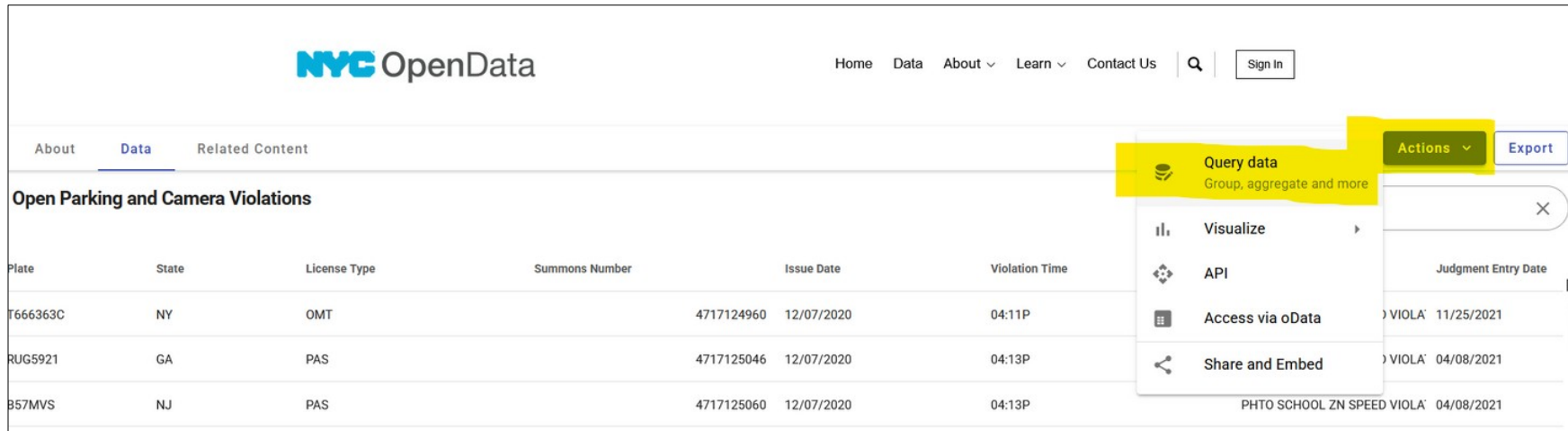


# ETL Process Using NYC OpenData, OpenRefine and Orange Data Mining

## Extract

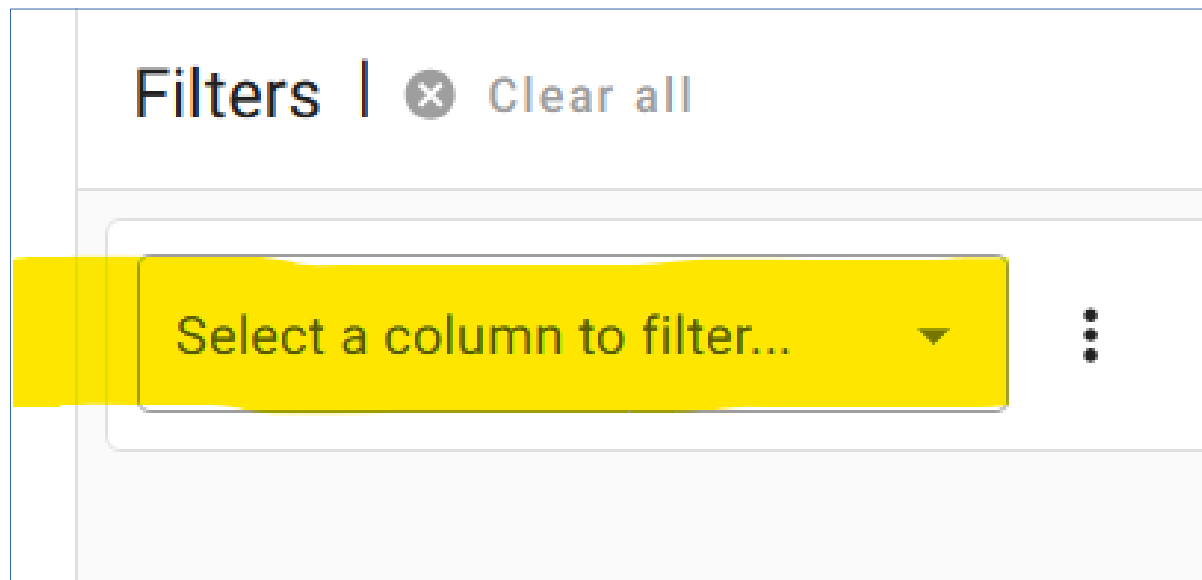
1. Go to NYC OpenData website > click on Actions > Query data



The screenshot shows the NYC OpenData website interface. The 'Data' tab is selected, and the dataset 'Open Parking and Camera Violations' is displayed. The 'Actions' menu is open, showing options: 'Query data' (highlighted), 'Visualize', 'API', 'Access via oData', and 'Share and Embed'. The 'Query data' option is described as 'Group, aggregate and more...'. The table below shows columns: Plate, State, License Type, Summons Number, Issue Date, and Violation Time.

