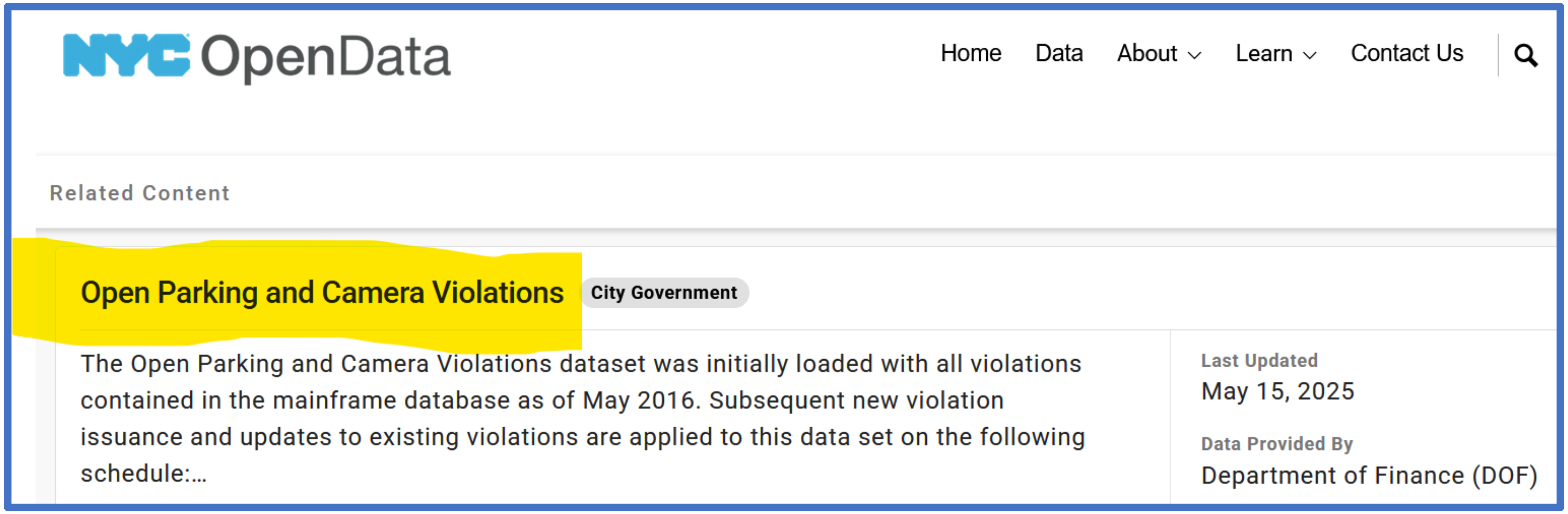
Group 9A – Matt R

Was separated from Group 9 at the ETL process milestone.
Took over KPI #2 which was exploring NYC parking ticket data.
Data was downloaded via manual query from NYC OpenData Open Parking and Camera Violations Database:
https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89/about_data



Related Content

Open Parking and Camera Violations City Government

The Open Parking and Camera Violations dataset was initially loaded with all violations contained in the mainframe database as of May 2016. Subsequent new violation issuance and updates to existing violations are applied to this data set on the following schedule:...

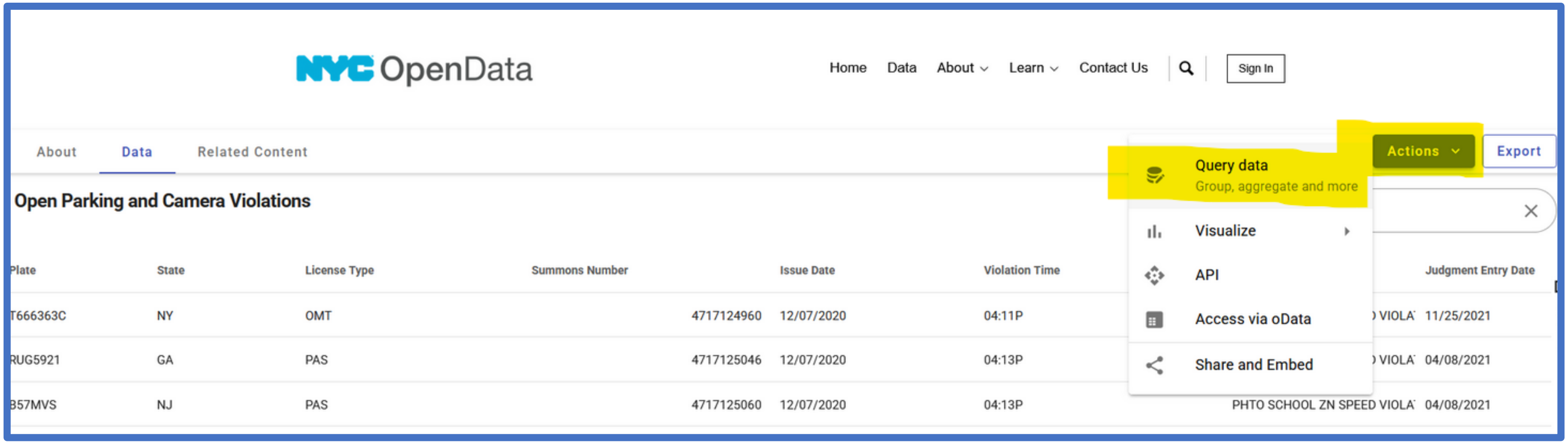
Last Updated
May 15, 2025

Data Provided By
Department of Finance (DOF)

Data was queried for the months June to December, 2023 in April and May of 2025.
Total amount of rows was approximately 70 million.
Each month contained between 9 million and 11 million rows.

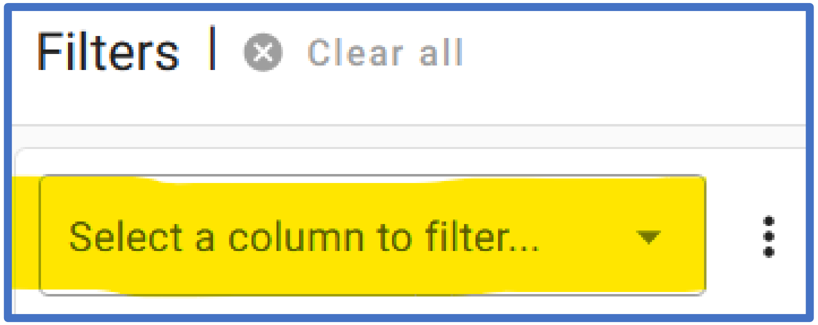
Tools used:

- Extraction: query tool on the NYC Open Data website

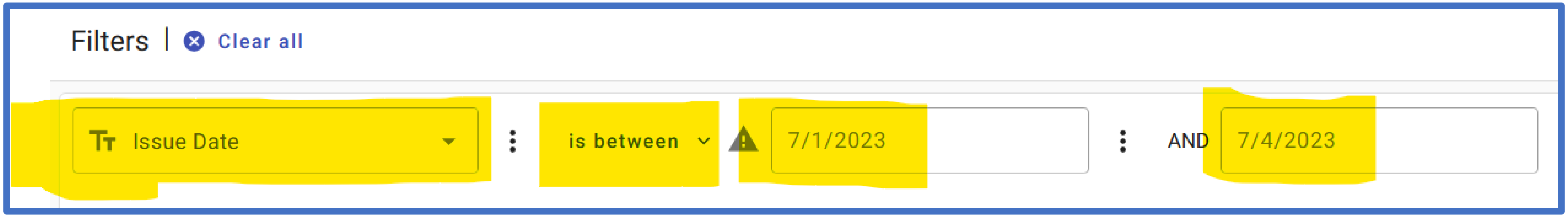


Open Parking and Camera Violations

Plate	State	License Type	Summons Number	Issue Date	Violation Time		Judgment Entry Date
T666363C	NY	OMT	4717124960	12/07/2020	04:11P	) VIOLA	11/25/2021
RUG5921	GA	PAS	4717125046	12/07/2020	04:13P	) VIOLA	04/08/2021
B57MVS	NJ	PAS	4717125060	12/07/2020	04:13P	PHTO SCHOOL ZN SPEED VIOLA	04/08/2021



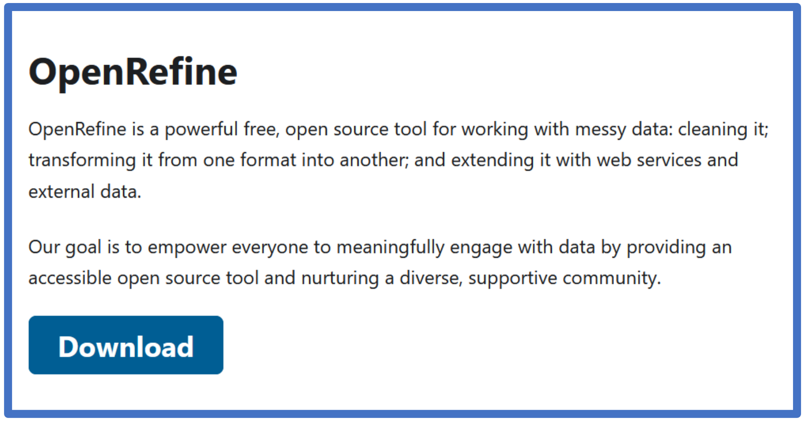
Select a column to filter...



Issue Date | is between | 7/1/2023 | AND | 7/4/2023

- Transform: OpenRefine

OpenRefine is an open-source, freely available tool that allows to cleaning of dirty and inconsistent data on thousands to millions of rows on a local desktop computer.

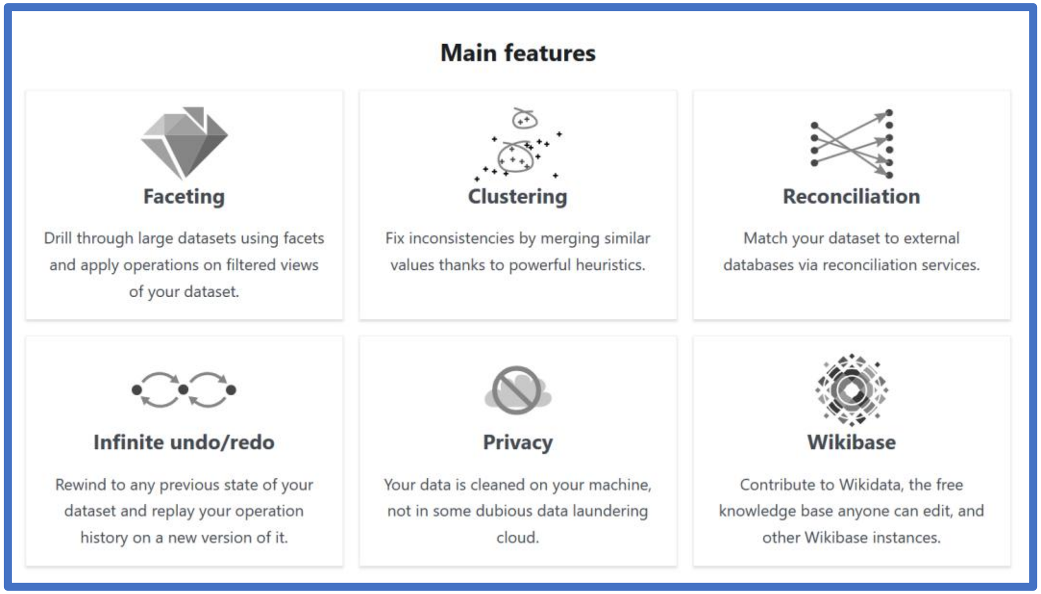


OpenRefine

OpenRefine is a powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

Our goal is to empower everyone to meaningfully engage with data by providing an accessible open source tool and nurturing a diverse, supportive community.

www.OpenRefine.org



Main features

Faceting

Drill through large datasets using facets and apply operations on filtered views of your dataset.

Clustering

Fix inconsistencies by merging similar values thanks to powerful heuristics.

Reconciliation

Match your dataset to external databases via reconciliation services.

Infinite undo/redo

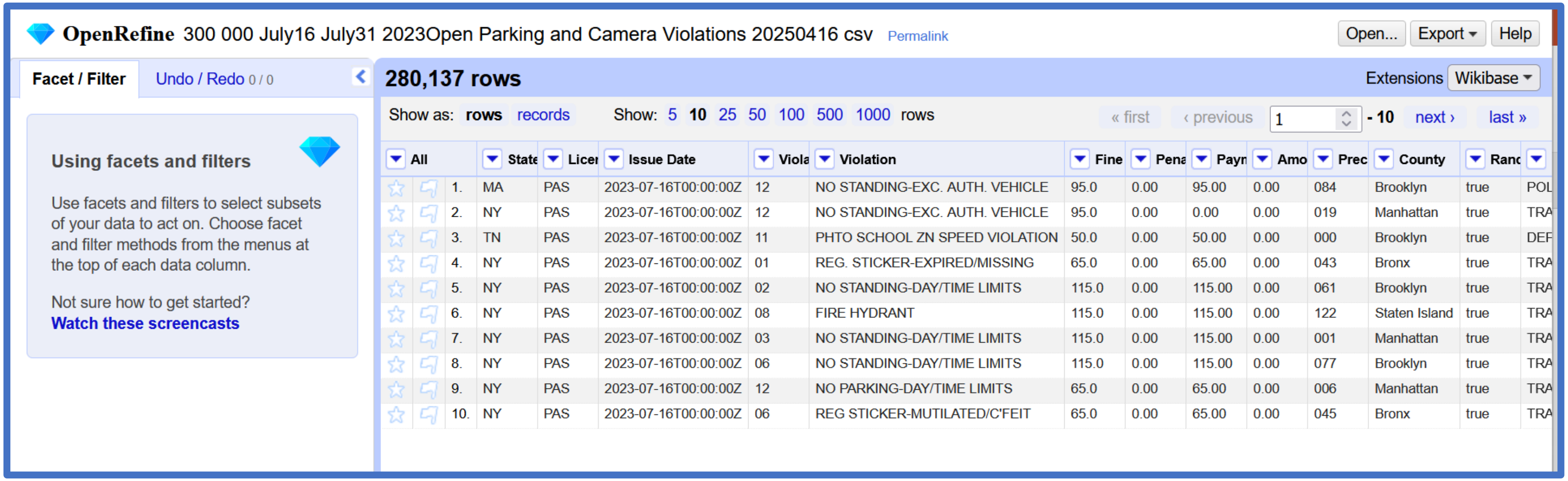
Rewind to any previous state of your dataset and replay your operation history on a new version of it.

Privacy

Your data is cleaned on your machine, not in some dubious data laundering cloud.

Wikibase

Contribute to Wikidata, the free knowledge base anyone can edit, and other Wikibase instances.



OpenRefine 300 000 July16 July31 2023Open Parking and Camera Violations 20250416 csv Permalink

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

280,137 rows

Show as: **rows** records   Show: 5 **10** 25 50 100 500 1000 rows   « first ‹ previous 1 - 10 next › last »

All	State	Licer	Issue Date	Viola	Violation	Fine	Pena	Payn	Amo	Prec	County	Ranc	
☆ 1.	MA	PAS	2023-07-16T00:00:00Z	12	NO STANDING-EXC. AUTH. VEHICLE	95.0	0.00	95.00	0.00	084	Brooklyn	true	POL
☆ 2.	NY	PAS	2023-07-16T00:00:00Z	12	NO STANDING-EXC. AUTH. VEHICLE	95.0	0.00	0.00	0.00	019	Manhattan	true	TRA
☆ 3.	TN	PAS	2023-07-16T00:00:00Z	11	PHTO SCHOOL ZN SPEED VIOLATION	50.0	0.00	50.00	0.00	000	Brooklyn	true	DEF
☆ 4.	NY	PAS	2023-07-16T00:00:00Z	01	REG. STICKER-EXPIRED/MISSING	65.0	0.00	65.00	0.00	043	Bronx	true	TRA
☆ 5.	NY	PAS	2023-07-16T00:00:00Z	02	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	061	Brooklyn	true	TRA
☆ 6.	NY	PAS	2023-07-16T00:00:00Z	08	FIRE HYDRANT	115.0	0.00	115.00	0.00	122	Staten Island	true	TRA
☆ 7.	NY	PAS	2023-07-16T00:00:00Z	03	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	001	Manhattan	true	TRA
☆ 8.	NY	PAS	2023-07-16T00:00:00Z	06	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	077	Brooklyn	true	TRA
☆ 9.	NY	PAS	2023-07-16T00:00:00Z	12	NO PARKING-DAY/TIME LIMITS	65.0	0.00	65.00	0.00	006	Manhattan	true	TRA
☆ 10.	NY	PAS	2023-07-16T00:00:00Z	06	REG STICKER-MUTILATED/C'FEIT	65.0	0.00	65.00	0.00	045	Bronx	true	TRA

- Loading/Visualizations: Orange Data Mining

Orange Data Mining is a powerful open source tool for visualizing various aspects of data without having to use complicated Python programming.

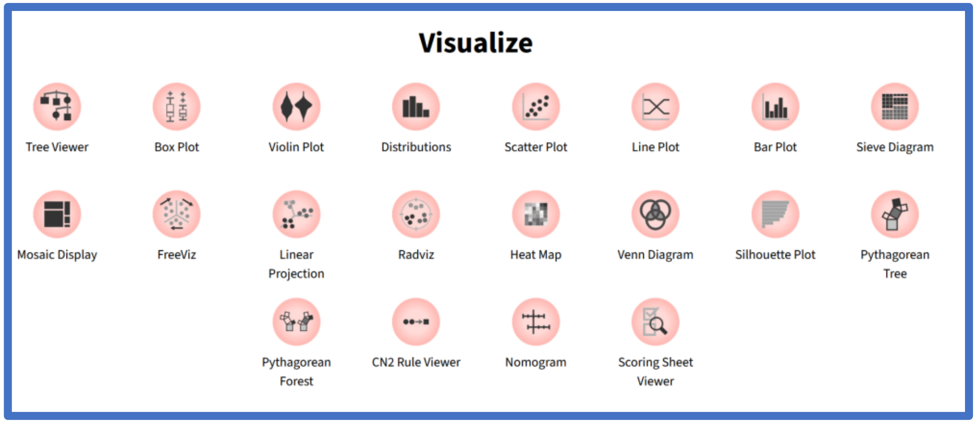
Data Mining
Fruitful and Fun

Open source machine learning and data visualization.

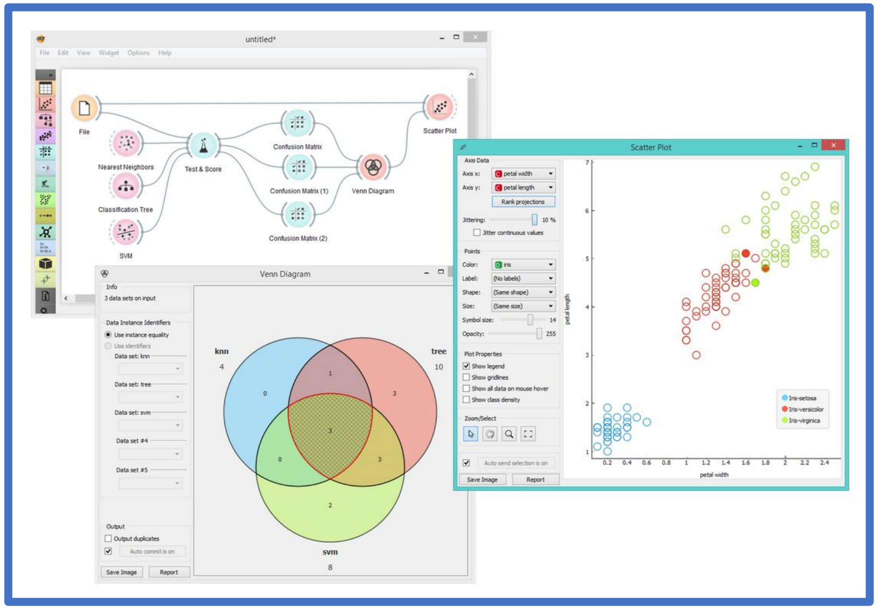
Download Orange 3.38.1

orange
DATA MINING

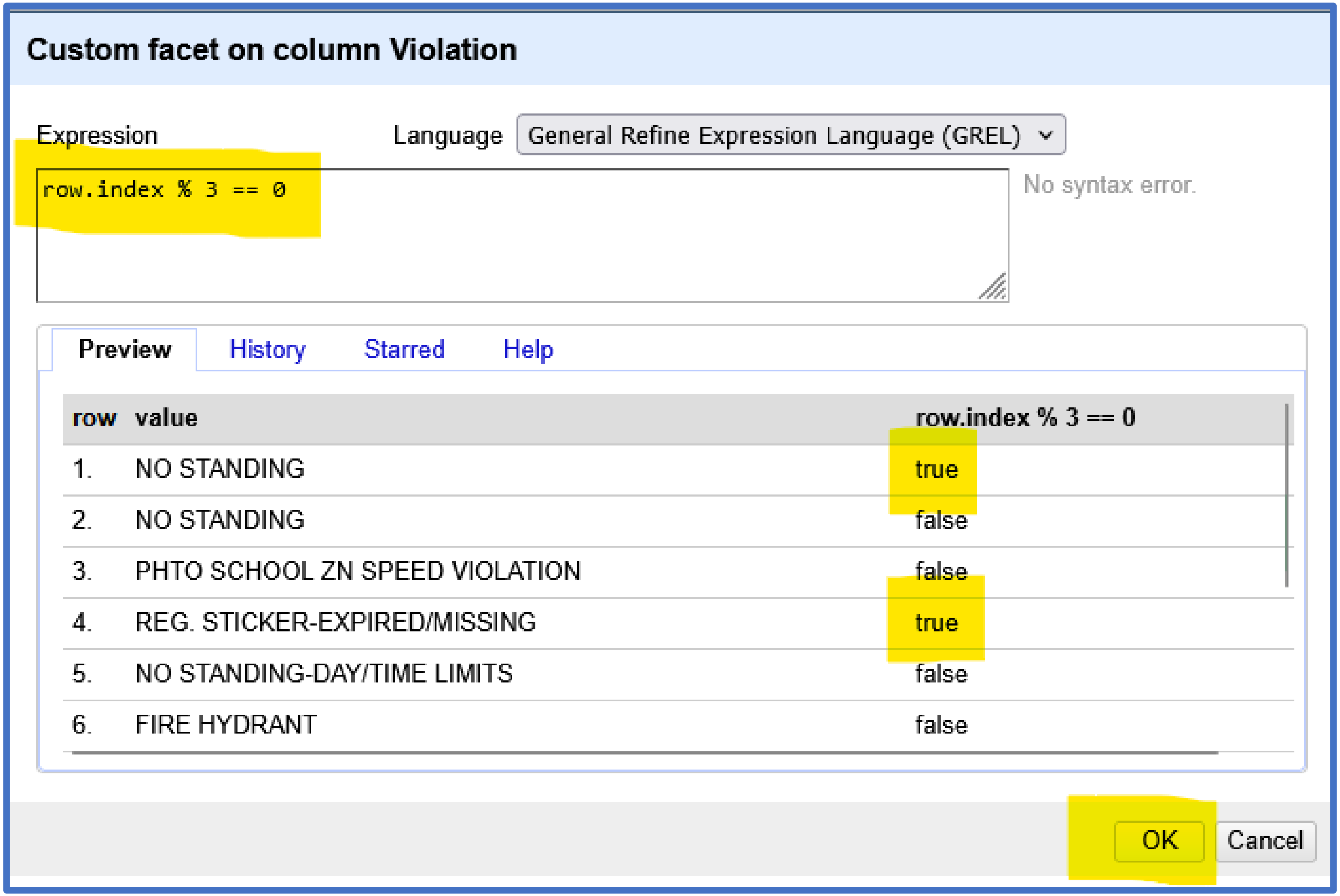
www.OrangeDataMining.com



Visualize



- Within OpenRefine, used based GREL [General Refine Expression Language]
GREL is a query language that can modify text and perform numerical calculations.



