

NYC Traffic Data Violation Analysis

1. Connect to the API:

Below is the python code used to pull recent 10,000 rows from the Data Source: https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89/about_data
And csv file was downloaded to local.

Snippet of Python code

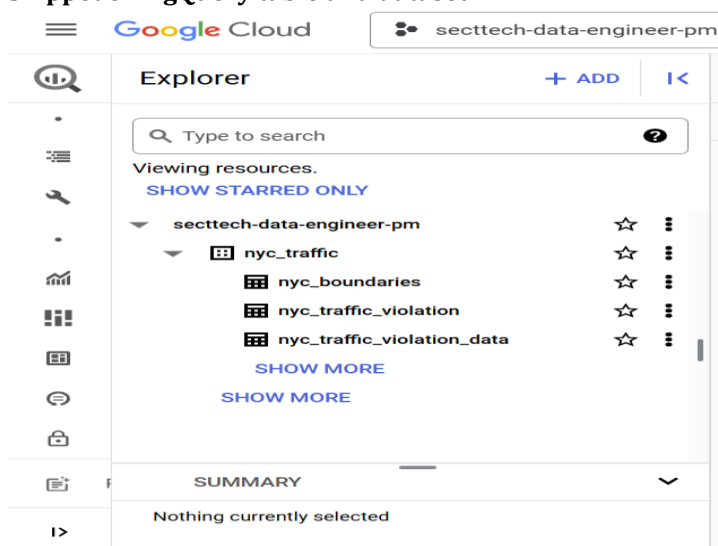
```
1  import pandas as pd
2  import requests
3
4  # Define the URL for the NYC Open Data API
5  dataset_url = "https://data.cityofnewyork.us/resource/nc67-uf89.json"
6
7  # Define parameters for sorting and selecting recent 10000 records
8  limit = 10000 # Number of records to fetch
9  params = {'$limit': limit, '$order': 'issue_date DESC'}
10
11 try:
12     # Fetch the data with sorting and selecting top records
13     response = requests.get(dataset_url, params=params)
14     # Check if the request was successful
15     response.raise_for_status()
16     # Attempt to parse the JSON data
17     data = response.json()
18     # Load data into a pandas DataFrame
19     df = pd.DataFrame(data)
20     # Display or save the result
21     print(df)
22     # Specify the directory where you want to save the file
23     save_directory = 'C:/documents/'
24     # Save the file with the specified directory
25     df.to_csv(f'{save_directory}recent_10000_records.csv', index=False)
26
27 except requests.exceptions.RequestException as e:
28     print(f"Request error: {e}")
```

2.Load data into BigQuery:

Created dataset inside given project with name `nyc_traffic` and created table `nyc_traffic_violation` from input csv file using bigquery upload option.

Created a lookup table inside this dataset namely `nyc_boundaries` as external source to get region's data.

Snippet of BigQuery table and dataset



Snippet of Table Structure nyc_boundaries

nyc_boundaries

QUERYSHARECOPYSNAPSHOT

SCHEMADETAILSPREVIEWLINEAGEDATA PROFILEDATA QUALITY

Filter

Enter property name or value

	Field name	Type	Mode	Key	Collation	Default value	Policy tags	Description
<input type="checkbox"/>	the_geom	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	BoroCode	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	BoroName	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	CountyFIPS	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	NTACode	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	NTAName	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Shape_Leng	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	Shape_Area	FLOAT	NULLABLE	-	-	-	-	-

EDIT SCHEMA

VIEW ROW ACCESS POLICIES

Snippet of Table Structure for nyc_traffic_violation

nyc_traffic_violation

QUERYSHARECOPYSNAPSHOTDELETEEXPORT

SCHEMADETAILSPREVIEWLINEAGEDATA PROFILEDATA QUALITY

	Field name	Type	Mode	Key	Collation	Default value	Policy tags	Description
<input type="checkbox"/>	plate	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	state	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	license_type	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	summons_number	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	violation	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	fine_amount	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	penalty_amount	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	interest_amount	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	reduction_amount	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	payment_amount	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	amount_due	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	precinct	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	issuing_agency	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	summons_image	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	violation_time	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	county	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	violation_status	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	issue_date	STRING	NULLABLE	-	-	-	-	-