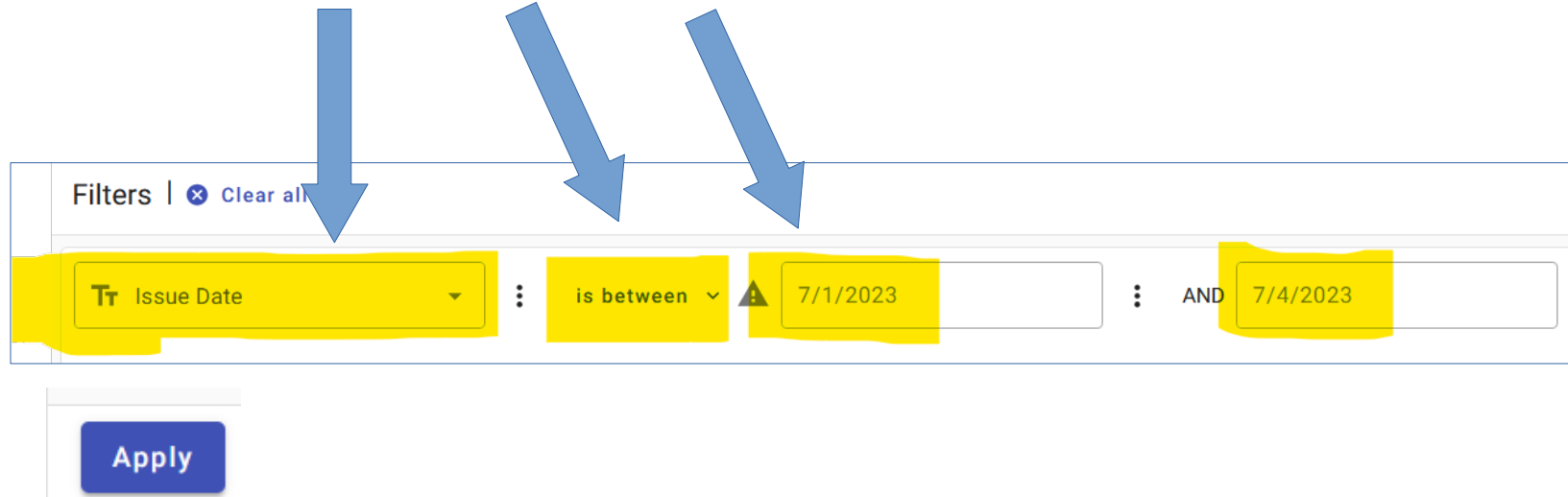
Plate	State	License Type	Summons Number	Issue Date	Violation Time
T666363C	NY	OMT	4717124960	12/07/2020	04:11P
RUG5921	GA	PAS	4717125046	12/07/2020	04:13P
B57MVS	NJ	PAS	4717125060	12/07/2020	04:13P

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



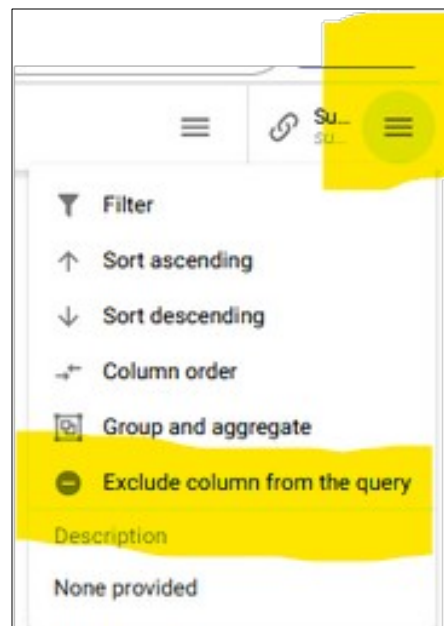
The screenshot shows the 'Filters' section of the NYC OpenData interface. The text 'Filters | Clear all' is visible. A yellow box highlights the dropdown menu labeled 'Select a column to filter...'. To the right of the dropdown is a vertical ellipsis icon.

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



The screenshot shows a filter bar with the text "Filters | Clear all". Below it, a filter is configured with the field "Issue Date", the boolean operator "is between", and the search terms "7/1/2023" and "7/4/2023". The field, boolean, and search term inputs are highlighted in yellow. Three blue arrows point from the text above to these inputs. Below the filter bar is a blue "Apply" button.

