Centre for Development of Advanced Computing, Mumbai

Report on

NYC Parking Tickets Analysis

PG-DBDA March 2022

Submitted by:

Project Team 07

Sarvesh Raut Amey Koli

Shubham Bhole Navnath Bhoskar

Prathamesh Satpute Darshan Ahire

Mr. Wasim Khan

Project Guide

1 INTRODUCTION

New York City is a thriving metropolis. Just like most other metros its size, one of the biggest problems its citizens face is parking. The classic combination of a huge number of cars and cramped geography leads to a huge number of parking tickets.

In an attempt to scientifically analyse this phenomenon, the NYC Police Department has collected data for parking tickets. Of these, the data files for multiple years are publicly available on NYC Open Data.

We will try and perform some ETL operation using AWS Glue. PySpark will allow us to analyse the full files at high speeds. AWS QuickSight will allow us to visualize the data. For the scope of this analysis, we will analyse the parking tickets over the year 2022.

1.1 Dataset:

The NYC Department of Finance collects data on every parking ticket issued in NYC (~10M per year). This data is made publicly available to aid in ticket resolution and to guide policymakers.

This dataset consists of nine files, covering April 2014-March 2022 sourced from NYC Open Data. Parking Violations Issuance datasets contain violations issued during the respective fiscal year. Each column contains information about the violation such as the vehicle ticketed, the type of ticket issued, location, and time. In total the dataset is almost contain 96M rows and total size 15GB in size.

2 PROBLEM STATEMENT

New York City is a thriving metropolis. Just like most other metros its size, one of the biggest problems its citizens face is parking. The classic combination of a huge number of cars and cramped geography leads to a huge number of parking tickets.

In an attempt to scientifically analyse this phenomenon, the NYC Police Department has collected data for parking tickets. For the scope of this analysis, we will analyse the parking tickets over the year 2022.

2.1 Objective:

Primarily, this project is meant as a deep dive into the usage of Spark and AWS services like AWS S3, AWS Glue, AWS QuickSight. One of the major objectives of this case study is to gain familiarity with how analysis works in PySpark as opposed to base Python.

Learning the basic idea behind AWS Glue as ETL tool and AWS QuickSight as business intelligence tool.

The purpose of this project is to conduct an exploratory data analysis that will understand the data and find out some insights from that data.

3 FUNCTIONAL REQUIREMENTS OVERVIEW

3.1 Modules Used:

1. Apache Spark
2. Pyspark
3. AWS EC2
4. AWS S3
5. AWS Glue
6. AWS Quick Sight

3.1.1 Apache Spark:

- Apache Spark is a lightning-fast cluster computing technology, designed for fast computation.
- It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing.
- The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application.
- Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming.
- Apart from supporting all these workloads in a respective system, it reduces the management burden of maintaining separate tools

3.1.2 PySpark:

- PySpark is the Python API for Apache Spark, an open source, distributed computing framework

- It is a Spark Python API and helps us to connect with Spark Dataframe.

- Through PySpark, we can write ETL transformation by using Python APIs. This interface also allows us to use PySpark Shell to analyze data in a distributed environment interactively.

3.1.3 AWS EC2:

- Amazon Elastic Compute Cloud

- Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) Cloud.

- Virtual computing environments, known as instances

- Storage volumes for temporary data that's deleted when you stop, hibernate, or terminate your instance, known as instance store volumes

3.1.4 AWS S3:

- AWS S3 can store unlimited amount of data.

- An object (file) can be as big as 5TB.

- AWS S3 will create multiple copy of your data and store these copy across different data centres. So, even if one data centre goes down, you can get your file from another data centre.