EDIT SCHEMA

VIEW ROW ACCESS POLICIES

Job history

REFRESH

Snippet of Table Structure for nyc_traffic_violation_data

nyc_traffic_violation_data								
SCHEMA								
Filter Enter property name or value								
<input type="checkbox"/>	Field name	Type	Mode	Key	Collation	Default value	Policy tags	Description
<input type="checkbox"/>	plate	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	state	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	license_type	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	summons_number	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	violation	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	penalty_amount	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	interest_amount	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	reduction_amount	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	payment_amount	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	amount_due	FLOAT	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	precinct	INTEGER	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	issuing_agency	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	summons_image	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	county	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	violation_status	STRING	NULLABLE	-	-	-	-	-
<input type="checkbox"/>	judgment_entry_date	DATE	NULLABLE	-	-	-	-	-
Job history								

3. Data Transformations:

Below are code snippets used to Analyse the data and found invalid entries in violation time column and time format was also not proper and while importing data fine_amount column was created as integer but actually should be integer.

Snippet of Data Analysis SQL Queries

```
1 SELECT * FROM `nyc_traffic.nyc_traffic_violation`
2 ##1)Analysis of input data
3 --count of violation based on type --> answer:7078 'PHTO SCHOOL ZN SPEED VIOLATION'
4 SELECT count(violation) no_of_violation, violation as violation_type FROM `nyc_traffic.nyc_traffic_violation`
5 group by violation_type
6 order by no_of_violation desc
7 ##2)max count of violation based on state?
8 --maximum violation happened in NY state--> 7017 NY
9 SELECT count(violation) no_of_violation,state FROM `nyc_traffic.nyc_traffic_violation`
10 group by state--,violation
11 order by no_of_violation desc
12 ##3)violation_time column analysis
13 --most common violation time 11:28A
14 ---Remove any invalid violation_time
15 ---This statement removed 7 rows from nyc_traffic.nyc_traffic_violation.
16 delete from `nyc_traffic.nyc_traffic_violation`
17 where violation_time not in (select violation_time from `nyc_traffic.nyc_traffic_violation`
18 where violation_time like('%A%') or violation_time like('%P%'))
19 ##4)Analysis violation_time column found that dates are invalid
20 most common date--12/31/2023
21 SELECT count(violation_time),violation_time
22 FROM `nyc_traffic.nyc_traffic_violation`
23 group by violation_time
24 order by 1 desc
```

Below transformation have been applied to the input data

- Column violation_time records were first converted to standard timings by using CASE statement and replace function.
- Column fine_amount data_type was changed to numeric as it cannot be negative (integer type) and renamed as fine_amount_final .
- REGEXP_CONTAINS was used to identify faulty/invalid issue date and replaced with most frequent appearing date.
- Concatenation of issue_date and violation_time to get Violation timestamp column,

- ### Snippet of Transformed Data Creation

Start your free trial with \$300 in credit. Don't worry - you won't be charged if you run out of credit. [Learn more](#)

DISMISS START FREE

Google Cloud sectech-data-engineer-pm bigquery X Search

Untitled query RUN SHARE SCHEDULE MORE SAVE DOWNLOAD Query completed

```
60 data_enrichment AS (  
61   SELECT  
62     c.*,  
63     CAST(CONCAT(issue_date_final, ' ', violation_timestamp_2) AS TIMESTAMP) violation_timestamp_final,  
64     #concatinating issue date and violation_timestamp to get actual datetime column  
65     DATE_DIFF(judgment_entry_date, issue_date_final, day) AS number_of_days_for_judgement  
66     #calculating difference days b/w issue_date_final and judgment_entry_date  
67   FROM  
68     data_type_convert_date_columns c  
69     WHERE EXTRACT(YEAR FROM issue_date_final) < 2024  
70     ##filtering dates higher than 2024  
71   SELECT  
72     * EXCEPT(  
73       Fine_Amount,  
74       issue_date,  
75       issue_date_1,  
76       violation_time,  
77       violation_timestamp1,  
78       violation_timestamp_2,  
79       NTACode,  
80       sm)  
81     ##removing extra unwanted columns after transformation  
82   FROM  
83     data_enrichment a  
84     ##joining Average_fine column to data_enrichment  
85   LEFT JOIN  
86     (SELECT AVG(Fine_Amount_final) Average_fine, summons_number as sm from data_enrichment  
87     GROUP BY violation, sm) b  
88   ON  
89     a.summons_number=b.sm  
90     ##joining NYC boundaries data to data_enrichment  
91   LEFT JOIN  
92     (SELECT distinct(SUBSTR(NTACode, 1, 2)) AS NTACode, BoroName as Borough  
93     FROM  
94       'nyc_traffic.nyc_boundaries') c  
95   ON  
96     a.county=c.NTACode  
97 )
```