Custom facet on column Violation

Expression

Language: General Refine Expression Language (GREL)

```
row.index % 3 == 0
```

No syntax error.

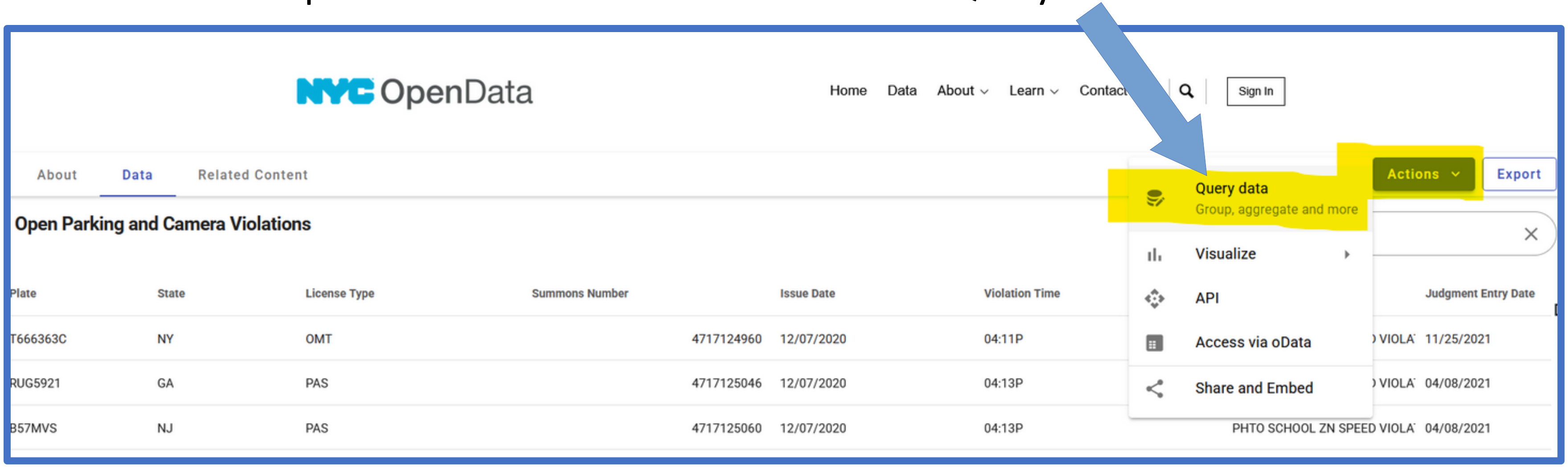
row	value	row.index % 3 == 0
1.	NO STANDING	true
2.	NO STANDING	false
3.	PHTO SCHOOL ZN SPEED VIOLATION	false
4.	REG. STICKER-EXPIRED/MISSING	true
5.	NO STANDING-DAY/TIME LIMITS	false
6.	FIRE HYDRANT	false

row.index % 3 == 0

ETL Process Using NYC OpenData, OpenRefine and Orange Data Mining

Extract Process

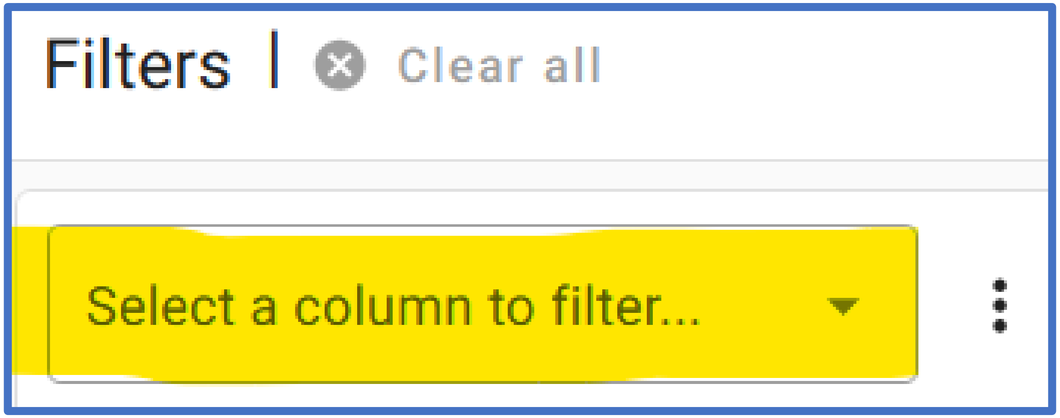
1. Go to NYC OpenData website > click on Actions > Query data



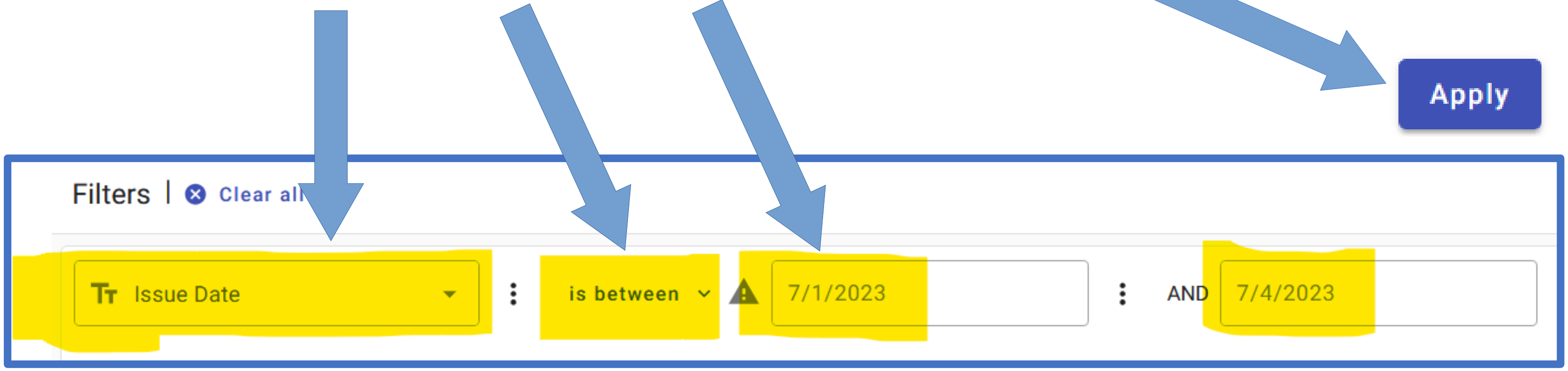
Open Parking and Camera Violations

Plate	State	License Type	Summons Number	Issue Date	Violation Time		Judgment Entry Date
T666363C	NY	OMT	4717124960	12/07/2020	04:11P	) VIOLA	11/25/2021
RUG5921	GA	PAS	4717125046	12/07/2020	04:13P	) VIOLA	04/08/2021
B57MVS	NJ	PAS	4717125060	12/07/2020	04:13P	PHTO SCHOOL ZN SPEED VIOLA	04/08/2021

2. Under Filters, click on “Select a column to filter”



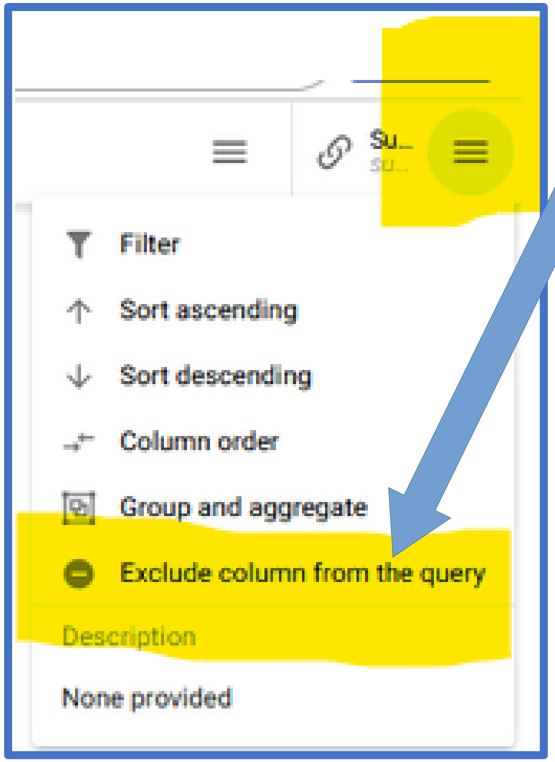
3. Select the Field > Boolean > search terms > Click “Apply”



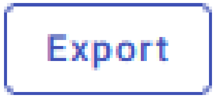
Filters | Clear all

Issue Date | is between | 7/1/2023 | AND | 7/4/2023

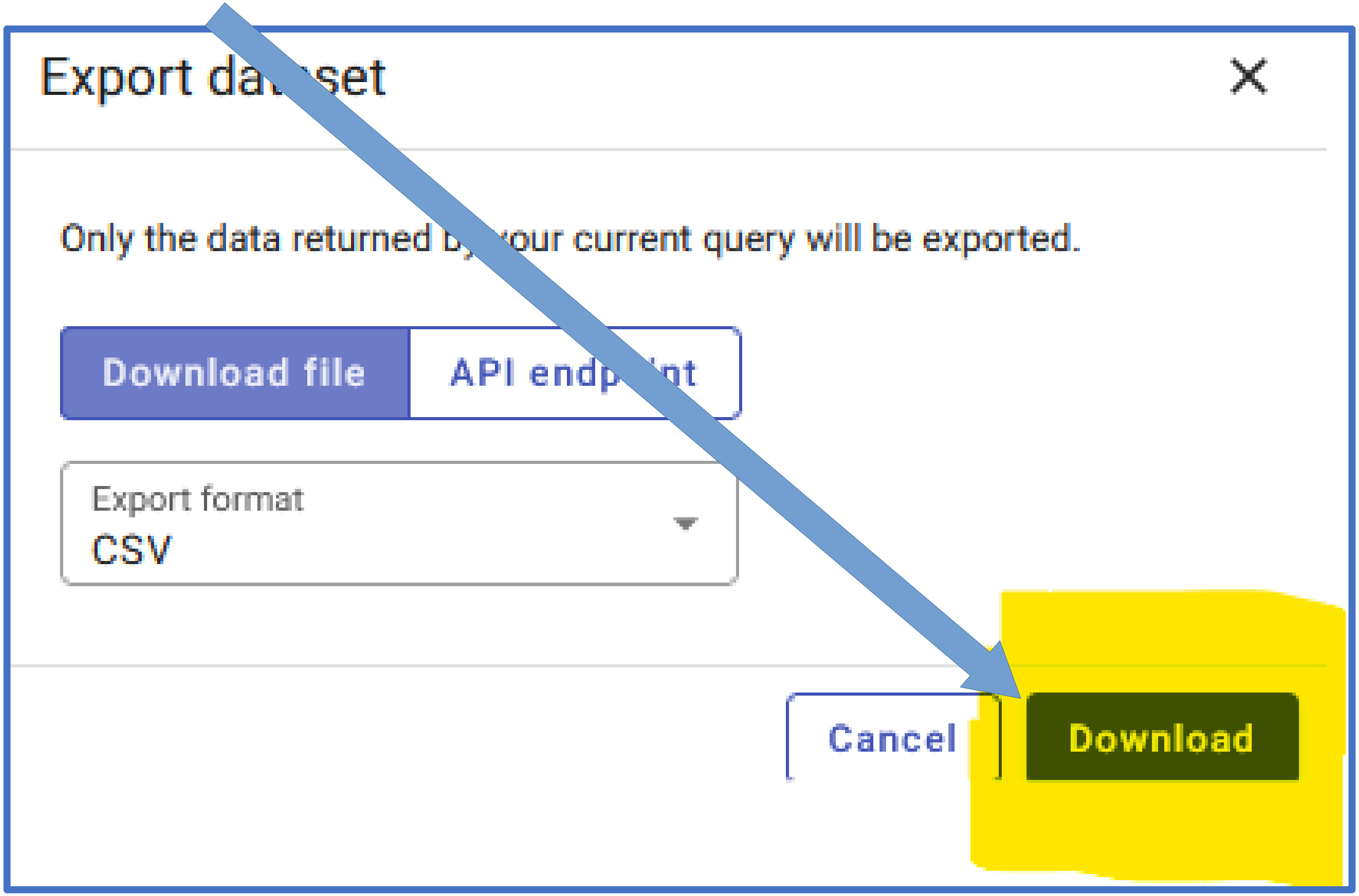
4. Remove columns that are not needed or have large amounts of data > find column > click on 3 horizontal lines > click on “Exclude column from the query”:



5. Click on “Export” button:



6. Click on “Download” to retrieve CSV file:



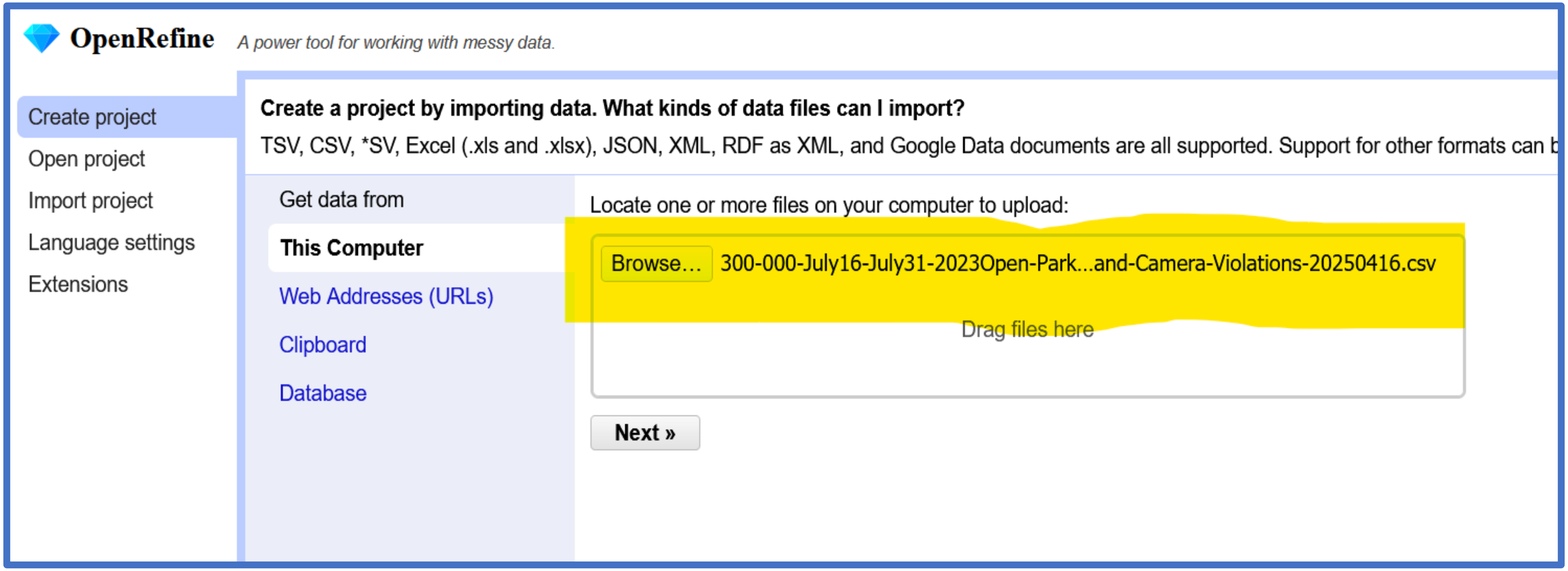
Export dataset

Only the data returned by your current query will be exported.

Export format
CSV

Transform Process

7. Import CSV file into OpenRefine > Browse to file > Next



OpenRefine *A power tool for working with messy data.*

Create a project by importing data. What kinds of data files can I import?

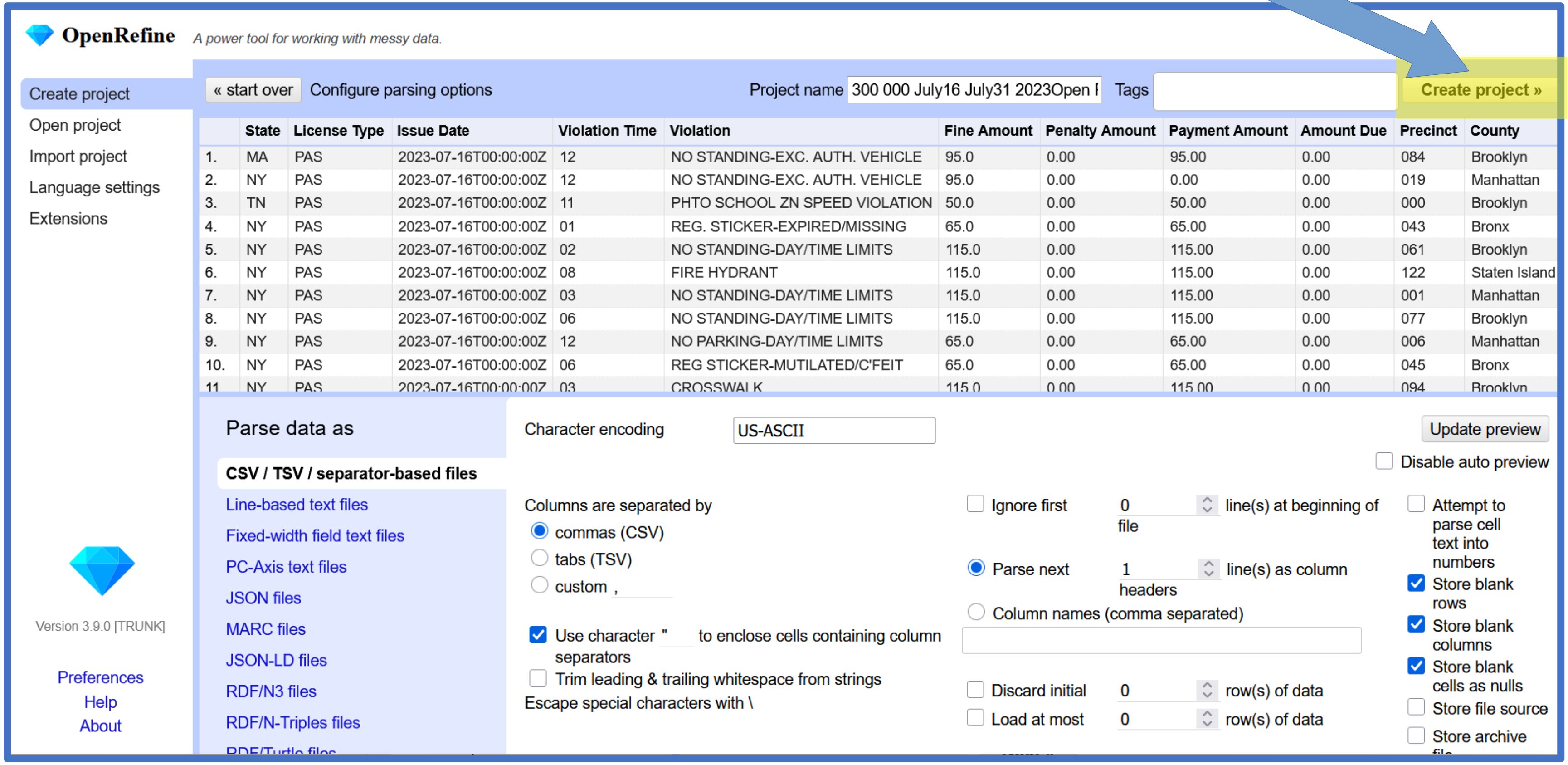
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can b