3.1.5 AWS Glue:

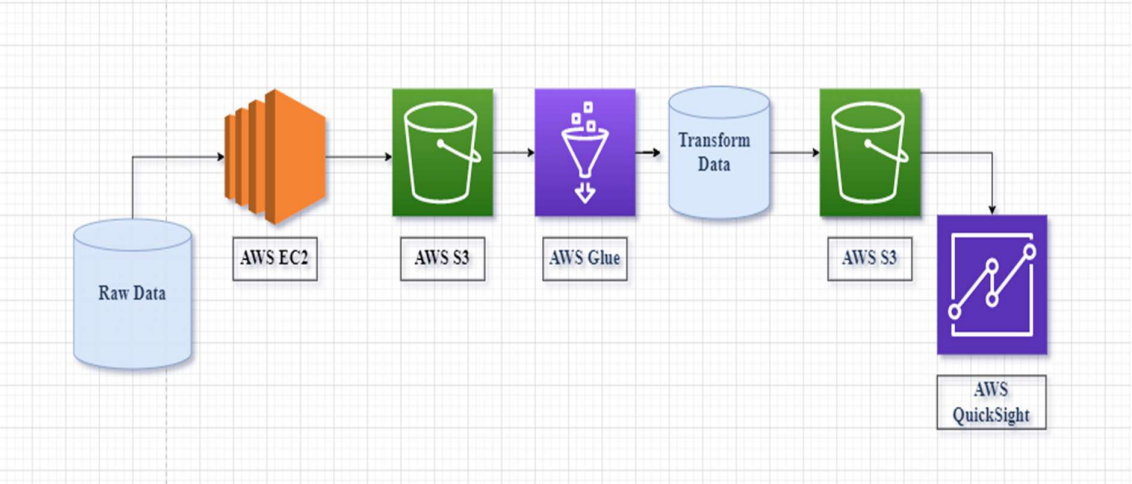
- AWS Glue is a fully managed service provided by Amazon for deploying ETL jobs. It reduces the cost, lowers the complexity, and decreases the time spent creating ETL jobs.
- Glue is a Serverless service. Amazon Glue ETL comes with crawlers that can create metadata to view the data stored in S3. This metadata comes very handy while authoring ETL jobs.
- With the use of Python scripts, Glue can translate one source format to another source format.

3.1.6 AWS QuickSight:

- AWS QuickSight is a Data Visualization and Business Intelligence tool that converts data from different data sources to interactive dashboards and BI reports.

- Amazon QuickSight is a fully managed fast performance cloud-scale business intelligence service offers an easy-to-use interface where users can easily connect their data, create analysis that is powered by machine learning and allows you to create data visualizations and dashboards.

4 PROJECT FLOW



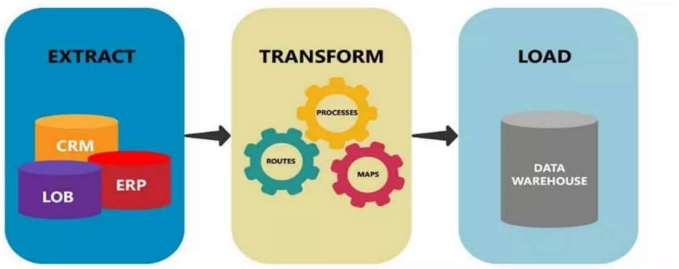
5 METHODOLOGY

5.1 Data Used

The data used in this project was parking violations issued in New York City over the year 2022. The data was in CSV format.

This dataset consists of 43 features of each ticket, with include information about the vehicle, the registration, the type of violation, the location, the borough and precinct, the street, the rank and division of the ticket issuer. This seemed especially rich for investigation.

