File structure

```
project01/
+-- Dockerfile
+-- requirements.txt
+-- src/
+-- +-- main.py
+-- assets/
+-- +-- kibanadashboard.png
+-- README
```

Python scripting

Packages used

argparse
math
os
sodapy
datetime
elseticsearch

Python script

```
import argparse
import math
import os
from sodapy import Socrata
from datetime import datetime
from elasticsearch import Elasticsearch

# set command line arguments
parser = argparse.ArgumentParser()
parser.add_argument('--page_size', type=int,
                    help='how many rows to get per page', required=True)
parser.add_argument('--num_pages',type=int,
                    help='how many pages to get in total')
args = parser.parse_args()

#set environment
DATASET_ID = os.environ["DATASET_ID"]
APP_TOKEN = os.environ["APP_TOKEN"]
ES_HOST = os.environ["ES_HOST"]
ES_USERNAME = os.environ["ES_USERNAME"]
ES_PASSWORD = os.environ["ES_PASSWORD"]

# connect to API  and count the number of rows
client = Socrata("data.cityofnewyork.us", APP_TOKEN ,timeout=50000)
number_of_rows=int(client.get(DATASET_ID, select='COUNT(*)')[0]['COUNT'])

# specify num_pages argument when it is 0
if args.num_pages == 0:
    args.num_pages= math.ceil(number_of_rows/args.page_size)
else:
    args.num_pages= args.num_pages

# connect to Elasticsearch and create an index (mapping is optional)
```

```
if __name__ == '__main__':
    try:
        es = Elasticsearch(ES_HOST,http_auth=(ES_USERNAME,ES_PASSWORD))
        es.indices.create(index='nycparking')

    except Exception:
        print("Index already exists! Skipping")

# get data, transform the format and load it to Elasticsearch
    for page in range(0,args.num_pages):
        offset= page* args.page_size
        results = client.get(DATASET_ID, limit=args.page_size, offset=offset)
        for result in results:
            try:
                result["issue_date"] = str(result["issue_date"])
                result["issue_date"] = datetime.strptime(result["issue_date"],"%m/%d/%Y").date()
                result["precinct"] = int(result["precinct"])
                result["fine_amount"] = float(result["fine_amount"])
                result["reduction_amount"] = float(result["reduction_amount"])
            except Exception as e:
                print(f"Error!: {e}, skipping row: {result}")
                continue
            try:
                es.index(index='nycparking',doc_type='parking', body=result)
            except Exception as e:
                print(f"Failed to insert in ES: {e}, skipping row: {result}")
                continue
```

Docker file

```
FROM python:3.7

WORKDIR /app

COPY requirements.txt /app

RUN pip install -r requirements.txt

COPY src/ /app

ENTRYPOINT ["python", "src/main.py"]
```

Terminal

Build the image:

```
docker build -t project01:1.0 .
```

Run the image:

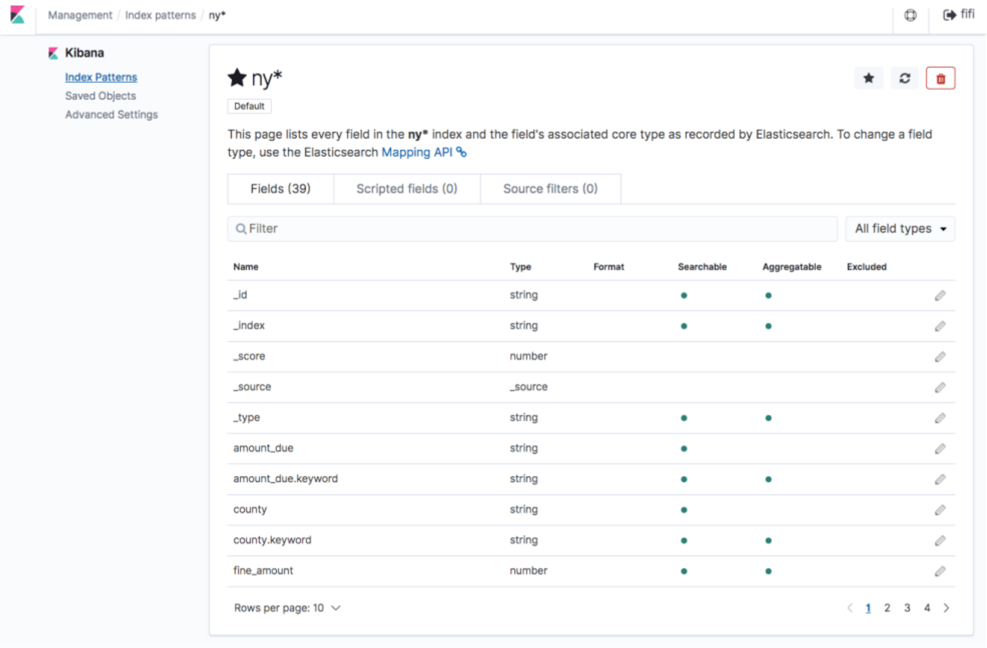
```
docker run -d -v $(pwd):/app -e DATASET_ID= {DATASET_ID} -e APP_TOKEN={APP_TOKEN} -e
ES_HOST={ ES_HOST} -e ES_USERNAME={ES_USERNAME} -e ES_PASSWORD={ ES_PASSWORD}
project01:1.0 --page_size=1000 --num_pages=1000
```

--page_size: This command line argument is required. It will ask for how many records to request from the API per call.

--num_pages: This command line argument is optional. If not provided, script should continue requesting data until the entirety of the content has been exhausted. If this argument is provided, continue querying for data num_pages times.

Visualizing and Analysis on Kibana

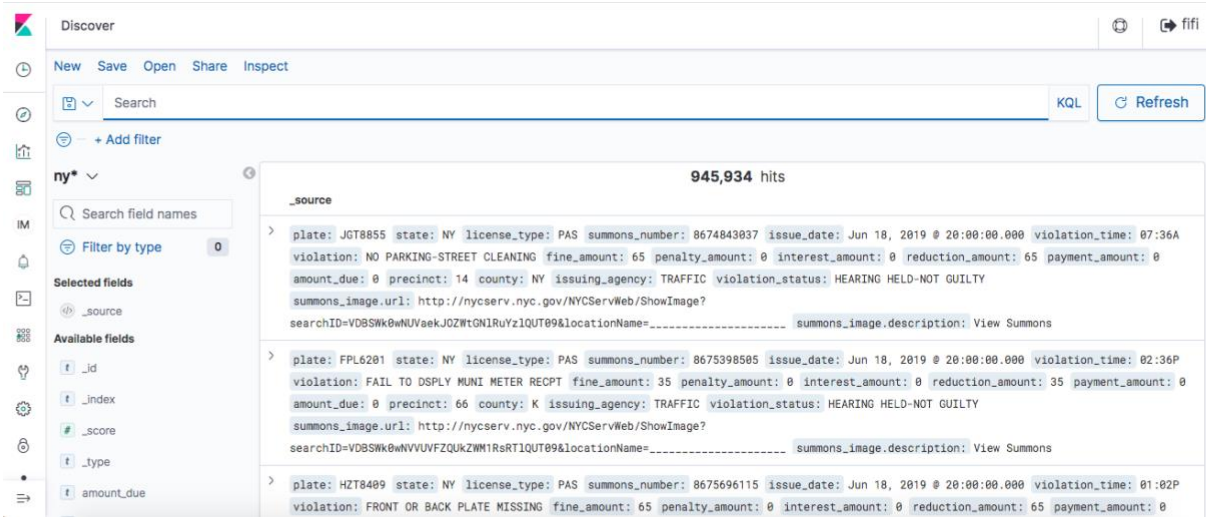
Define index pattern



This page lists every field in the **ny*** index and the field's associated core type as recorded by Elasticsearch. To change a field type, use the Elasticsearch Mapping API

Name	Type	Format	Searchable	Aggregatable	Excluded
_id	string		●	●	
_index	string		●	●	
_score	number				
_source	_source				
_type	string		●	●	
amount_due	string		●		
amount_due.keyword	string		●	●	
county	string		●		
county.keyword	string		●	●	
fine_amount	number		●	●	

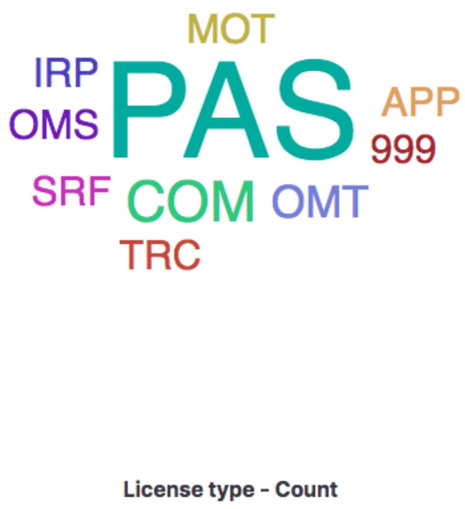
Overall, 945,934 records are loaded into the Kibana because there are some missing values.



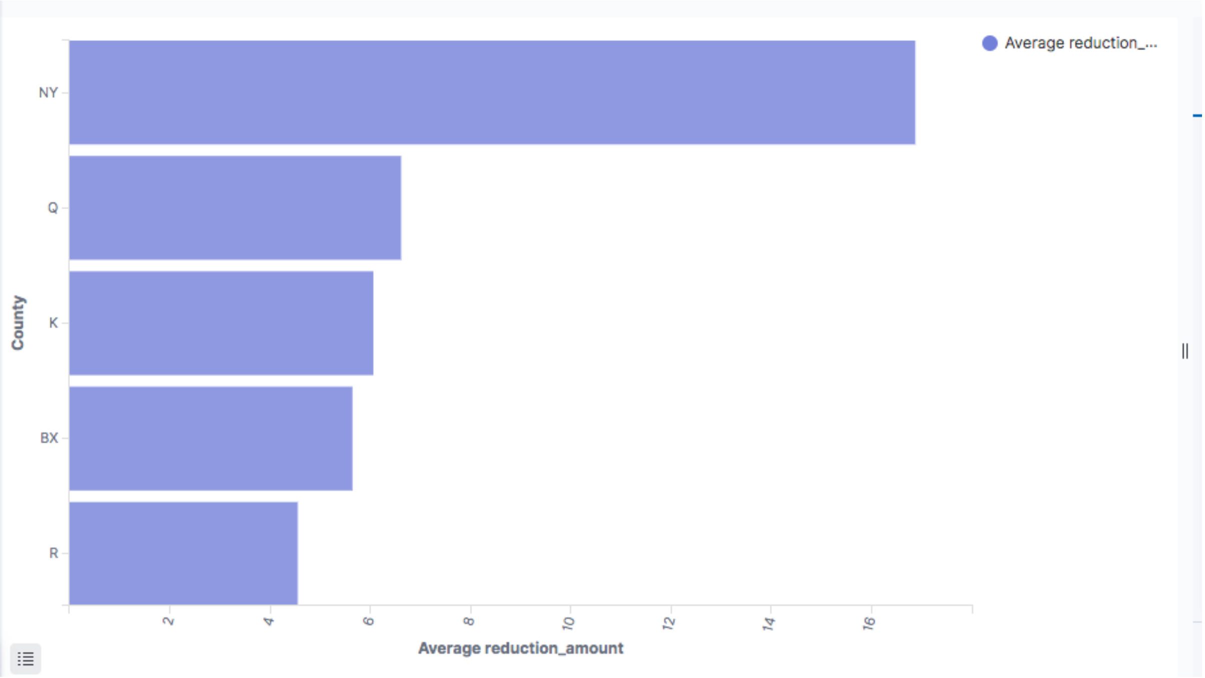
945,934 hits

```
_source
plate: JGT8855 state: NY license_type: PAS summons_number: 8674843037 issue_date: Jun 18, 2019 @ 20:00:00.000 violation_time: 07:36A
violation: NO PARKING-STREET CLEANING fine_amount: 65 penalty_amount: 0 interest_amount: 0 reduction_amount: 65 payment_amount: 0
amount_due: 0 precinct: 14 county: NY issuing_agency: TRAFFIC violation_status: HEARING HELD-NOT GUILTY
summons_image.url: http://nycserv.nyc.gov/NYCServWeb/ShowImage?
searchID=VDBSWk0wNUVaekJOZWtGNlRuYzlQUT09&locationName=_____________________ summons_image.description: View Summons

plate: FPL6201 state: NY license_type: PAS summons_number: 8675398505 issue_date: Jun 18, 2019 @ 20:00:00.000 violation_time: 02:36P
violation: FAIL TO DSPLY MUNI METER RECPT fine_amount: 35 penalty_amount: 0 interest_amount: 0 reduction_amount: 35 payment_amount: 0
amount_due: 0 precinct: 66 county: K issuing_agency: TRAFFIC violation_status: HEARING HELD-NOT GUILTY
summons_image.url: http://nycserv.nyc.gov/NYCServWeb/ShowImage?
searchID=VDBSWk0wNVVUVFZQUkZWM1RsRT1QUT09&locationName=_____________________ summons_image.description: View Summons

plate: HZT8409 state: NY license_type: PAS summons_number: 8675696115 issue_date: Jun 18, 2019 @ 20:00:00.000 violation_time: 01:02P
violation: FRONT OR BACK PLATE MISSING fine_amount: 65 penalty_amount: 0 interest_amount: 0 reduction_amount: 65 payment_amount: 0
```