Press Alt+F1 for accessibility options

Query results SAVE RESULTS EXPLORE DATA

Job history REFRESH

DISMISS [START FREE](#)

Untitled query [RUN](#) [SHARE](#) [SCHEDULE](#) [MORE](#) [SAVE](#) [DOWNLOAD](#) [Query completed.](#)

60 data_enrichment AS (
61 SELECT

Query results [SAVE RESULTS](#) [EXPLORE DATA](#)

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS		EXECUTION GRAPH						
Row	plate	state	license_type	summons_number	violation	penalty_amount	interest_amount	reduction_amount	payment_amount	amount_due	precinct	issuing_agency	summons_image
1	15968...	99	999	8368523084	null	null	null	null	null	null	null	null	{url: http://nycserv.nyc.gov/NYCServWeb/ShowImage?searchID=VDBStk1rOUVWVGxOZWfMFRrRTIQT09
2	1GGT...	99	999	7771325415	null	null	null	null	null	null	null	null	{url: http://nycserv.nyc.gov/NYCServWeb/ShowImage?searchID=VG5wak0wMVVUWGxPVkZGNFRrRTIQT09
3	1CHA...	99	999	7390772700	null	null	null	null	null	null	null	null	{url: http://nycserv.nyc.gov/NYCServWeb/ShowImage?searchID=VG5wTk5VMUJZek5OYW10M1RVRTIQT09

Results per page: 50 1 – 50 of 9906 [REFRESH](#)

Job history

DISMISS [START FREE](#)

Untitled query [RUN](#) [SHARE](#) [SCHEDULE](#) [MORE](#) [SAVE](#) [DOWNLOAD](#)

60 data_enrichment AS (
61 SELECT

Query results [SAVE RESULTS](#) [EXPLORE DATA](#)

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS		EXECUTION GRAPH		
Row	county	violation_status	Judgment_entry_date	Fine_Amount_final	issue_date_final	violation_timestamp_final	number_of_days_for_judgement	Average_fine	Borough
1	null	null	null	null	2023-12-31	null	null	null	null
2	null	null	null	null	2023-12-31	null	null	null	null

Results per page: 50 1 – 50 of 9906 [REFRESH](#)

Job history

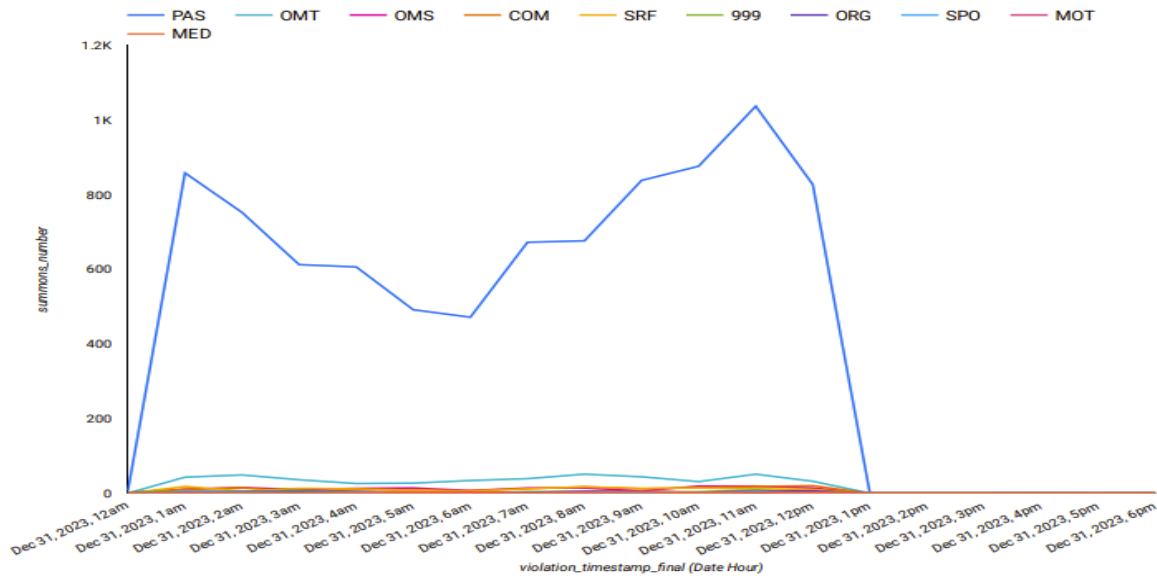
4.Data Visualization:

Looker Studio has been used to visualize and explore insights from transformed data.