5.2 ETL Process Implementation:



ETL - Extract, Transform, Load

5.2.1 Data Extraction

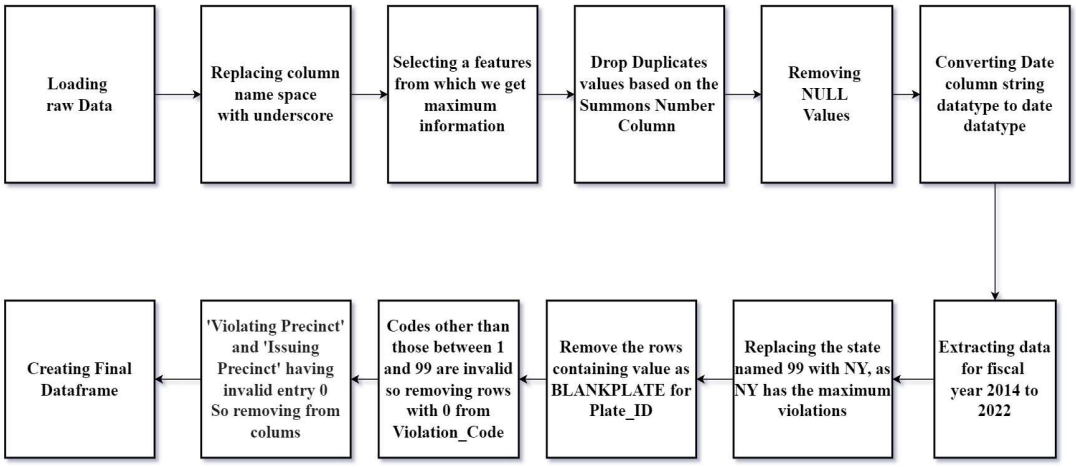
Before we can begin organizing our data, the first step in the data ETL process is to pull or extract the data from all the relevant sources and compile it. This ETL requirement and gathering process will include the necessary preparation for carrying out data integration.

Raw data was dumped into AWS S3. It is an Object Storage i.e., AWS S3 was used for storing the large amount of data.

5.2.2 Data Transformation

Data Transformation is the second step of the ETL process. In the first step, the ETL deployment was carried out. Now, in the second ETL phase, the ETL transformation is carried out: data extracted from the sources are compiled, converted, reformatted, and cleansed in the staging area to be fed into the target database in the next step.

Transformations:



5.2.3 Data Loading

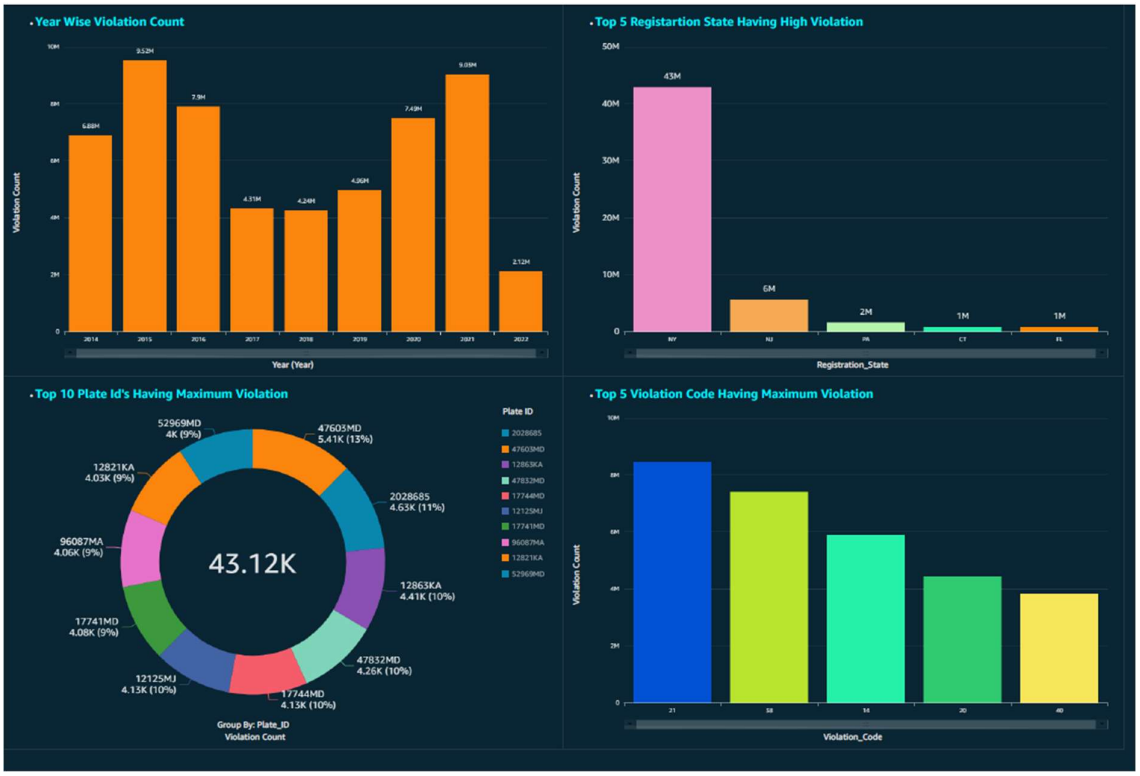
The concluding step in the three-step data ETL process is loading the datasets that have been extracted and transformed earlier into the target database.