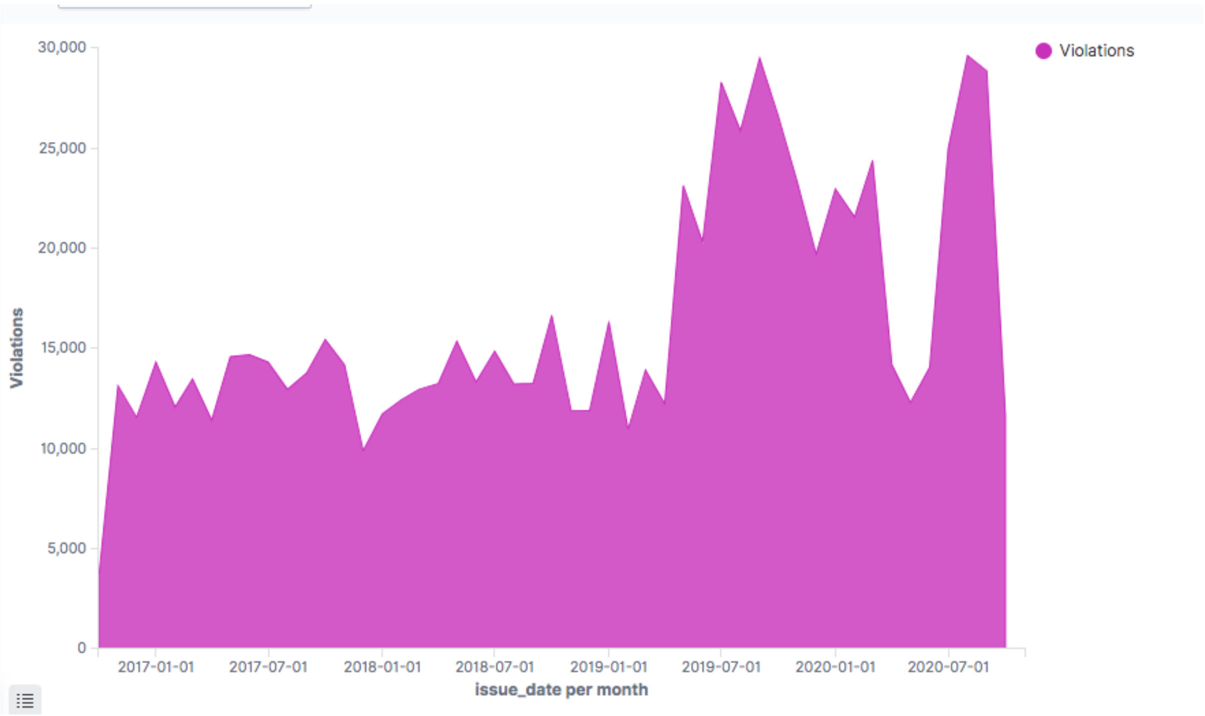
This word cloud shows the most frequently seen license type for violations. The bigger the word, the higher the frequency.



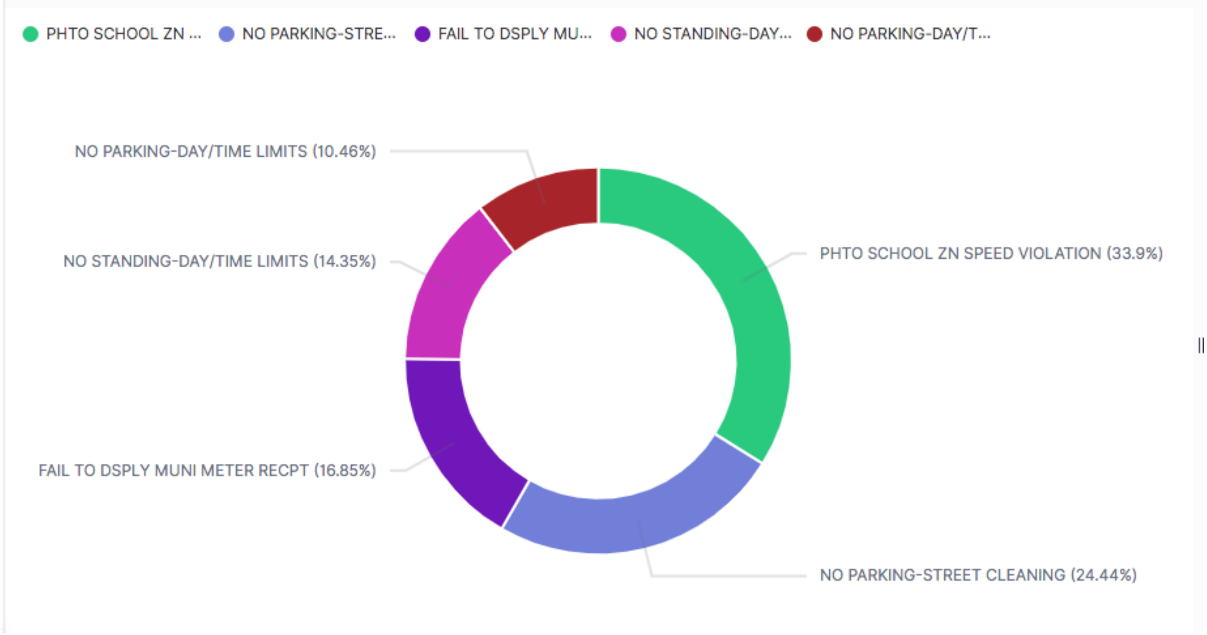
This horizontal bar chart shows the top 5 average reduction amount by county.



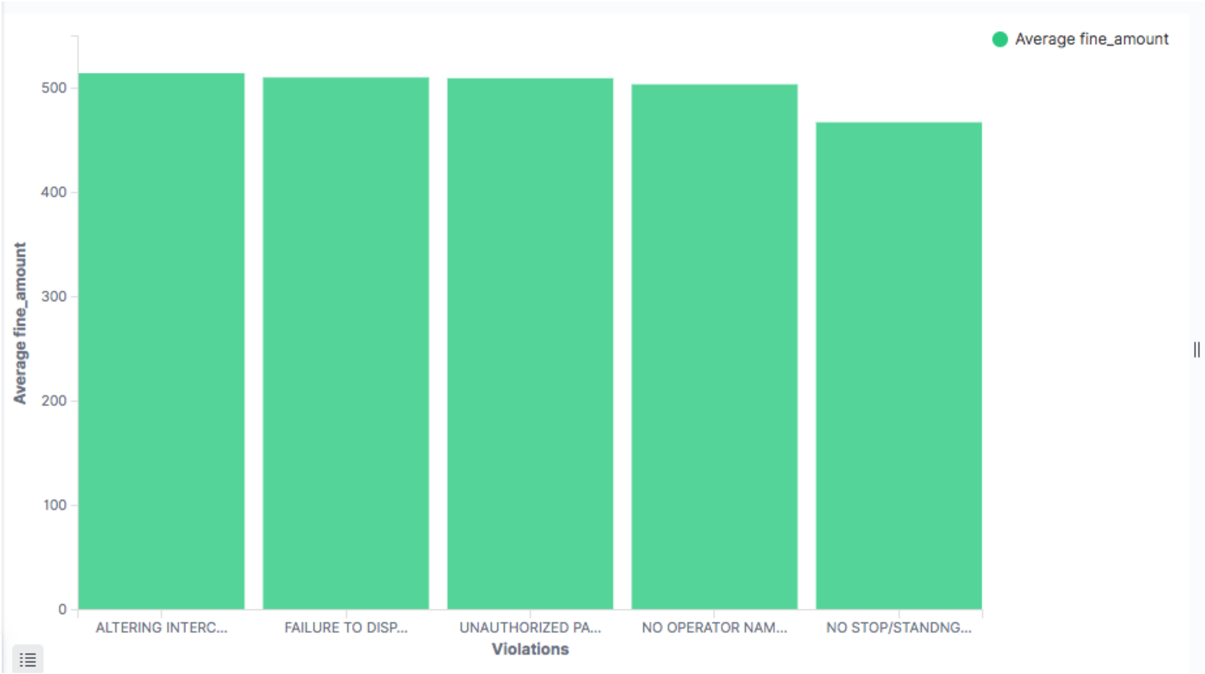
Number of violations by month over the last 3 years. More violations were recorded since 2019.



Top 5 violations are in the pie chart. PHTO school zone speed violations is the highest, accounting for 33.9%.



Top 5 violations with most average fine amount.


● Average fine_amount
Average fine_amount
ALTERING INTERC...
FAILURE TO DISP...
UNAUTHORIZED PA...
NO OPERATOR NAM...
NO STOP/STANDNG...
Violations