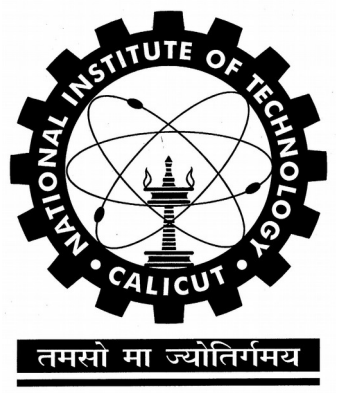


National Institute Of Technology – Calicut

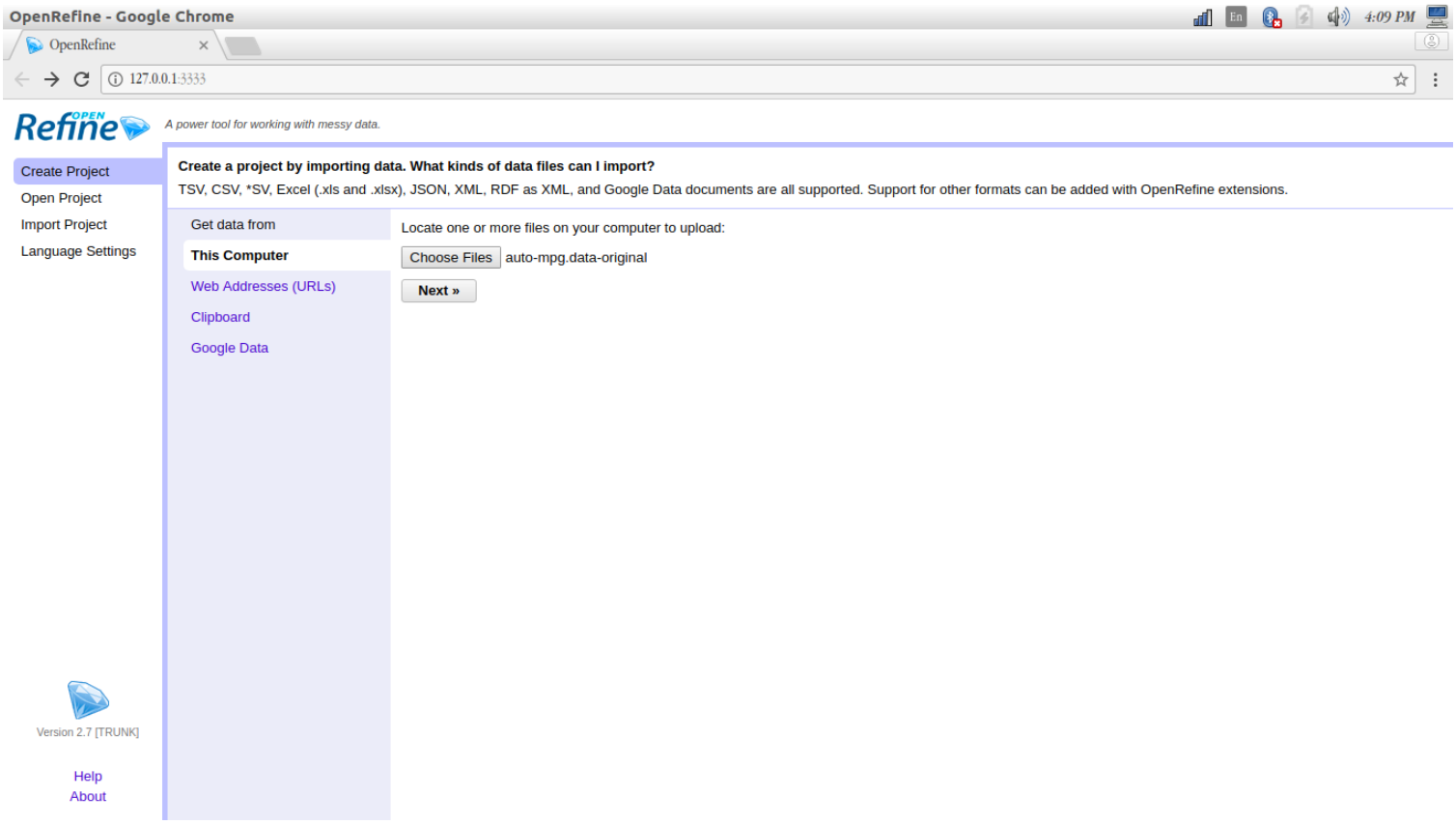
Data Mining

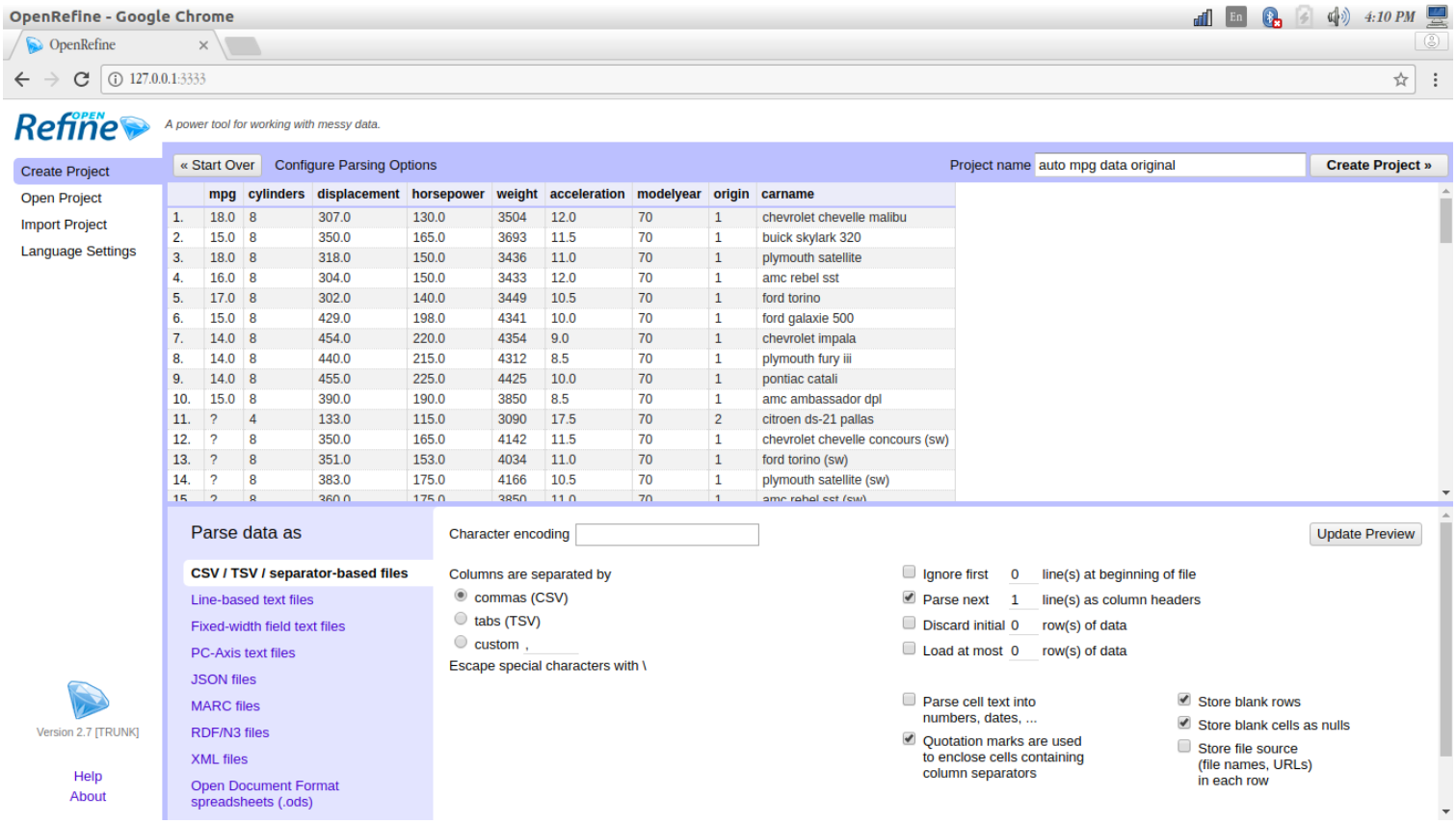
(Data pre-processing assignment)

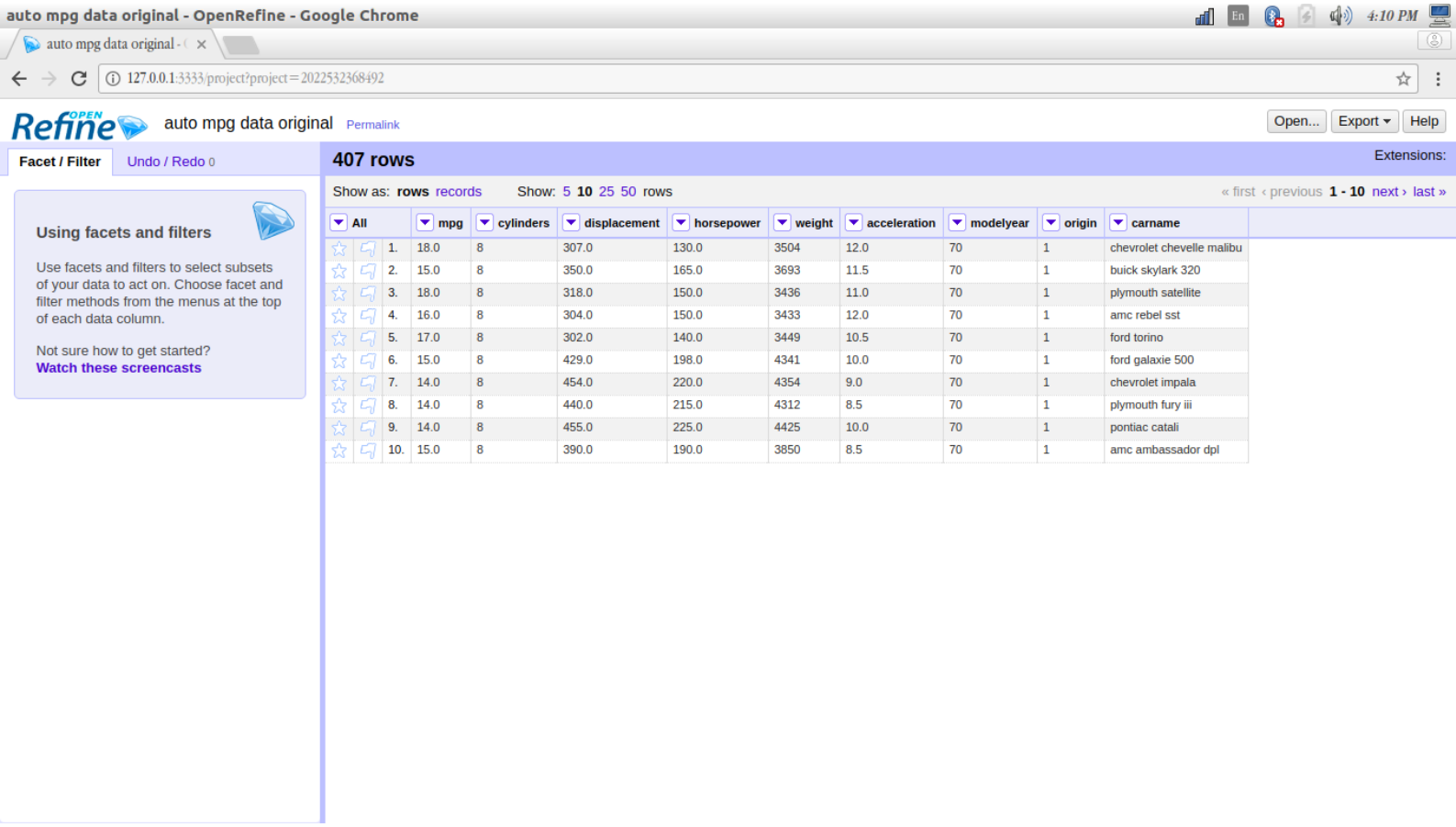
Que 01.
(AUTO MPG data set)
Data Pre-processing Open refine

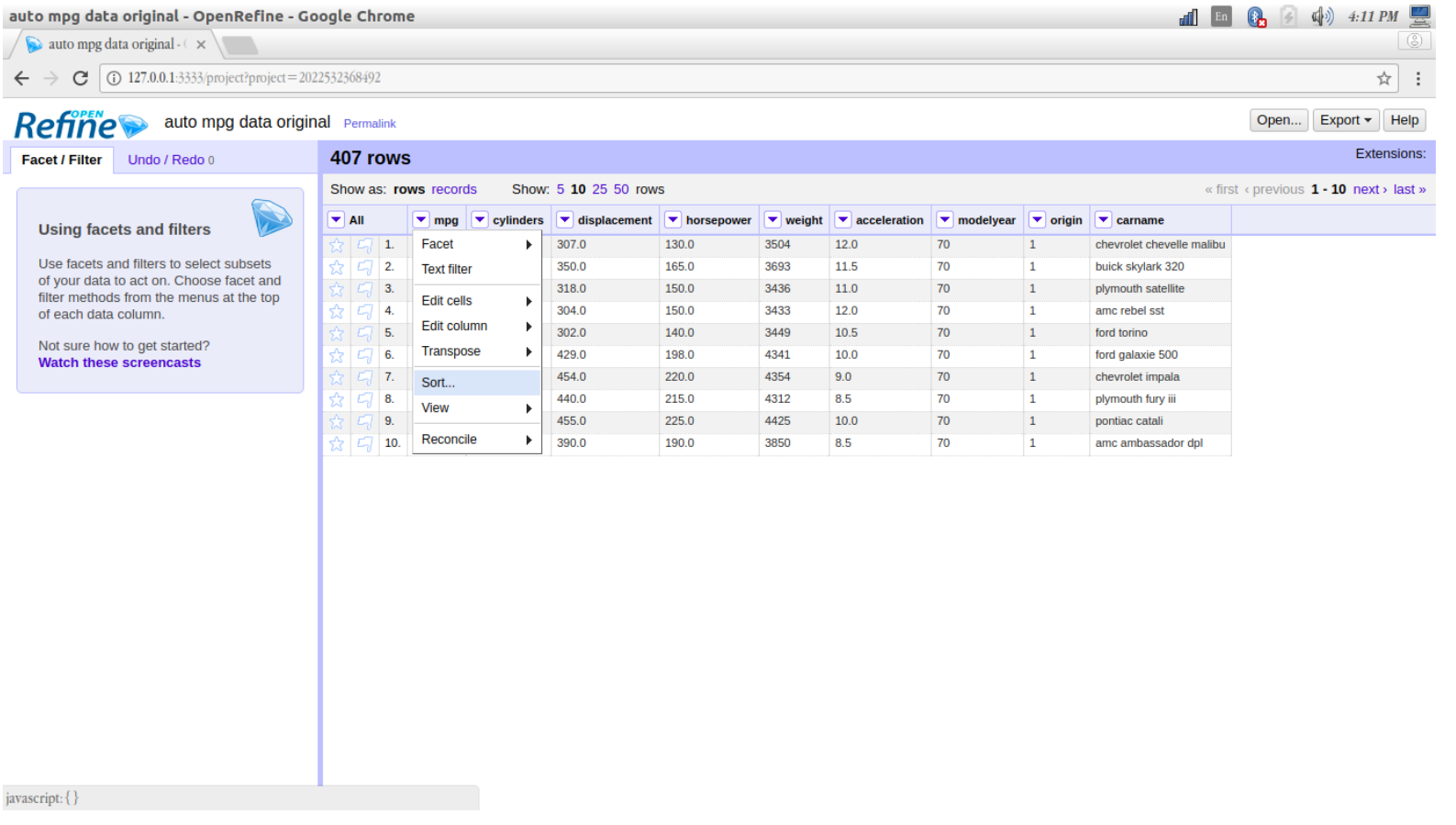
T.G. Deshan K. Sumanathilaka
B150413CS
Computer Science and Engineering
(B.Tech)

1)Loading the Data Set to Open Refine

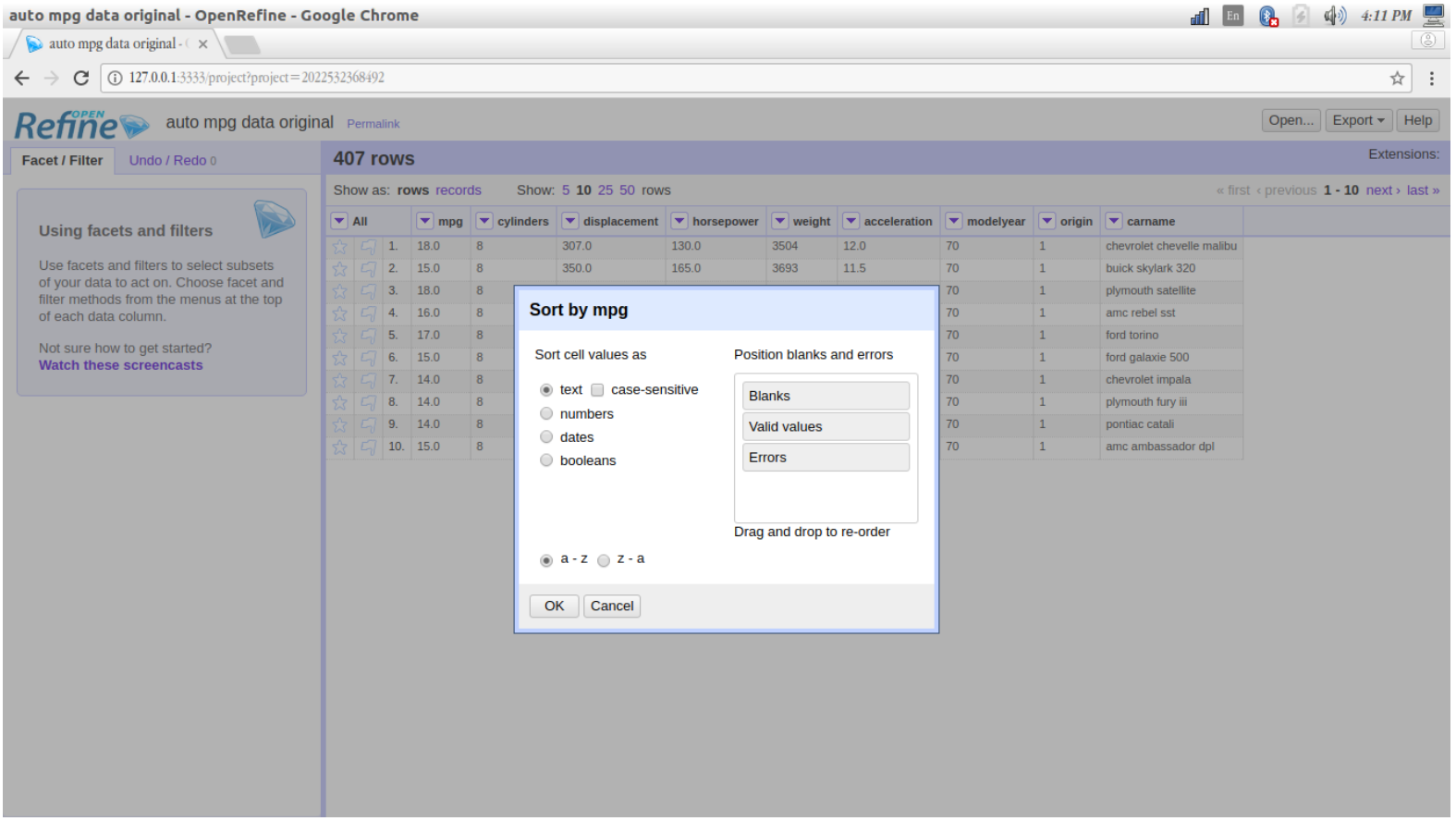


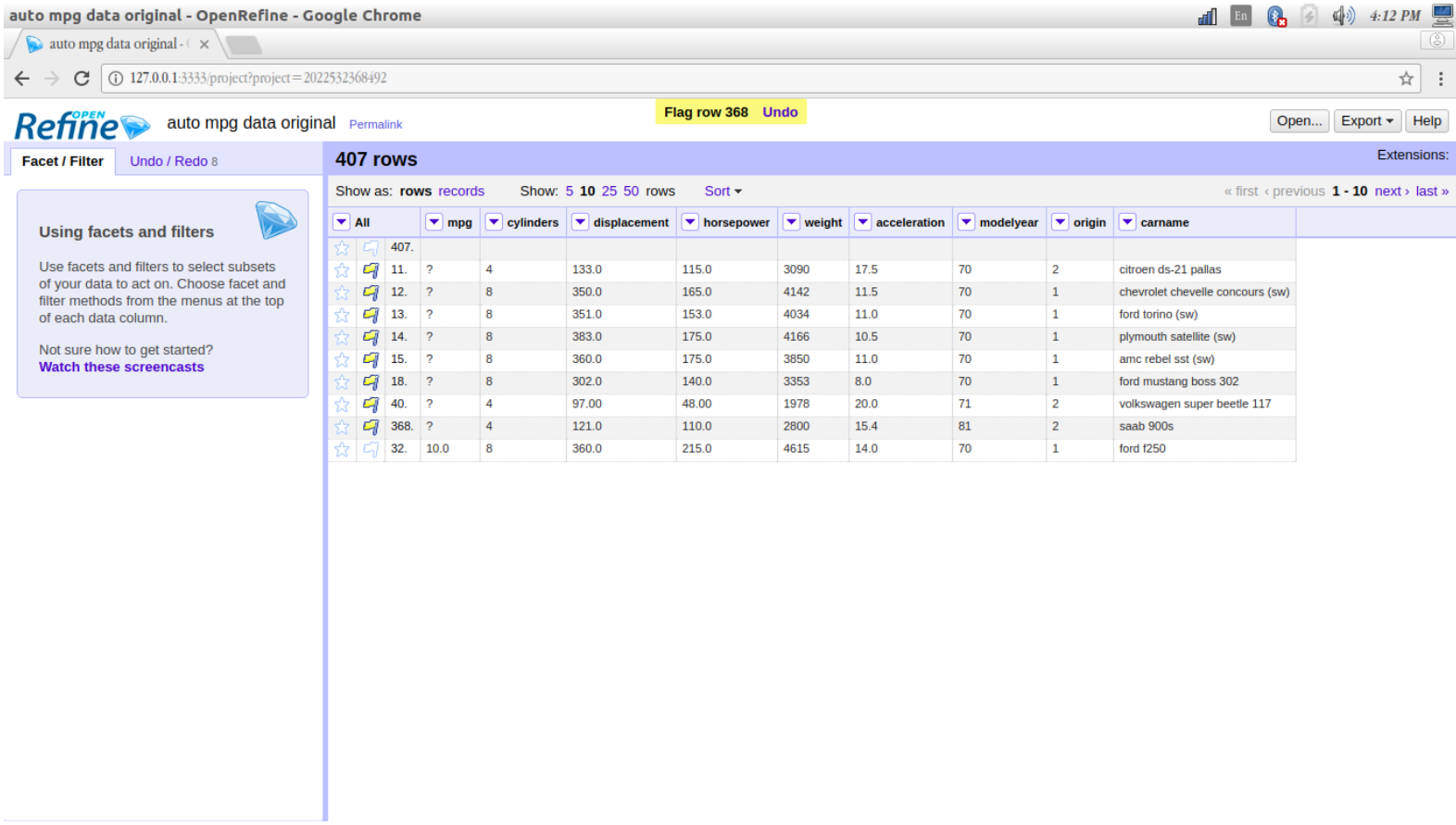



auto mpg data original - OpenRefine - Google Chrome
4:10 PM
auto mpg data original
127.0.0.1:3333/project?project=2022532368492
auto mpg data original
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 0
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
407 rows
Extensions:
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 10 next › last »
All
mpg
cylinders
displacement
horsepower
weight
acceleration
modelyear
origin
carname
1. 18.0 8 307.0 130.0 3504 12.0 70 1 chevrolet chevelle malibu
2. 15.0 8 350.0 165.0 3693 11.5 70 1 buick skylark 320
3. 18.0 8 318.0 150.0 3436 11.0 70 1 plymouth satellite
4. 16.0 8 304.0 150.0 3433 12.0 70 1 amc rebel sst
5. 17.0 8 302.0 140.0 3449 10.5 70 1 ford torino
6. 15.0 8 429.0 198.0 4341 10.0 70 1 ford galaxie 500
7. 14.0 8 454.0 220.0 4354 9.0 70 1 chevrolet impala
8. 14.0 8 440.0 215.0 4312 8.5 70 1 plymouth fury iii
9. 14.0 8 455.0 225.0 4425 10.0 70 1 pontiac catali
10. 15.0 8 390.0 190.0 3850 8.5 70 1 amc ambassador dpl

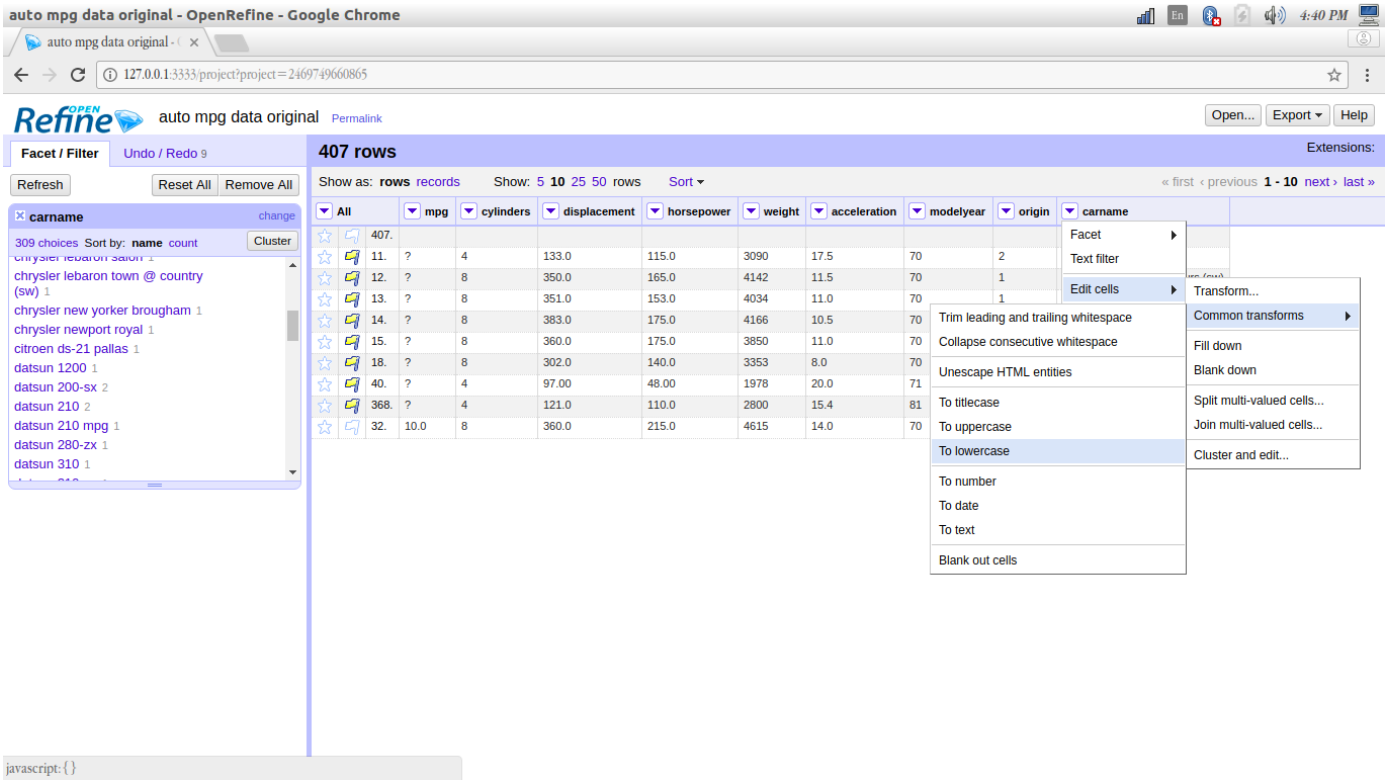

auto mpg data original - OpenRefine - Google Chrome
4:11 PM
auto mpg data original
127.0.0.1:3333/project?project=2022532368492
auto mpg data original
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 0
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
407 rows
Extensions:
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 10 next › last »
All
mpg
cylinders
displacement
horsepower
weight
acceleration
modelyear
origin
carname
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
1. 307.0 130.0 3504 12.0 70 1 chevrolet chevelle malibu
2. 350.0 165.0 3693 11.5 70 1 buick skylark 320
3. 318.0 150.0 3436 11.0 70 1 plymouth satellite
4. 304.0 150.0 3433 12.0 70 1 amc rebel sst
5. 302.0 140.0 3449 10.5 70 1 ford torino
6. 429.0 198.0 4341 10.0 70 1 ford galaxie 500
7. 454.0 220.0 4354 9.0 70 1 chevrolet impala
8. 440.0 215.0 4312 8.5 70 1 plymouth fury iii
9. 455.0 225.0 4425 10.0 70 1 pontiac catali
10. 390.0 190.0 3850 8.5 70 1 amc ambassador dpl
javascript:{}

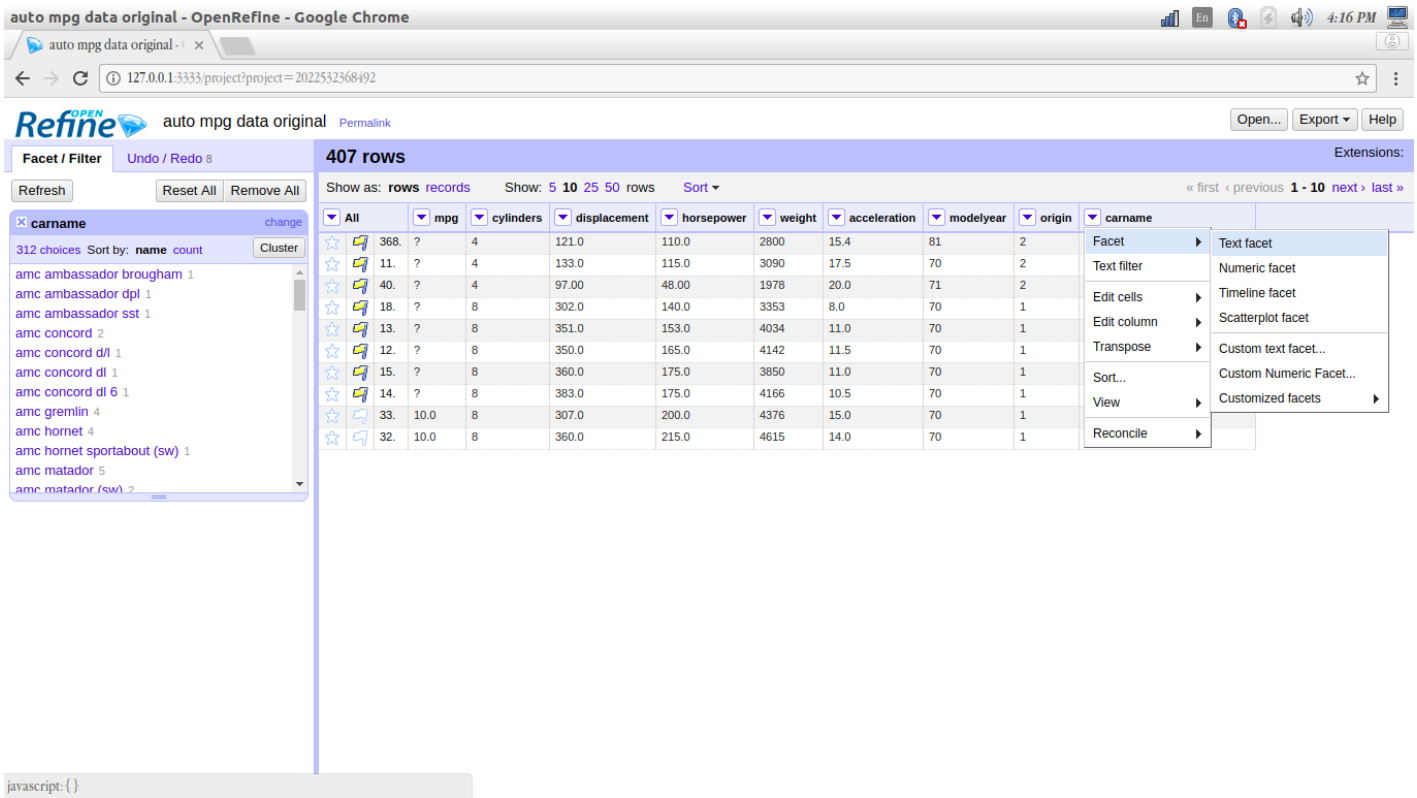
2) Sort the position of the blank ,error data tuples to a order , and flag the data tuples with missing values.

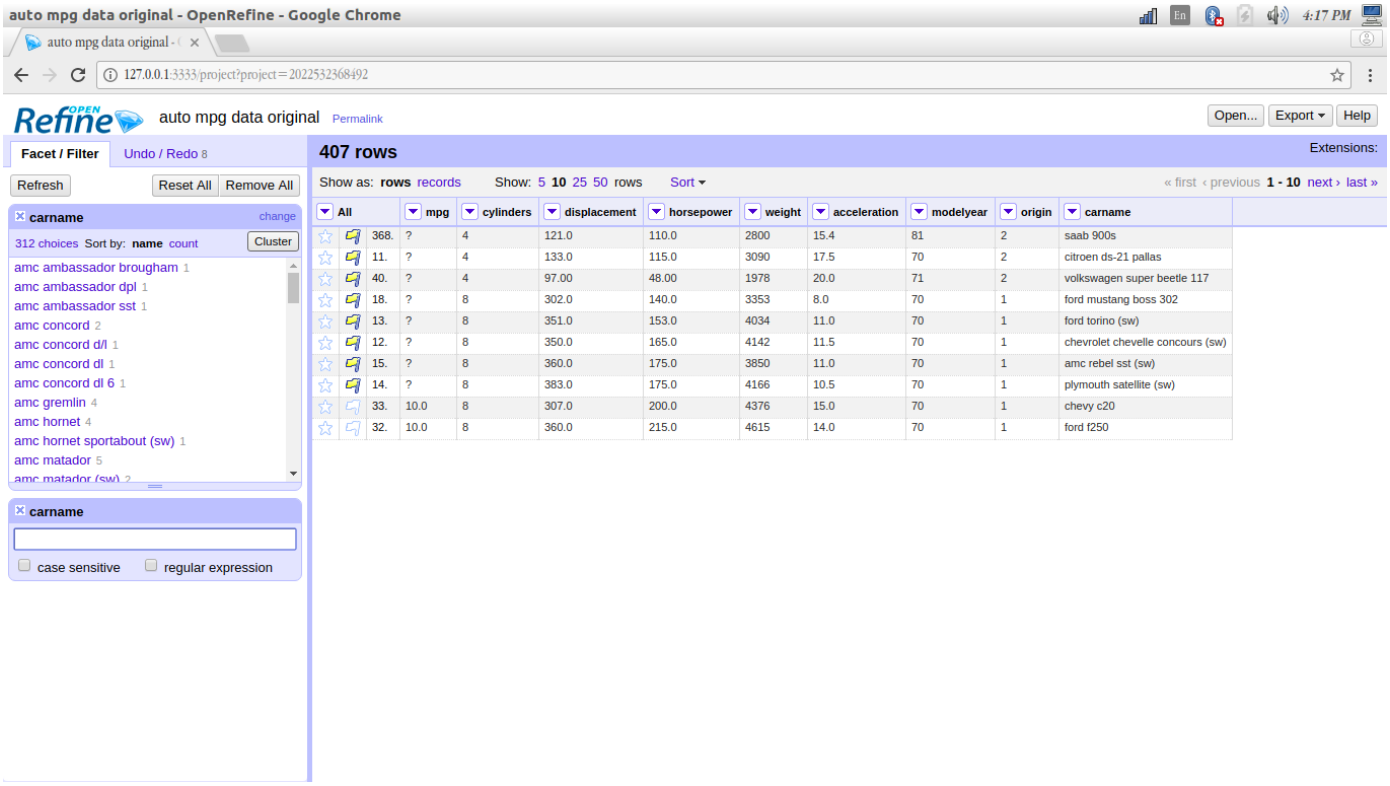




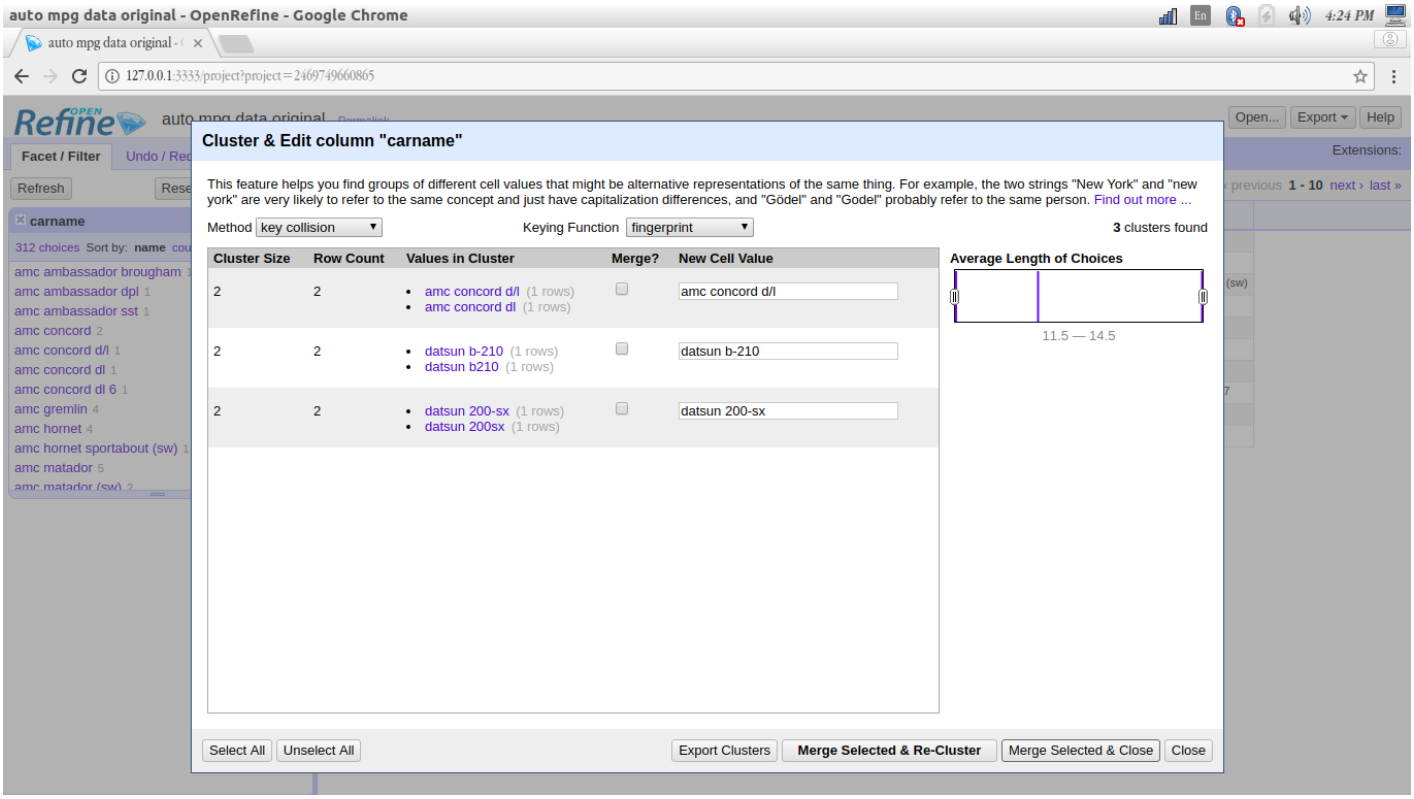
3) Change the non-numeric attribute to lowercase before it begins to cluster.



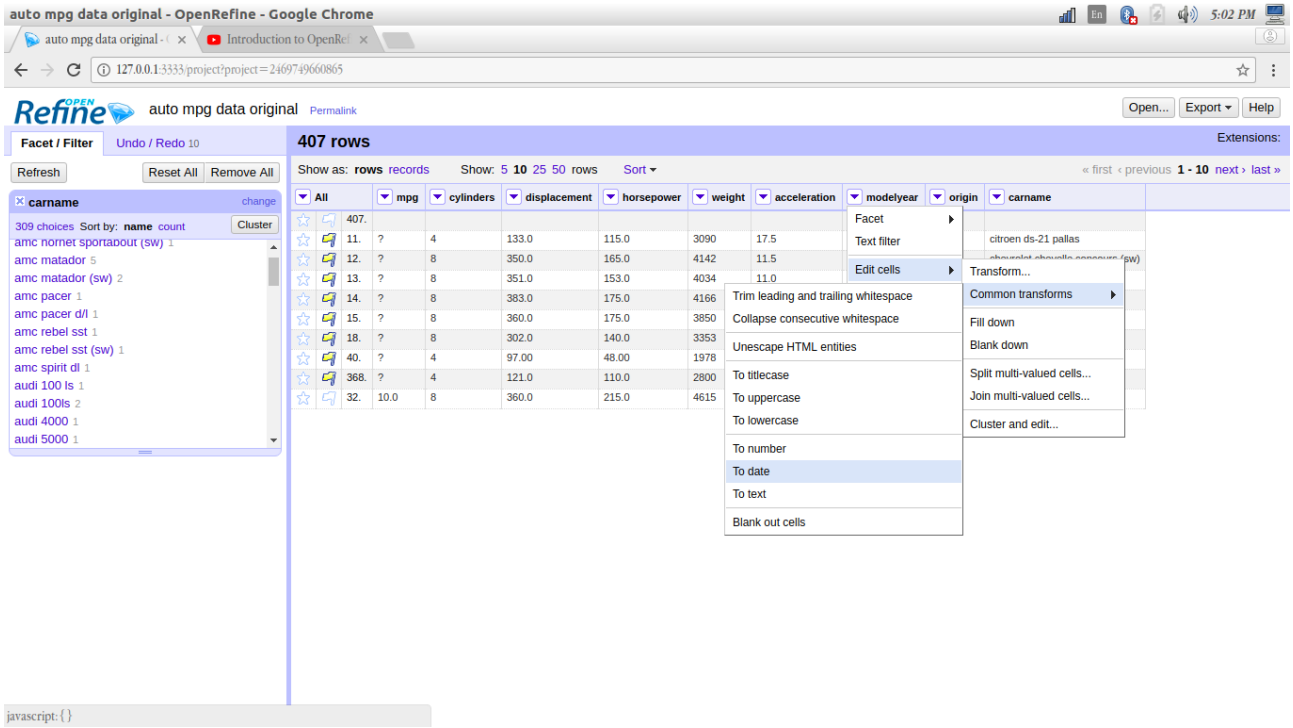




4)Cluster it and merger the related cells to remove inconsistent spellings.

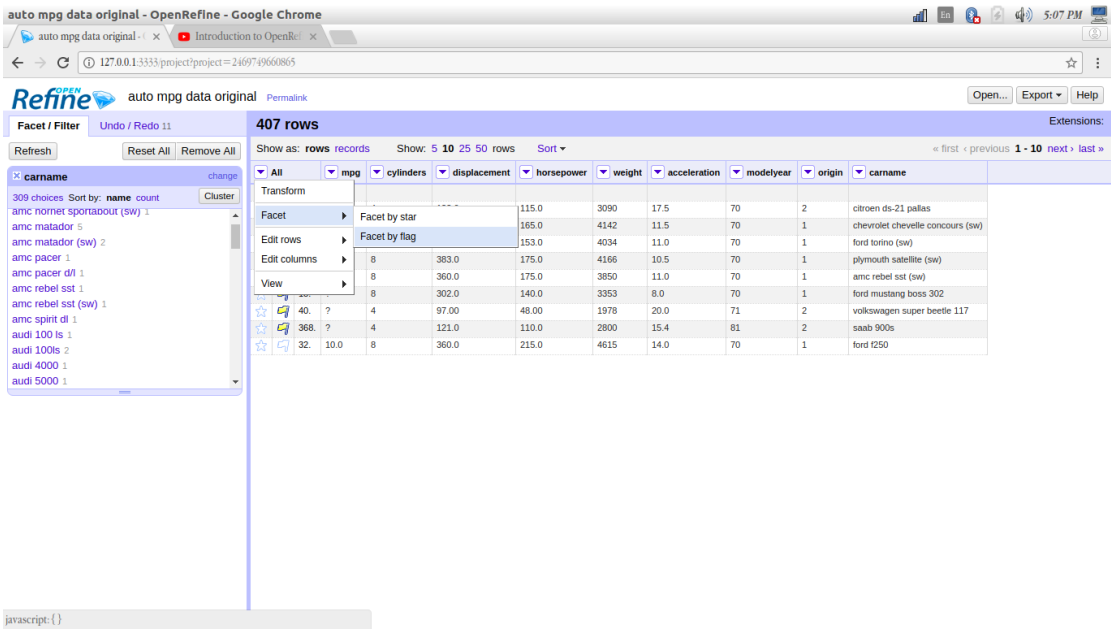


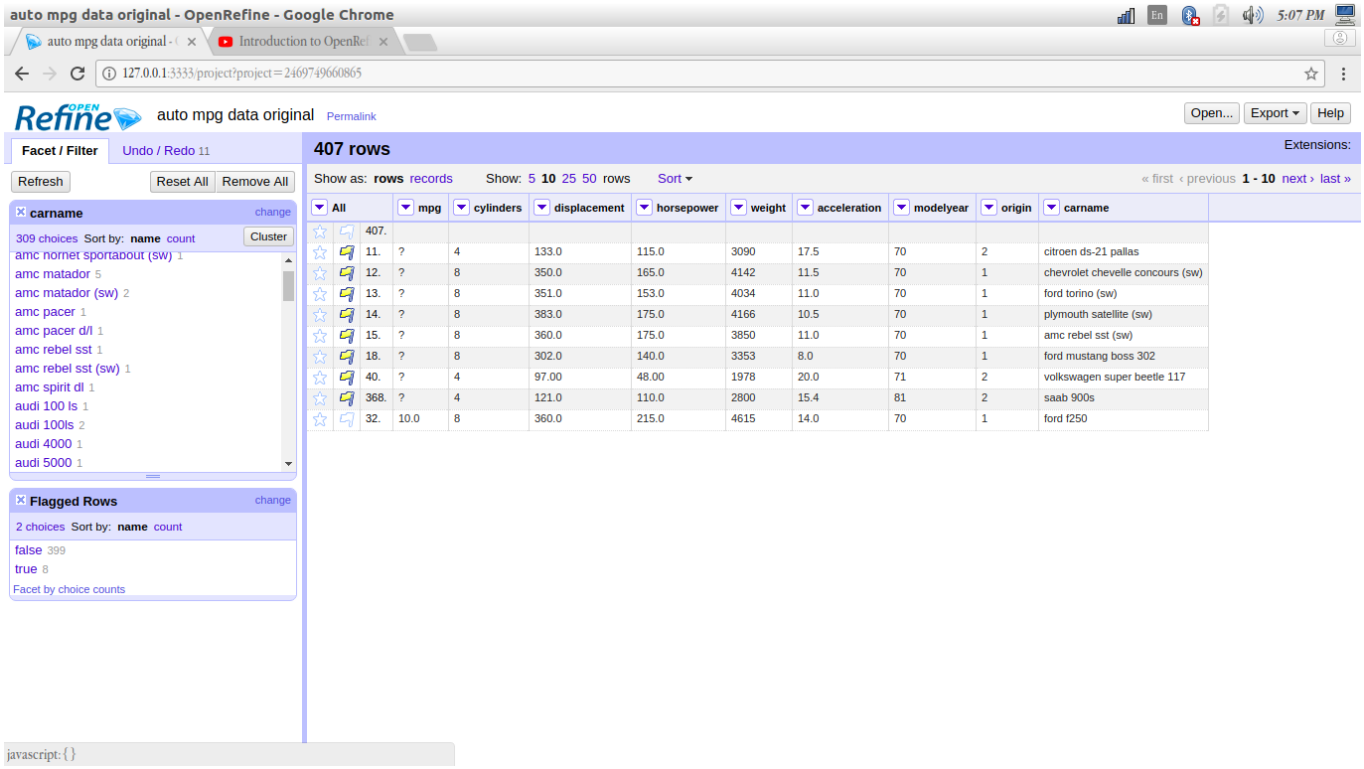
5) Extracting and cleaning values for dates

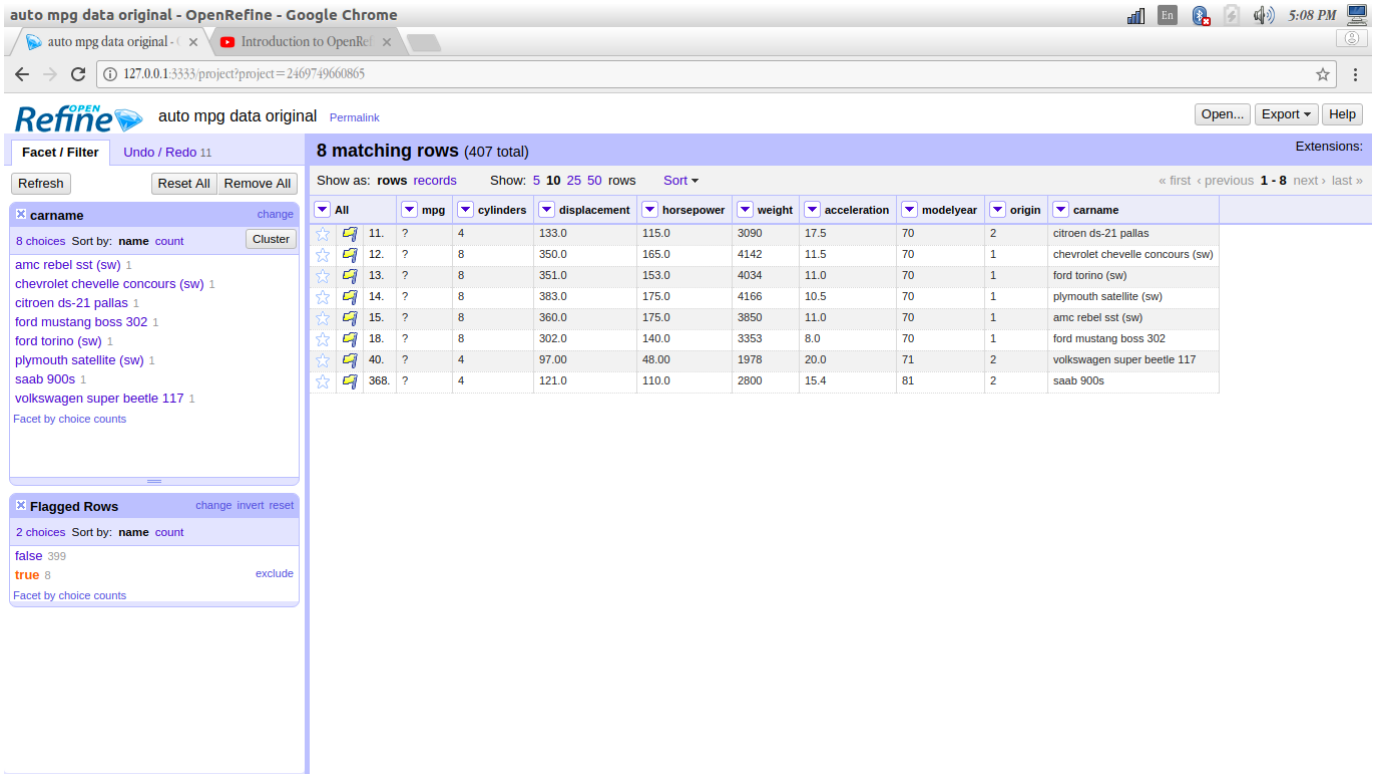


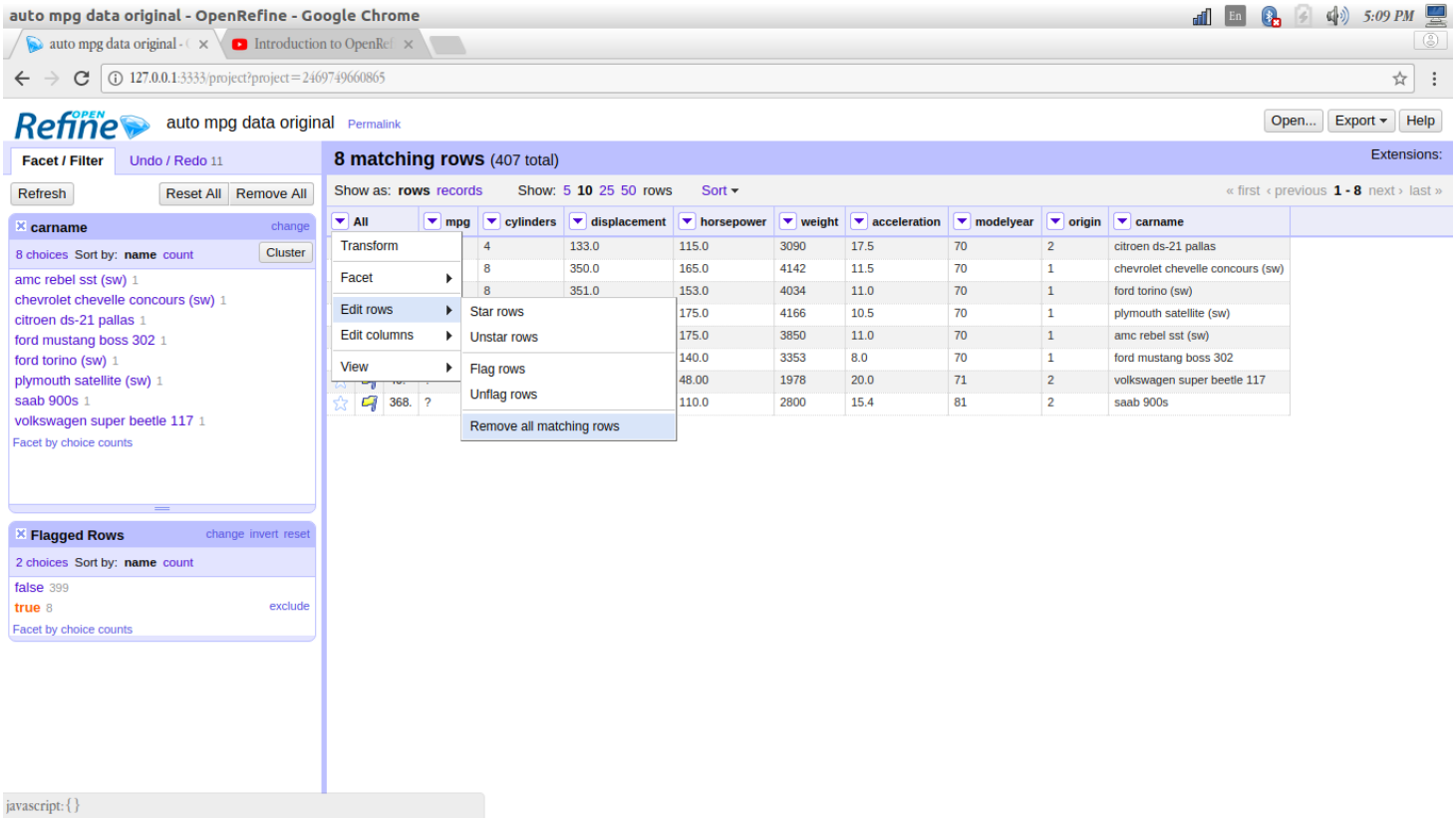
(but in this data set changing the date to a specific order is not required.)

6) Fill in the missing values by the various options supported by Open refine

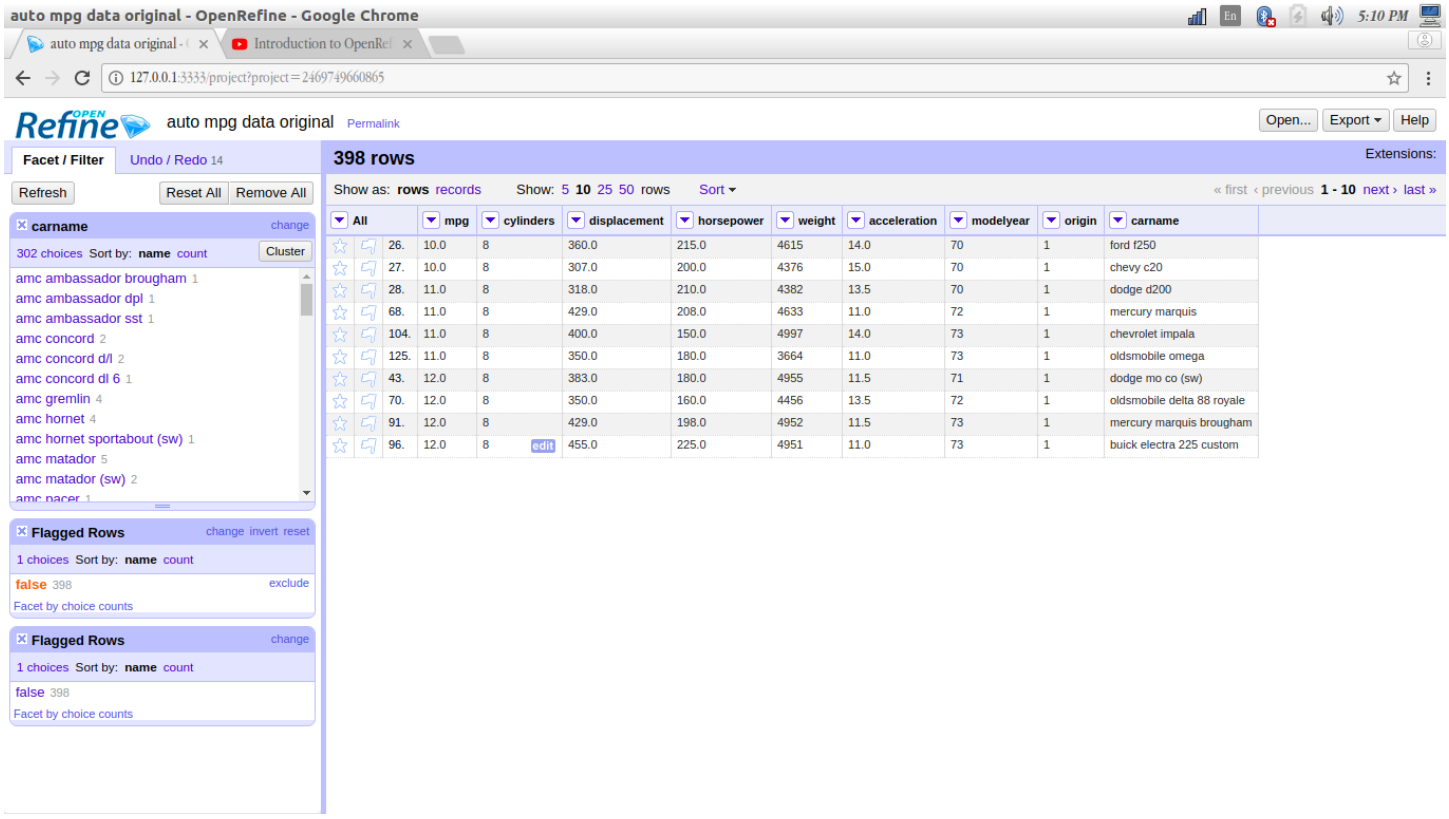



auto mpg data original - OpenRefine - Google Chrome
5:07 PM
auto mpg data original - OpenRefine
Introduction to OpenRef
127.0.0.1:3333/project?project=2469749660865
auto mpg data original
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 11
Refresh
Reset All
Remove All
carname
change
309 choices Sort by: name count
Cluster
amc hornet sportabout (sw) 1
amc matador 5
amc matador (sw) 2
amc pacer 1
amc pacer d/l 1
amc rebel sst 1
amc rebel sst (sw) 1
amc spirit dl 1
audi 100 ls 1
audi 100ls 2
audi 4000 1
audi 5000 1
Flagged Rows
change
2 choices Sort by: name count
false 399
true 8
Facet by choice counts
407 rows
Extensions:
Show as: rows records
Show: 5 10 25 50 rows
Sort
« first ‹ previous 1 - 10 next › last »
All
mpg
cylinders
displacement
horsepower
weight
acceleration
modelyear
origin
carname
407.
11. ? 4 133.0 115.0 3090 17.5 70 2 citroen ds-21 pallas
12. ? 8 350.0 165.0 4142 11.5 70 1 chevrolet chevelle concours (sw)
13. ? 8 351.0 153.0 4034 11.0 70 1 ford torino (sw)
14. ? 8 383.0 175.0 4166 10.5 70 1 plymouth satellite (sw)
15. ? 8 360.0 175.0 3850 11.0 70 1 amc rebel sst (sw)
18. ? 8 302.0 140.0 3353 8.0 70 1 ford mustang boss 302
40. ? 4 97.00 48.00 1978 20.0 71 2 volkswagen super beetle 117
368. ? 4 121.0 110.0 2800 15.4 81 2 saab 900s
32. 10.0 8 360.0 215.0 4615 14.0 70 1 ford f250
javascript:{}

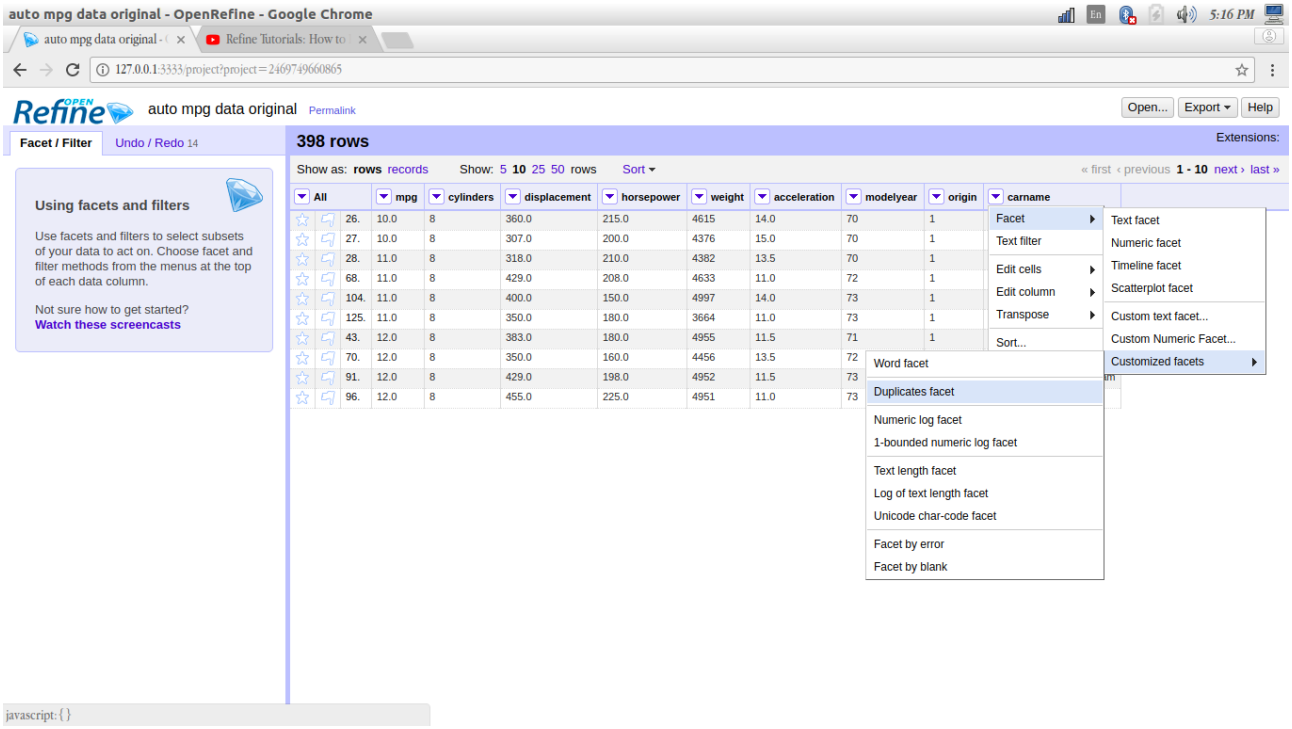

auto mpg data original - OpenRefine - Google Chrome
5:08 PM
auto mpg data original - OpenRefine
Introduction to OpenRef
127.0.0.1:3333/project?project=2469749660865
auto mpg data original
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 11
Refresh
Reset All
Remove All
carname
change
8 choices Sort by: name count
Cluster
amc rebel sst (sw) 1
chevrolet chevelle concours (sw) 1
citroen ds-21 pallas 1
ford mustang boss 302 1
ford torino (sw) 1
plymouth satellite (sw) 1
saab 900s 1
volkswagen super beetle 117 1
Facet by choice counts
Flagged Rows
change invert reset
2 choices Sort by: name count
false 399
true 8
exclude
Facet by choice counts
8 matching rows (407 total)
Extensions:
Show as: rows records
Show: 5 10 25 50 rows
Sort
« first ‹ previous 1 - 8 next › last »
All
mpg
cylinders
displacement
horsepower
weight
acceleration
modelyear
origin
carname
11. ? 4 133.0 115.0 3090 17.5 70 2 citroen ds-21 pallas
12. ? 8 350.0 165.0 4142 11.5 70 1 chevrolet chevelle concours (sw)
13. ? 8 351.0 153.0 4034 11.0 70 1 ford torino (sw)
14. ? 8 383.0 175.0 4166 10.5 70 1 plymouth satellite (sw)
15. ? 8 360.0 175.0 3850 11.0 70 1 amc rebel sst (sw)
18. ? 8 302.0 140.0 3353 8.0 70 1 ford mustang boss 302
40. ? 4 97.00 48.00 1978 20.0 71 2 volkswagen super beetle 117
368. ? 4 121.0 110.0 2800 15.4 81 2 saab 900s

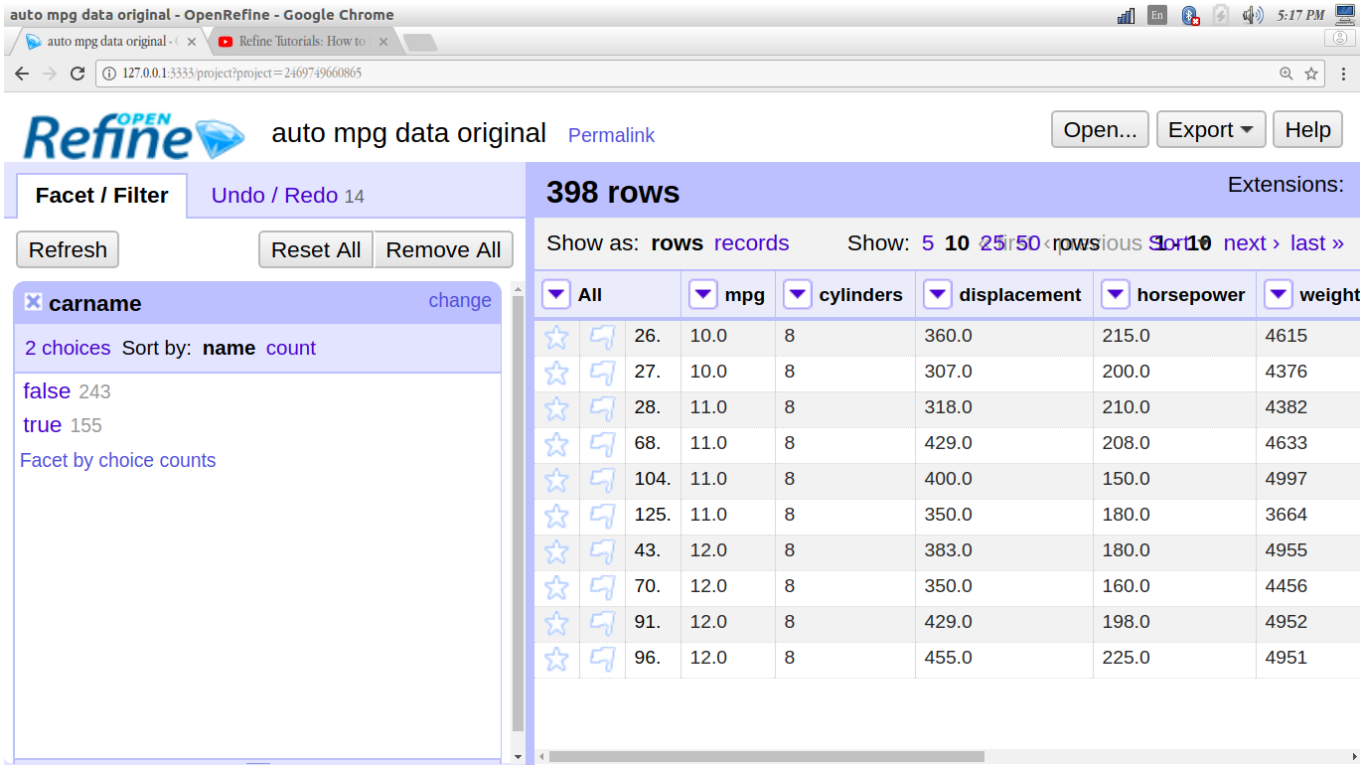


(by removing the missing values tuples, no of rows changed.)



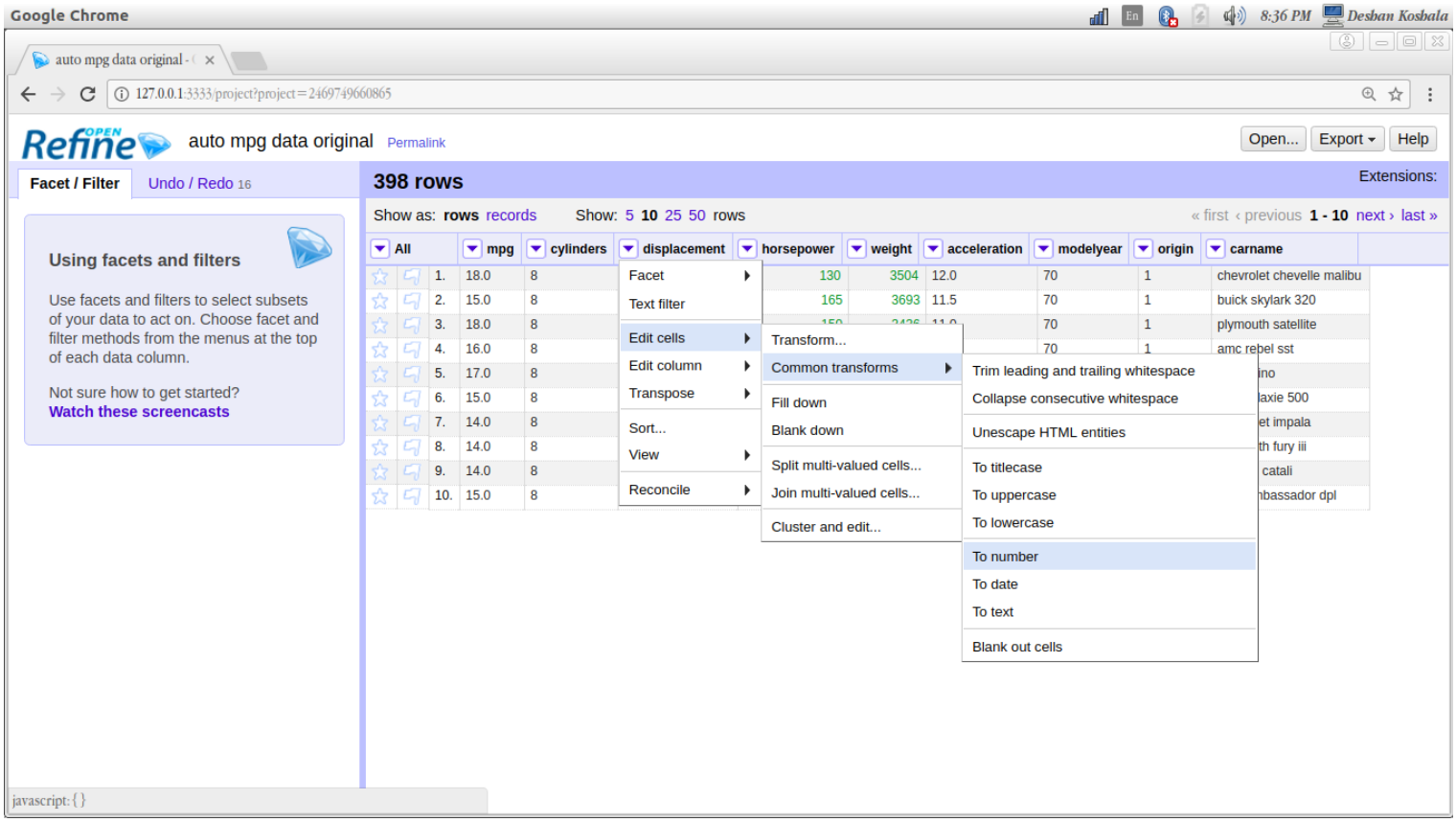
7) Removing duplicate columns(in some cases)

