BigData Project 01

Sai Pavani Cheruku

1. Created the ES domain and the resulting master username and password were used to access the domain endpoint and the opensearch dashboard URL.

2. Connected to EC2 via the web browser. Used this to create the necessary folder structure and the files for the project

3. I had decided the visualizations that I wanted to create beforehand and keeping that in mind I mapped only the columns that I selected.

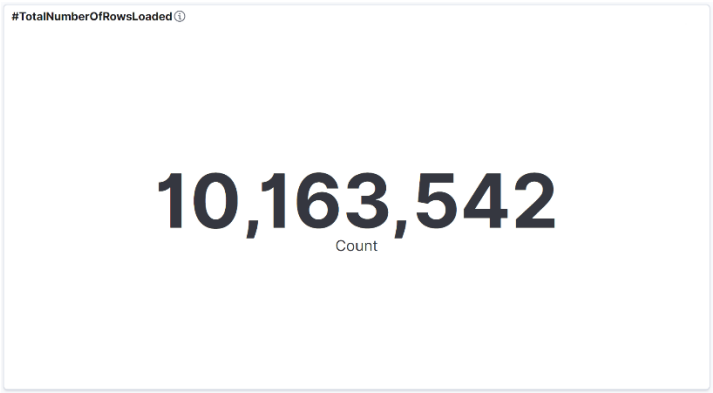
4. Tested that the data was getting fetched using small number of rows using page_size arg

5. Once I got the bulk data API to working, I had to test few iterations to see how to split the no of rows I want to load in terms of num_pages and offset.
   Iteration1:
   10000 per page, 100 pages → Successful
   Iteration2:
   100000 per page, 200 pages → Instance froze and stopped loading data after a while
   Iteration3:
   50000 per page, 240 pages → Successfully was able to load 10.16M rows of data with this.

6. The command used for running the docker file is –
   docker run -e INDEX_NAME="parking01" -e DATASET_ID="nc67-uf89" -e APP_TOKEN="bcL7B5CfP7xg7Vzk8kd0vho8F" -e ES_HOST="https://search-cis9760-saipavani-cheruku-azdlz2okyls4n226zikrcy5aua.us-east-1.es.amazonaws.com" -e ES_USERNAME="spcproj" -e ES_PASSWORD="SPCproj1$" bigdataproject1:1.0 --page_size=50000 --num_pages=240

7. The screenshots that show the no of rows that were downloaded are below –
   Checking the no of rows count in the index created

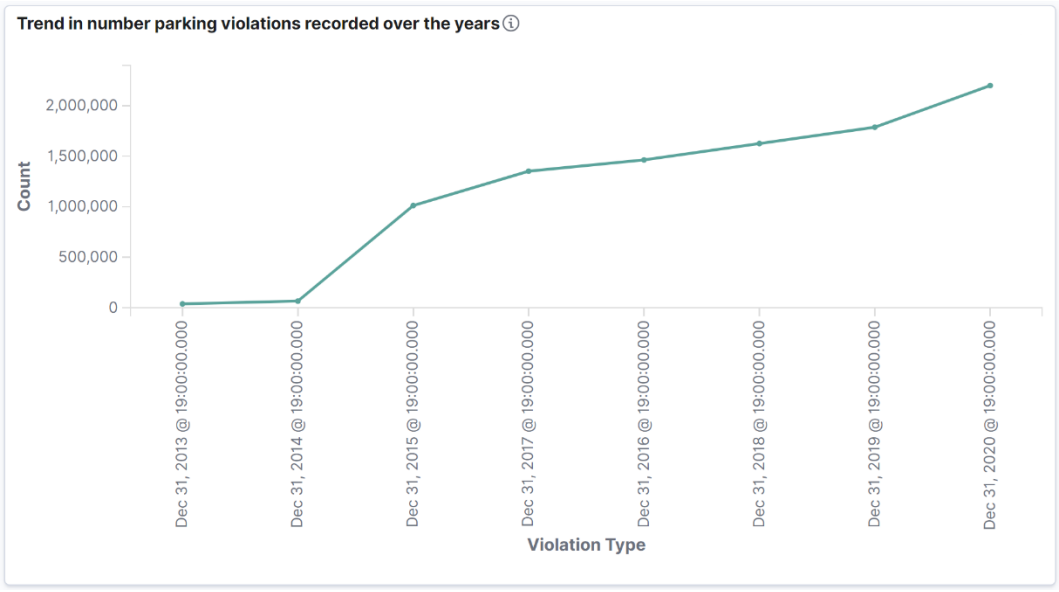
Checking the no of rows count using Kibana's visualization