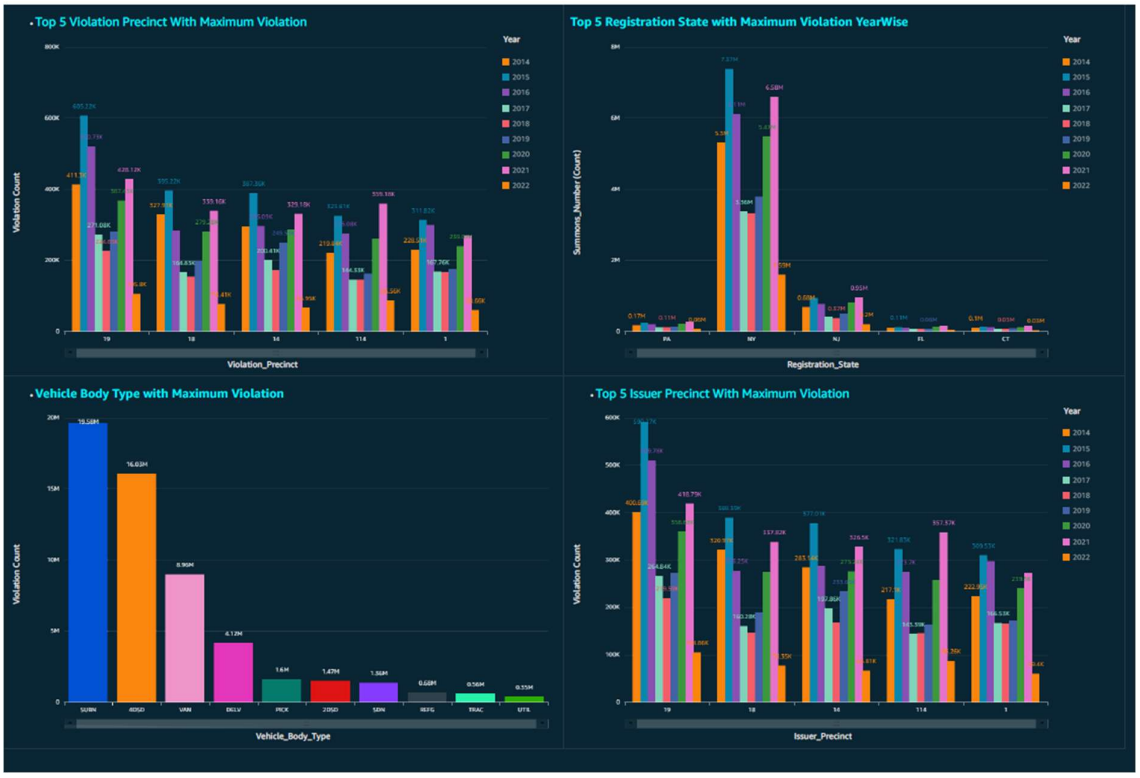
We load transformed data into AWS S3 bucket.

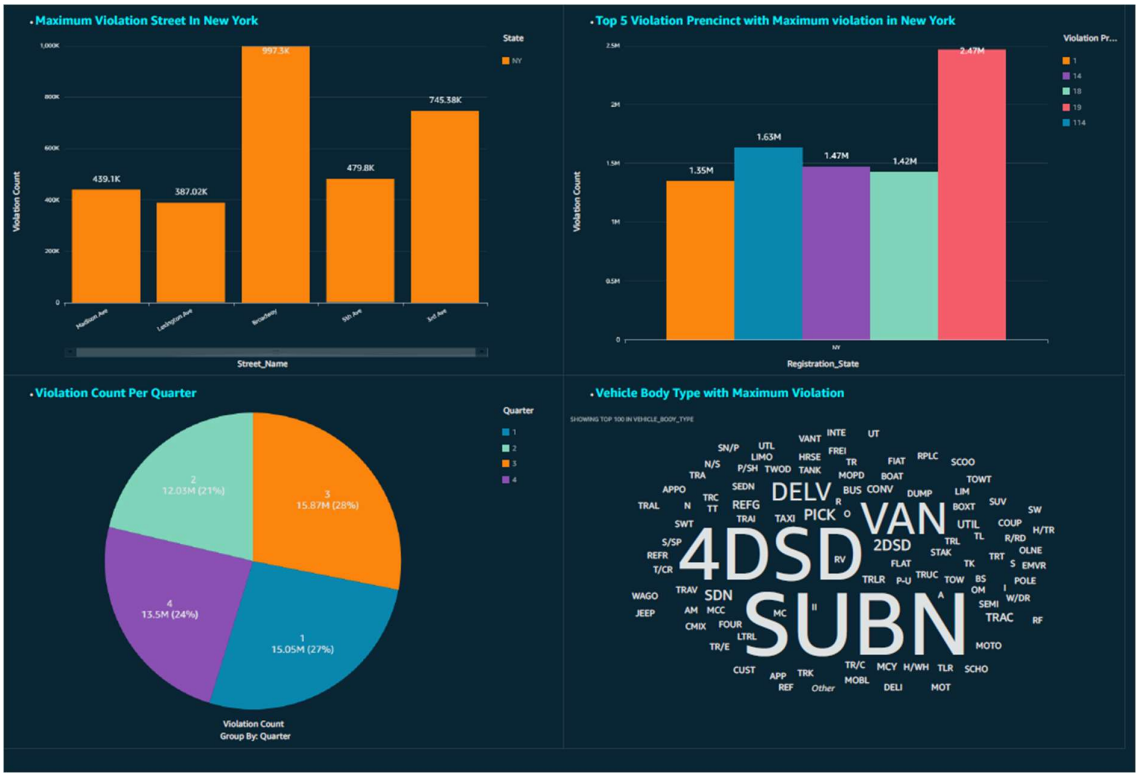
6 Data Analysis

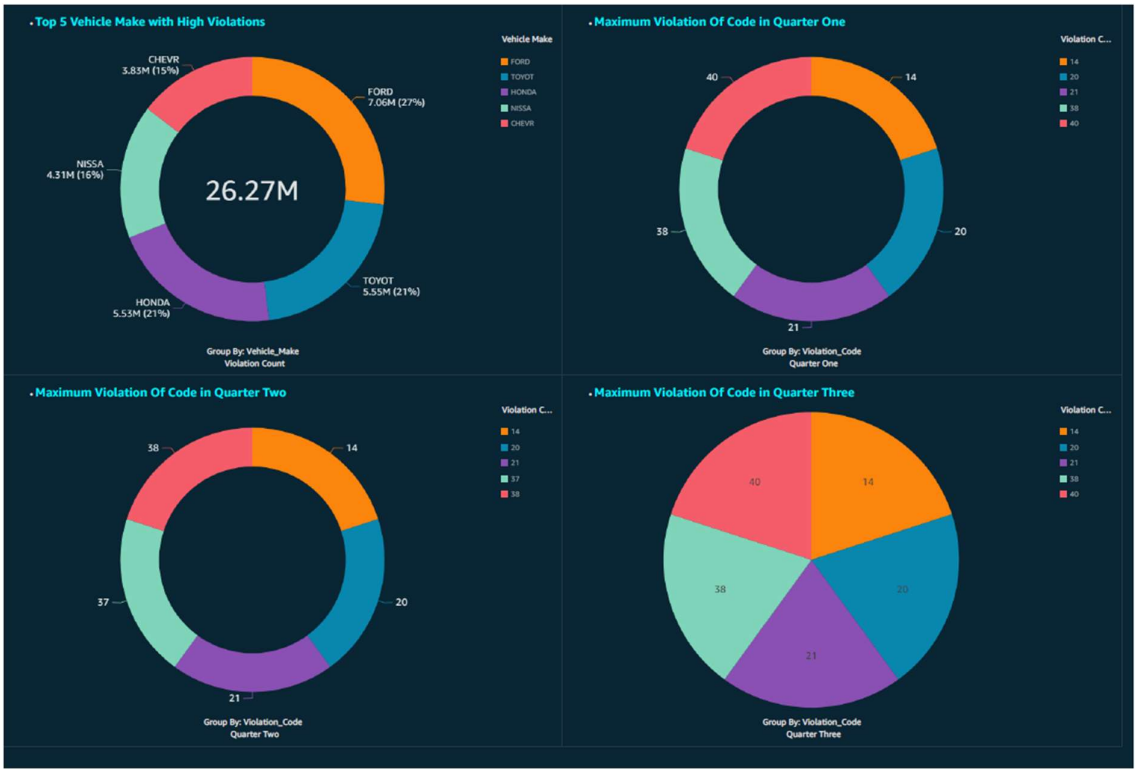
The refined data is used to perform sentiment analysis and tried to visualized appropriate conclusions from the performed data analysis.

