update 8 juni 2018

- Moeite met het terugzetten van de csv in JSON, verschillende manieren geprobeerd
- Kolommen met *URL*’s zijn er uitgehaald, deze gaan we niet gebruiken.
- Was maar 1 kleine aanpassing nodig
- Hebben we bepaalde data extra nodig of juist niet nodig.

 dataset.json	13 jun. 2018 15:17	202,4 MB	JSON
--	--------------------	----------	------

- De JSON file was te groot om op *GitHub* te zetten, dus moest er apart bij gezet worden

## Update 11 juni 2018

- Bepaalde data was echt onnodig en vervolgens uit de set gehaald.
- Lege *entries* vervangen met “NaN”
- Leeftijden die ongeloofwaardig zijn (311 en 201) zijn er uitgehaald.
- In regel 997511 was er een character dat niet gelezen kon worden csv parser, dit is eruit gehaald.

```
{'0': 'Julian Sims'}
{'0': 'Bernard Gillis'}
{'0': 'Damien Bell', '1': 'Desmen Noble', '2': 'Herman Seagers', '3': 'Ladd Tate Sr', '4': 'Tallis Moore'}
{'0': 'Stacie Philbrook', '1': 'Christopher Ratliffe', '2': 'Anthony Ticali', '3': 'Sonny Arculeta'}
{'0': 'Danielle Imani Jamelison', '1': 'Maurice Eugene Edmonds, Sr.', '2': 'Maurice Edmonds I', '3': 'Sandra Palmer'}
{'0': 'Rebeika Powell', '1': 'Kayetie Melchor', '2': 'Misty Nunley', '3': 'Julie Jackson', '4': 'James Poore', '5': 'Cedric Poore'}
{'0': 'Greg Griego', '1': 'Sara Griego', '2': 'Zephania Griego', '3': 'Jael Griego', '4': 'Angelina Griego', '5': 'Nehemiah Griego'}
NaN
NaN
{'0': 'Deshaun Jones'}
```

## Update 13 juni 2018

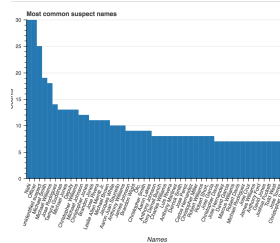
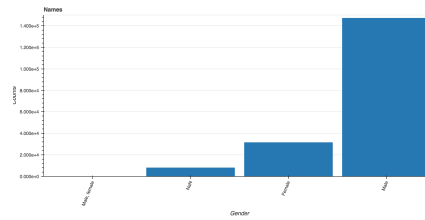
- Oplossing gezocht voor het omzetten van participant data naar dictionaries

```
In [11]: df_p = pd.read_json('dataset_incidents.json', lines=True)
          df_p[:10]
```

```
Out[11]:
```

	incident_id	participant_age	participant_age_group	participant_gender	participant_name	participant_relationship	participant_
0	461105	20.0	Adult 18+	Male	Julian Sims	NaN	Ar
1	461105	NaN	Adult 18+	Male	None	None	I
2	461105	NaN	Adult 18+	Male	None	None	I
3	461105	NaN	Adult 18+	Female	None	None	I
4	461105	NaN	Adult 18+	None	None	None	I
5	460726	20.0	Adult 18+	Male	Bernard Gillis	NaN	
6	460726	NaN	Adult 18+	None	None	None	I
7	460726	NaN	Adult 18+	None	None	None	I
8	460726	NaN	Adult 18+	None	None	None	I
9	460726	NaN	None	None	None	None	

- Uiteindelijk gekozen voor een extra DataFrame voor de participants, dit duurde een nacht en kostte enorm veel stroom.
- Begonnen met de eerste plotjes.



```
df_suspect = df_p.loc[(df_p['participant_type'] == "Subject-Suspect")]
df_relationships = df_suspect.groupby(["participant_relationship"])
# df_relationships = df_relationship.drop(df_relationship.index[11])

df_relationships.describe()

# df_relation_age = df_relationship.set_index('participant_age')
df_relations = df_relationship.groupby('participant_relationship')
df_relations.describe()
## for index, row in df_relationship.iterrows():

##     print(row)

# df_aquaintance = df_relationship.loc[df_relationship['participant_relation'] == 'acquaintance']
# # print(df_aquaintance)
# df_aquaintance
# counter = 0
# counter1 = 0
# incidentcount = 0

zus = []
dit = []
dat = []
for line in df_p['participant_type']:
    if line == 'Victim':
        counter += 1
        dit.append(counter)
    else:
        dit.append(counter)
for line in df_p['participant_type']:
    if line == 'Subject-Suspect':
        counter1 += 1
```