8. **Parking Violation data analysis – Kibana (Data considered from 2013 to 2020)**

Graph 1: Trend in number of parking violations recorded over the years

Inference1: This graph shows us a trend of the number of parking violations recorded from 2013 to 2020. We clearly see an upward trend that shows that the people violating the parking rules are increasing each year.

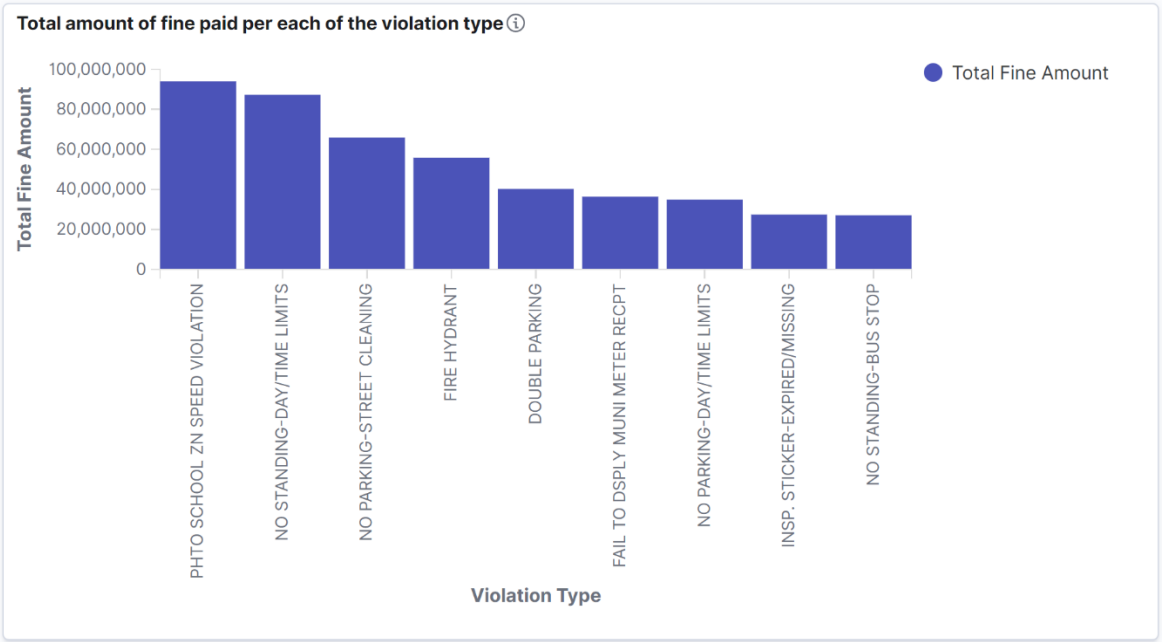


Graph 2: Which states have the highest number of parking violations?