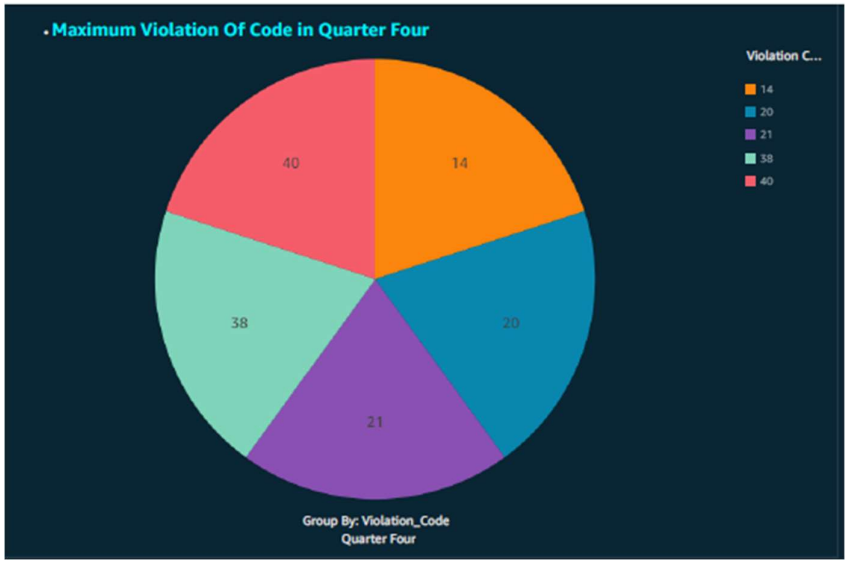
For data visualization we used AWS QuickSight. Amazon QuickSight is a fast, cloud-powered business intelligence service that delivers insights to everyone in organization. Amazon QuickSight connects to data in the cloud and combines data from many different sources.




Top 5 Violation Precinct With Maximum Violation
Violation Count
Violation_Precinct
Year
Top 5 Registration State with Maximum Violation YearWise
Summons_Number (Count)
Registration_State
Vehicle Body Type with Maximum Violation
Violation Count
Vehicle_Body_Type
SUBN
4DSD
VAN
DELV
PICK
2DSD
SDN
REFG
TRAC
UTIL
Top 5 Issuer Precinct With Maximum Violation
Violation Count
Issuer_Precinct