Snippet of Visualization done in Looker Studio

NYC Traffic Violation Analysis

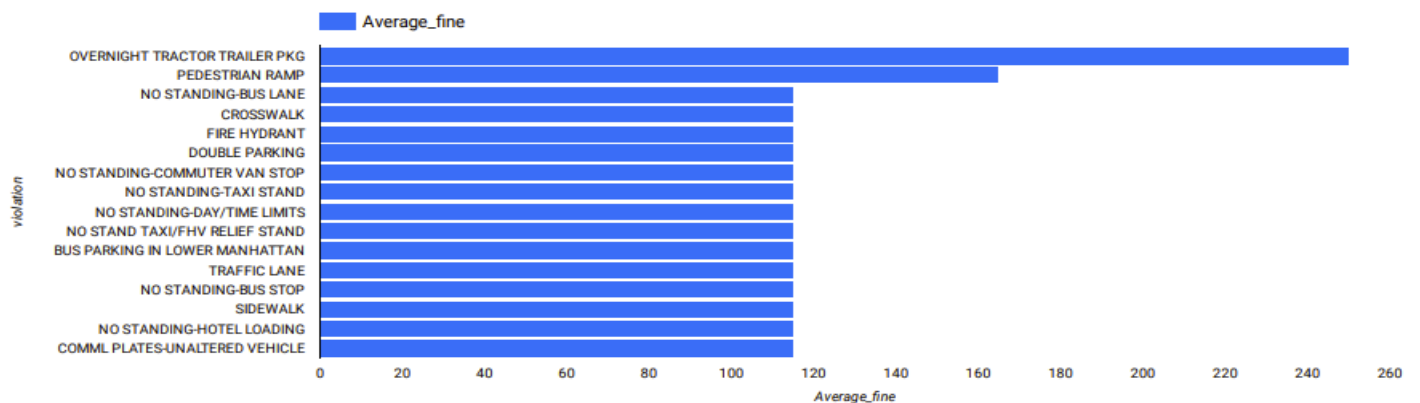
Time series Analysis of violation by License type



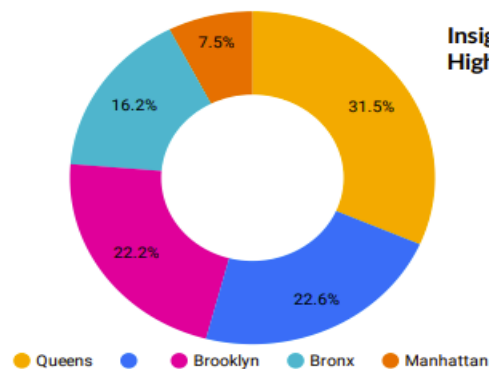
Insights-

- 1) Most of the violations happened between the time 12AM TO 1PM
- 2) Highest number of violations are made by license type PAS

Average fine violation wise



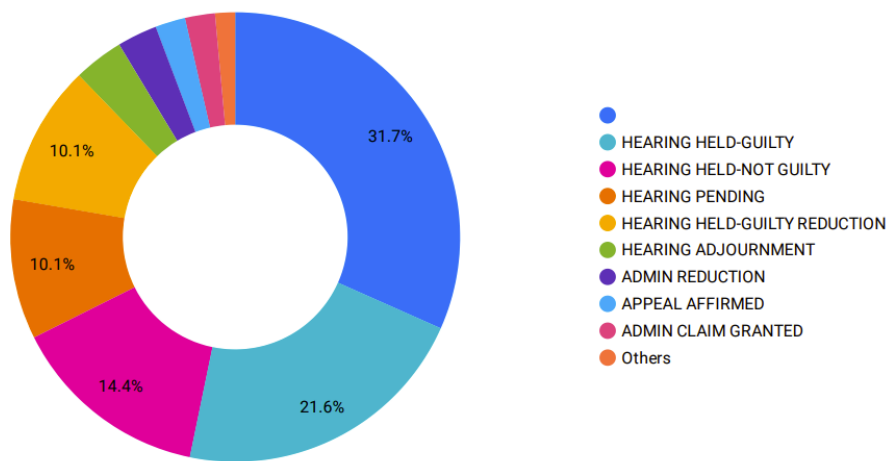
Average fine violation borough wise



Insights-

Highest number of violations are in Queens region

Violation Status



Insights-
18% of violations has violation status as Hearing Held-not Guilty