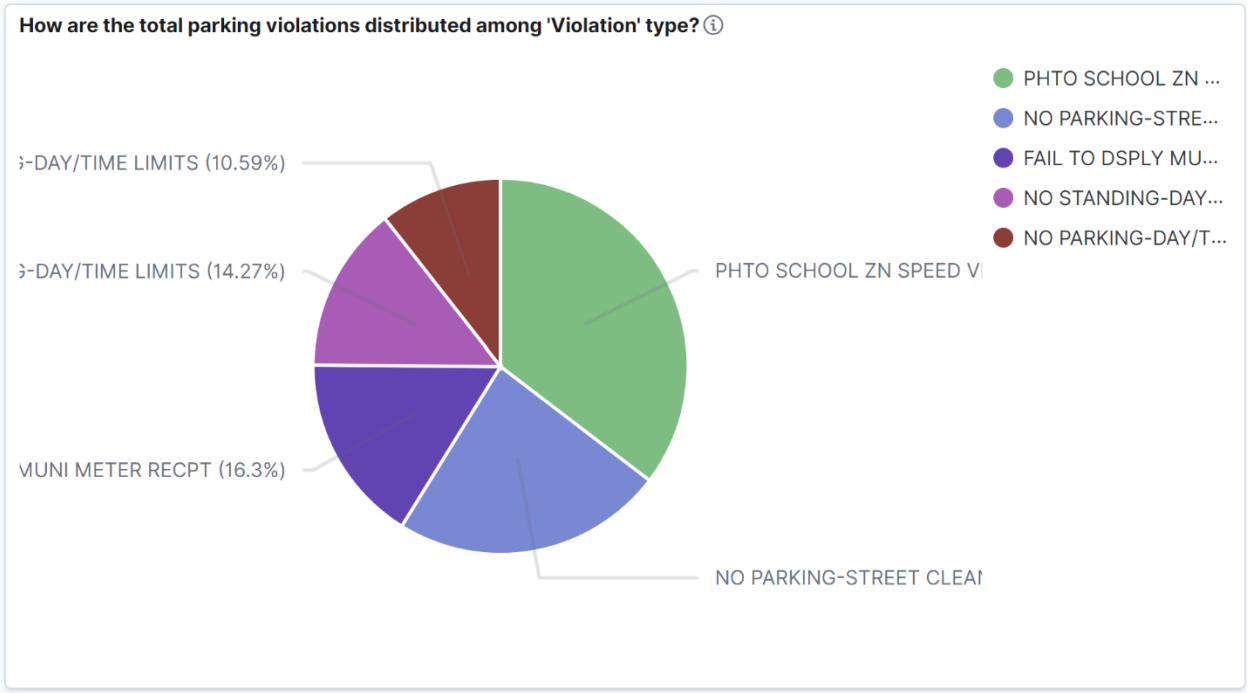
Inference 2: This bar graph below shows us the state wise count of the parking violation summons recorded. From the graph we see that the state of New York has the all time highest record in the years considered.

Graph 3: Total amount of fine paid per each of the violation type.

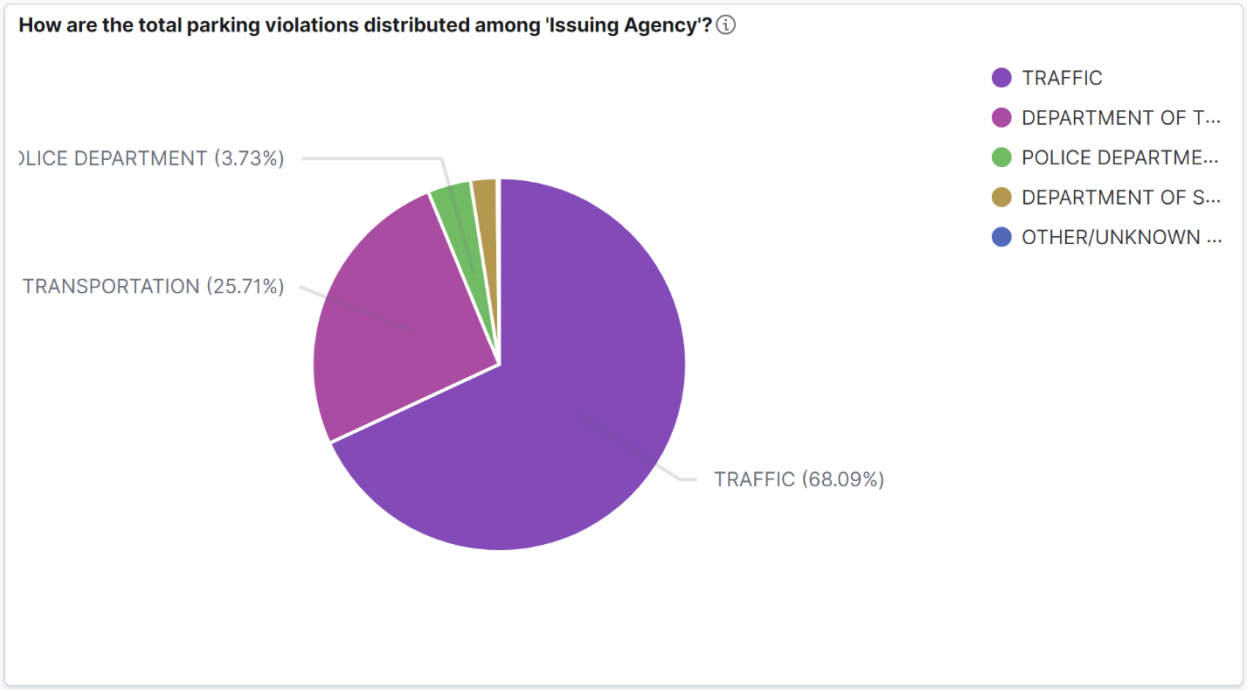
Inference 3: This bar graph shows us the cumulative fine amount paid under each of the violation type category.