Locate one or more files on your computer to upload:

300-000-July16-July31-2023Open-Park...and-Camera-Violations-20250416.csv

OpenRefine

8. Initial upload of data > click on “Create project” button:



Project name: 300 000 July16 July31 2023Open I

Tags:

	State	License Type	Issue Date	Violation Time	Violation	Fine Amount	Penalty Amount	Payment Amount	Amount Due	Precinct	County
1.	MA	PAS	2023-07-16T00:00:00Z	12	NO STANDING-EXC. AUTH. VEHICLE	95.0	0.00	95.00	0.00	084	Brooklyn
2.	NY	PAS	2023-07-16T00:00:00Z	12	NO STANDING-EXC. AUTH. VEHICLE	95.0	0.00	0.00	0.00	019	Manhattan
3.	TN	PAS	2023-07-16T00:00:00Z	11	PHTO SCHOOL ZN SPEED VIOLATION	50.0	0.00	50.00	0.00	000	Brooklyn
4.	NY	PAS	2023-07-16T00:00:00Z	01	REG. STICKER-EXPIRED/MISSING	65.0	0.00	65.00	0.00	043	Bronx
5.	NY	PAS	2023-07-16T00:00:00Z	02	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	061	Brooklyn
6.	NY	PAS	2023-07-16T00:00:00Z	08	FIRE HYDRANT	115.0	0.00	115.00	0.00	122	Staten Island
7.	NY	PAS	2023-07-16T00:00:00Z	03	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	001	Manhattan
8.	NY	PAS	2023-07-16T00:00:00Z	06	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	077	Brooklyn
9.	NY	PAS	2023-07-16T00:00:00Z	12	NO PARKING-DAY/TIME LIMITS	65.0	0.00	65.00	0.00	006	Manhattan
10.	NY	PAS	2023-07-16T00:00:00Z	06	REG STICKER-MUTILATED/C'FEIT	65.0	0.00	65.00	0.00	045	Bronx
11.	NY	PAS	2023-07-16T00:00:00Z	03	CROSSWALK	115.0	0.00	115.00	0.00	094	Brooklyn

Character encoding: US-ASCII

□ Disable auto preview

Columns are separated by
● commas (CSV)
○ tabs (TSV)
○ custom ,

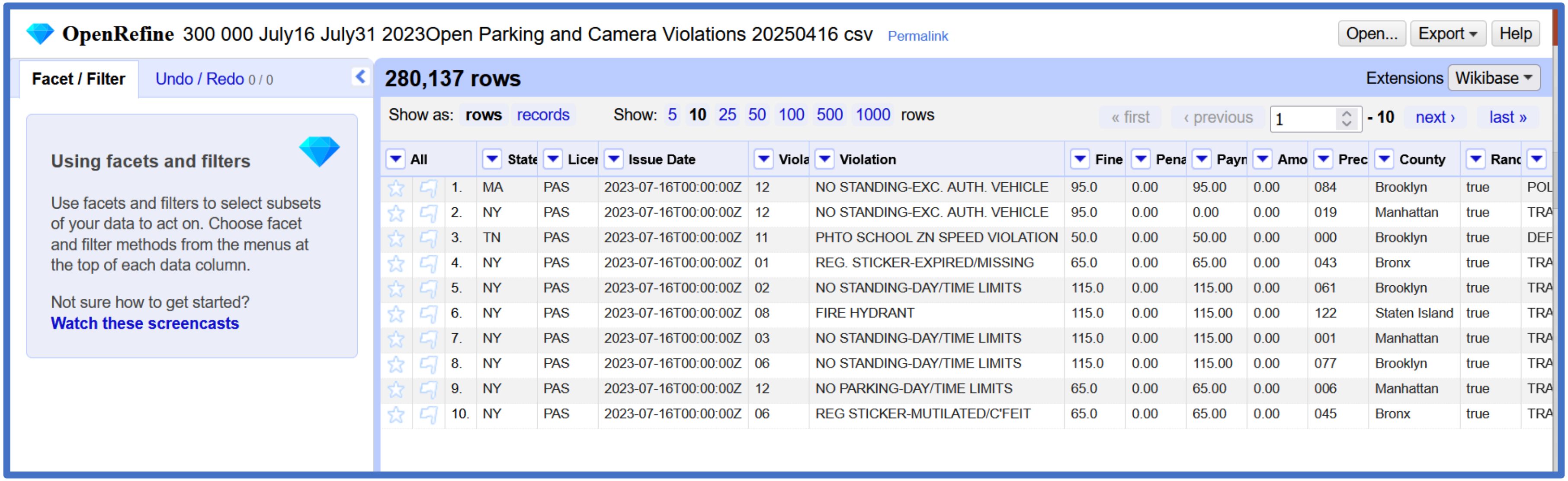
☑ Use character " to enclose cells containing column separators
□ Trim leading & trailing whitespace from strings
Escape special characters with \

□ Ignore first 0 line(s) at beginning of file
● Parse next 1 line(s) as column headers
○ Column names (comma separated)

□ Discard initial 0 row(s) of data
□ Load at most 0 row(s) of data

□ Attempt to parse cell text into numbers
☑ Store blank rows
☑ Store blank columns
☑ Store blank cells as nulls
□ Store file source
□ Store archive file

9. Creation of Project:



OpenRefine 300 000 July16 July31 2023Open Parking and Camera Violations 20250416 csv Permalink
Open... Export Help
Facet / Filter
Undo / Redo 0 / 0
Extensions Wikibase

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

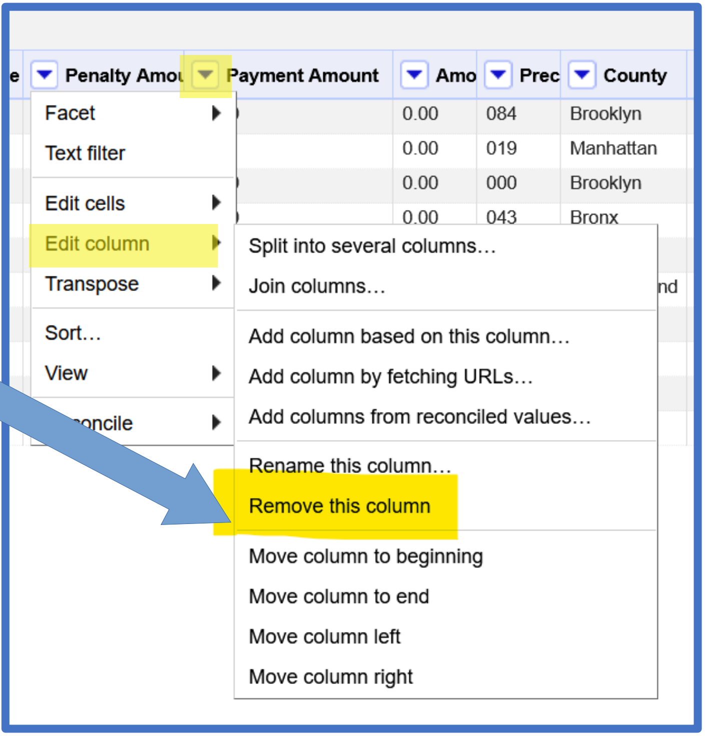
Not sure how to get started?
Watch these screencasts

280,137 rows

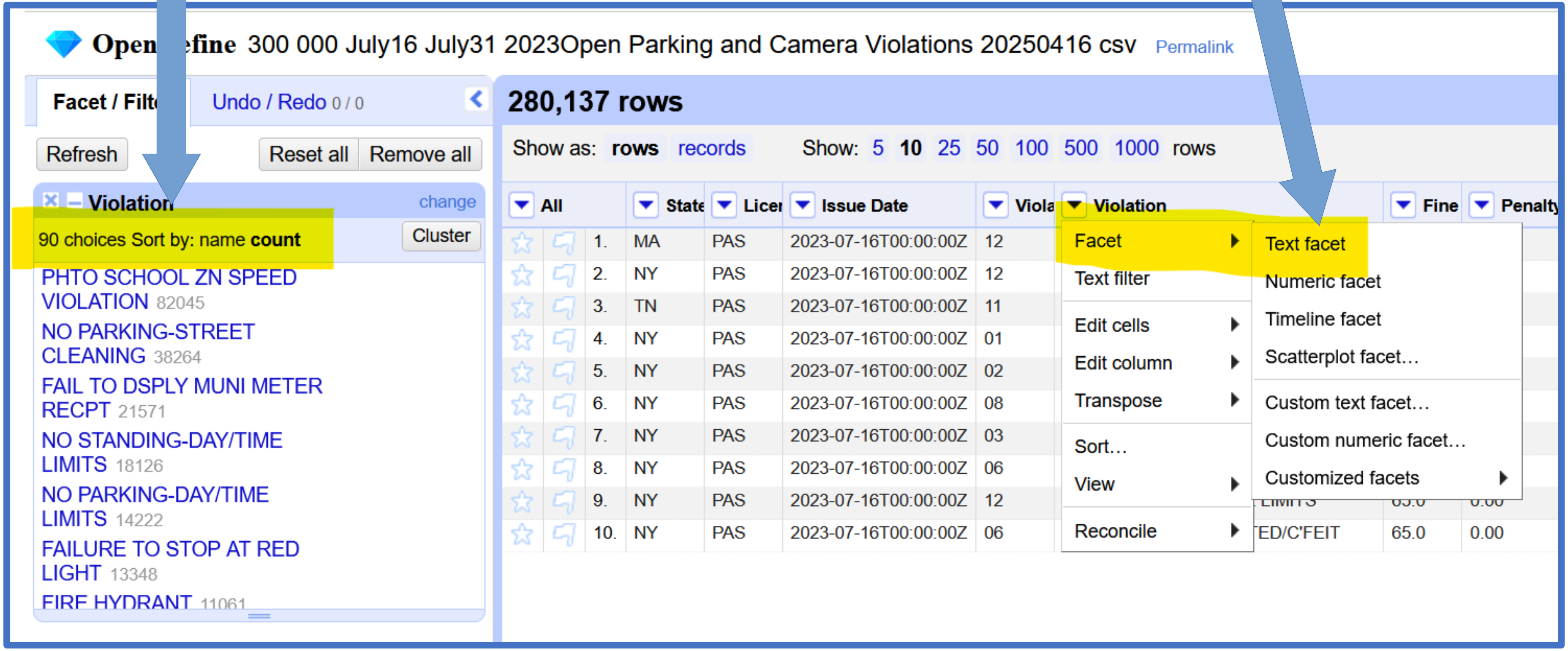
Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows « first ‹ previous 1 - 10 next › last »

All	State	Licer	Issue Date	Viola	Violation	Fine	Pena	Payn	Amo	Prec	County	Ranc	
1.	MA	PAS	2023-07-16T00:00:00Z	12	NO STANDING-EXC. AUTH. VEHICLE	95.0	0.00	95.00	0.00	084	Brooklyn	true	POL
2.	NY	PAS	2023-07-16T00:00:00Z	12	NO STANDING-EXC. AUTH. VEHICLE	95.0	0.00	0.00	0.00	019	Manhattan	true	TRA
3.	TN	PAS	2023-07-16T00:00:00Z	11	PHTO SCHOOL ZN SPEED VIOLATION	50.0	0.00	50.00	0.00	000	Brooklyn	true	DEF
4.	NY	PAS	2023-07-16T00:00:00Z	01	REG. STICKER-EXPIRED/MISSING	65.0	0.00	65.00	0.00	043	Bronx	true	TRA
5.	NY	PAS	2023-07-16T00:00:00Z	02	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	061	Brooklyn	true	TRA
6.	NY	PAS	2023-07-16T00:00:00Z	08	FIRE HYDRANT	115.0	0.00	115.00	0.00	122	Staten Island	true	TRA
7.	NY	PAS	2023-07-16T00:00:00Z	03	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	001	Manhattan	true	TRA
8.	NY	PAS	2023-07-16T00:00:00Z	06	NO STANDING-DAY/TIME LIMITS	115.0	0.00	115.00	0.00	077	Brooklyn	true	TRA
9.	NY	PAS	2023-07-16T00:00:00Z	12	NO PARKING-DAY/TIME LIMITS	65.0	0.00	65.00	0.00	006	Manhattan	true	TRA
10.	NY	PAS	2023-07-16T00:00:00Z	06	REG STICKER-MUTILATED/C'FEIT	65.0	0.00	65.00	0.00	045	Bronx	true	TRA

10. Remove columns that are not needed > click on down arrow of column > Edit column > Remove this column:



11. Generate text facets to understand the different types of data that are in the column > click on arrow of column data that you are interested in > Facet > Text facet > analysis box on left side:



300 000 July16 July31 2023Open Parking and Camera Violations 20250416 csv Permalink

280,137 rows

Violation change

90 choices Sort by: name count Cluster

PHTO SCHOOL ZN SPEED VIOLATION 82045

NO PARKING-STREET CLEANING 38264

FAIL TO DSPLY MUNI METER RECPT 21571

NO STANDING-DAY/TIME LIMITS 18126

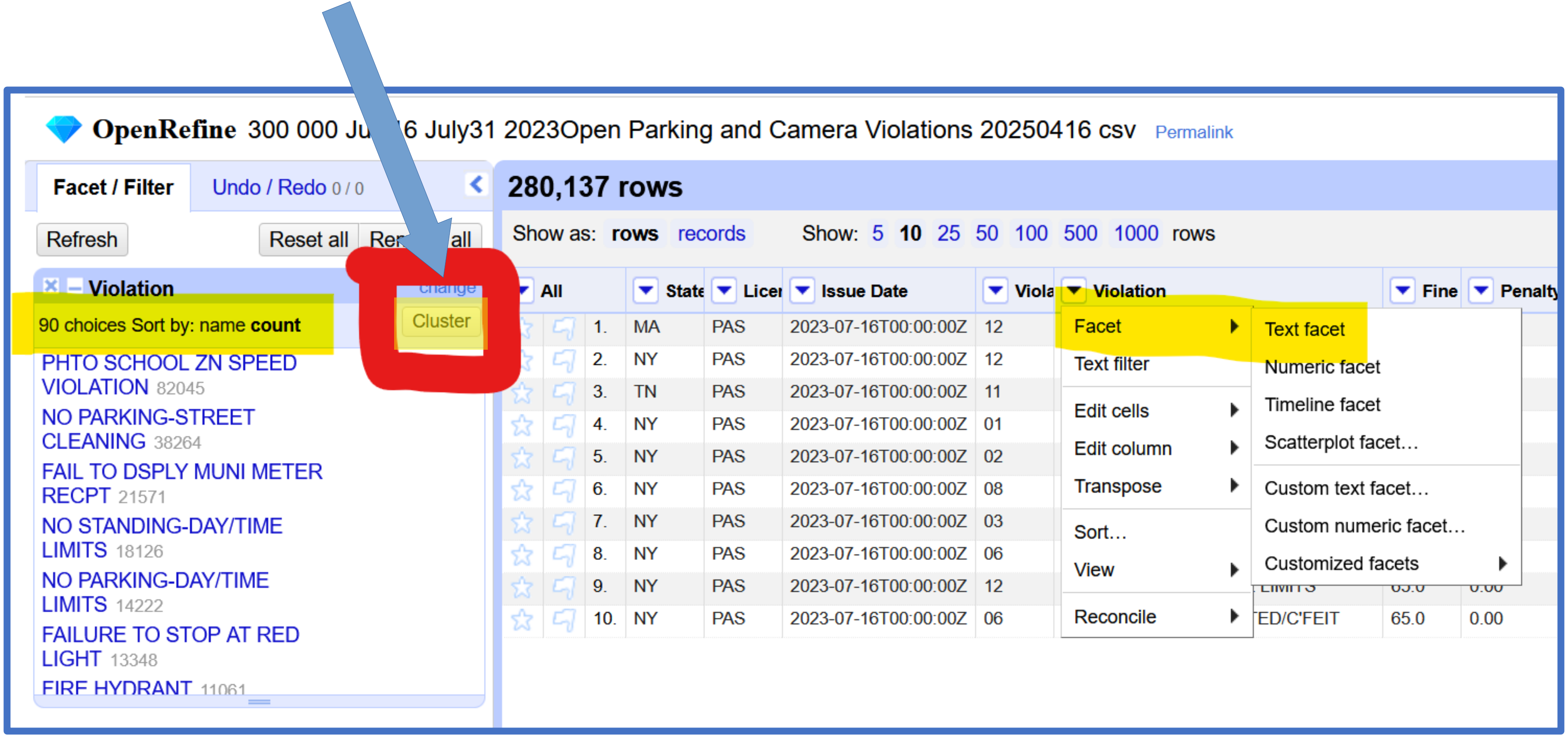
NO PARKING-DAY/TIME LIMITS 14222

FAILURE TO STOP AT RED LIGHT 13348

FIRE HYDRANT 11061

All	State	Licen	Issue Date	Viola	Violation	Fine	Penalty
1.	MA	PAS	2023-07-16T00:00:00Z	12			
2.	NY	PAS	2023-07-16T00:00:00Z	12			
3.	TN	PAS	2023-07-16T00:00:00Z	11			
4.	NY	PAS	2023-07-16T00:00:00Z	01			
5.	NY	PAS	2023-07-16T00:00:00Z	02			
6.	NY	PAS	2023-07-16T00:00:00Z	08			
7.	NY	PAS	2023-07-16T00:00:00Z	03			
8.	NY	PAS	2023-07-16T00:00:00Z	06			
9.	NY	PAS	2023-07-16T00:00:00Z	12	LIMITS	65.0	0.00
10.	NY	PAS	2023-07-16T00:00:00Z	06	TED/C'FEIT	65.0	0.00

12. Click on “Cluster” button to begin clustering process:



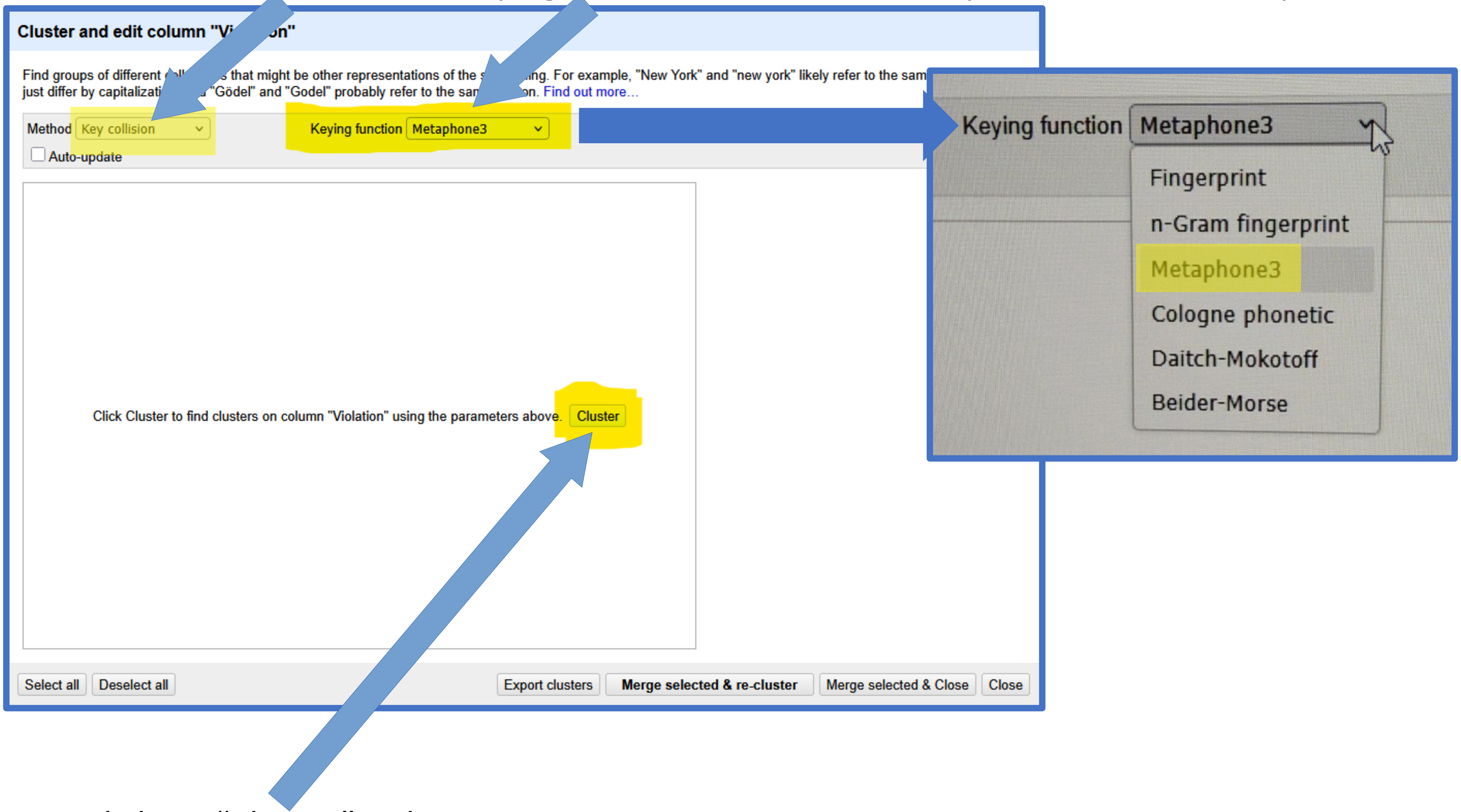
90 choices Sort by: name **count**

PHTO SCHOOL ZN SPEED VIOLATION 82045
NO PARKING-STREET CLEANING 38264
FAIL TO DSPLY MUNI METER RECPT 21571
NO STANDING-DAY/TIME LIMITS 18126
NO PARKING-DAY/TIME LIMITS 14222
FAILURE TO STOP AT RED LIGHT 13348
FIRE HYDRANT 11061

All	State	Licer	Issue Date	Viola	Violation	Fine	Penalty
1.	MA	PAS	2023-07-16T00:00:00Z	12			
2.	NY	PAS	2023-07-16T00:00:00Z	12			
3.	TN	PAS	2023-07-16T00:00:00Z	11			
4.	NY	PAS	2023-07-16T00:00:00Z	01			
5.	NY	PAS	2023-07-16T00:00:00Z	02			
6.	NY	PAS	2023-07-16T00:00:00Z	08			
7.	NY	PAS	2023-07-16T00:00:00Z	03			
8.	NY	PAS	2023-07-16T00:00:00Z	06			
9.	NY	PAS	2023-07-16T00:00:00Z	12	LIMITS	65.0	0.00
10.	NY	PAS	2023-07-16T00:00:00Z	06	ED/C'FEIT	65.0	0.00

Clustering refers to finding similar data entries that could be combined under one topic.

13. Select “Method” and “Keying function” [in this case, Key Collision and Metaphone3:



Cluster and edit column "Vi   on"

Find groups of different c   s that might be other representations of the s   ing. For example, "New York" and "new york" likely refer to the sam
just differ by capitalizati   "Gödel" and "Godel" probably refer to the sar   on. Find out more…

Method Key collision

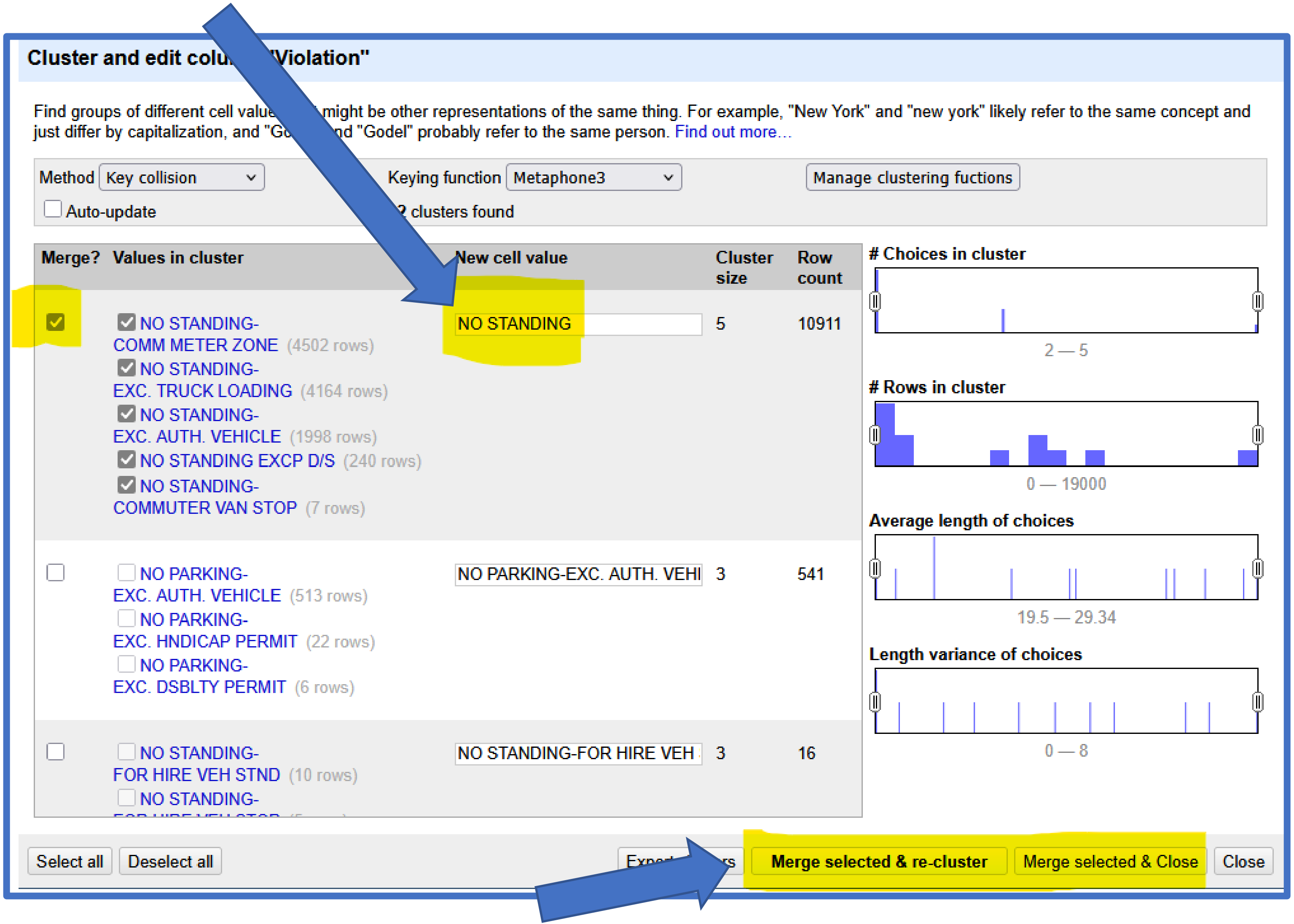
Keying function Metaphone3

□ Auto-update

Click Cluster to find clusters on column "Violation" using the parameters above.

7. Click on “Cluster” to begin process

14. Find values that can be consolidated, merge and type in replacement value in “New Cell Value”:



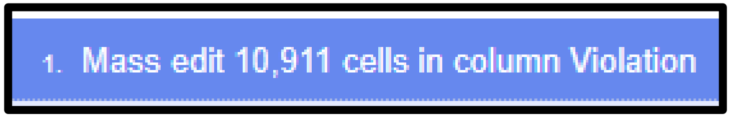
Cluster and edit colu "Violation"

Find groups of different cell value might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Go nd "Godel" probably refer to the same person. Find out more…

Method: Key collision
Keying function: Metaphone3
Manage clustering fuctions
☐ Auto-update
2 clusters found

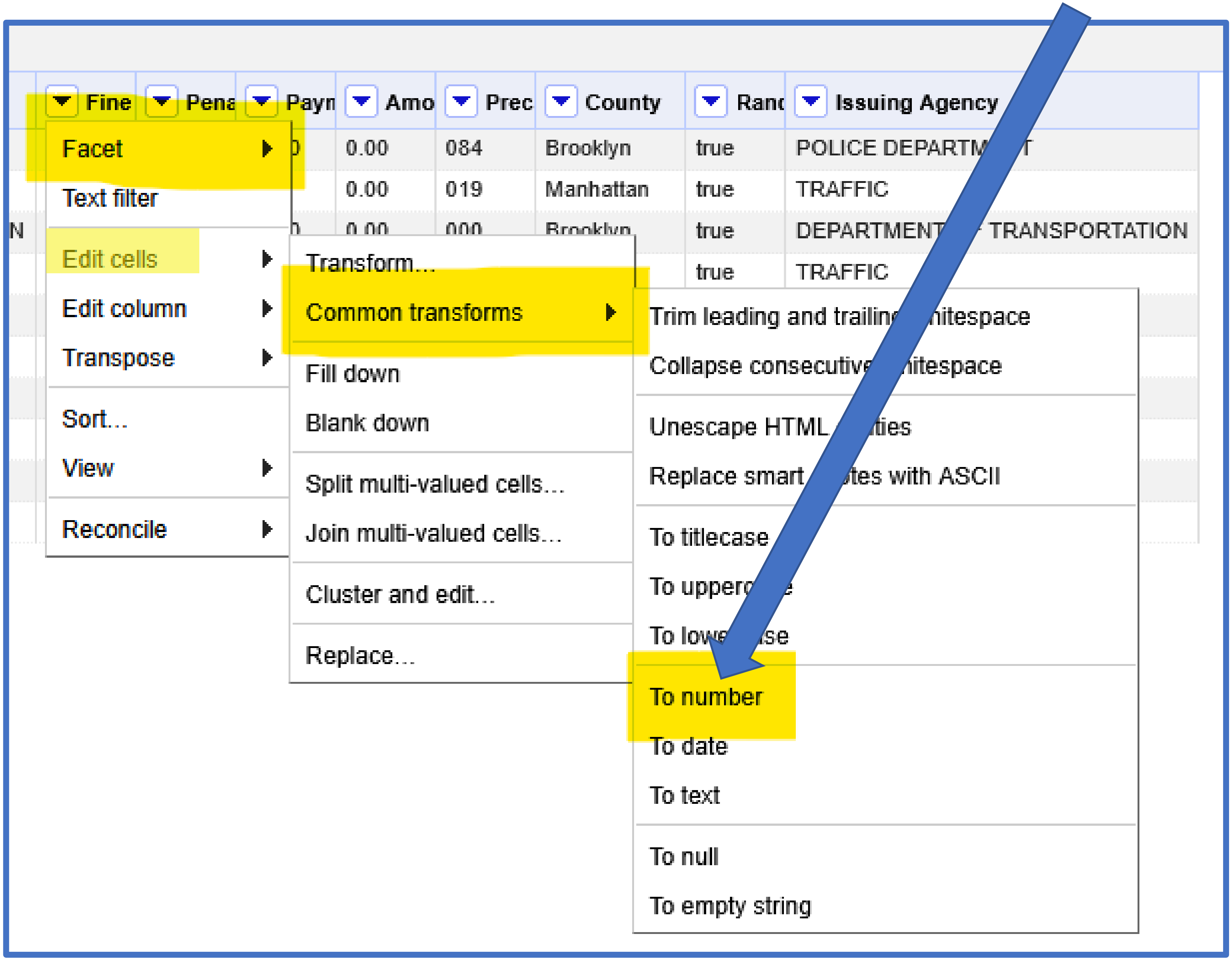
Merge?	Values in cluster	New cell value	Cluster size	Row count
☑	☑ NO STANDING-COMM METER ZONE (4502 rows) ☑ NO STANDING-EXC. TRUCK LOADING (4164 rows) ☑ NO STANDING-EXC. AUTH. VEHICLE (1998 rows) ☑ NO STANDING EXCP D/S (240 rows) ☑ NO STANDING-COMMUTER VAN STOP (7 rows)	NO STANDING	5	10911
☐	☐ NO PARKING-EXC. AUTH. VEHICLE (513 rows) ☐ NO PARKING-EXC. HNDICAP PERMIT (22 rows) ☐ NO PARKING-EXC. DSBLTY PERMIT (6 rows)	NO PARKING-EXC. AUTH. VEHI	3	541
☐	☐ NO STANDING-FOR HIRE VEH STND (10 rows) ☐ NO STANDING-	NO STANDING-FOR HIRE VEH	3	16

15. Click on one of the “Merge Selected” to begin process:



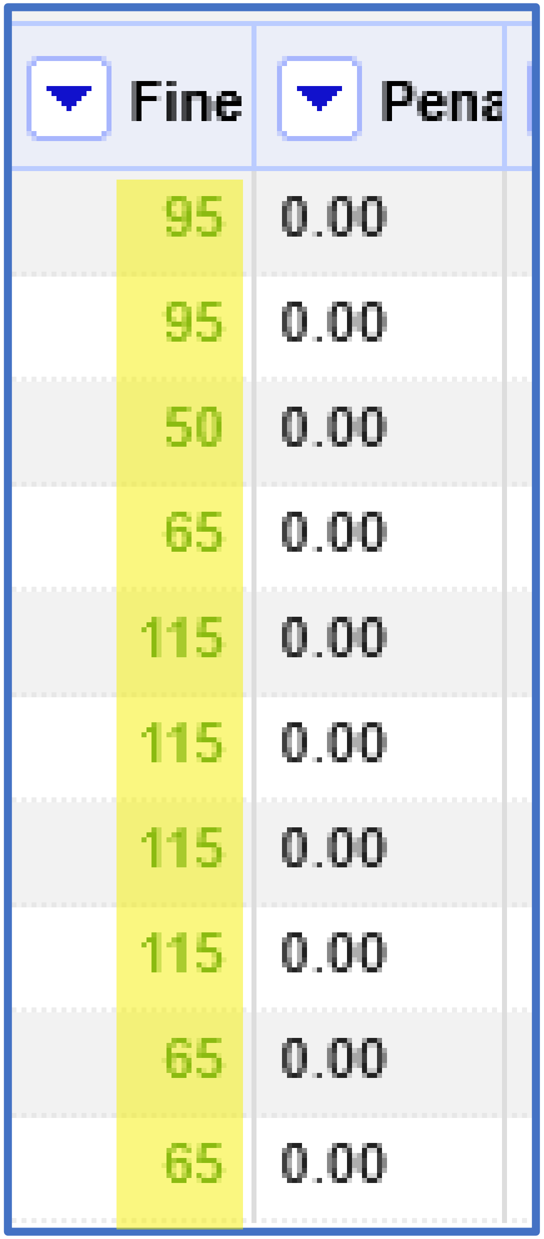
1. Mass edit 10,911 cells in column Violation

16. Convert text to number: Facet > Edit cells > Common transforms > To Number:



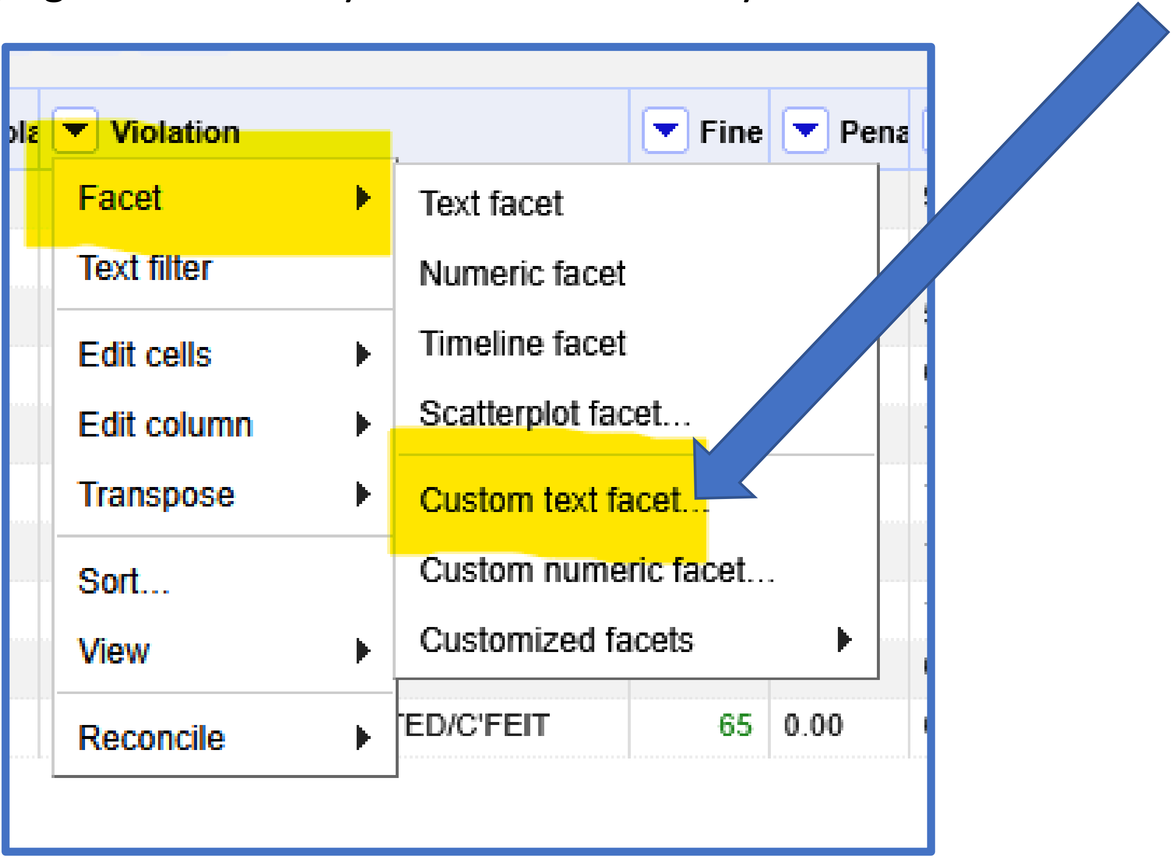
Fine	Pena	Payn	Amo	Prec	County	Ranc	Issuing Agency
		0	0.00	084	Brooklyn	true	POLICE DEPARTM T
			0.00	019	Manhattan	true	TRAFFIC
N		0	0.00	000	Brooklyn	true	DEPARTMENT TRANSPORTATION
						true	TRAFFIC

17. Text is converted into numbers [turns green]:



Fine	Pena
95	0.00
95	0.00
50	0.00
65	0.00
115	0.00
115	0.00
115	0.00
115	0.00
65	0.00
65	0.00

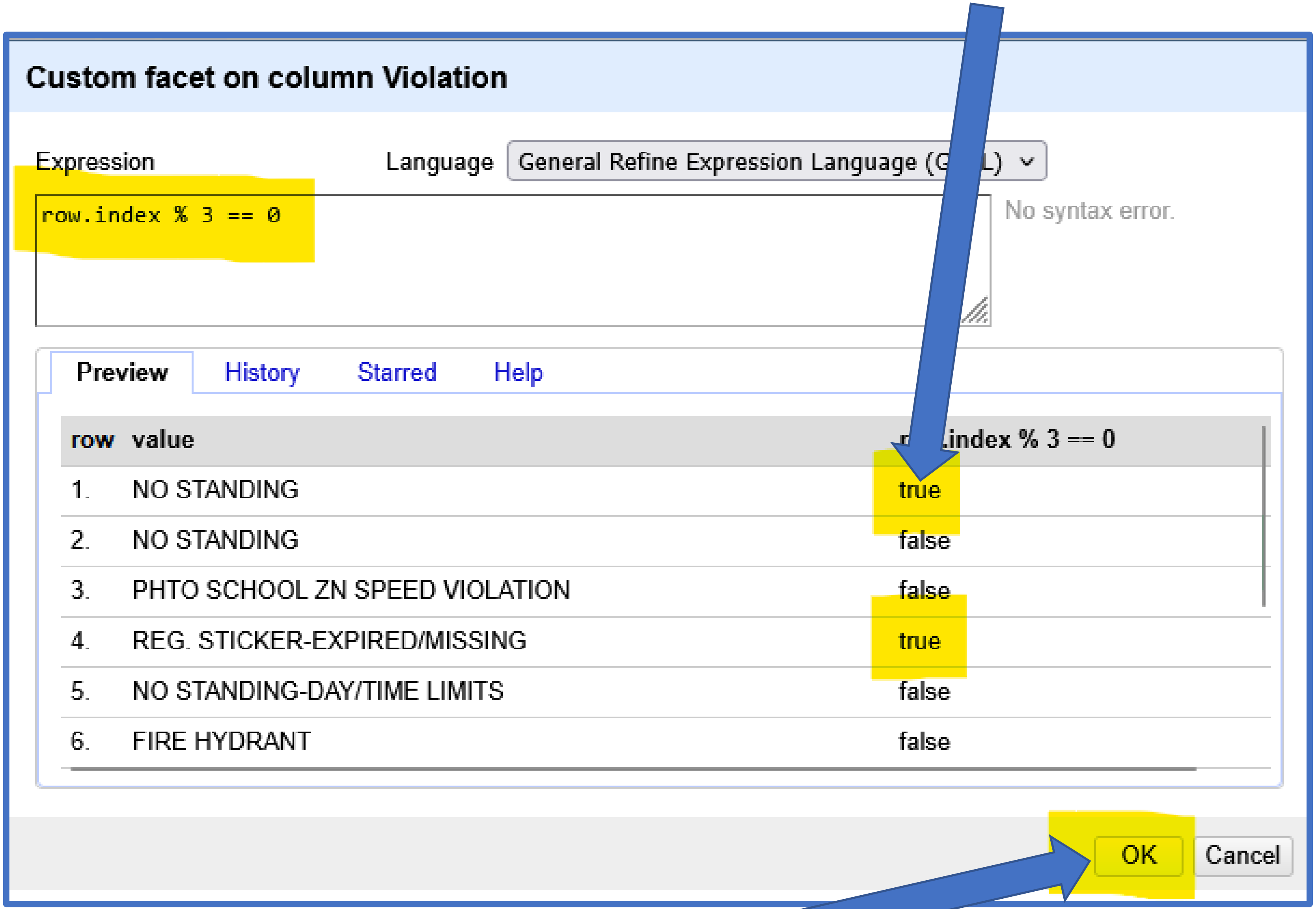
18. Extract sample size, eg: 1 out of every 3 rows. Choose any columne > Facet > Custom text facet:



19. In the “Expression” box, type:

```
row.index % 3 == 0
```

20. This will create a new column that will set every 3rd row to “True”:



Custom facet on column Violation

Expression

```
row.index % 3 == 0
```

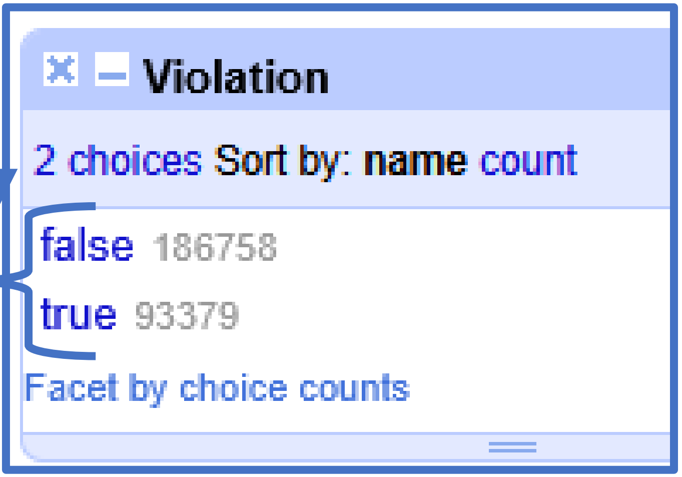
Language: General Refine Expression Language (GREL)

No syntax error.

row	value	row.index % 3 == 0
1.	NO STANDING	true
2.	NO STANDING	false
3.	PHTO SCHOOL ZN SPEED VIOLATION	false
4.	REG. STICKER-EXPIRED/MISSING	true
5.	NO STANDING-DAY/TIME LIMITS	false
6.	FIRE HYDRANT	false

21. Click “OK” on lower right hand corner:

22. A text facet with “True” and “False” will be created:



Violation

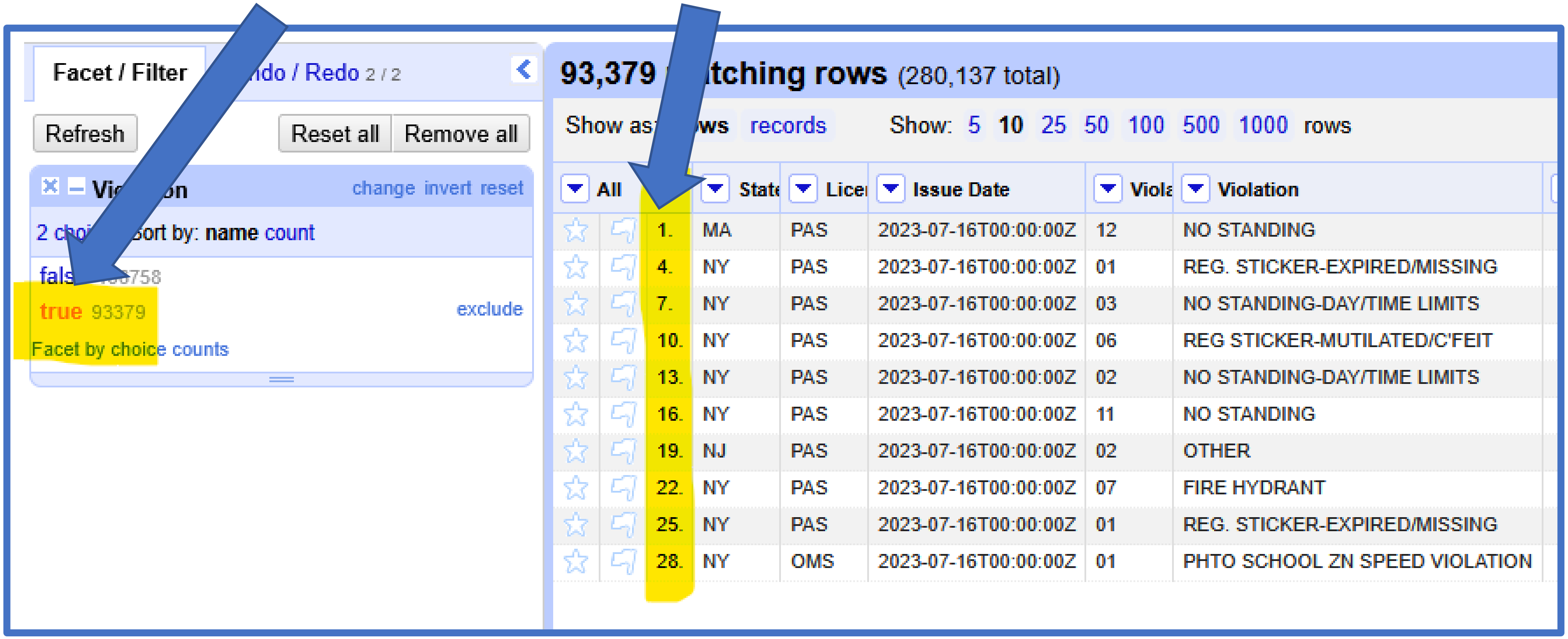
2 choices Sort by: name count

false 186758

true 93379

Facet by choice counts

23. Click on “True” to extract every 3rd value:



Facet / Filter   ndo / Redo 2 / 2

Refresh   Reset all   Remove all

Vi on   change invert reset

2 choi Sort by: name count

fals 6758

true 93379   exclude

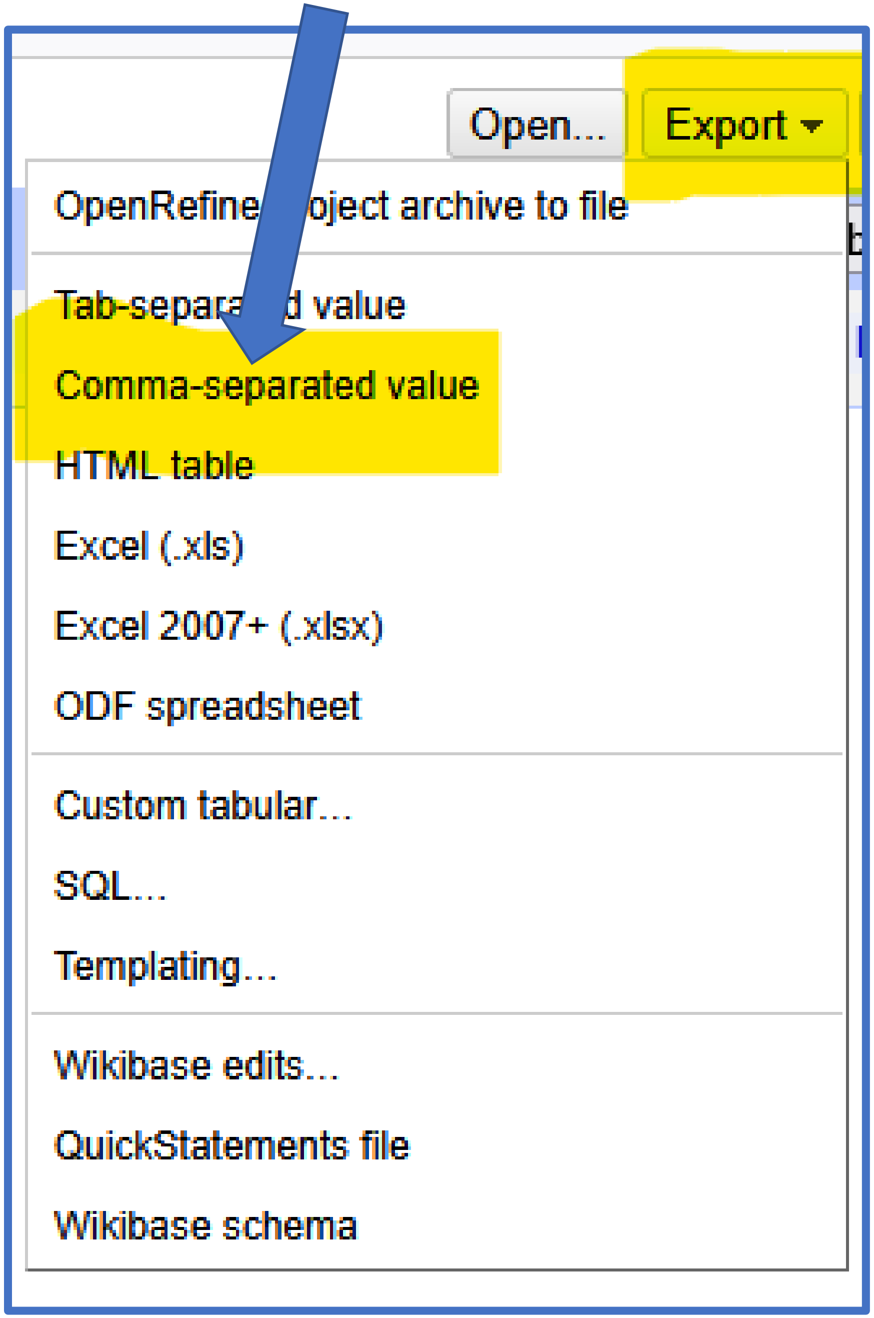
Facet by choice counts

93,379 tching rows (280,137 total)

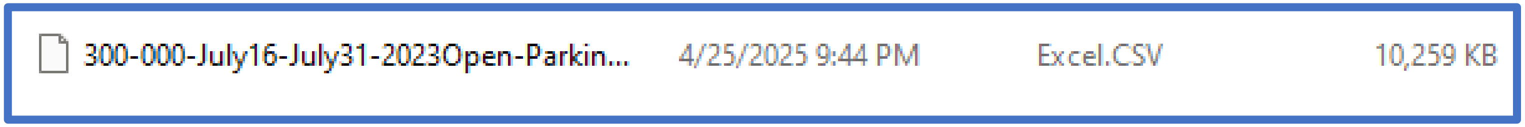
Show as: ws records   Show: 5 10 25 50 100 500 1000 rows

All	State	Licei	Issue Date	Viola	Violation
1.	MA	PAS	2023-07-16T00:00:00Z	12	NO STANDING
4.	NY	PAS	2023-07-16T00:00:00Z	01	REG. STICKER-EXPIRED/MISSING
7.	NY	PAS	2023-07-16T00:00:00Z	03	NO STANDING-DAY/TIME LIMITS
10.	NY	PAS	2023-07-16T00:00:00Z	06	REG STICKER-MUTILATED/C'FEIT
13.	NY	PAS	2023-07-16T00:00:00Z	02	NO STANDING-DAY/TIME LIMITS
16.	NY	PAS	2023-07-16T00:00:00Z	11	NO STANDING
19.	NJ	PAS	2023-07-16T00:00:00Z	02	OTHER
22.	NY	PAS	2023-07-16T00:00:00Z	07	FIRE HYDRANT
25.	NY	PAS	2023-07-16T00:00:00Z	01	REG. STICKER-EXPIRED/MISSING
28.	NY	OMS	2023-07-16T00:00:00Z	01	PHTO SCHOOL ZN SPEED VIOLATION

24. Extract sample size: Export > Comma-separated value:

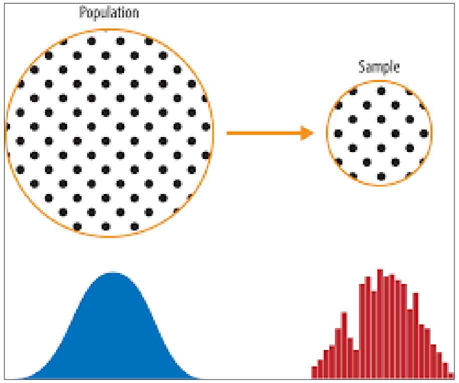


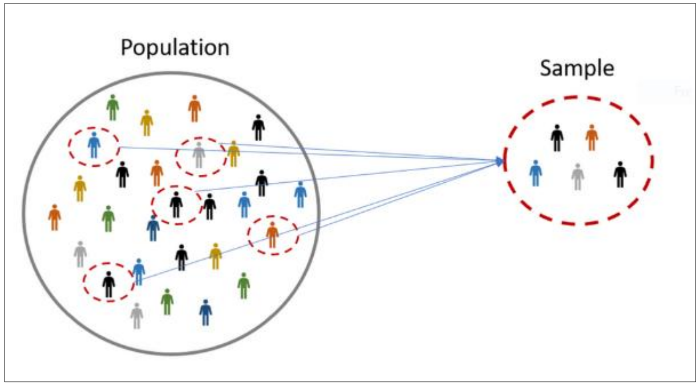
25. CSV file will be 1/3 the size of the original file:

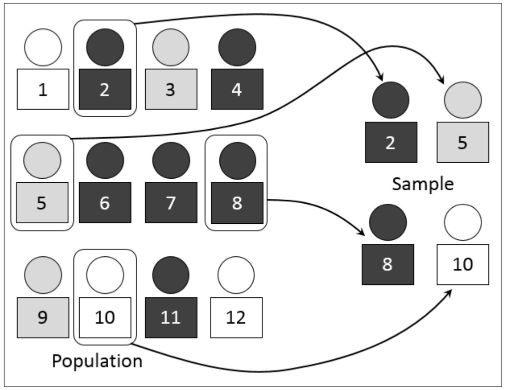


300-000-July16-July31-2023Open-Parkin... 4/25/2025 9:44 PM Excel.CSV 10,259 KB

Overall, was able to reduce 70 million rows covering 7 months to 439,000 rows.







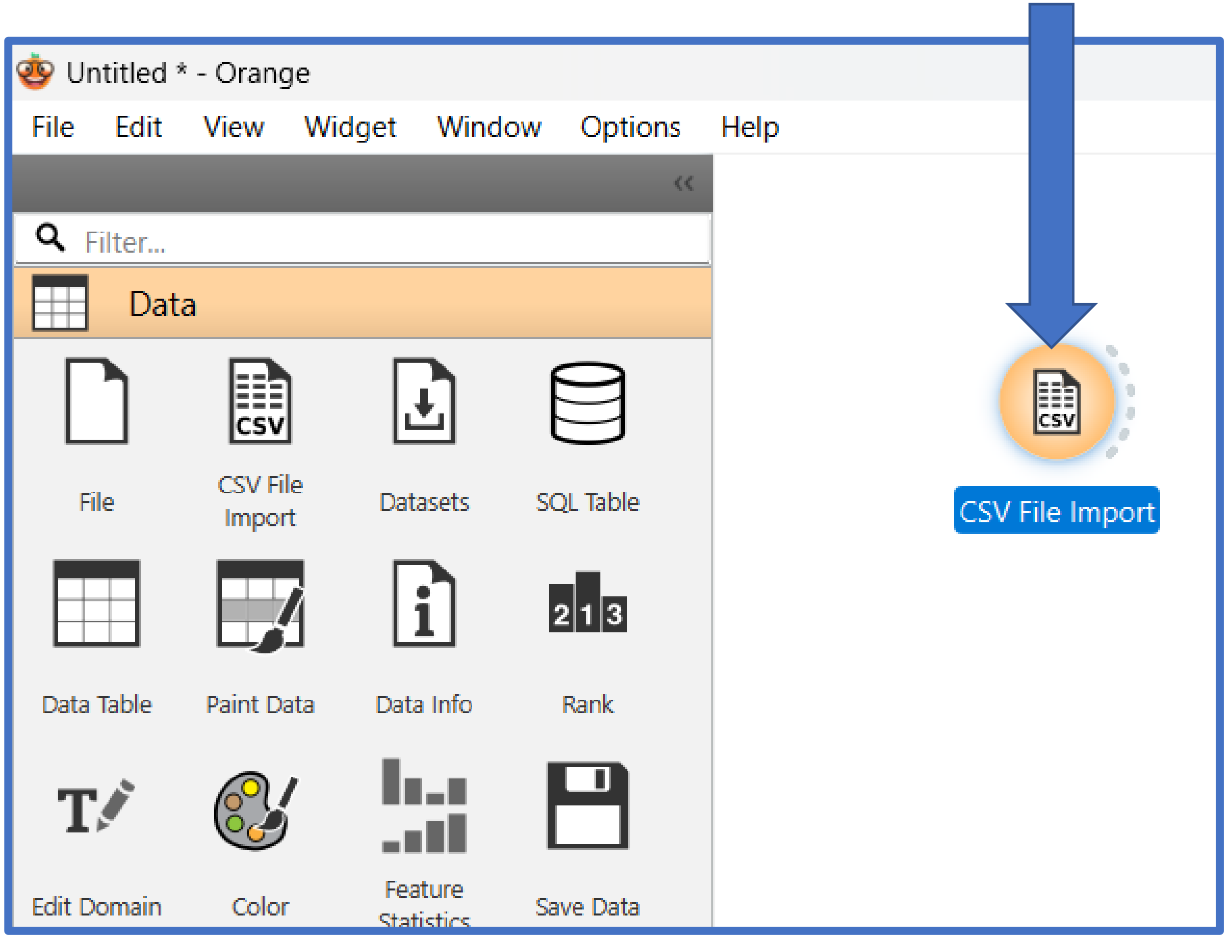
Load Process

Load file into Orange Data Miner:

Data Mining Fruitful and Fun

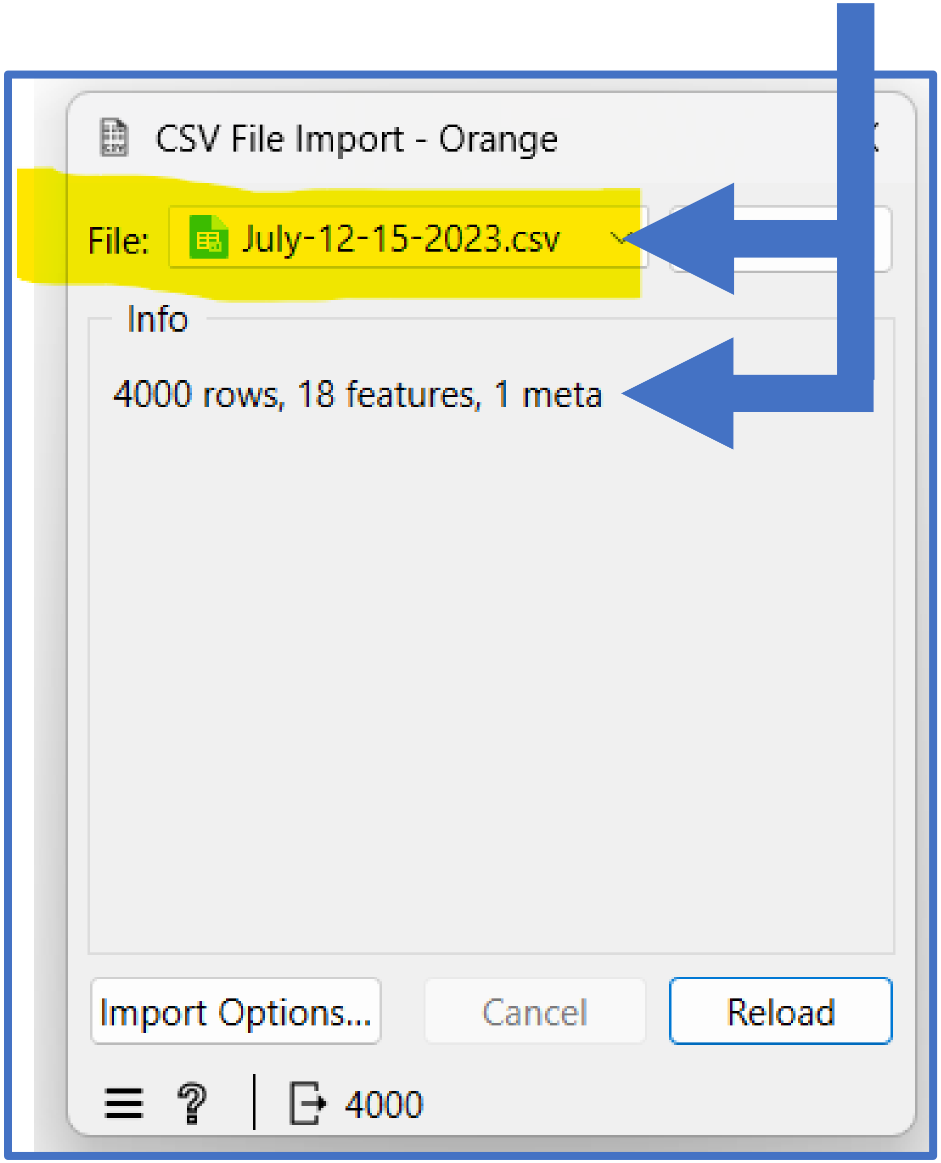
Open source machine learning and data visualization.