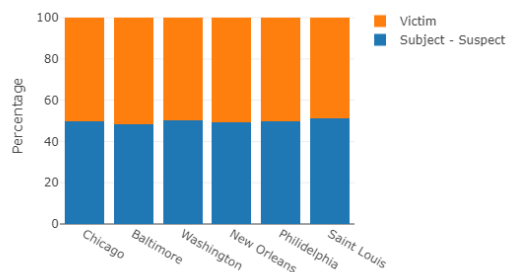
## Update 15 juni 2018

- Gekeken naar de beste manier om een groupby DataFrame om te zetten naar een grafiek
- Moeite met hoe precies data van een bepaalde DF te verwerken
- Probleem dat waarden te vaak uitgerekend werd, waardoor de plot steeds andere waarden had. De berekening in een eigen *cell* gezet waardoor het probleem werd verholpen

Out[9]:

	count					participant_age								
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	
participant_relationship														
Aquaintance	65.0	13.030769	11.468666	1.0	4.00	9.0	22.00	38.0	65.0	41.953846	19.505953	1.0	26.00	
Armed Robbery	85.0	51.764706	69.337279	1.0	3.00	24.0	67.00	313.0	85.0	44.070588	25.388781	0.0	23.00	
Co-worker	44.0	2.590909	1.435727	1.0	1.00	3.0	3.00	6.0	44.0	41.386364	14.231011	18.0	29.75	
Drive by - Random victims	20.0	1.650000	0.745160	1.0	1.00	1.5	2.00	3.0	20.0	31.200000	13.813037	9.0	21.75	
Family	94.0	33.829787	19.766181	1.0	15.25	37.0	47.75	77.0	94.0	46.500000	27.279418	0.0	23.25	
Friends	72.0	13.277778	14.863911	1.0	3.00	5.5	18.25	60.0	72.0	39.680556	21.292614	3.0	21.75	
Gang vs Gang	50.0	9.900000	13.286866	1.0	1.00	4.0	13.50	52.0	50.0	31.680000	17.523675	1.0	18.25	
Home Invasion - Perp Does Not Know Victim	76.0	7.947368	9.102226	1.0	1.75	3.5	11.25	37.0	76.0	42.526316	22.735273	2.0	23.75	
Home Invasion - Perp Knows Victim	56.0	6.875000	7.163195	1.0	1.75	4.5	11.25	31.0	56.0	40.714286	19.139407	2.0	25.75	
Mass shooting - Perp Knows Victims	10.0	1.100000	0.316228	1.0	1.00	1.0	1.00	2.0	10.0	43.400000	24.404918	16.0	21.75	
Mass shooting - Random victims	9.0	1.222222	0.440959	1.0	1.00	1.0	1.00	2.0	9.0	35.444444	23.335119	14.0	19.00	
Neighbor	77.0	7.610390	5.582293	1.0	2.00	7.0	12.00	25.0	77.0	45.896104	23.239930	2.0	27.00	
Significant others - current or former	91.0	33.494505	33.256035	1.0	3.50	21.0	58.00	105.0	91.0	47.000000	26.886593	0.0	24.50	

- Nog veel “test code” tussen de echte code die we gaan gebruiken



```
#####
# testabit:
#####

# first 5 lines
dataset[15]

# count
# dataset['n_killed'].value_counts()

# # column
dataset[dataset['state'] == 'California']

# # check if contains
dataset[dataset['state'].str.contains("a")]['state'].unique()

#doden per staat, met gemiddelden etc.
State_kill = dataset["n_killed"].groupby(dataset['state'])
State_kill.describe()
```

## Update 18 juni 2018

- Meerdere manieren gebruikt om te checken dat de 50/50 verdeling van suspect en victim over alle incidenten klopt.
- Er waren wat problemen met het toekennen van kleuren aan bepaalde variabelen of lijnen.

	State	Census	Estimation	2010	2011	2012	2013	2014	2015	2016	2017
0	Alabama	4.779.736	4.780.135	4.785.579	4.798.649	4.813.946	4.827.660	4.840.037	4.850.858	4.860.545	4.874.747
1	Alaska	710.231	710.249	714.015	722.259	730.825	736.760	736.759	737.979	741.522	739.795
2	Arizona	6.392.017	6.392.309	6.407.002	6.465.488	6.544.211	6.616.124	6.706.435	6.802.262	6.908.642	7.016.270
3	Arkansas	2.915.918	2.916.031	2.921.737	2.938.640	2.949.208	2.956.780	2.964.800	2.975.626	2.988.231	3.004.279
4	California	37.253.956	37.254.518	37.327.690	37.672.654	38.019.006	38.347.383	38.701.278	39.032.444	39.296.476	39.536.653
5	Colorado	5.029.196	5.029.325	5.048.029	5.116.411	5.186.330	5.262.556	5.342.311	5.440.445	5.530.105	5.607.154