Graph 4: How are the total parking violations distributed among ‘Violation Type’?
Inference 4: This pie chart shows us which are the top 5 parking violations in terms of the max number of summonses recorded for.

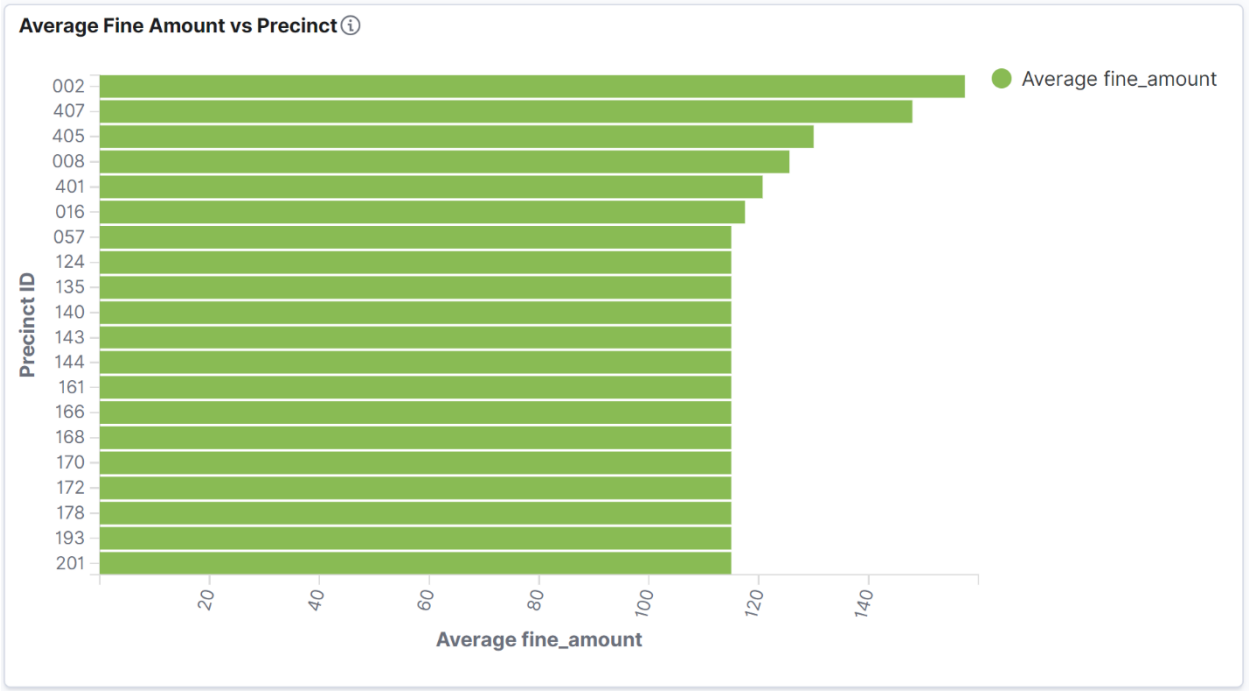


Graph 5: How are the total parking violations distributed among ‘Issuing Agency’?
Inference 5: This pie chart shows us the distribution of the total number of parking violations among the Issuing agencies that recorded them.



Graph 6: Average fine amount per precinct?

Inference 6: This graph shows us the top 20 precincts based on their average fine amount values.



Dashboard created in Kibana with all of the above graphs: