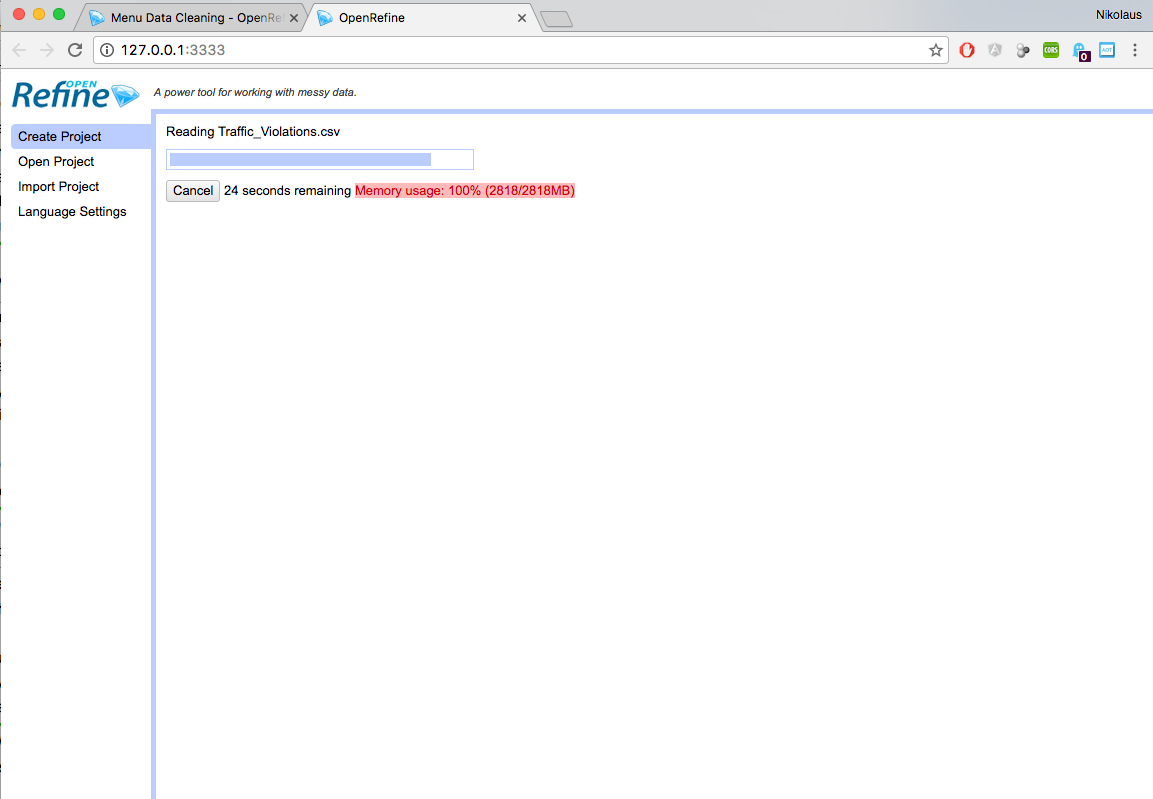
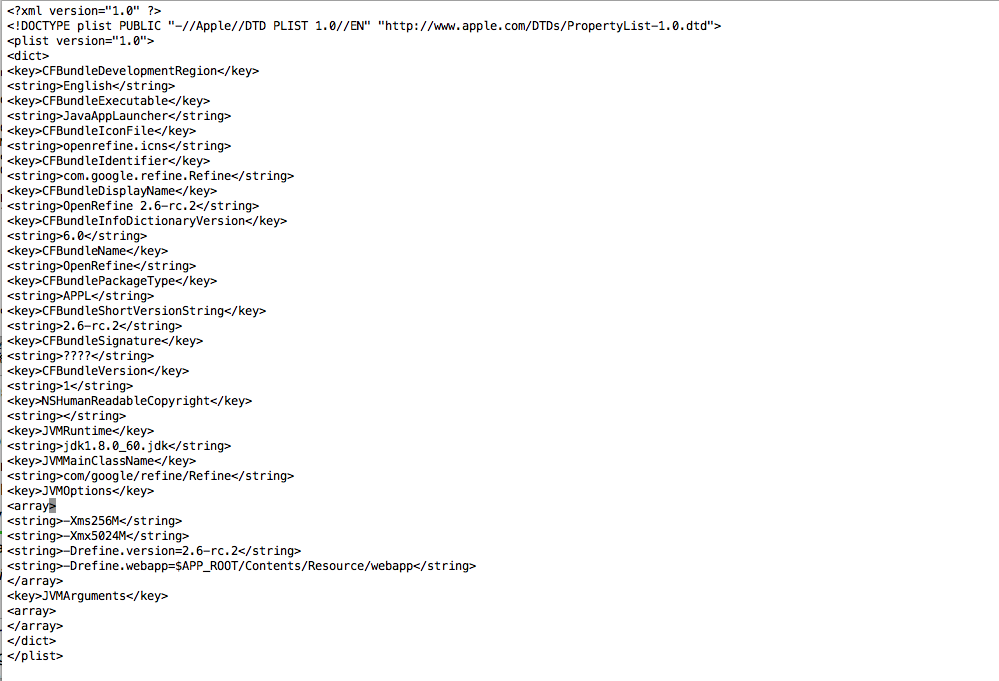
Just opening the csv file with openrefine we already faced Out of Memory problem,