Open Orange Data Mining and drag the CSV File Import widget into the work area:



orange DATA MINING

Double click, find file and load file into Orange Data Miner:



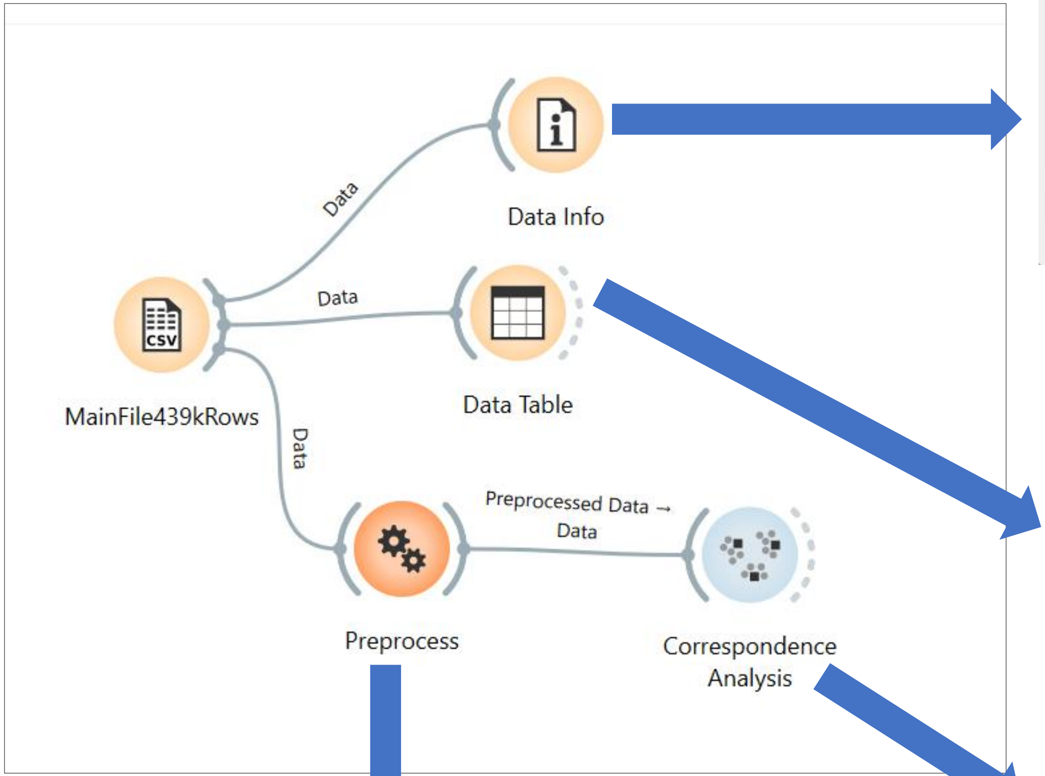
CSV File Import - Orange

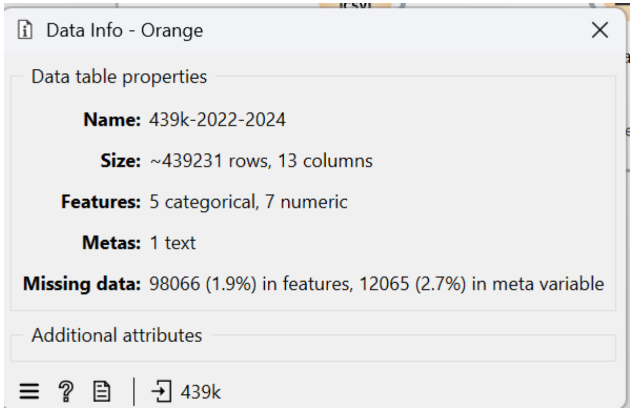
File: July-12-15-2023.csv

Info

4000 rows, 18 features, 1 meta

Main File with 439,000 rows with basic analysis:





Data table properties

Name: 439k-2022-2024

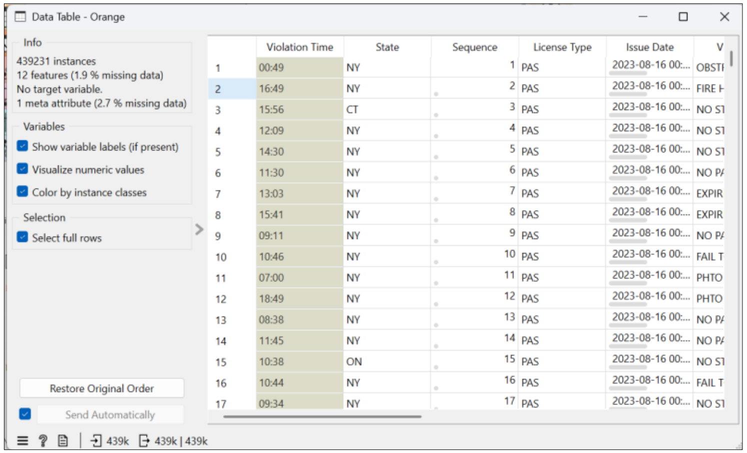
Size: ~439231 rows, 13 columns

Features: 5 categorical, 7 numeric

Metas: 1 text

Missing data: 98066 (1.9%) in features, 12065 (2.7%) in meta variable

Additional attributes



439231 instances
12 features (1.9 % missing data)
No target variable.
1 meta attribute (2.7 % missing data)

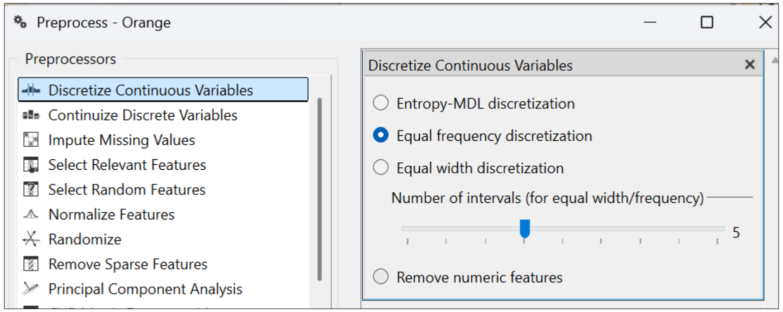
Variables

- ☑ Show variable labels (if present)
- ☑ Visualize numeric values
- ☑ Color by instance classes

Selection

- ☑ Select full rows

	Violation Time	State	Sequence	License Type	Issue Date	V
1	00:49	NY	1	PAS	2023-08-16 00:...	OBSTI
2	16:49	NY	2	PAS	2023-08-16 00:...	FIRE H
3	15:56	CT	3	PAS	2023-08-16 00:...	NO ST
4	12:09	NY	4	PAS	2023-08-16 00:...	NO ST
5	14:30	NY	5	PAS	2023-08-16 00:...	NO ST
6	11:30	NY	6	PAS	2023-08-16 00:...	NO PA
7	13:03	NY	7	PAS	2023-08-16 00:...	EXPIR
8	15:41	NY	8	PAS	2023-08-16 00:...	EXPIR
9	09:11	NY	9	PAS	2023-08-16 00:...	NO PA
10	10:46	NY	10	PAS	2023-08-16 00:...	FAIL T
11	07:00	NY	11	PAS	2023-08-16 00:...	PHTO
12	18:49	NY	12	PAS	2023-08-16 00:...	PHTO
13	08:38	NY	13	PAS	2023-08-16 00:...	NO PA
14	11:45	NY	14	PAS	2023-08-16 00:...	NO PA
15	10:38	ON	15	PAS	2023-08-16 00:...	NO ST
16	10:44	NY	16	PAS	2023-08-16 00:...	FAIL T
17	09:34	NY	17	PAS	2023-08-16 00:...	NO ST



Preprocessors

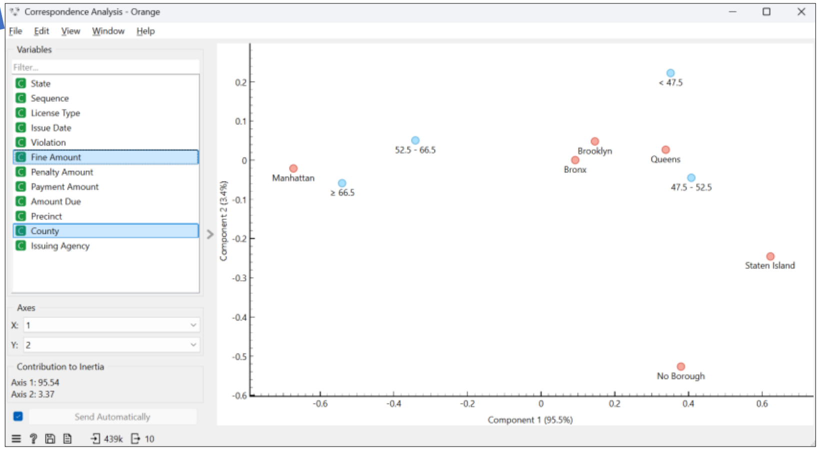
- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Remove Sparse Features
- Principal Component Analysis

Discretize Continuous Variables

- ○ Entropy-MDL discretization
- ● Equal frequency discretization
- ○ Equal width discretization

Number of intervals (for equal width/frequency): 5

- ○ Remove numeric features



- State
- Sequence
- License Type
- Issue Date
- Violation
- Fine Amount
- Penalty Amount
- Payment Amount
- Amount Due
- Precinct
- County
- Issuing Agency

Axes

X: 1

Y: 2

Contribution to Inertia

Axis 1: 95.54
Axis 2: 3.37

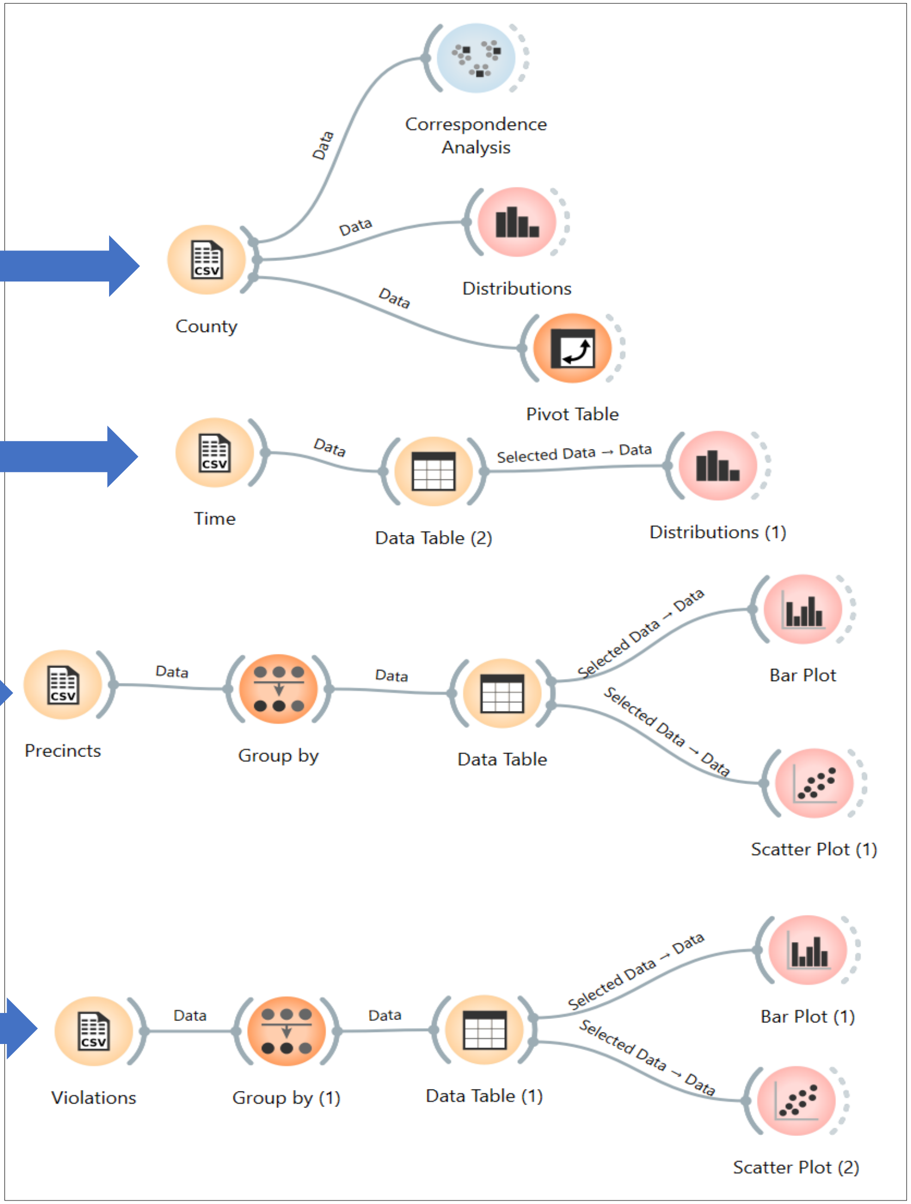
More detailed analysis workflow by breaking out individual topics by CSV file:

County CSV file:

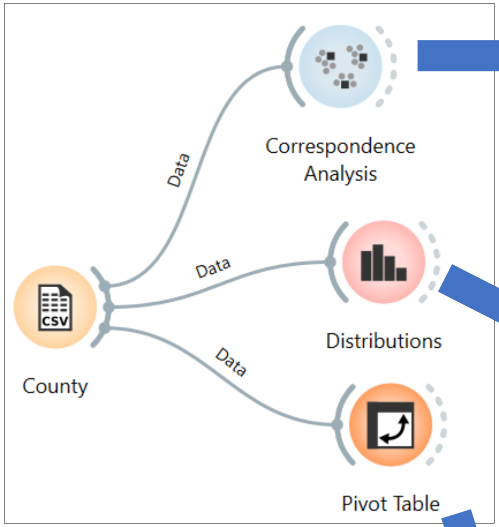
Time CSV file:

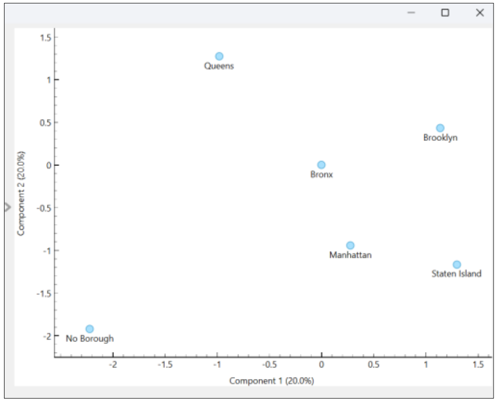
Precincts CSV file:

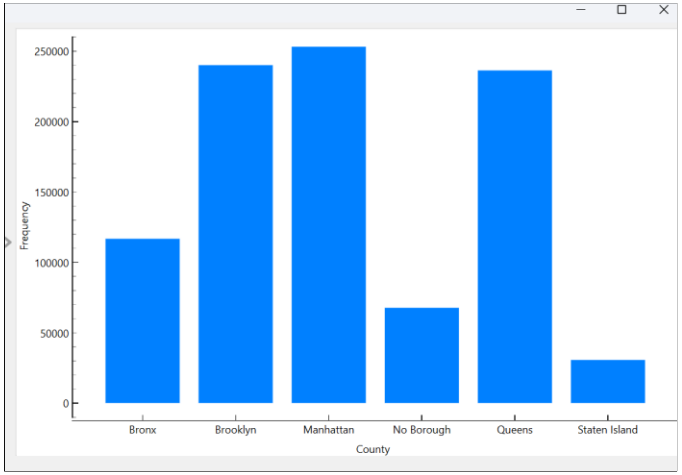
Violations CSV file:



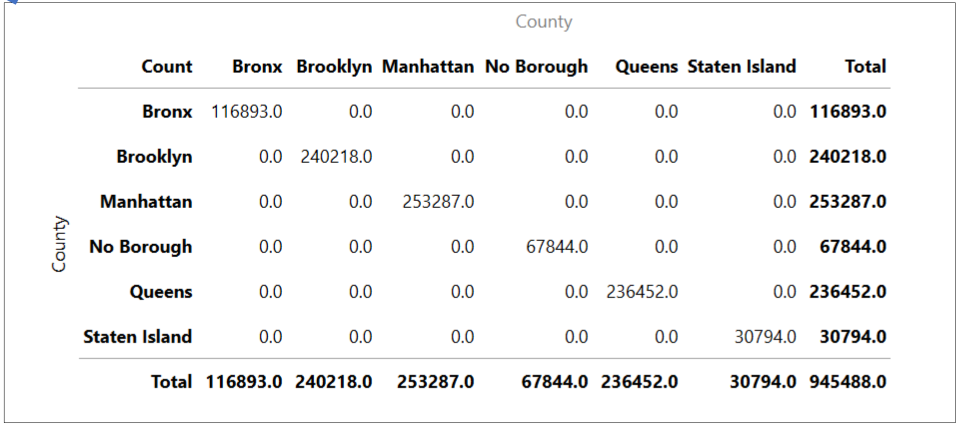
County analysis:





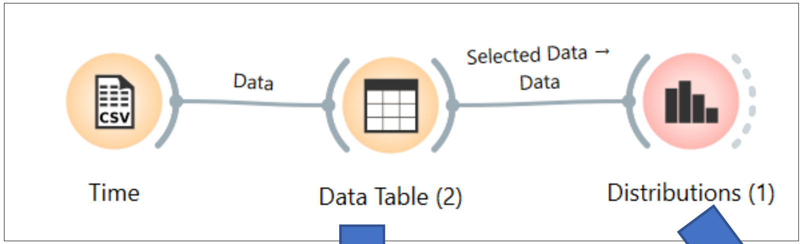


Note that Manhattan accounts for the highest number of tickets.

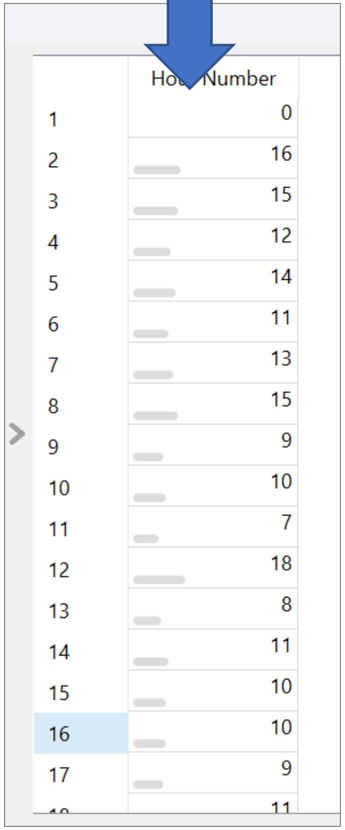


		County						
	Count	Bronx	Brooklyn	Manhattan	No Borough	Queens	Staten Island	Total
County	Bronx	116893.0	0.0	0.0	0.0	0.0	0.0	**116893.0**
	Brooklyn	0.0	240218.0	0.0	0.0	0.0	0.0	**240218.0**
	Manhattan	0.0	0.0	253287.0	0.0	0.0	0.0	**253287.0**
	No Borough	0.0	0.0	0.0	67844.0	0.0	0.0	**67844.0**
	Queens	0.0	0.0	0.0	0.0	236452.0	0.0	**236452.0**
	Staten Island	0.0	0.0	0.0	0.0	0.0	30794.0	**30794.0**
	Total	**116893.0**	**240218.0**	**253287.0**	**67844.0**	**236452.0**	**30794.0**	**945488.0**

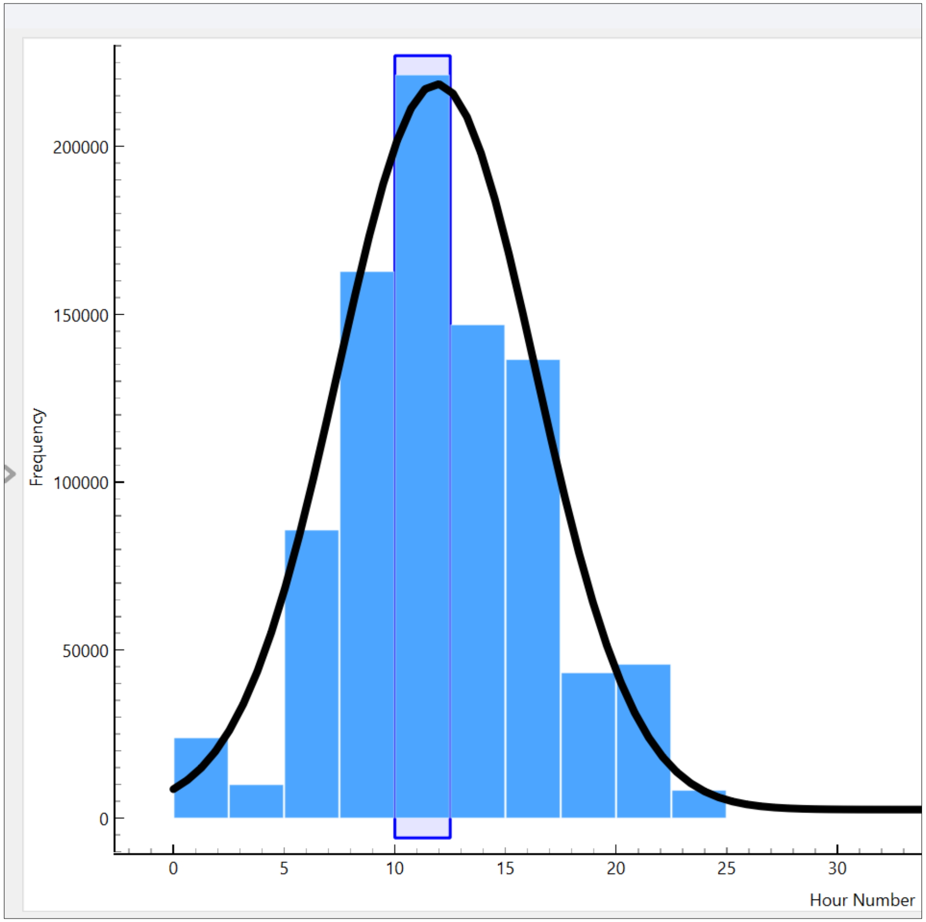
Time analysis by the hour:



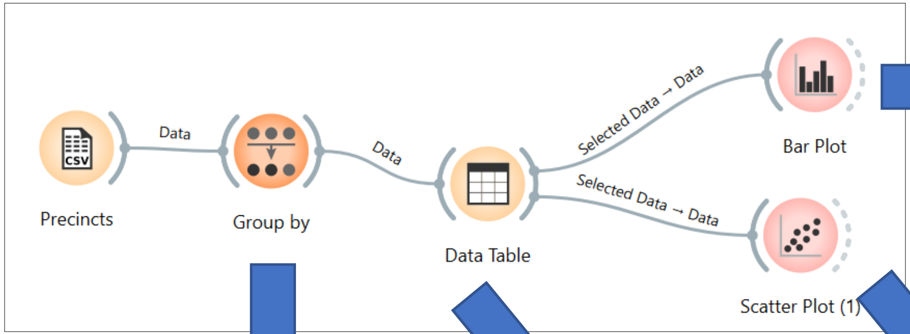
The time of day with the most tickets is from 10 AM to 12 PM