Maximum Violation Street In New York
Violation Count
Street_Name
439.1K
387.02K
997.3K
479.8K
745.38K
Top 5 Violation Prencinct with Maximum violation in New York
Violation Count
Registration_State
1.35M
1.63M
1.47M
1.42M
2.47M
Violation Count Per Quarter
Quarter
2
12.03M (21%)
3
15.87M (28%)
4
13.5M (24%)
1
15.05M (27%)
Violation Count
Group By: Quarter
Vehicle Body Type with Maximum Violation
SHOWING TOP 100 IN VEHICLE_BODY_TYPE
4DSD
VAN
SUBN
DELV
2DSD
PICK
SDN
REFG
TRAC
UTIL


Top 5 Vehicle Make with High Violations
Vehicle Make
FORD
TOYOT
HONDA
NISSA
CHEVR
CHEVR
3.83M (15%)
FORD
7.06M (27%)
NISSA
4.31M (16%)
26.27M
TOYOT
5.55M (21%)
HONDA
5.55M (21%)
Group By: Vehicle_Make
Violation Count
Maximum Violation Of Code in Quarter One
Violation C...
14
20
21
38
40
40
14
38
20
21
Group By: Violation_Code
Quarter One
Maximum Violation Of Code in Quarter Two
Violation C...
14
20
21
37
38
38
14
37
20
21
Group By: Violation_Code
Quarter Two
Maximum Violation Of Code in Quarter Three
Violation C...
14
20
21
38
40
40
14
38
20
21
Group By: Violation_Code
Quarter Three


Maximum Violation Of Code in Quarter Four
Violation C...
14
20
21
38
40
40
14
38
20
21
Group By: Violation_Code
Quarter Four

7 Conclusion

- Total Number of tickets issued for the year

Year	Violation Count
2014	6.88M
2015	9.52M
2016	7.9M
2017	4.3M
2018	4.2M
2019	4.96M
2020	7.49M
2021	9.03M
2022	2.12M

- In year 2017 had the highest parking tickets issued followed by years 2021, 2020.
- There are 64 unique states where the cars that got parking tickets came from.
- New York has the highest parking violations followed by New Jersey and Pennsylvania
- Plate Id 47603MD had the maximum parking violations.
- **The top 5 violation codes are 21, 38, 14, 20,40**

  Code 21 - No parking where parking is not allowed by sign, street marking or traffic control device.

  Code 38 - Parking Meter - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.

Code 14 - General No Standing: Standing or parking where standing is not allowed by sign, street marking or; traffic control device.

Code 20 - General No Parking: No parking where parking is not allowed by sign, street marking or traffic control device.

Code 40 - Stopping, standing or parking closer than 15 feet of a fire hydrant. Between sunrise and sunset, a passenger vehicle may stand alongside a fire hydrant as long as a driver remains behind the wheel and is ready to move the vehicle if required to do so.

- For Vehicle Body Type, maximum parking violations happen for Suburban (SUBN) followed by four door sedan(4DSD), Van, Delv and Pick

- For Vehicle Make, maximum parking violations happen for Ford followed by Toyota, Honda, Nissa, Chevr

- The top 3 violation precincts and Issuer Precincts where maximum parking violations happen are 19, 18,14,114,1
- Maximum Ticket Frequency occur in Quarter 3(July-Sept) followed by Quarter 1 (Jan-March), Quarter 4 (Oct-Dec). Quarter 4 (April- June). has the least Ticket Frequency.
- The highest fine amount of $548,586,025 was for violation code 21.
- A total fine amount of $1,703,847,085 was collected for the three violation codes 21, 38 and 14

8 Reference

"NYC Parking Tickets" [Online] Available:

https://www.kaggle.com/datasets/new-york-city/nyc-parking-tickets"Amazon Learner Lab Guide" [Online] Available:

https://awsacademy.instructure.com/courses/24007/modules/items/1969629

Nick Cox "Learn Python Data Analytics by Example: NYC Parking Violations" [Online] Available:

https://towardsdatascience.com/learn-python-data-analytics-by-example-ny-parking-violations-e1ce1847fa2

"Violation Codes, Fines, Rules & Regulations" [Online] Available:

https://www1.nyc.gov/site/finance/vehicles/services-violation-codes.page

QuickSight Manifest files
https://docs.aws.amazon.com/quicksight/latest/user/supported-manifest-file-format.html