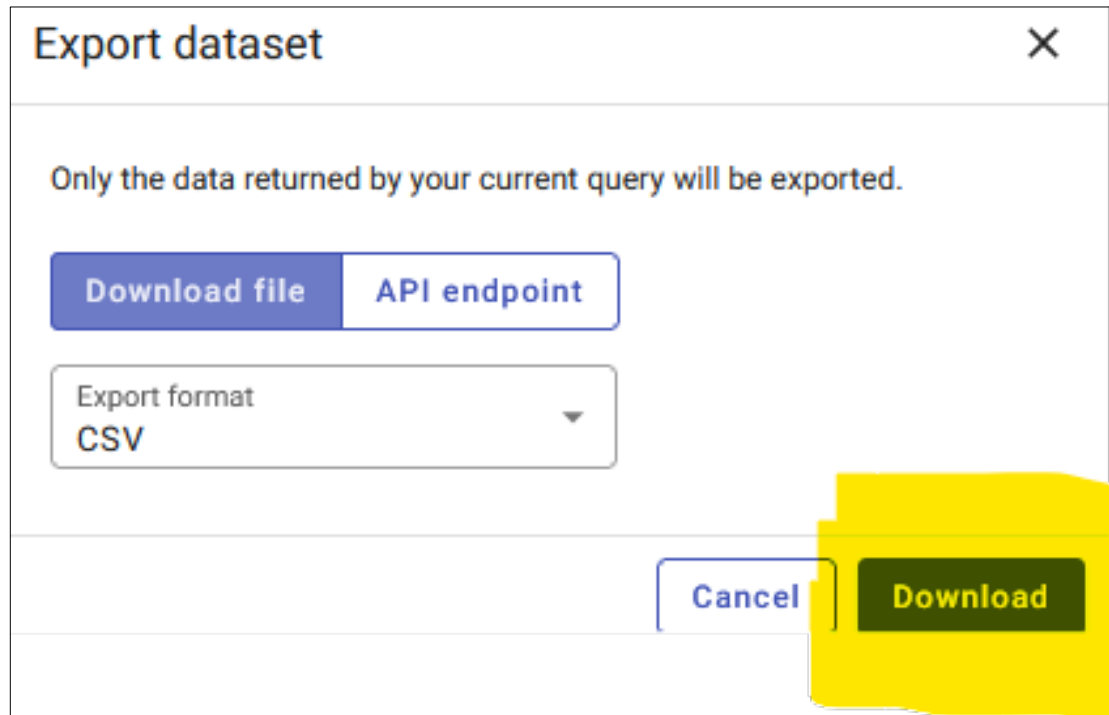
4. Remove columns that are not needed or have large amounts of data > find column > click on 3 horizontal lines > click on “Exclude column from the query”:



5. Click on “Export” button:




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



# Transform

## 7. Import CSV file into OpenRefine

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

Create project  
Open project  
Import project  
Language settings  
Extensions

**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 be found in the documentation.  

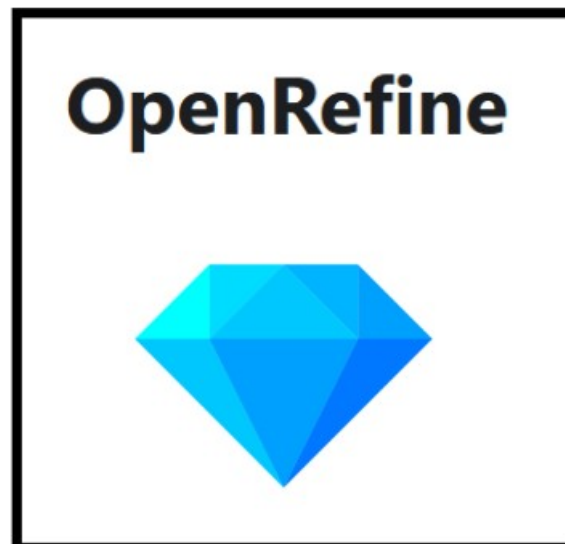
Get data from  
**This Computer**  
Web Addresses (URLs)  
Clipboard  
Database

Locate one or more files on your computer to upload:  


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

Drag files here

Next »



## 8. Initial upload of data:

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

Create project

Open project

Import project

Language settings

Extensions

« start over »

Configure parsing options

Project name  Tags

Create project »

	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	PHOTO 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

Parse data as

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

JSON-LD files

RDF/N3 files

RDF/N-Triples files

RDF/Turtle files

Character encoding

Update preview

☐ 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  line(s) at beginning of file

☒ Parse next  line(s) as column headers

☐ Column names (comma separated)

☐ Attempt to parse cell text into numbers

☒ Store blank rows

☒ Store blank columns

☒ Store blank cells as nulls

☐ Store file source

☐ Store archive file

☐ Discard initial  row(s) of data

☐ Load at most  row(s) of data

Version 3.9.0 [TRUNK]

Preferences

Help

About

## 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 < **280,137 rows** Extensions Wikibase ▾

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

**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](#)

		▼ All	▼ State	▼ Licen	▼ Issue Date	▼ Viol	▼ Violation	▼ Fine	▼ Pens	▼ 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	PHOTO 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:

▼ Penalty Amou	▼ Payment Amount	▼ Amo	▼ Prec	▼ County
Facet	0.00	084	Brooklyn	
Text filter	0.00	019	Manhattan	
Edit cells	0.00	000	Brooklyn	
Edit column	0.00	043	Bronx	
Transpose	Split into several columns...			
Sort...	Join columns...			
View	Add column based on this column...			
Reconcile	Add column by fetching URLs...			
	Add columns from reconciled values...			
	Rename this column...			
	<b>Remove this column</b>			
	Move column to beginning			
	Move column to end			
	Move column left			
	Move column right			

## 11. Generate text facets to understand different data that are in the column:

 **OpenRefine** 300 000 July16 July31 2023Open Parking and Camera Violations 20250416 csv [Permalink](#)

**Facet / Filter** Undo / Redo 0 / 0 **280,137 rows**

Refresh Reset all Remove all Show as: rows records Show: 5 10 25 50 100 500 1000 rows

**Violation** change 90 choices Sort by: name count Cluster

PHOTO 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	Violat	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				
☆	🗨	10.	NY	PAS	2023-07-16T00:00:00Z	06				

Facet

- Text facet
- Text filter
- Numeric facet
- Timeline facet
- Scatterplot facet...
- Custom text facet...
- Custom numeric facet...
- Customized facets

Edit cells

Edit column

Transpose


Sort...