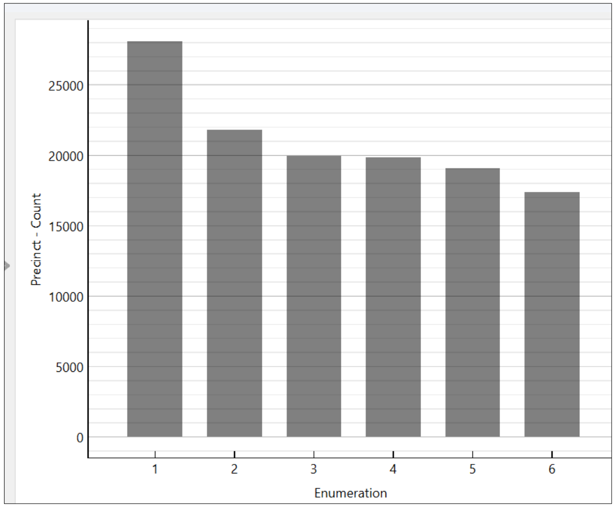


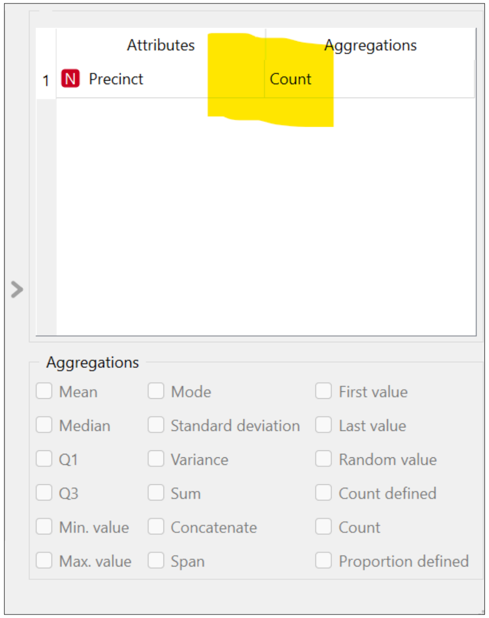
	Hour Number
1	0
2	16
3	15
4	12
5	14
6	11
7	13
8	15
9	9
10	10
11	7
12	18
13	8
14	11
15	10
16	10
17	9
18	11



Precinct analysis:



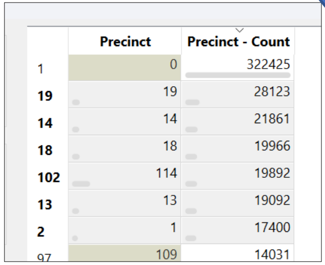




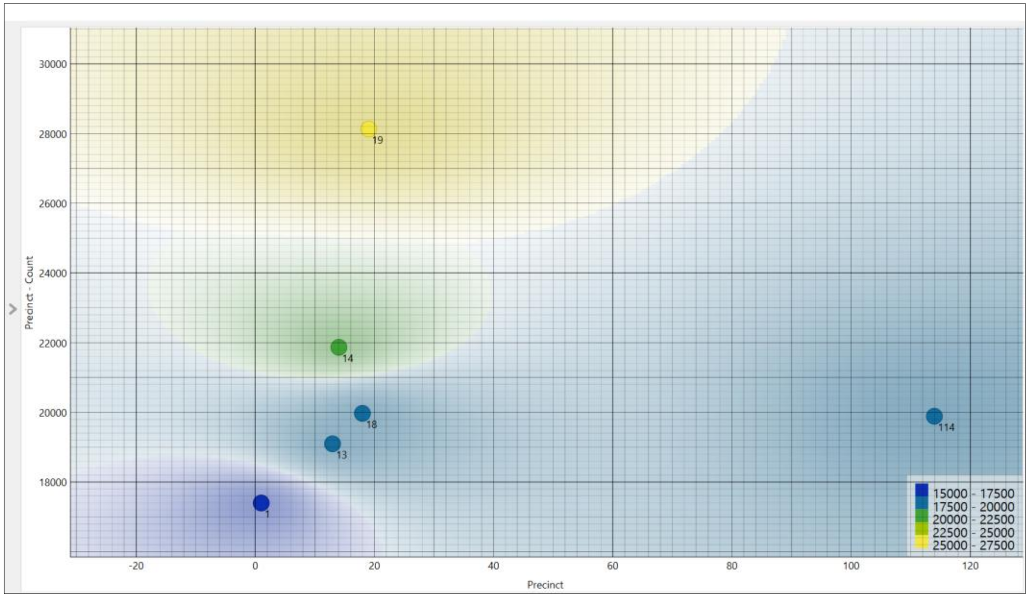
	Attributes	Aggregations
1	N Precinct	Count

Aggregations

□ Mean □ Mode □ First value
□ Median □ Standard deviation □ Last value
□ Q1 □ Variance □ Random value
□ Q3 □ Sum □ Count defined
□ Min. value □ Concatenate □ Count
□ Max. value □ Span □ Proportion defined



	Precinct	Precinct - Count
1	0	322425
19	19	28123
14	14	21861
18	18	19966
102	114	19892
13	13	19092
2	1	17400
97	109	14031

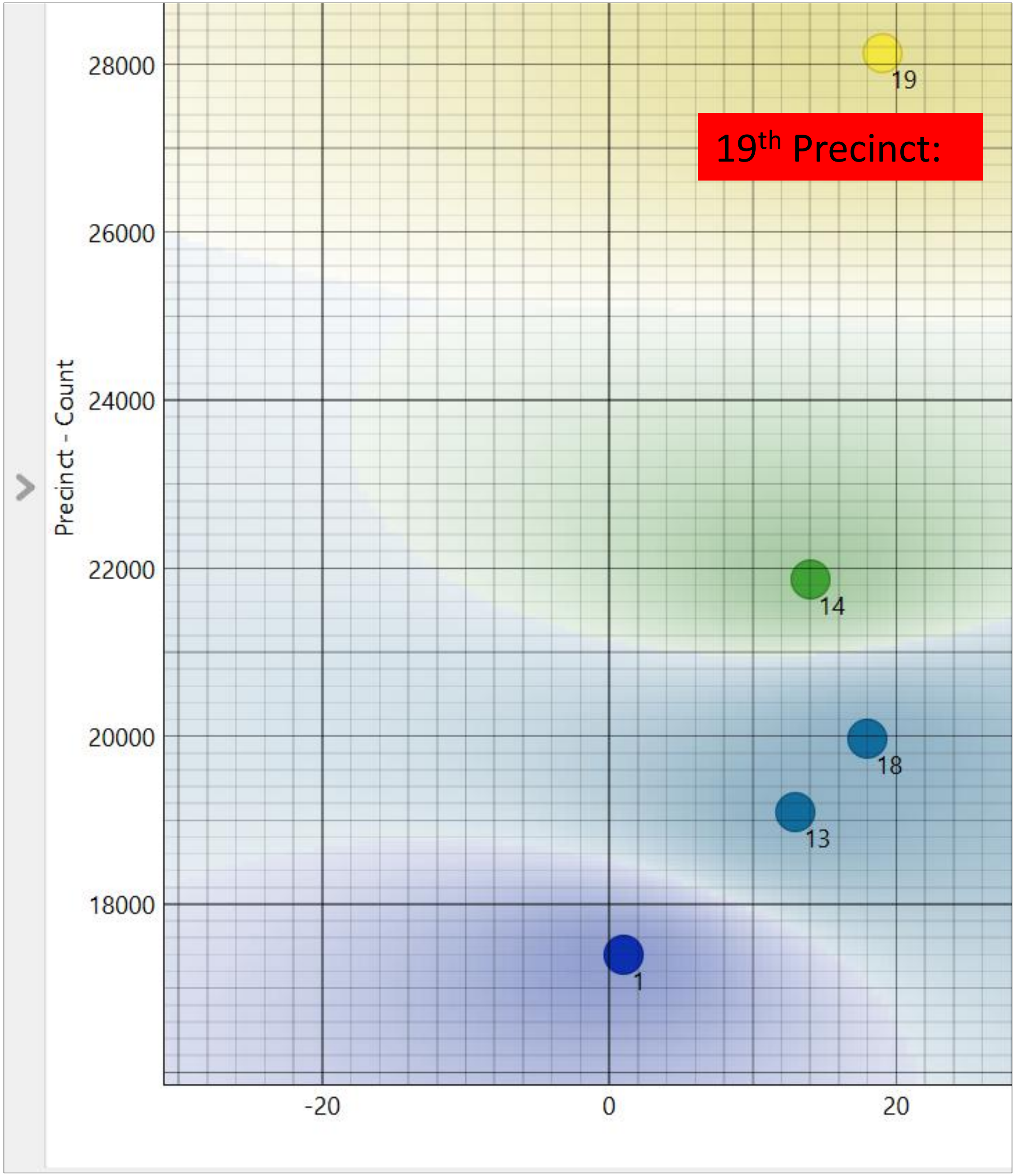


Close up of Precinct scatter plot.

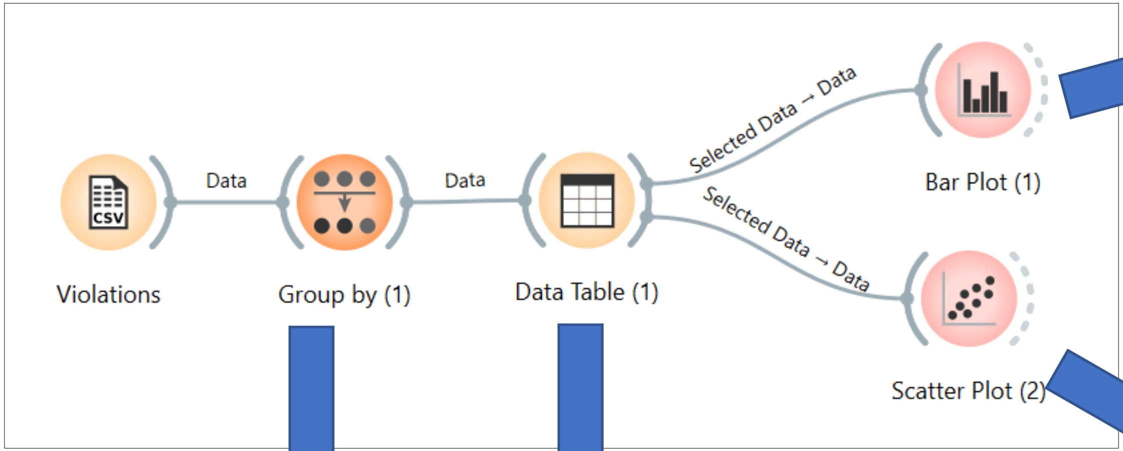
Top Precincts:

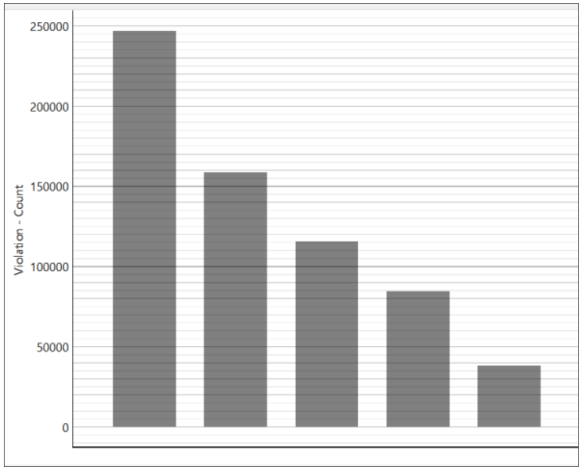
a. 19th
b. 14th
c. 18th
d. 13th
e. 1st

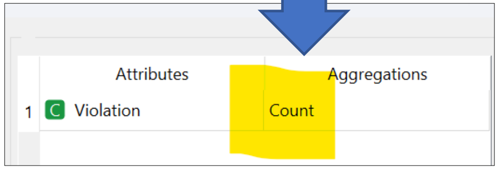
The 19th Precinct is on the Upper East Side has the highest number of tickets, followed by the 14th [Midtown South] and the 18th [Midtown North]. These areas have some of the highest traffic volumes in NYC.



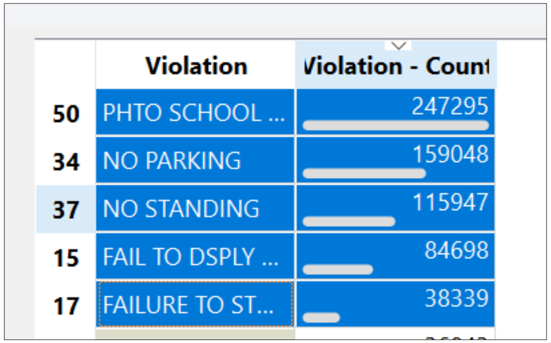
Violations analysis:





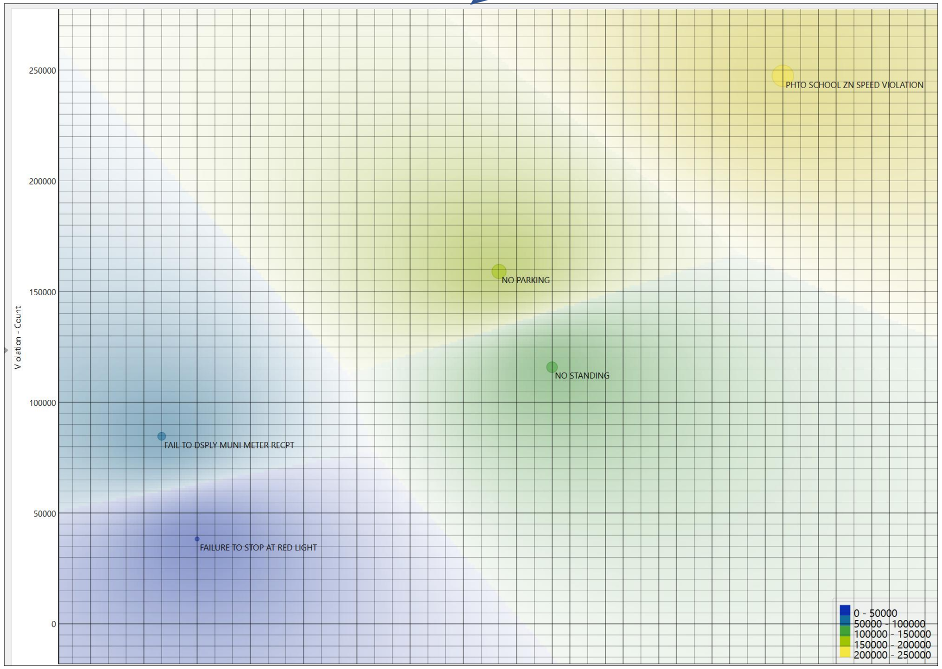


	Attributes	Aggregations
1	C Violation	Count



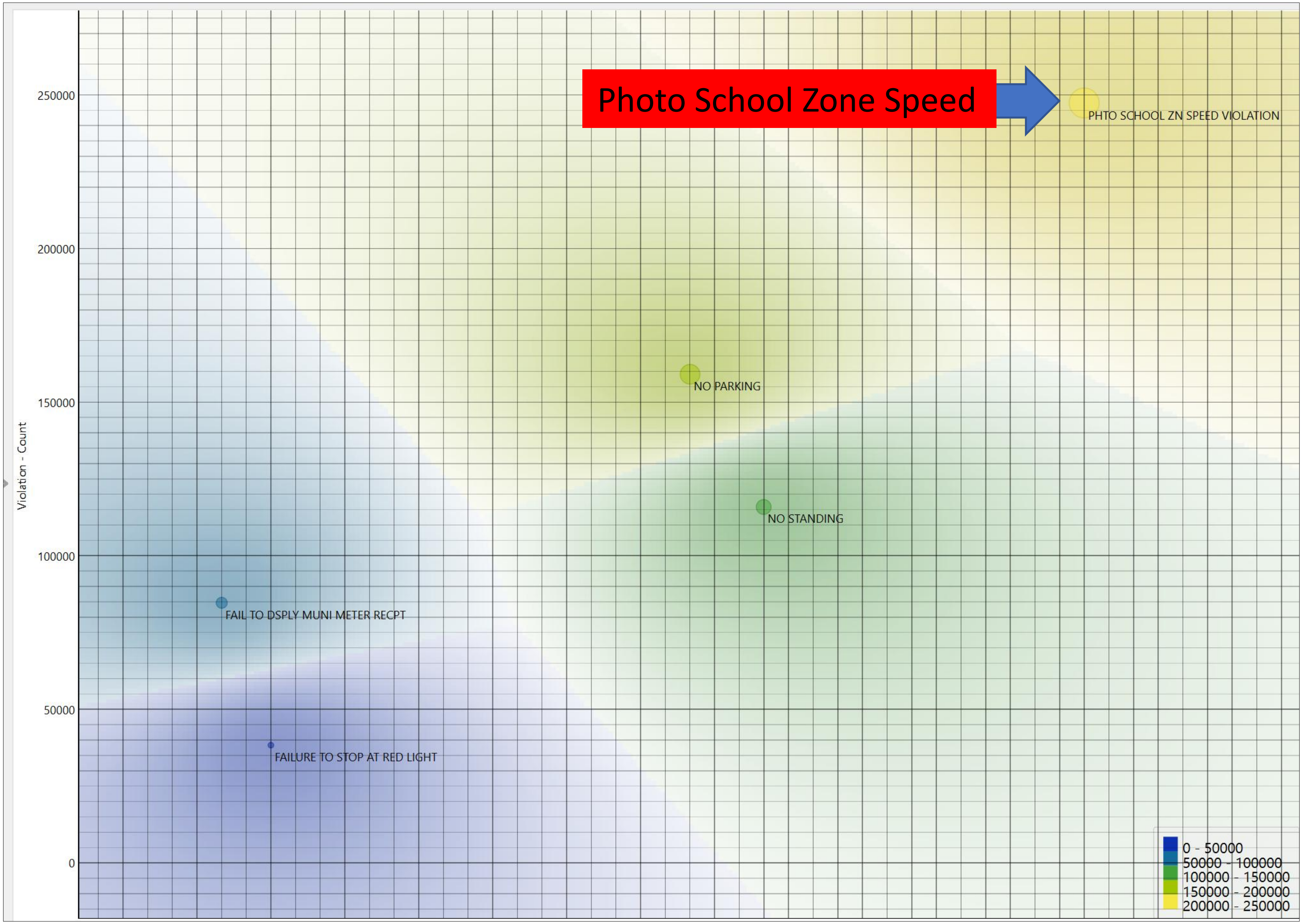
	Violation	Violation - Count
50	PHTO SCHOOL ...	247295
34	NO PARKING	159048
37	NO STANDING	115947
15	FAIL TO DSPLY ...	84698
17	FAILURE TO ST...	38339

Selected Top 5 in Data Table

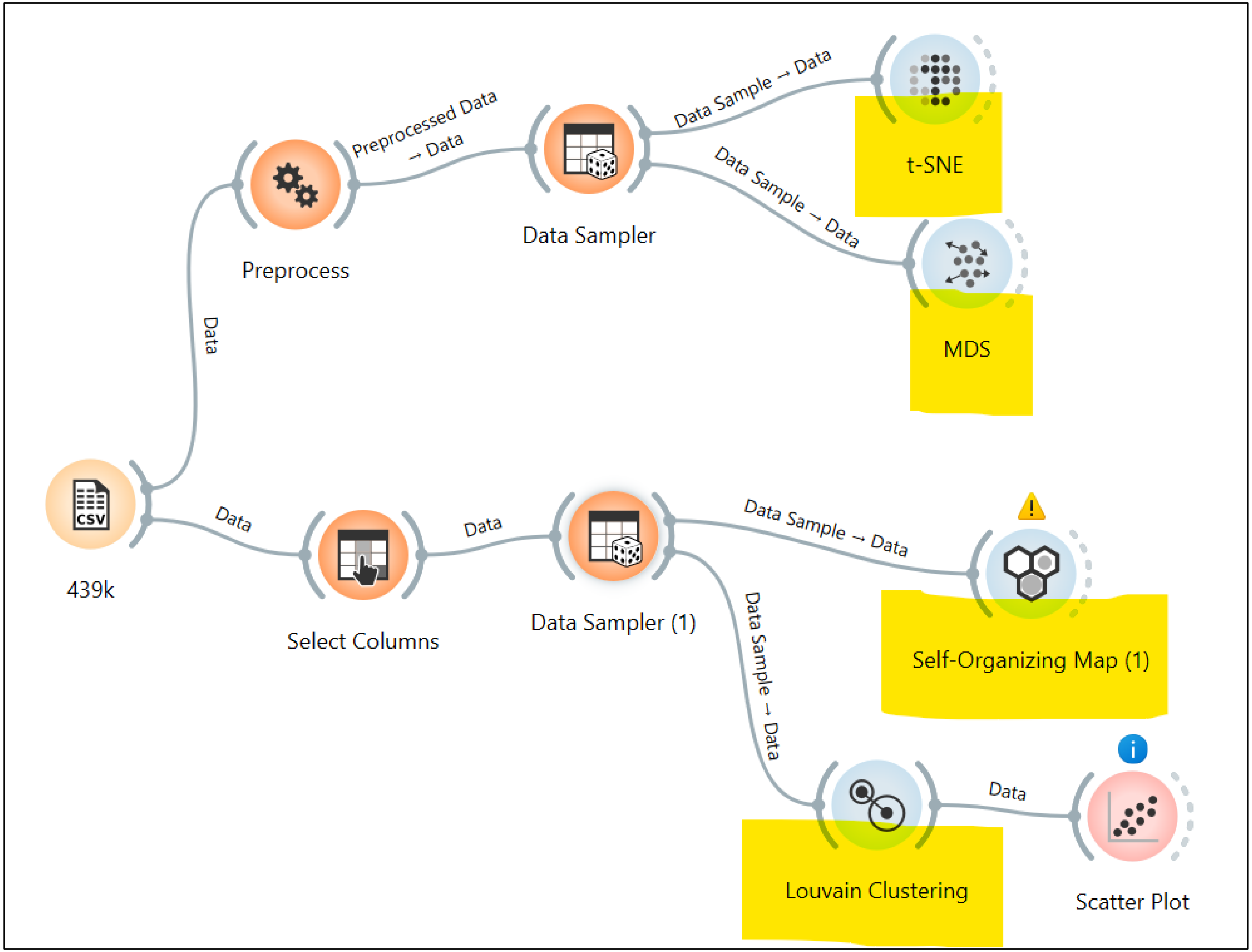


Top 5 Violations scatter plot :