Turn out there is limitation on how much memory openrefine can use. This one reason is why we consider using python / programming to clean up the data. However, to give more understanding about what we can clean in the dataset, openrefine may be a good start.



Therefore, we need to find a solution for this out of memory problem first and we found out that because Openrefine runs on java platform, we can configure memory usage by doing a little tweak in the openrefine configuration file (mac version)

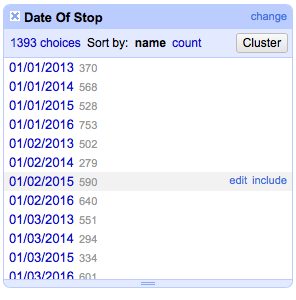
/Applications/OpenRefine.app/Contents/Info.plist



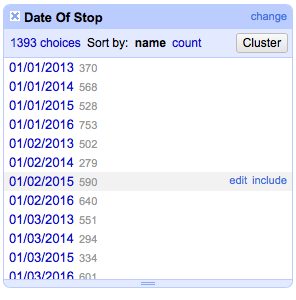
We change the memory usage into 5024M (5GB), be careful to do this one must have at least 10GB memory.

* Potential null value checking: Date of Stop, Time of Stop

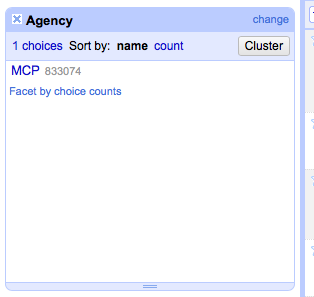
Date of Stop Field



Date of stop

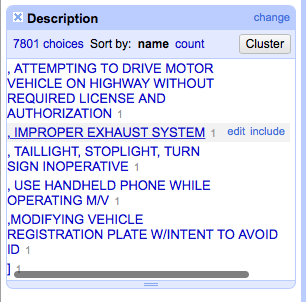


Agency

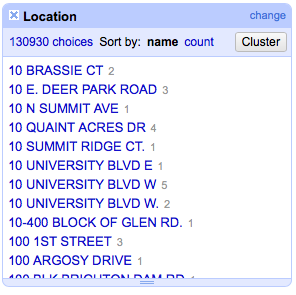


* Potential regular expression / string cleaning:

Description



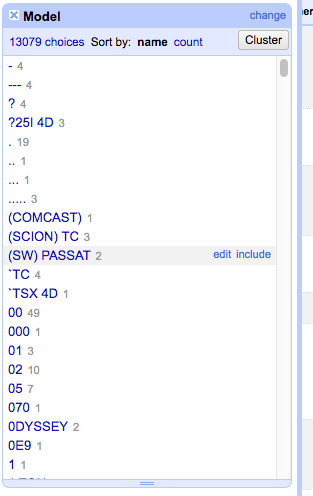
Location



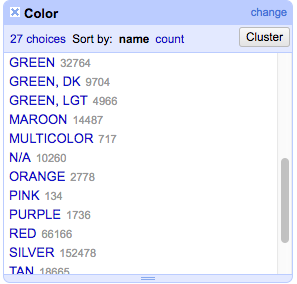
Make (unwanted values) : in addition we can do merging as well



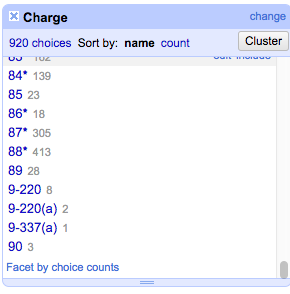
Model (Unwanted values)



Color (NA value and additional description with ,)



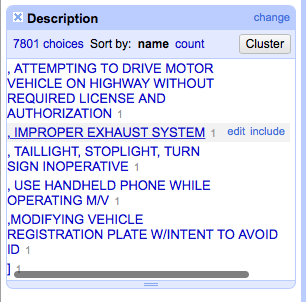
Charge (Unwanted value \*)



Driver City

* Potential clustering + merging operation:

Description

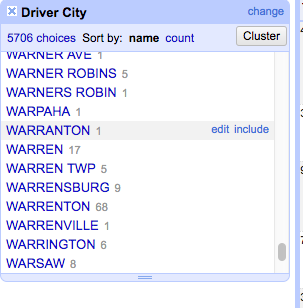


Location,

Make,

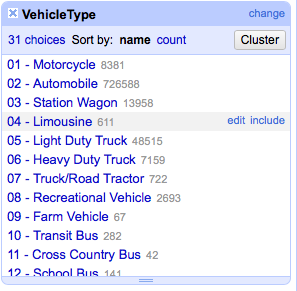
Model,

Driver City

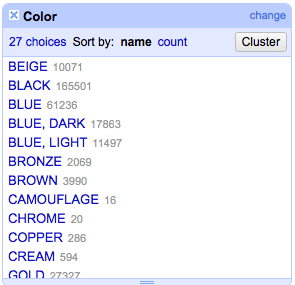


* Potential lookup value: State, SubAgency,

VehicleType,



Color,



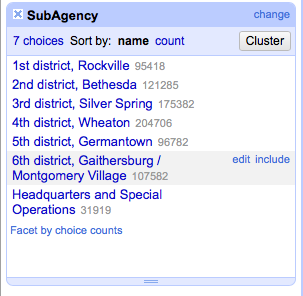
Charge (this should define the charge for the driver but somehow, the variance is so big, probably we can divide the charge as law number and section),

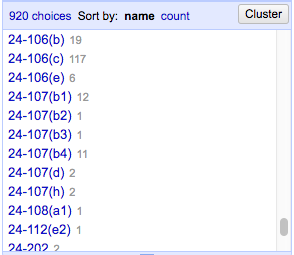
Article,

Race,

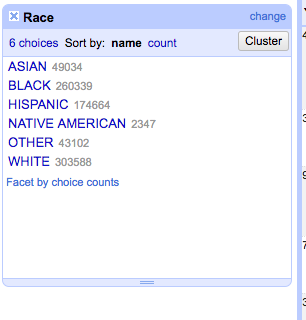
and Fields that have yes / no value (Accidents – Workzone, Contributed to Accident)

SubAgency:

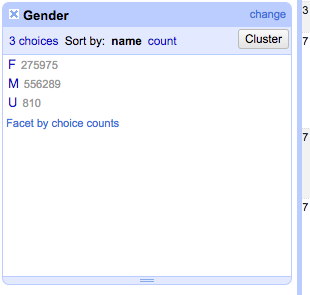




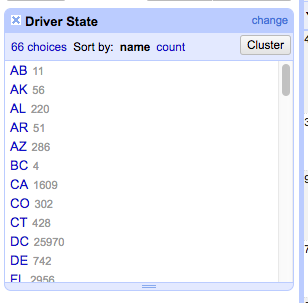
Race



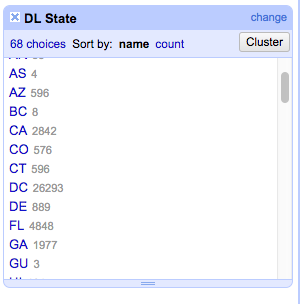
Gender (What is U mean -> unknown?) we accept these 3 values



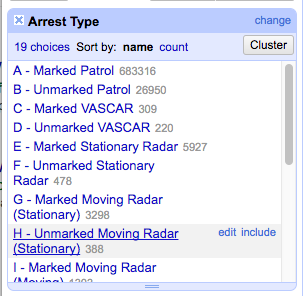
Driver State, beware of XX value -> what state is this?



DL State

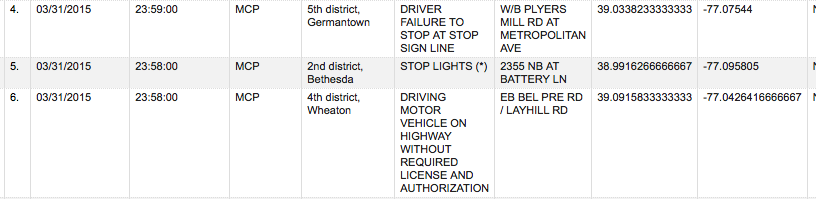


Arrest Type



Potential format checking: Date of stop, Time of Stop,

Latitude, Longitude => we can clean the value for at max 6 digits precision.



Potential Numeric Constraint Checking: Year (ridiculous number), we clean value for year more than this year (2016) and less than 1882



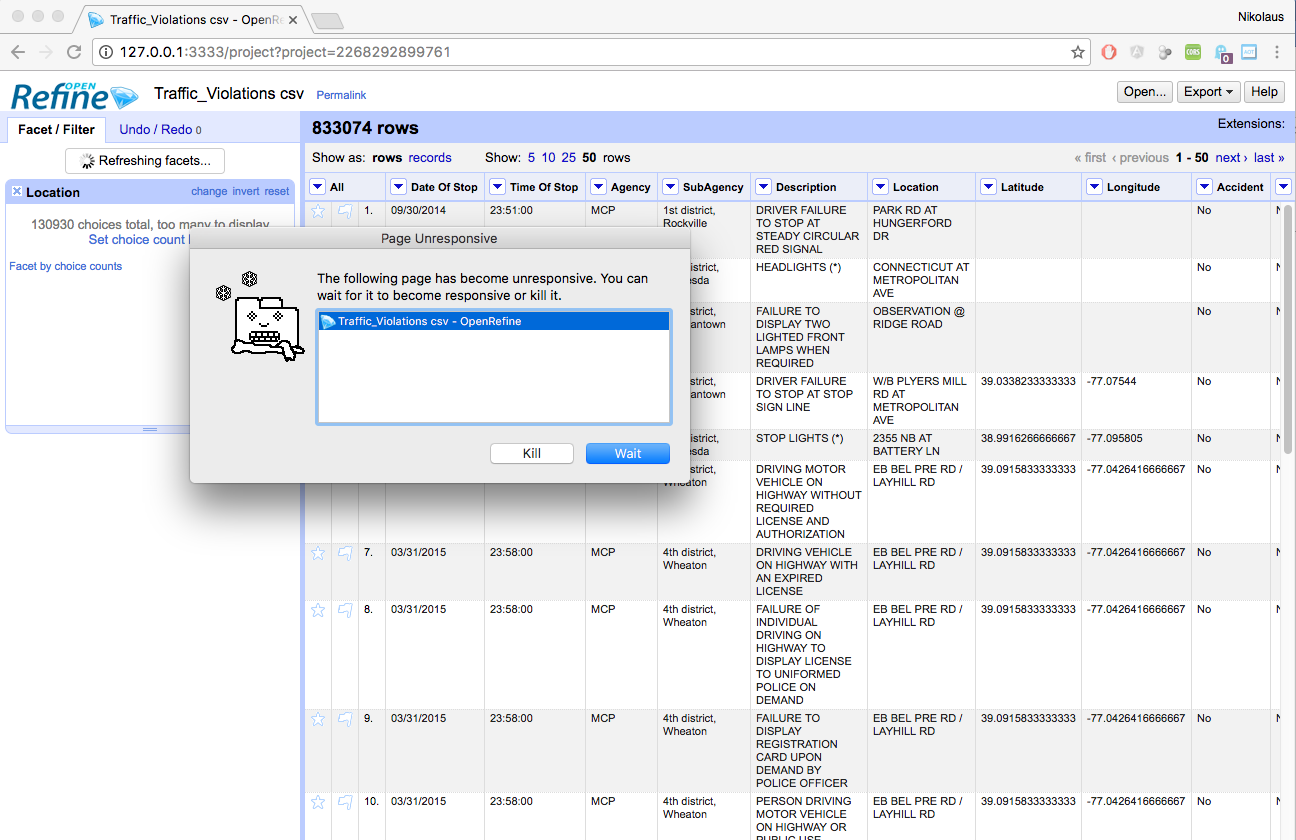


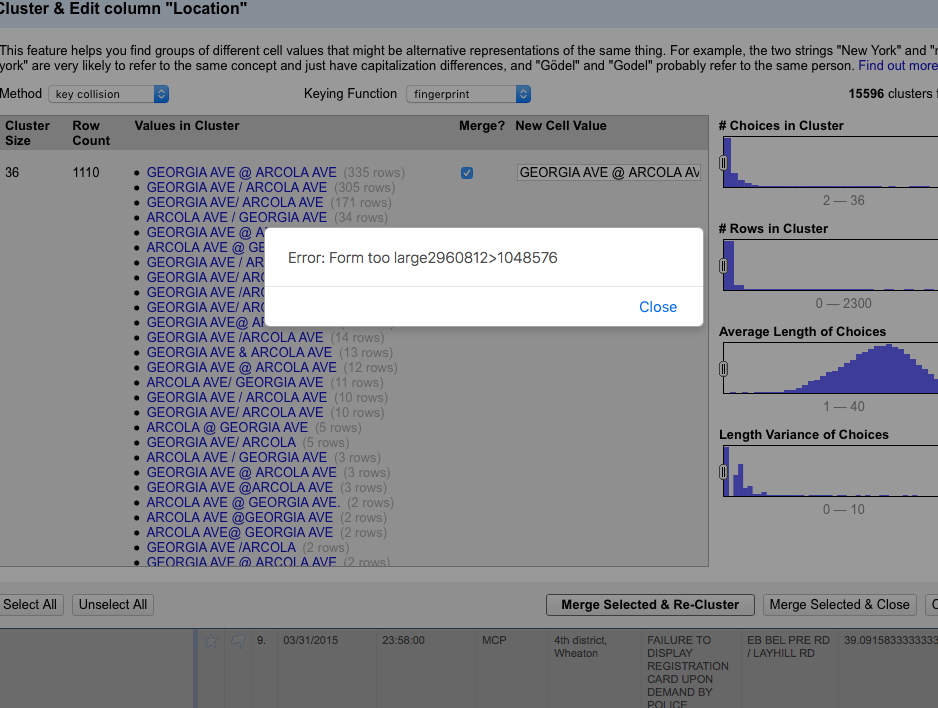
Article has blank values



Geolocation is duplicate value for latitude and longitude

OpenRefine hang when trying to make facet for Location





Need to find better way / technique

Function: clean / replace, custom function, ltrim, rtrim, uppercase, lowercase, lookup

Github project for source code and documentation:

<https://github.com/nikolausn/lab-data-cleaning-project>

Action:

Field description

* Replace improper characters: , ] ( ) \* to space
* Trim the value
* Collapse white space
* Uppercase the value