View

Reconcile

LIMITS 65.0 0.00  
ED/C/FEIT 65.0 0.00

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

 **OpenRefine** 300 000 July16 July31 2023Open Parking and Camera Violations 20250416 csv [Permalink](#)

Facet / Filter Undo / Redo 0 / 0

Refresh Reset all Remove all

280,137 rows

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

**Violation** 90 choices Sort by: name count

PHOTO SCHOOL ZN SPEED VIOLATION 82045

NO PARKING-STREET CLEANING 38264

FAIL TO DSNLY 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

Cluster

Change

All State Licen Issue Date Viola Violation Fine Penalty

Facet Text facet

Text filter Numeric facet

Edit cells Timeline facet

Edit column Scatterplot facet...

Transpose Custom text facet...

Sort... Custom numeric facet...

View Customized facets

Reconcile

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			
10.	NY	PAS	2023-07-16T00:00:00Z	06			

LIMITS	65.0	0.00
ED/C/FEIT	65.0	0.00



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

#### Cluster and edit column “Violation”

Find groups of different cell values that 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 “Gödel” and “Godel” probably refer to the same person. [Find out more...](#)

Method **Key collision** ▼

Keying function **Metaphone3** ▼

☐ Auto-update

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

**Cluster**

Keying function **Metaphone3** ▼

Fingerprint

n-Gram fingerprint

**Metaphone3**

Cologne phonetic

Daitch-Mokotoff

Beider-Morse

Select all

Deselect all

Export clusters

**Merge selected & re-cluster**

Merge selected & Close

Close

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 column "Violation"**

Find groups of different cell values that 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 "Godel" and "Godel" probably refer to the same person. [Find out more...](#)

Method: Key collision Keying function: Metaphone3 Manage clustering functions

☐ Auto-update 2 clusters found

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

**# Choices in cluster**  
2 — 5

**# Rows in cluster**  
0 — 19000

**Average length of choices**  
19.5 — 29.34

**Length variance of choices**  
0 — 8

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

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

The screenshot shows a data table with columns: Fine, Pena, Payn, Amo, Prec, County, Ranc, and Issuing Agency. A menu is open from the 'Facet' column header, showing options: Facet, Text filter, Edit cells, Edit column, Transpose, Sort..., View, and Reconcile. The 'Edit cells' option is selected, opening a sub-menu with: Transform..., Common transforms, Fill down, Blank down, Split multi-valued cells..., Join multi-valued cells..., Cluster and edit..., and Replace... The 'Common transforms' option is selected, opening a third-level menu with: Trim leading and trailing whitespace, Collapse consecutive whitespace, Unescape HTML entities, Replace smart quotes with ASCII, To titlecase, To uppercase, To lowercase, To number, To date, To text, To null, and To empty string. The 'To number' option is highlighted.

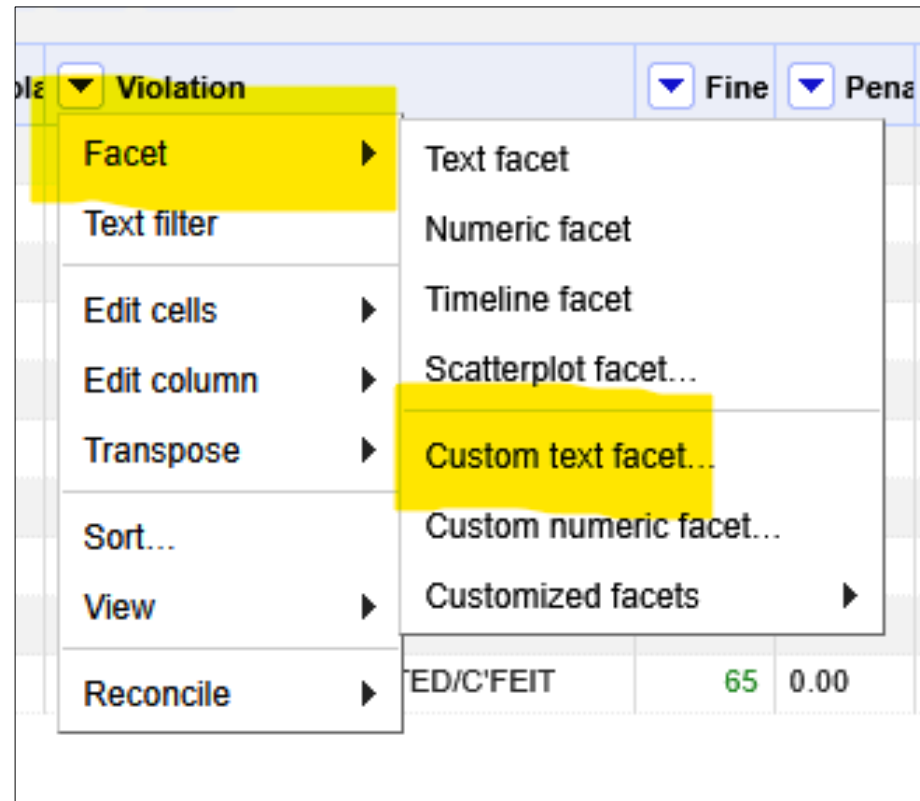
▼ Fine	▼ Pena	▼ Payn	▼ Amo	▼ Prec	▼ County	▼ Ranc	▼ Issuing Agency
Facet		0	0.00	084	Brooklyn	true	POLICE DEPARTMENT
Text filter			0.00	019	Manhattan	true	TRAFFIC
N		0	0.00	000	Brooklyn	true	DEPARTMENT OF TRANSPORTATION
						true	TRAFFIC