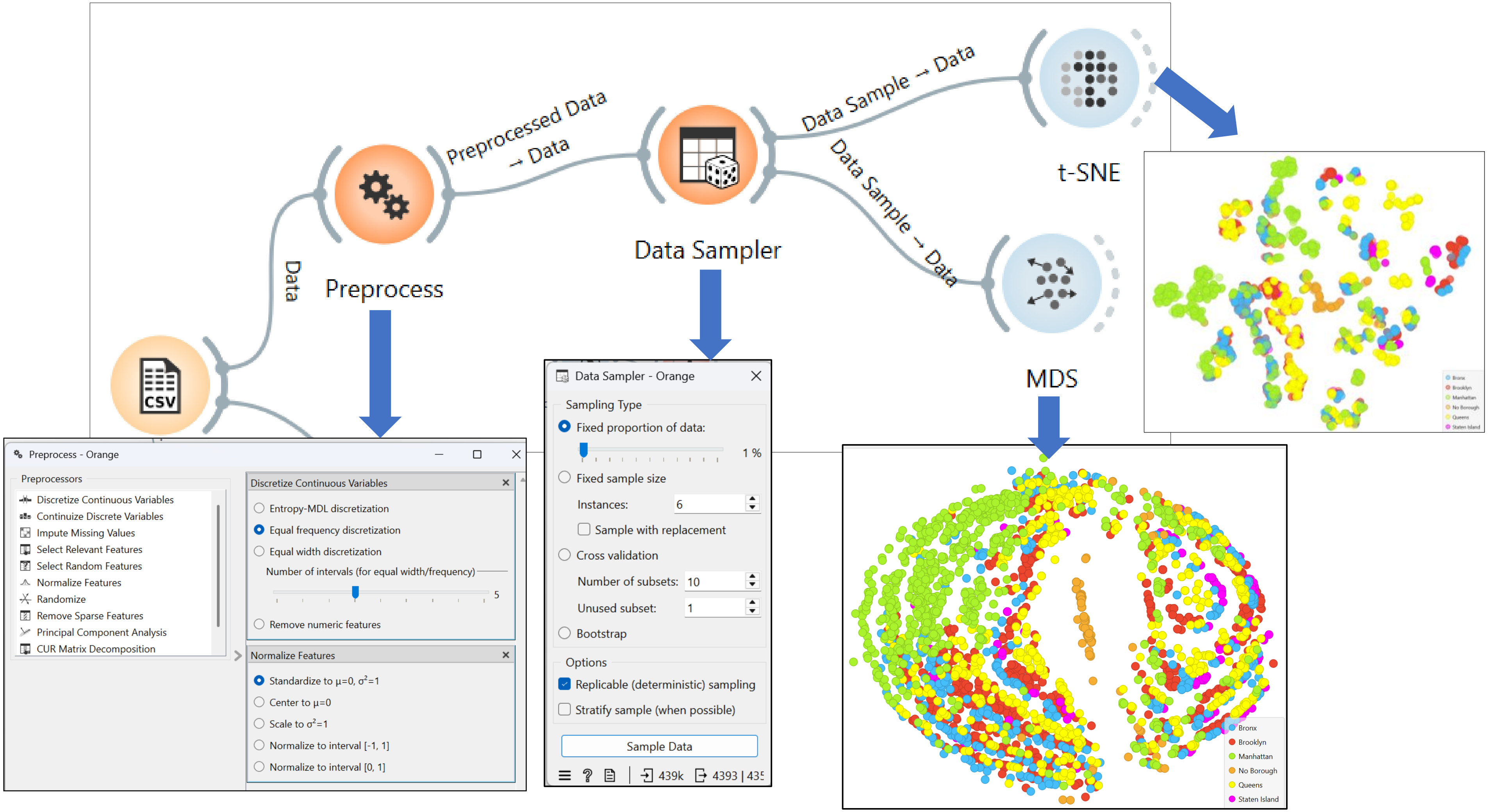
Top violation is Speeding in a School Zone taken by an automatic speed camera.



Unsupervised learning refers to algorithms that discover patterns and relationships



Unsupervised learning refers to algorithms that discover patterns and relationships



Preprocess - Orange

Preprocessors
- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Remove Sparse Features
- Principal Component Analysis
- CUR Matrix Decomposition

Discretize Continuous Variables
- ○ Entropy-MDL discretization
- ● Equal frequency discretization
- ○ Equal width discretization

Number of intervals (for equal width/frequency) 5
- ○ Remove numeric features

Normalize Features
- ● Standardize to μ=0, σ^2=1
- ○ Center to μ=0
- ○ Scale to σ^2=1
- ○ Normalize to interval [-1, 1]
- ○ Normalize to interval [0, 1]

Data Sampler - Orange

Sampling Type
- ● Fixed proportion of data: 1 %
- ○ Fixed sample size

  Instances: 6

  □ Sample with replacement
- ○ Cross validation

  Number of subsets: 10

  Unused subset: 1
- ○ Bootstrap

Options
- ☑ Replicable (deterministic) sampling
- □ Stratify sample (when possible)

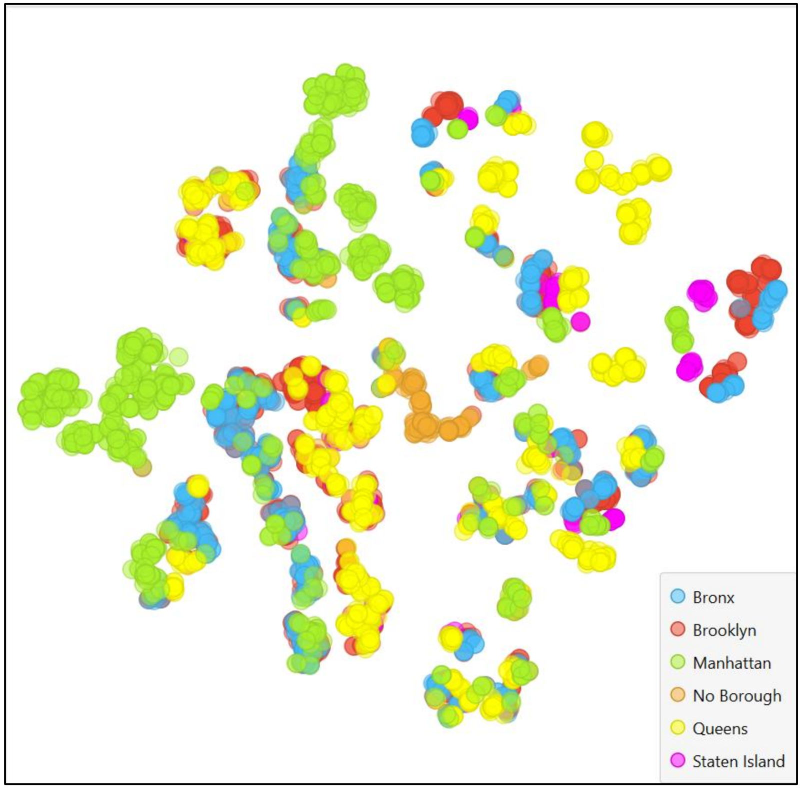
Sample Data

439k 4393 | 435

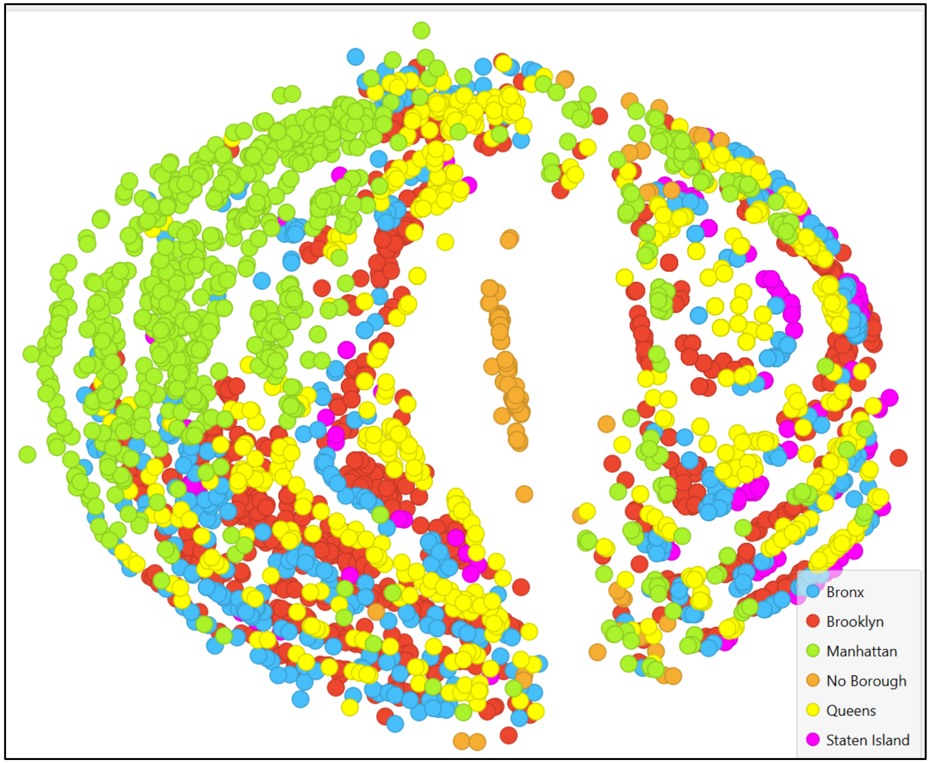
<u>t-SNE:</u> [t-Stochastic Neighbor Embedding] reduces a large number of dimensions to 2D or 3D visualizations. Uncovers clusters and patterns in data.

<u>MDS:</u> [Mulitdimensional Scaling] also maps high dimensional data to lower dimensions but is less intensive with regards to computer processing.

t-SNE: [t-Stochastic Neighbor Embedding]

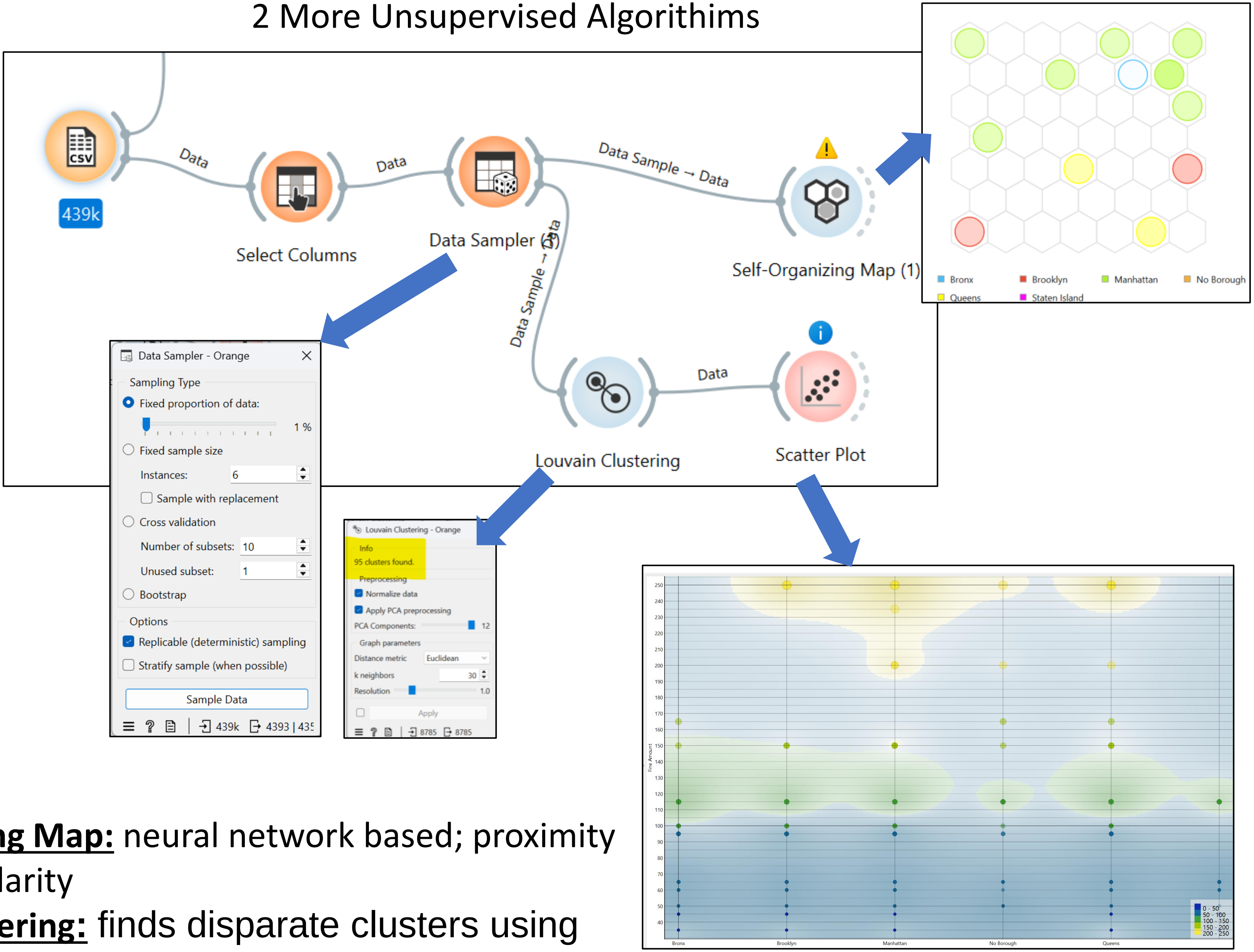


MDS: [Mulitdimensional Scaling]



Same dataset but some difference in clustering patterns, where t-SNE is preserving local relationships, while MDS maintains global structures and overall data geometry.

2 More Unsupervised Algorithims



Data Sampler - Orange

Sampling Type

● Fixed proportion of data: 1 %

○ Fixed sample size

Instances: 6

□ Sample with replacement

○ Cross validation

Number of subsets: 10

Unused subset: 1

○ Bootstrap

Options

☑ Replicable (deterministic) sampling

□ Stratify sample (when possible)

Sample Data

439k 4393 | 435

Louvain Clustering - Orange

Info

95 clusters found.

Preprocessing

☑ Normalize data

☑ Apply PCA preprocessing

PCA Components: 12

Graph parameters

Distance metric: Euclidean

k neighbors: 30

Resolution: 1.0

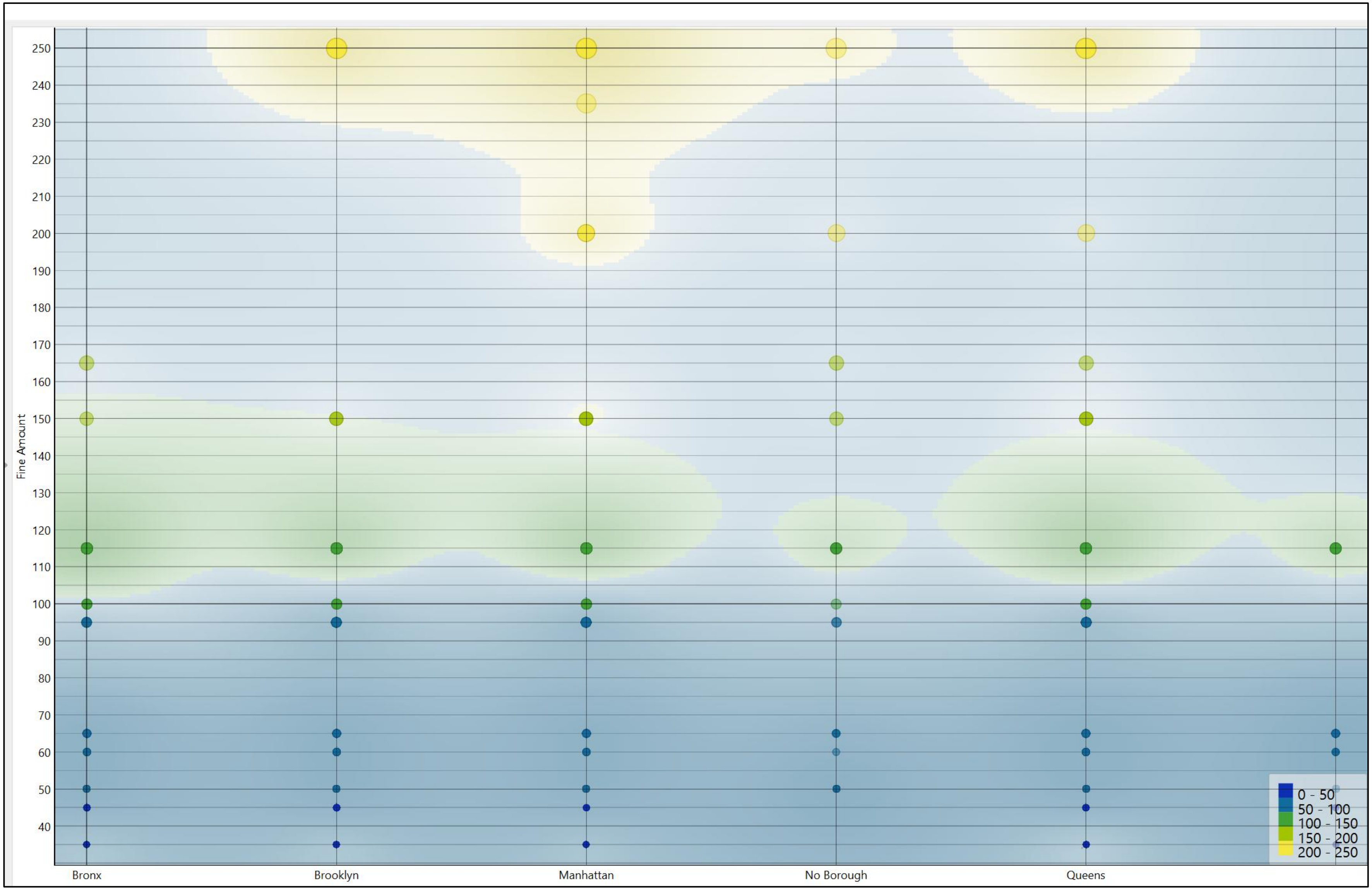
□ Apply

8785 8785

Self-Organizing Map: neural network based; proximity indicates similarity

Louvain Clustering: finds disparate clusters using community detection

Louvain Clustering Feeding Into Scatter Plot



Brighter color indicates higher Fine Amount Along With Heat Map
Manhattan has the highest concentration of the highest ticket fines.

Conclusion

Tools Used:

- Extraction: Query tool on NYC OpenData for Parking Tickets
- Transform: OpenRefine
- Load/Visualizations: Orange Data Mining - Additional transformation work done with GREL [General Refine Expression Language]

Project Review:

Having worked previously with data from the GAIA Space Observatory where the data was precise and standardized, it was an intereseting learning process to work with NYC OpenData. By contrast, NYC OpenData was much less precise and chaotic due to humans recording and performing data entry with little quality control. However, this challenge was an opportunity to learn how to clean and modify the data using OpenRefine. The Clustering Function in OpenRefine was quite fascinating, as were the results produced by the different algorithimic clustering functions. Likewise, using Orange Data Mining was a wonderful chance to learn about the different visualizations that data was capable of being presented. In addition, this application introduced me to the large number of Unsupervised Algorithims that are available to further reveal hidden patterns and trends within large amounts of data.

Both programs required a fairly steep learning curve in order to understand their basic functions. In doing so, I was able to get a taste of the true power of these programs.

In retrospect, I would have allocated more time towards the download process from NYC OpenData, as I did not anticipate that there would be upwards of 10 million rows of data per month. However, this gave me the chance to learn the true power and usefulness of OpenRefine in handling massive levels of data.

Benefits of using OpenRefine and Orange Data Mining:

- Do not have to create customized Python code from scratch
- Both tools are refined and well-tested applications
- Data does not have to be uploaded to the cloud, saving time and money
- Since all data is processed locally, the user's privacy is enhanced
- After using both applications, the skills gained will allow users to take those skills into the field to work on data quickly without having to pay for use of the cloud

After working with both tools in the course of this project, it seems that Baruch could develop a whole course on how to use all aspects of these substantial programs. In doing so, this would allow students to perform sophisticated data manipulations and robust modifications without learning the intracacies of Python programming.

References:

- Extraction: Query tool on NYC OpenData for Parking Tickets

Website: https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89/about_data

- Transform: OpenRefine: https://openrefine.org/
- Load/Visualizations: Orange Data Mining: https://orangedatamining.com/
- Additional transformation work: GREL [General Refine Expression Language]: https://openrefine.org/docs/manual/grel