Reproducible Research for crimes datset

Name Surname
2020.02.08

Synopsis

The main aim of this notebook is going to be a very high-level Exploratory Data A

Loading the data

Firstly let's import necessary packages.

```
# modules we'll use
import numpy as np
import pandas as pd
import seaborn as sns
from matplotlib import pyplot as plt
from wordcloud import WordCloud, STOPWORDS
%matplotlib inline


def load_data(file):
    url = 'https://raw.githubusercontent.com/HikkaV/VNTU-ML-Courses/master/assignme
    try:
        df = pd.read_csv('../files/{}'.format(file))
    except:
        df = pd.read_csv(url)
    return df


# read in our data
crime = load_data('crime.csv')

# set seed for reproducibility
np.random.seed(0)
```

Results

```
crime.head()
crime.head()
```

	Dates	Category	Descript	DayOfWeek	PdDistrict
0	5/13/2015 23:53	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN
1	5/13/2015 23:53	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN
2	5/13/2015 23:33	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN
3	5/13/2015 23:30	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN
4	5/13/2015 23:30	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	PARK

```
len(crime)
```

```
835
```

It's quite a small and concise dataset but there should be quite a lot that can b
So let's get started from the left to the right with the "Category" column.

Most common categories of crime committed:

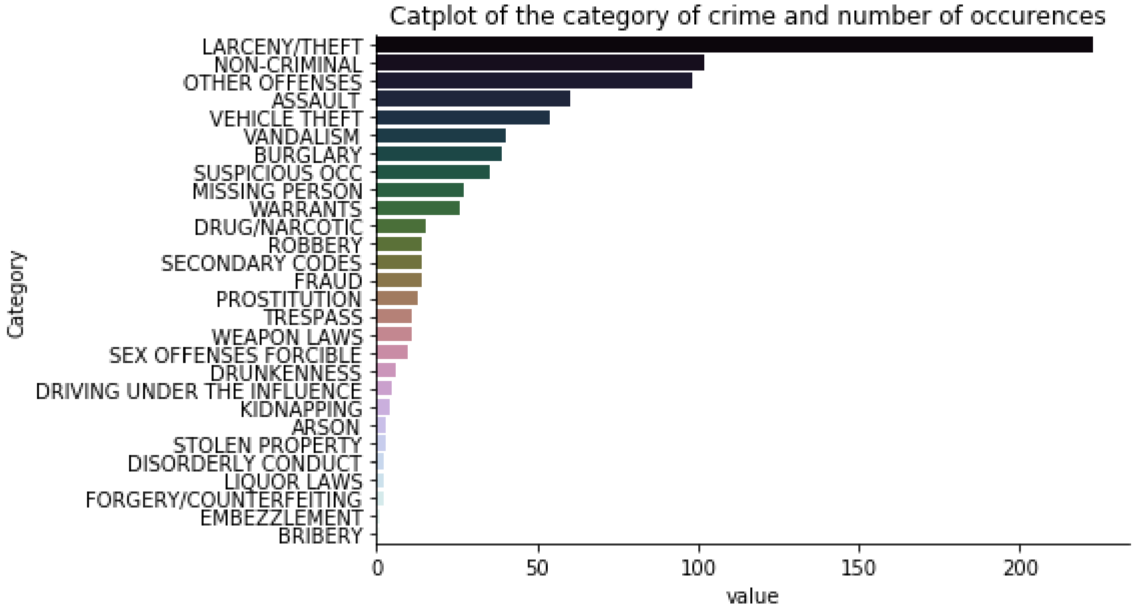
```
crime.Category.value_counts()
```

```
LARCENY/THEFT                 223
NON-CRIMINAL                  102
OTHER OFFENSES                 98
ASSAULT                        60
VEHICLE THEFT                  54
VANDALISM                      40
BURGLARY                       39
SUSPICIOUS OCC                 35
MISSING PERSON                 27
WARRANTS                       26
DRUG/NARCOTIC                  15
ROBBERY                        14
SECONDARY CODES                14
FRAUD                          14
PROSTITUTION                   13
TRESPASS                       11
WEAPON LAWS                    11
SEX OFFENSES FORCIBLE          10
DRUNKENNESS                     6
DRIVING UNDER THE INFLUENCE     5
KIDNAPPING                      4
ARSON                           3
STOLEN PROPERTY                 3
DISORDERLY CONDUCT              2
LIQUOR LAWS                     2
FORGERY/COUNTERFEITING          2
EMBEZZLEMENT                    1
BRIBERY                         1
Name: Category, dtype: int64
```

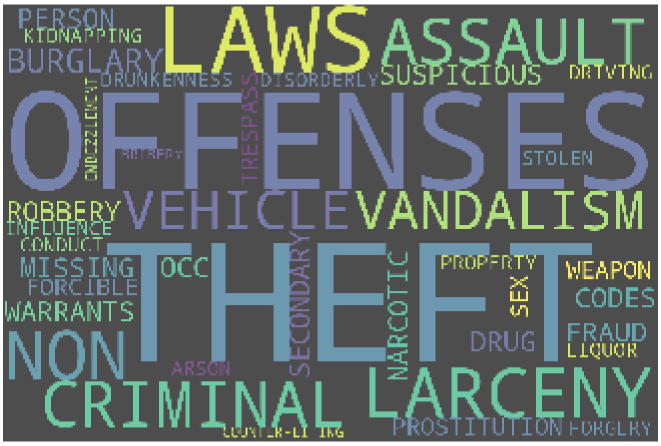
```
category = pd.DataFrame(list(zip(crime.Category.value_counts().index,crime.Categor
```

```
sns.catplot(x='value', y = 'Category', data=category, kind="bar", height=4.25, asp
plt.title('Catplot of the category of crime and number of occurences ')
```

```
Text(0.5, 1, 'Catplot of the category of crime and number of occurences ')
```



```
wordcloud = WordCloud(
                    stopwords=STOPWORDS,
                    background_color='black',
                    width=1200,
                    height=800
                    ).generate(" ".join(category['Category'].values))

plt.imshow(wordcloud, alpha=0.7)
plt.axis('off')
plt.show()
```



From all data visualizations result is the same: THEFT is most common crime in th

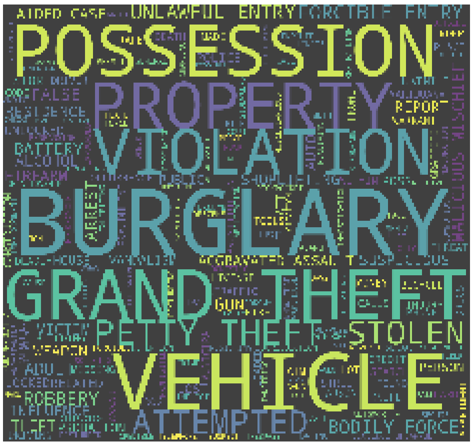
Most common crimes carried out per it's description:

```
crime.Descript.value_counts()
```

```
GRAND THEFT FROM LOCKED AUTO             76
STOLEN AUTOMOBILE                        30
PETTY THEFT OF PROPERTY                  30
AIDED CASE, MENTAL DISTURBED             30
BATTERY                                  22
                                         ..
ATTEMPTED GRAND THEFT PURSESNATCH         1
EMBEZZLED VEHICLE                         1
PROBATION VIOLATION, DV RELATED           1
ATTEMPTED THEFT FROM UNLOCKED VEHICLE     1
TRANSPORTATION OF MARIJUANA               1
Name: Descript, Length: 165, dtype: int64
```

This column contains a lot more detailed information about the type of the crime
And right-away we can observe that Grand Theft Auto is the most common crime desc
Again, we create another dataframe which will make it convenient for the plotting

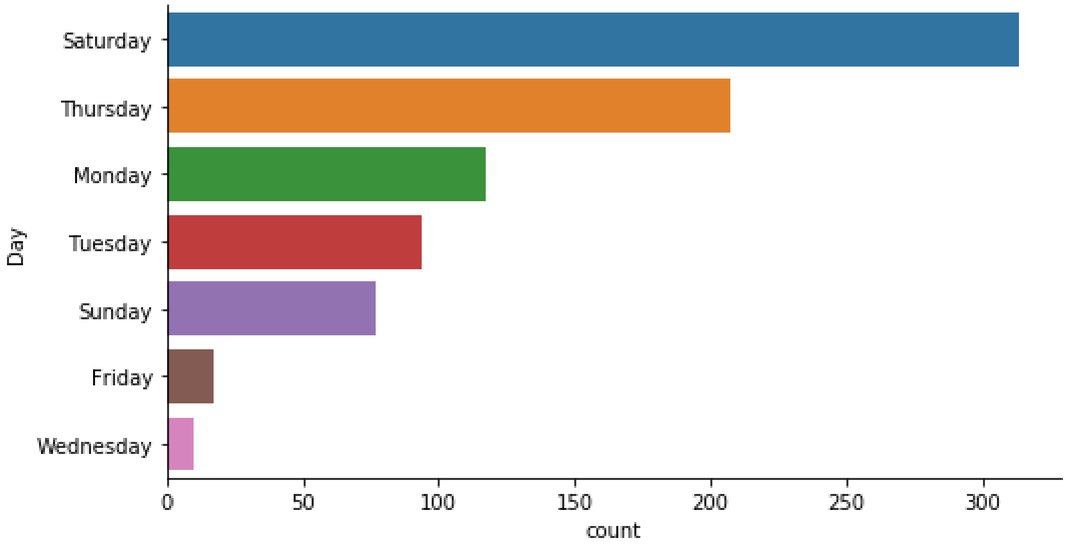
```
descript = pd.DataFrame(list(zip(crime.Descript.value_counts().index,crime.Descrip
descript_cloud = WordCloud(
                          stopwords=STOPWORDS,
                          background_color='black',
                          width=1500,
                          height=1400
                         ).generate(" ".join(descript['Description'].values))


plt.imshow(descript_cloud,alpha=0.75)
plt.axis('off')
plt.show()
```



Day on which there is most crimes:

```
DOW = pd.DataFrame(list(zip(crime.DayOfWeek.value_counts(),crime.DayOfWeek.value_c

sns.catplot(x="count", y="Day", data = DOW, kind="bar", height=4, aspect=1.9)
```

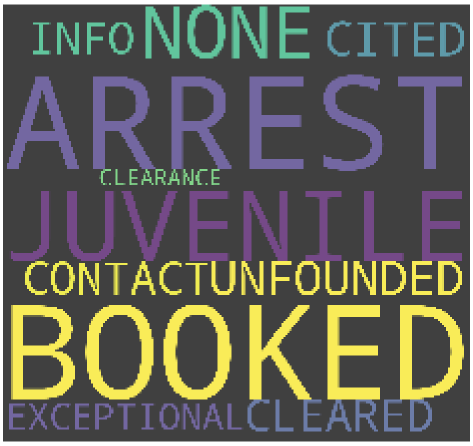
```
<seaborn.axisgrid.FacetGrid at 0x7f5eec72e828>
```



Saturday is most dangerous day according to this dataset.

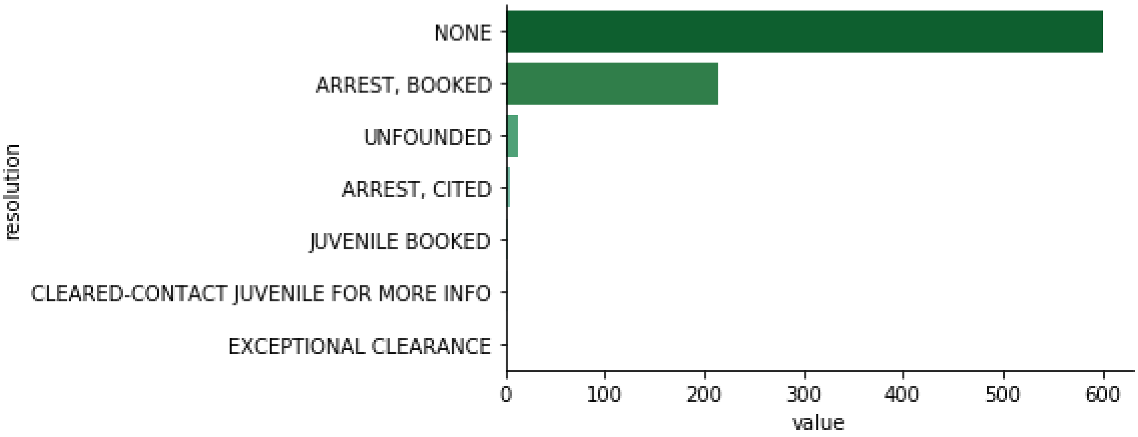
How good crimes are being resolved:

```
Resolution = pd.DataFrame(list(zip(crime.Resolution.value_counts().index,crime.Res

rescloud = WordCloud(
                    stopwords=STOPWORDS,
                    background_color='black',
                    width=1500,
                    height=1400
                    ).generate(" ".join(Resolution['resolution'].values))

plt.imshow(rescloud, alpha=0.75)
plt.axis('off')
plt.show()
```



```
sns.catplot(x='value' , y = 'resolution', data=Resolution, kind="bar", height=3.25
```

```
<seaborn.axisgrid.FacetGrid at 0x7f5eec52fac8>
```



NONE means that most of crimes in this dataset are not resolved...

SUMMARY

```
We have performed multiple actions in order to test and understand crime dataset
Even few simple steps discover many interesting facts regarding the data, for exa
And it's not a good idea to have a walk at Saturday.
```