- Facet
  - Text filter
  - Edit cells
    - Transform...
    - Common transforms
      - Trim leading and trailing whitespace
      - Collapse consecutive whitespace
      - Unescape HTML entities
      - Replace smart quotes with ASCII
      - To titlecase
      - To uppercase
      - To lowercase
      - To number
      - To date
      - To text
      - To null
      - To empty string
    - Fill down
    - Blank down
    - Split multi-valued cells...
    - Join multi-valued cells...
    - Cluster and edit...
    - Replace...
  - Edit column
  - Transpose
  - Sort...
  - View
  - Reconcile

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

	 Fine	 Penal
	95	0.00
	95	0.00
N	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. Facet > Custom text facet:



19. In the “Expression” box, type:

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

20. This will set every 3<sup>rd</sup> row to “True”:

### Custom facet on column Violation

Expression

Language General Refine Expression Language (GREL) ▾

`row.index % 3 == 0`

No syntax error.

Preview

History

Starred

Help

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

OK

Cancel

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 3<sup>rd</sup> value:

Facet / Filter Undo / Redo 2 / 2

Refresh Reset all Remove all

✕

Violation

2 choices Sort by: name count

false 186758

**true 93379**

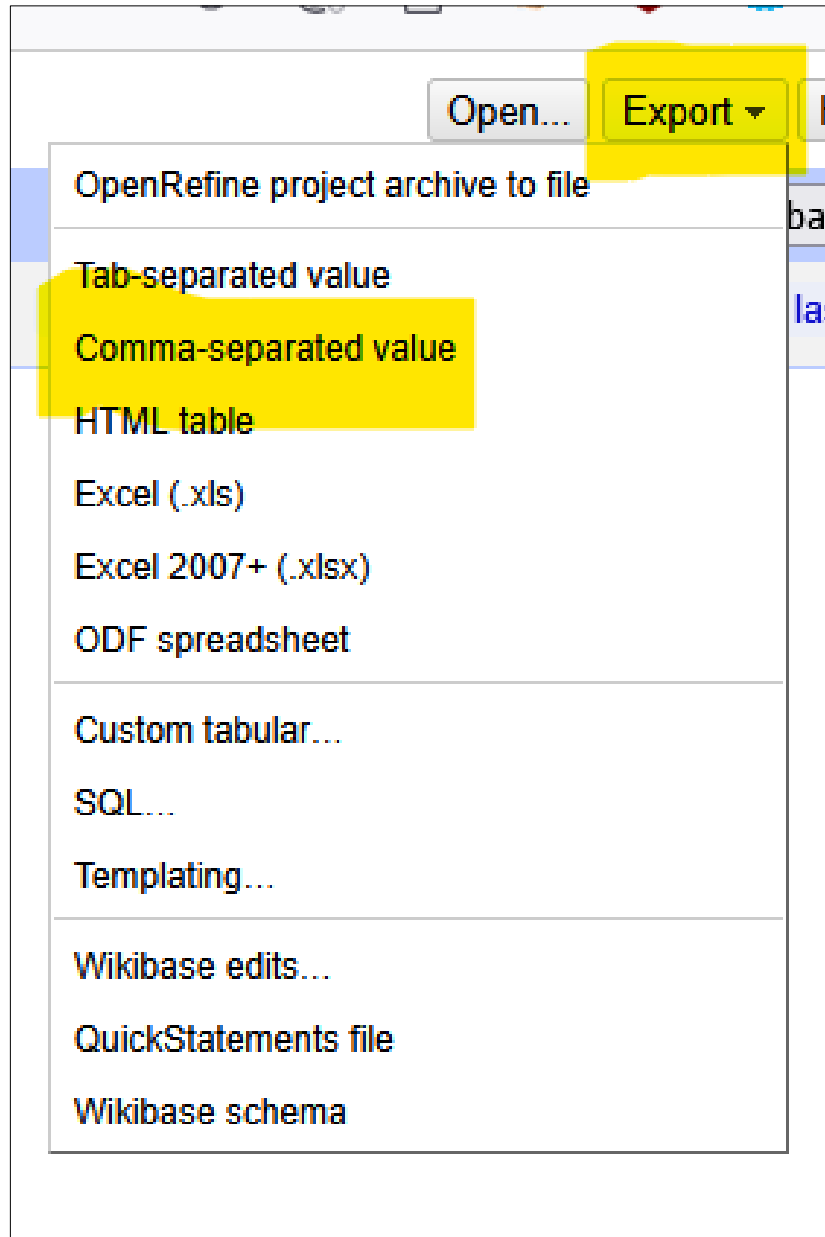
Facet by choice counts

93,379 matching rows (280,137 total)

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


		State	Lic	Issue Date	Viol	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	PHOTO 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
--	-------------------	-----------	-----------

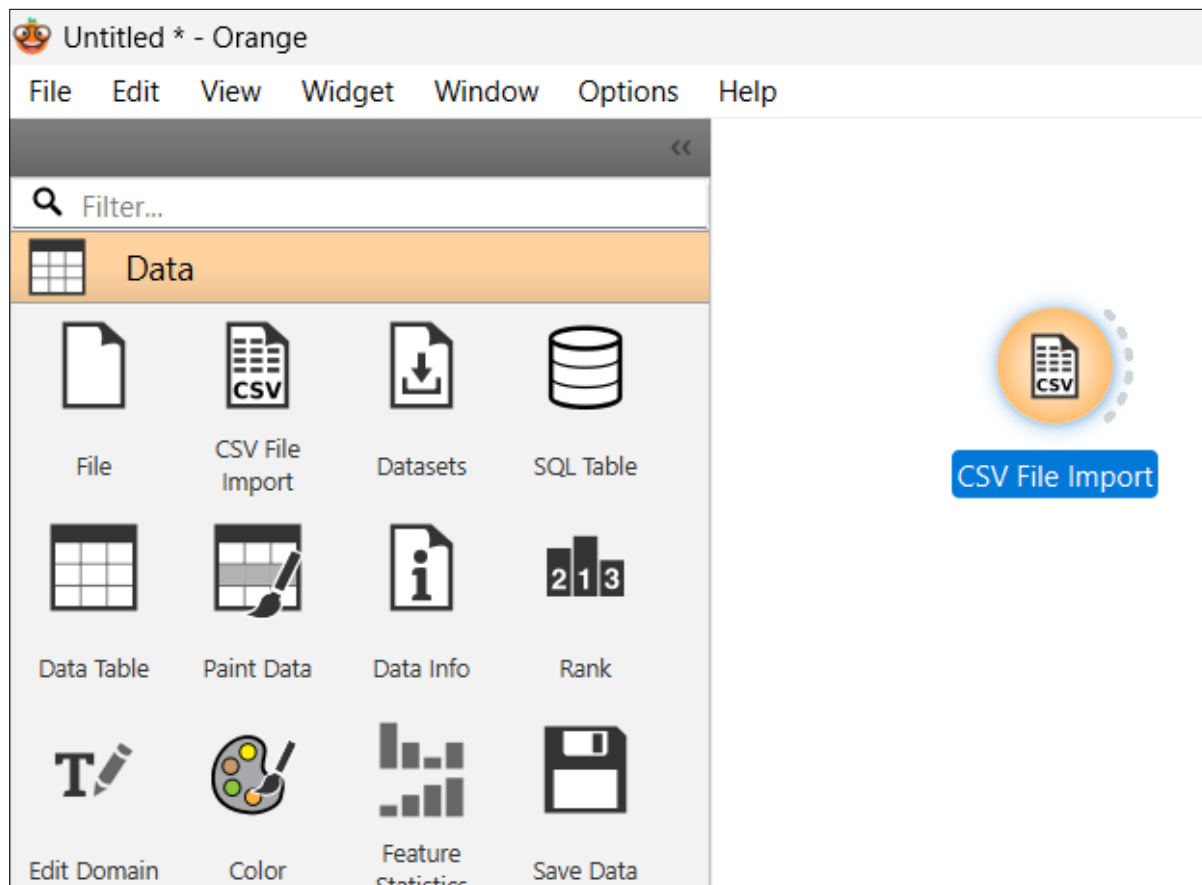
### Load

26. Load file into Orange Data Miner:

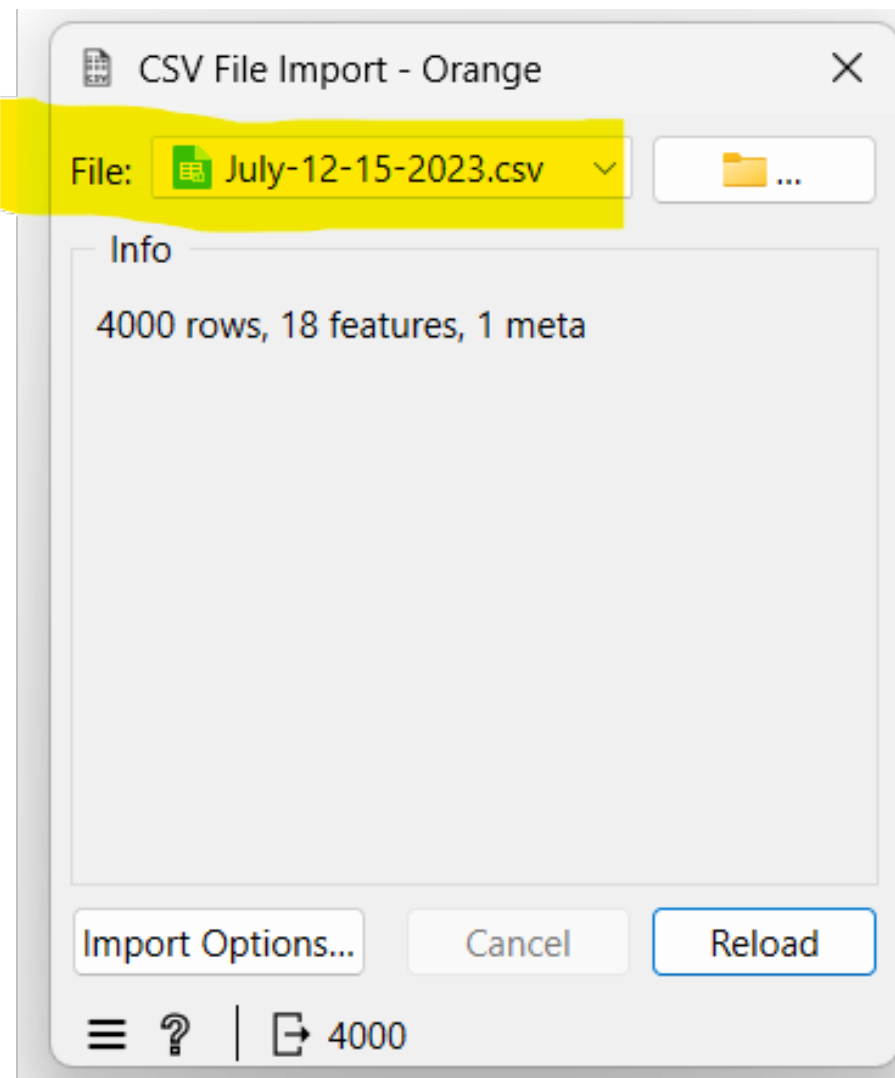


The banner features the Orange Data Mining logo on the left, which includes the word "orange" in a bold, lowercase font with "DATA MINING" in a smaller, uppercase font below it. To the right of the logo, the text "Data Mining Fruitful and Fun" is displayed in a large, white, sans-serif font. Below this, a smaller line of text reads "Open source machine learning and