- Extra dataset toegevoegd met de bevolkingen van elke staat, met deze dataset konden we plots genereren per inwoner. Voor percentuale resultaten.

```
df['month'] = df['date'].dt.month
df['year'] = df['date'].dt.year
```

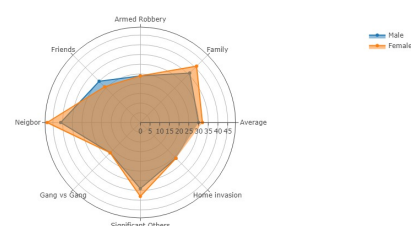
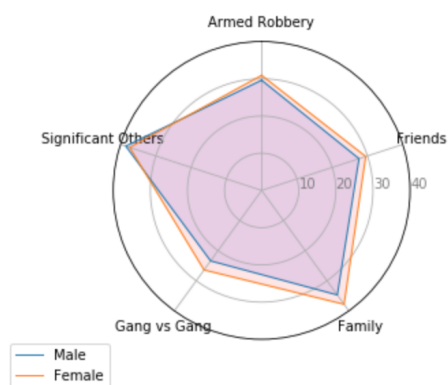
## Update 20 juni 2018

year	month	day
2013.0	1.0	1.0

- Legenda toevoegen aan de kaart van Amerika was lastig, dit hebben we ernaast gefotoshopt.
- Martijn zijn knie werd verbrijzeld tijdens boksen, waardoor afspreken op Science Park moeilijker werd
- Apart de dagen, maanden en jaren toegevoegd zodat we konden zoeken op alleen dag of maand of jaar.

## Update 22 juni 2018

- Tegen een aantal beperkingen van Bokeh aangelopen, dit vervolgens met Matplotlib opgelost
- Helaas was Matplotlib ook niet erg goed, en niet van hoge kwaliteit.



# VS.

## Update 25 juni 2018

- Vanuit Matplotlib kon geen output HTML file gemaakt worden, we hebben dus screenshots gemaakt van de plot.
- Wat standaard typefoutjes, die het css/html file wat vertraging opleverde, drie keer goed lezen en het was opgelost.



Month\_incidents.html

HTML document - 105,2 MB  
Aanmaak dinsdag 26 juni 2018 om 14:09  
Bewerking woensdag 27 juni 2018 om 13:10  
Geopend --

## Update 27 juni 2018

- De HTML files van de plotjes waren erg groot, dit vertraagde de website aanzienlijke. Van een aantal van deze plotjes zijn screenshots gemaakt wat minder werkgeheugen vraagt.
- Andere HTML files zijn zelfs weggelaten omdat de bestanden te groot werden. 105.2 MB !!!



## Update 29 juni 2018

- Te horen gekregen dat plotly beter werkt dan Bokeh of Matplotlib
- Moeite met het creëren van een folder op Git

## Research idea's

- 1) Correlation between surname's and incidents.
- 2) Incidents correlated to universities/campus/students with the help of keywords like: 'campus', 'university', 'school', 'student', 'professor', 'teacher'.
- 3) Participant relationships; family, friends, co-workers, gang vs gang, acquaintance, significant others.
- 4) Potential forecasting of incidents, the course of gun violence in the future.
- 5) Woman/man distribution in family incidents.

## Steps taken

- 1) **From CSV to JSON:** The provided dataset was in a CSV Format, but the discovery was made that Panda<sup>1</sup> would work better with a JSON format.
- 2) **Deleting columns:** Columns containing url's were deleted, because it looked messy and they were not needed.
- 3) **Tidy up:** By removing unnecessary enters and punctuation the data set got a better appearance.
- 4) **Empty cells:** Because empty cells could cause some problems during plotting and programming, we replaced them all with 'NaN' or 'Unknown'.
- 5) **Removing ages:** We removed all the age information above a 100 years. These were only a few, including one where someone was stated to be 209 and 311.

## Questions

We will be focusing on the level of federal involvement in gun violence incidents. Federal involvement includes the rate of suspects that include police officers, deputies and agents.

---

<sup>1</sup><https://pandas.pydata.org/>



- 1) How has federal involvement developed over time since 2013, per state? Where are they concentrated? Could we find any correlation between these concentrations and other incident characteristics?
- 2) Does federal involvement show any correlation with age? State? Gender? The amount of mass shootings? Accidental shootings?
- 3) Is there a correlation between the percentage of suspects that are male or female in a state, and the percentage of times the victim in an incident with federal involvement is male or female?
- 4)