# Data Cleaning and Profiling

The unfiltered dataset has 120,878,667 rows.

Below is a complete schema of the dataset

| | | |
|---|---|---|
| Summons Number | UNIQUE IDENTIFIER OF SUMMONS | Number |
| Plate ID | REGISTERED PLATE ID | Plain Text |
| Registration State | STATE OF PLATE REGISTRATION | Plain Text |
| Plate Type 3 | TYPE OF PLATE | Plain Text |
| Issue Date | ISSUE DATE | Date & Time |
| Violation Code | | Number |
| Vehicle Body Type | VEHICLE BODY TYPE WRITTEN ON | Plain Text |

| | | |
|---|---|---|
| | SUMMONS (SEDAN, ETC.) | |
| Vehicle Make | | Plain Text |
| Issuing Agency | | Plain Text |
| Street Code1 | | Number |
| Street Code2 | | Number |
| Street Code3 | | Number |
| Vehicle Expiration Date | | Number |
| Violation Location 13 | | Plain Text |
| Violation Precinct 14 | | Number |
| Issuer Precinct | | Number |
| Issuer Code | | Number |
| Issuer Command | | Plain Text |

| | |
|---|---|
| Issuer Squad | Plain Text |
| Violation Time 19 | Plain Text |
| Time First Observed 20 | Plain Text |
| Violation County 21 | Plain Text |
| Violation In Front Of Or Opposite 22 | Plain Text |
| House Number | Plain Text |
| Street Name 24 - street process | Plain Text |
| Intersecting Street - street process | Plain Text |
| Date First Observed 26 | Number |
| Law Section | Number |
| Sub Division | Plain Text |
| Violation Legal Code | Plain Text |

| | |
|---|---|
| Days Parking In Effect 30 | Plain Text |
| From Hours In Effect 31 | Plain Text |
| To Hours In Effect 32 | Plain Text |
| Vehicle Color 33 | Plain Text |
| Unregistered Vehicle? 34 | Plain Text |
| Vehicle Year 35 | Number |
| Meter Number | Plain Text |
| Feet From Curb 37 | Number |
| Violation Post Code | Plain Text |
| Violation Description | Plain Text |
| No Standing or Stopping Violation 40 | Plain Text |
| Hydrant Violation 41 | Plain Text |
| Double Parking Violation 42 | Plain Text |

# Cleaning and profiling of important columns:

## State:

This column describes the state where the vehicle is registered to.

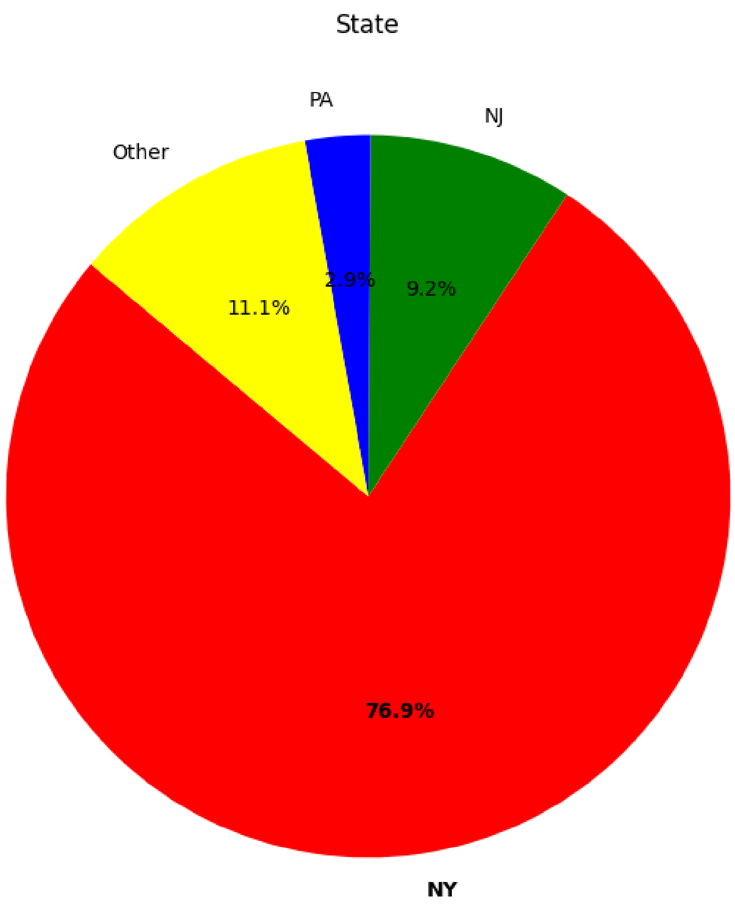
### Profiling:

The raw data contains arbitrary strings for many entries. These data entries are removed from the dataset. These include entries arbitrary strings like “[PP[_[" or strings that do not correspond to any state like “ZW”. However, these are relatively low so these rows can be removed. The overal quality of the data is HIGH.

### Cleaning:

The entry is check to be corresponding to the codes of the 50 states in USA. If there is a match the state is taken else removed. Minor trailing dashes and quotes are removed.

The most common state is NY, followed by NJ, then PA. Other states are shown as other for the sake of brevity.



## Issue date:

This column corresponds to the date for the issuance of the parking ticket

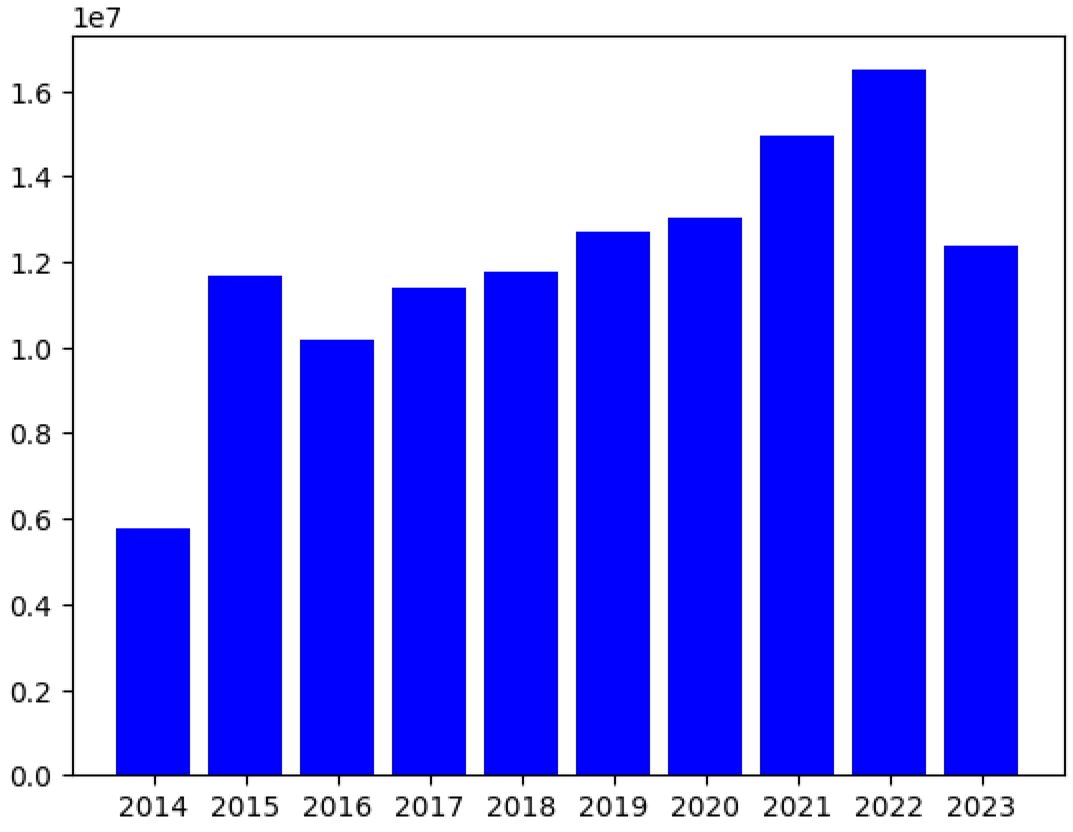
**Profiling:**

The raw data contains dates in the format YYYY/MM/DD. The only problem with the data is that some strings have trailing characters.

**Cleaning:**

Trailing characters are chopped off for consistency. Some data entries are missing an issue date. These data entries are removed from dataset.

The bar plot of issue date years post cleaning is as follows. Note the number of parking violations is in tens of millions



## Vehicle make and Bodytype:

These columns corresponds to the make of the vehicle and bodytype

**Profiling:**

The data for these categories is free flowing text and low signal to noise ratio

**Cleaning:**

Since these columns are not relevant to analysis and are of low quality they are removed from dataset

## Voilation/Issuer Precinct:

**Profiling:**

On profiling the data we see that the data has some precincts beyond 123 which is a cause of concern. However, these instances are limited to around a 1000 rows so this column is still useful in case we want to do aggregate analysis

**Cleaning:**

If the precinct number is more than 123 the precinct is removed from the data

## Vehicle Expiration date:

**Profiling:**

Similar to issue date

**Cleaning:**

Similar to issue date

## Violation Time and Time first issued:

These are columns indicating at what time the violation occurred and first time the ticket was issued.

**Profiling:**

These columns are in the format, HHMM AM/PM. Some rows are missing AM/PM. Some time stamps are in 24-hour and others in 12-hour format. The data is valuable and can glean important insights even though data quality is low due to high noise in data collection.

**Cleaning:**

To clean the data if a PM string is detected 12 hours are added to the timestamp if the hour value is less than 12. All timestamps are converted into 24-hours

## Street Name and Intersecting Street:

This column indicates the street location where the violation takes place.

**Profiling:**

These columns are present for almost all data and are valuable for analysis. The main issue is that multiple names are used for each street. For example, sometimes nd is added at the end of the street name and so on.

**Cleaning:**

For cleaning these columns suffixes at the end of road names are removed. Furthermore, common terms like avenue, road are shortened to av and rd for simplicity.

## Feet from curb

The number of feet from curb is important because if the car is too far away from the curb it’s considered a parking violation.

**Profiling:**

The column has lot of erroneous values. Some rows are filled with years like 2018,2019 etc. Other rows are filled with values like “125-3006”. However, most rows are filled with sensible values between 1-10.

**Cleaning:**

Since most rows have relevant values this column is retained. Every string is converted to integer to check it’s validity. Rows that cannot be converted to integers are removed

The variable varies from 0-10 feet. However, most of the dataset has 0 feet as entry. The mean of this variable is 0.01 feet.

## From hours in effect/To hours in effect

These columns mention the from and to hours. Their difference can be used to infer the duration of parking violation making them very important.

**Profiling:**

Similar to Violation Time and Time first issued

**Cleaning:**

Similar to Violation Time and Time first issued

## No Standing or Stopping Violation/Hydrant Violation/Double Parking Violation

These three columns are important for analysis and can lend insights into the type of violation. These are boolean variables which indicate the type of violation.

**Profiling:**

The data for most row entries is not present. In fact, data is present for only some 1000 data entries out of more than 100,000,000. This means that this column is mostly missing.

**Cleaning:**

Since the column is mostly missing it is removed from the dataset.

## Post filtering

Post filtering a row of the dataset is converted to the following format

```
AK,01/01/2018,67,83,83,,,s/w c/o knickerbocke,schaefer st,0,,,0,0
```

Total number of rows post removal is **120348742** which means around 500,000 rows have been removed

***Note that having all columns is not necessary for the datapoint only columns with poor quality are removed from data***

## Code guide

Two codes are mainly used for profiling and filtering:

1. WordCount - this code takes as input the file for analysis and then runs WordCount using MapReduce. This program is mainly used for profiling the data to get an idea of the different terms in different columns with their frequencies. Remember to change the WordCountMapper.java file's index to the column you want a profile of
2. Filter - This code filters the raw dataset to a simpler version with only the relevant columns

Sample.csv is provided since the whole dataset is too large to store in submission which has 1000 lines of the dataset