data visualization." At the bottom left, there is an orange button with the text "Download Orange 3.38.1". On the right side of the banner, a cartoon orange character with large, round glasses and a green leaf on its head is depicted. The character is holding several strings that connect to various data-related icons, including a lightbulb, a bar chart, a scatter plot, a network diagram, a pie chart, and a line graph. The character is also holding a tablet that displays a bar chart and a line graph.

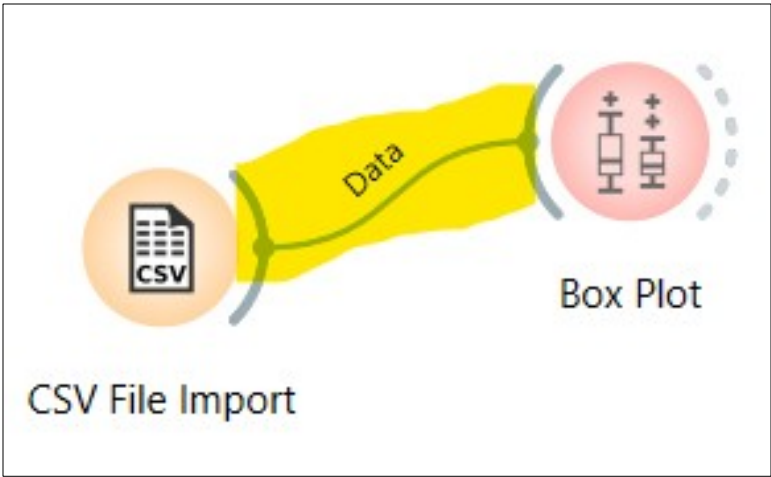
27. Open Orange Data Mining and drag the CSV File Import widget into the work area:



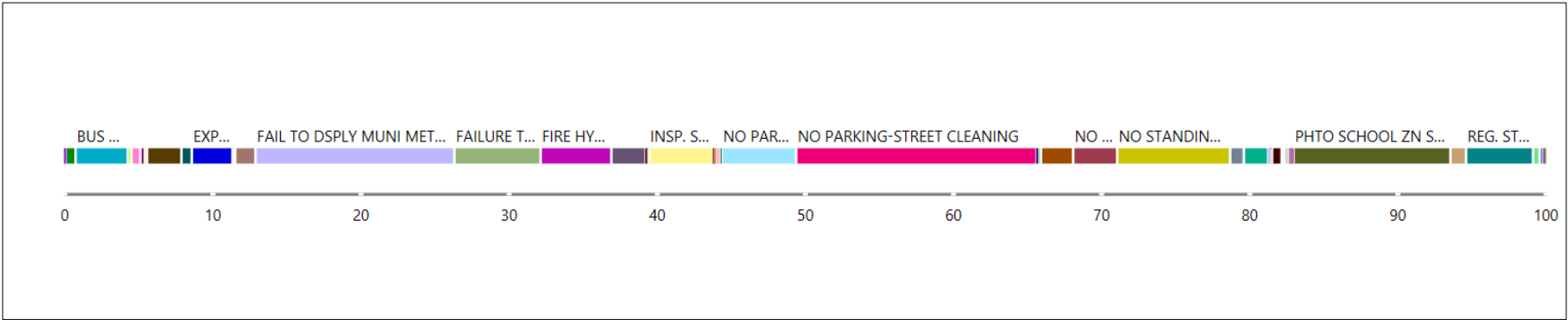
28. Double click, find file and load file into Orange Data Miner:



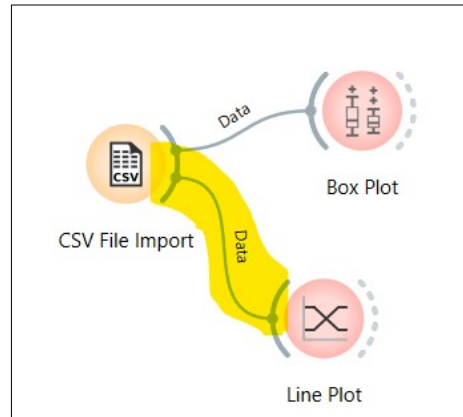
29. Drag Box Plot widget into work area and connect from CSV File to Box Plot:



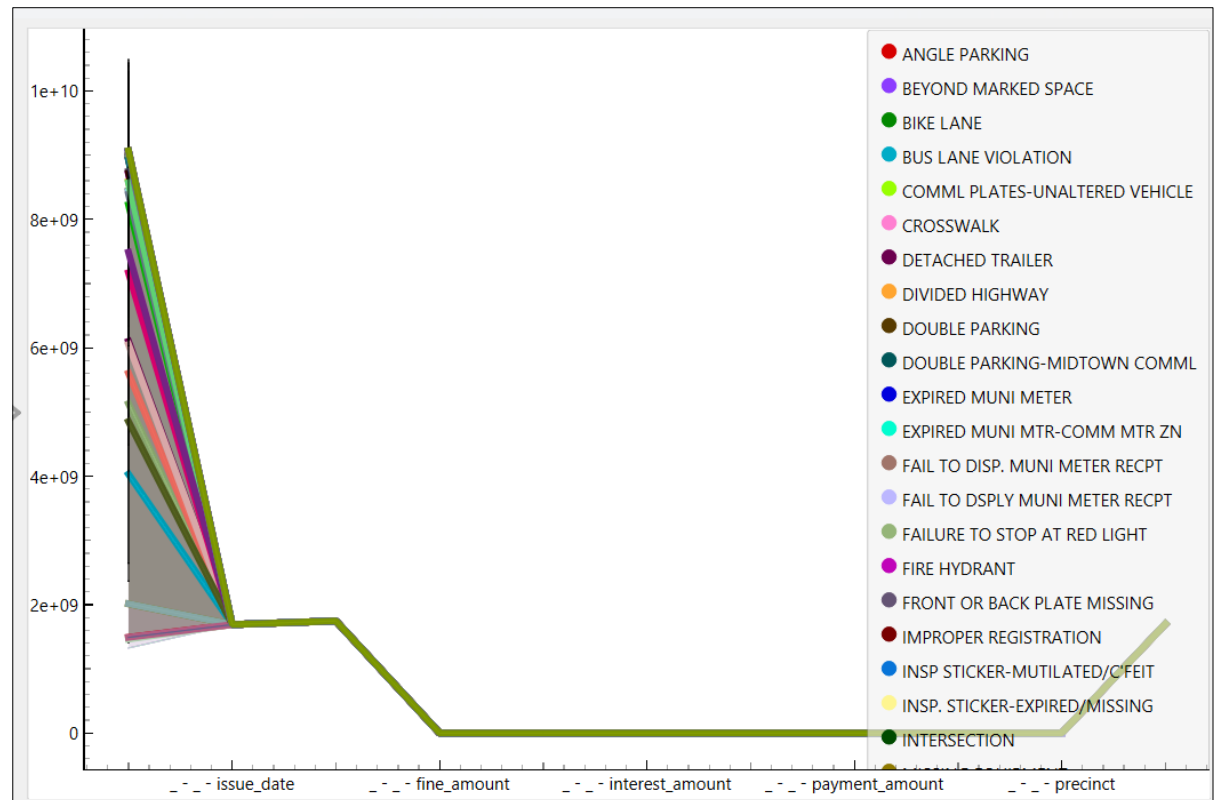
30. Double click on Box Plot and select field to graph, eg: Violation Type:



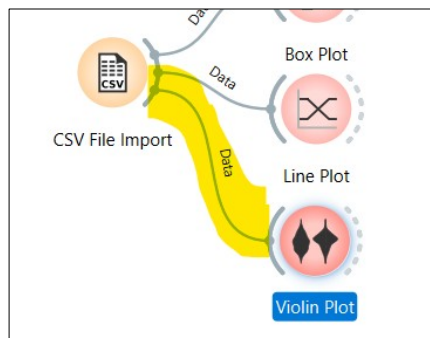
31. Drag Line Plot widget into work area and connect from CSV File to Line Plot:



32. Double click on Line Plot and select field to graph. Eg: Violation Type:



33. Drag Violin Plot widget into work area and connect from CSV File to Box Plot:



34. Double click on Violin Plot and select field to graph. Eg: Precinct:

