FACULTY OF COMPUTER AND INFORMATION SCIENCE
UNIVERISITY OF LJUBLJANA

# Project

Student name: *Aljaž Mur Eržen, Jakob Gaberc Artenjak*

Course: *Big data*
Date: *June 18, 2022*

## 1. Problem

New York City Open Data is a website that contains many different datasets that are produced or in use by the city of New York, USA. One of the datasets, titled "Parking Violations Issued" [1] contains information about 15M fines issued for parking violations for each of the years from 2014 onwards.
Our task was to handle the dataset, explore its contents and apply basics of machine learning, all while using big data approaches.

## 2. Task 1: data formats

To get started, we converted the dataset from its original CSV format to a few different formats and applied a few different compression algorithms to see, which format is the most suitable for use further on.

| Format | Compression | Size |
|---|---|---|
| parquet | gzip | 339 MB |
| parquet | brotli | 398 MB |
| avro | snappy | 662 MB |
| parquet | | 671 MB |
| avro | | 2.2 GB |
| HDF | zlib, comp. level=9 | 2.2 GB |
| CSV (original) | | 2.2 GB |
| HDF | | 2.6 GB |

Table 1: Comparison of sizes produced by different storage formats

One can see that HDF is not efficient at all. Without compression, the size of the dataset actaully grew when converted from plain-text CSV.
Most efficient format was parquet with gzip compression, which produced not only the smallest output, but has also been comparable to other formats in terms of time required for the compression.
Parquet is a file format developed by Apache Fundation. It stores tabular data, organized in rows and columns, but contrary to CSV and most relational databases, it does not store data by rows, but by columns. This approach is called columnar storage and is beneficial because of the fact that usually, values in the same column are more similar than values in the same row, which may differ even in their data type.

## 3. Task 2: augemtation

## 4. Task 3: exploration

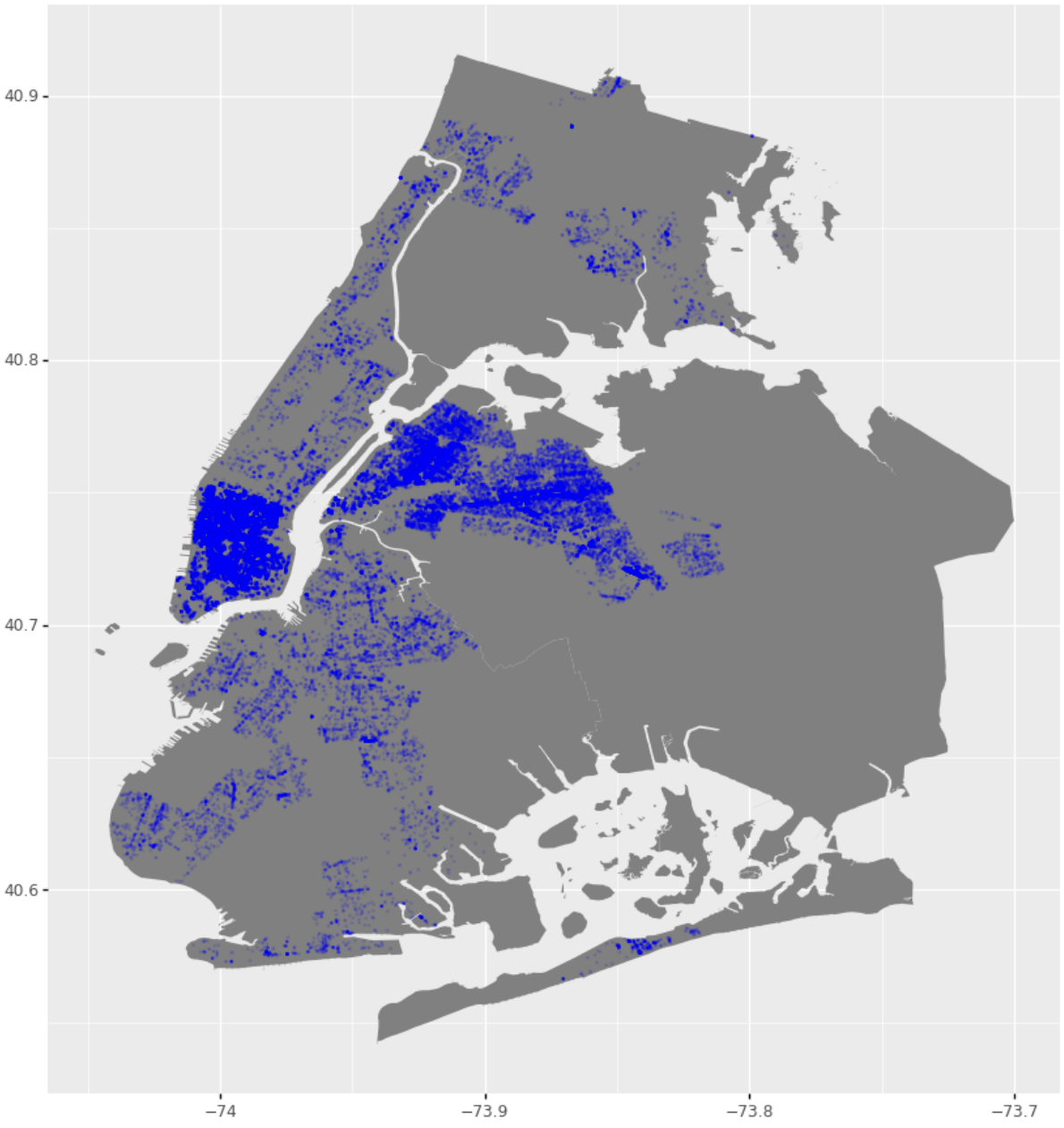


Figure 1: Locations of violations in the year 2022 in the months January to May

## References

[1] City of New York, Open Data, *Parking Violations Issued - Fiscal Year 2021*,

[2] jq, *Lightweight and flexible command-line JSON processor*,

[3] *JSON path*,

[4] GraphQL, *A query language for your API*,

[5] PartiQL, *SQL-compatible access to relational, semi-structured, and nested data.*,

[6] *json-query*,

[7] Graham Cormode; Muthu Muthukrishnan, *Approximating Data with the Count-Min Sketch*.