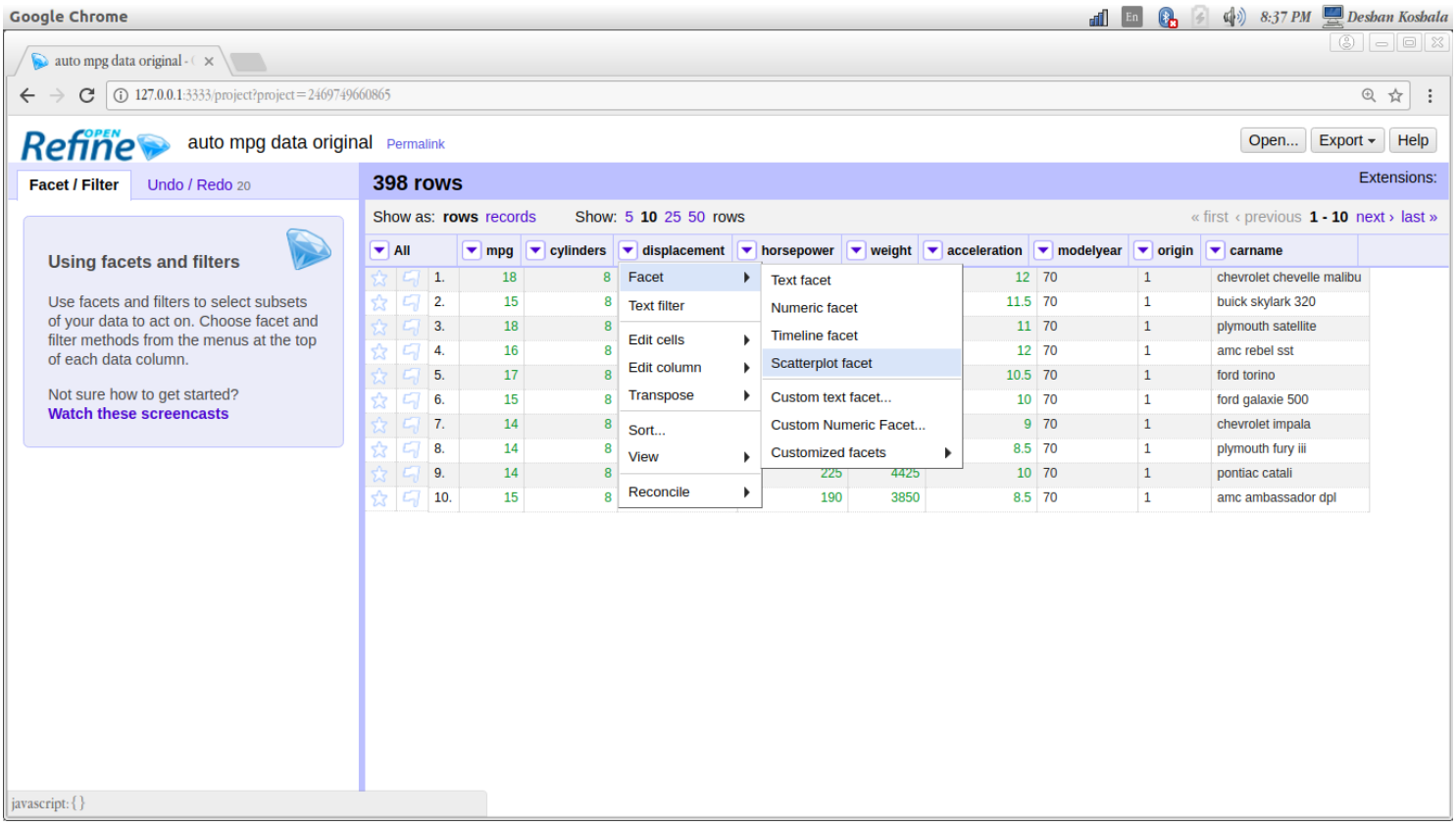


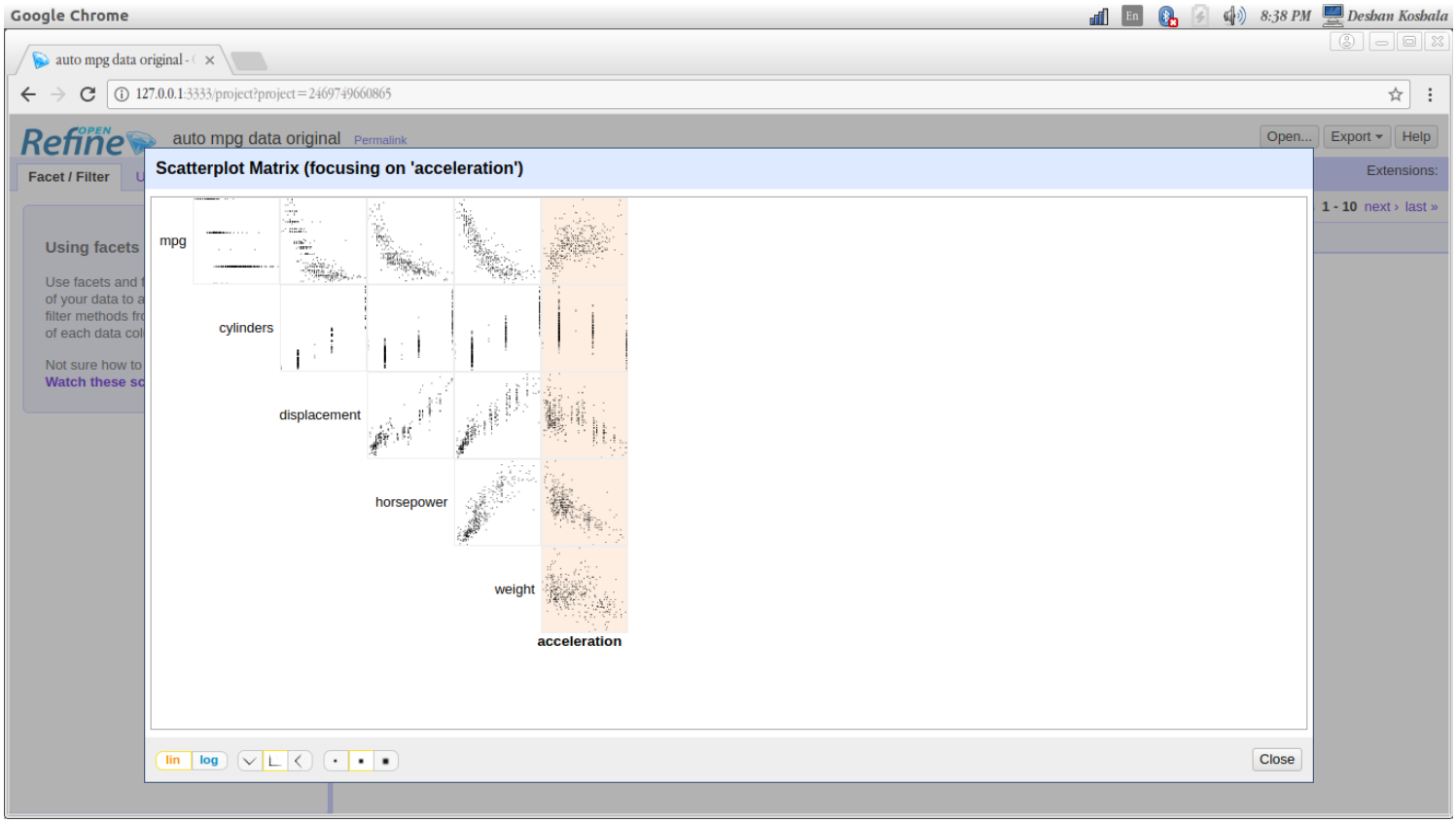


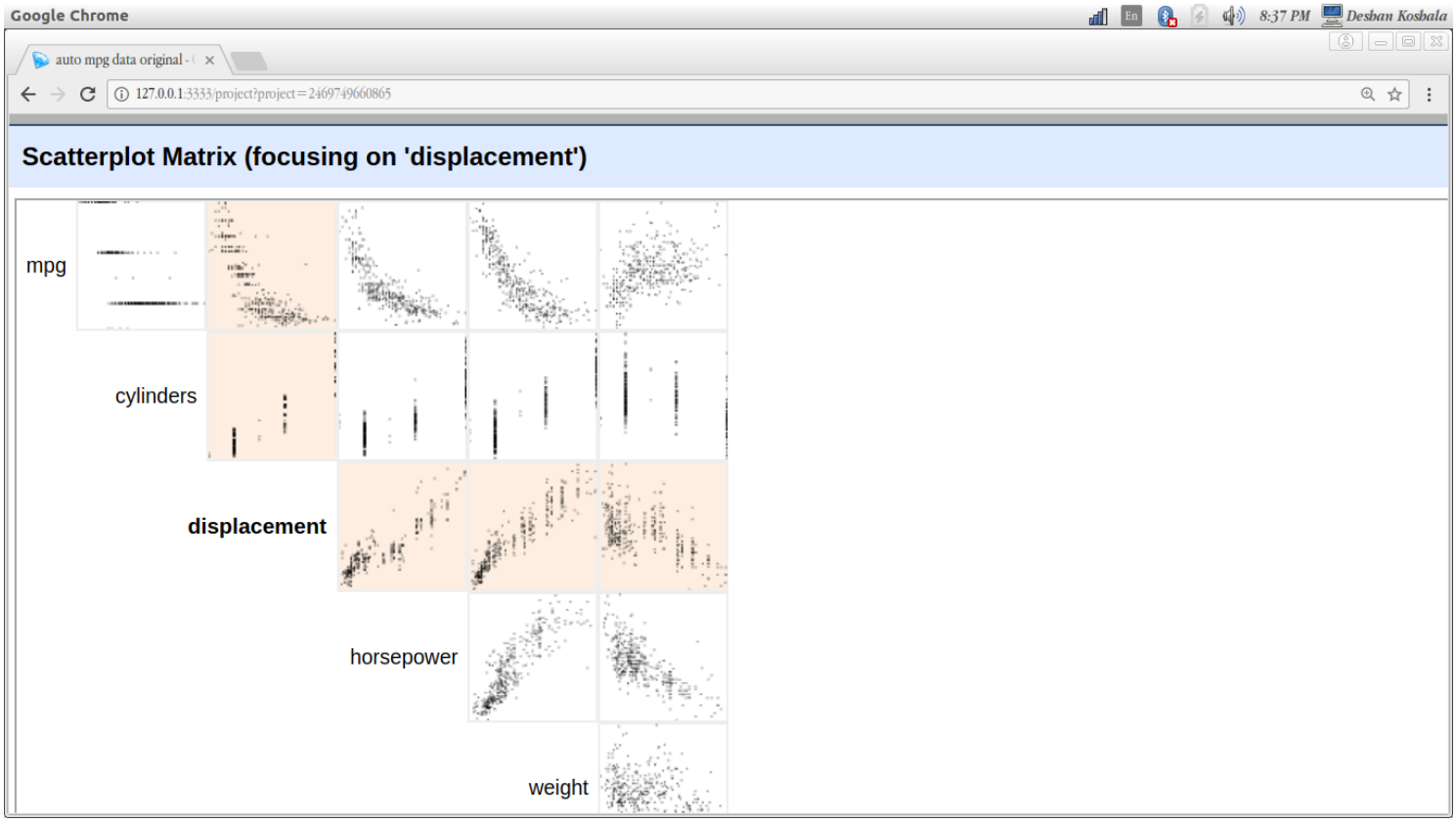
(In the dataset Removing column Duplicate value will be cause for a miss prediction because column)

8.Scatter plot

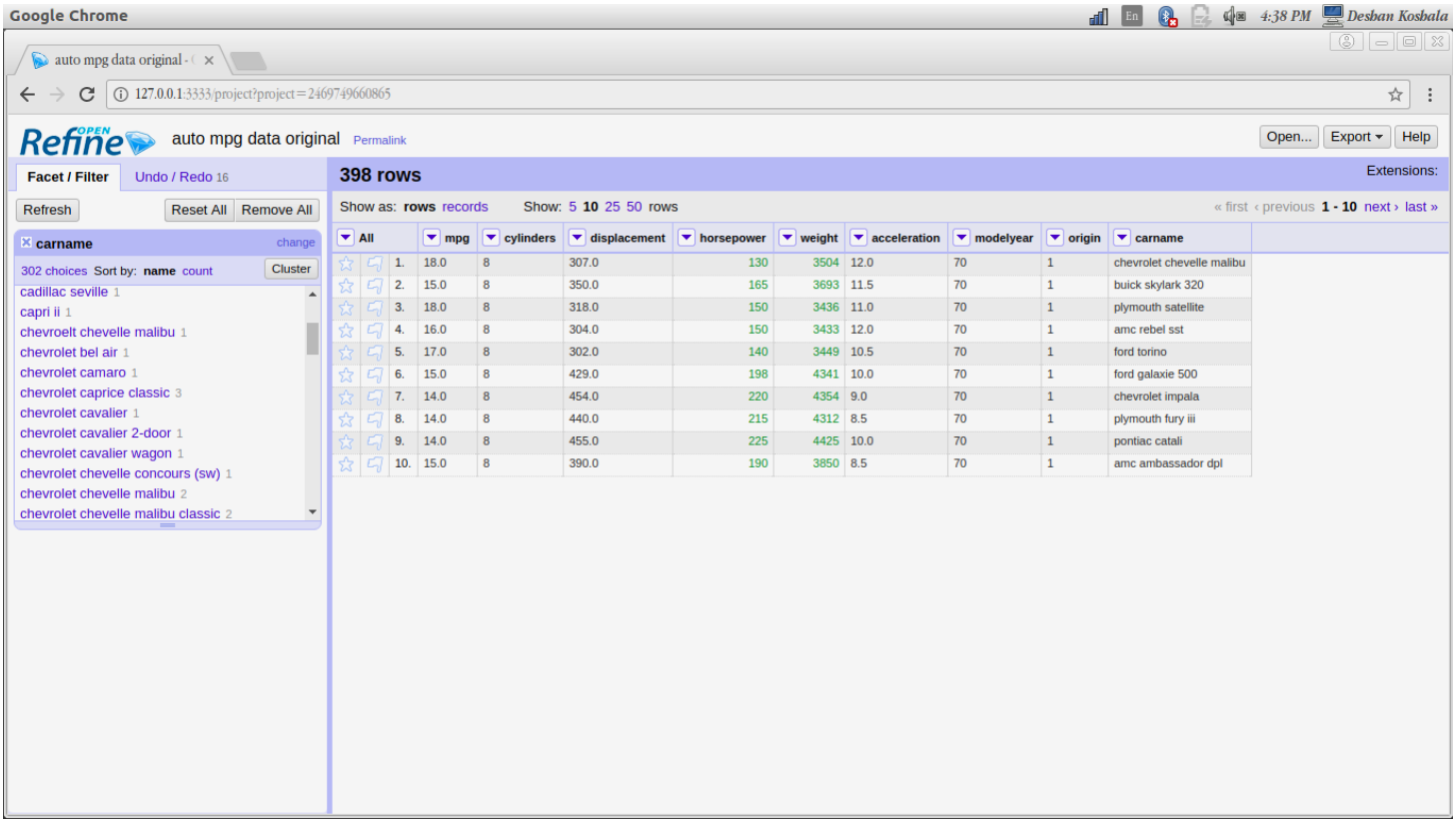





Google Chrome
8:38 PM
Desban Kosbala
auto mpg data original
127.0.0.1:3333/project?project=2469749660865
OpenRefine
auto mpg data original
Permalink
Open...
Export
Help
Facet / Filter
Extensions:
1 - 10 next › last »
Using facets
Scatterplot Matrix (focusing on 'acceleration')
mpg
cylinders
displacement
horsepower
weight
acceleration
lin
log
Close


Google Chrome
8:37 PM
Desban Kosbala
auto mpg data original
127.0.0.1:3333/project?project=2469749660865
Scatterplot Matrix (focusing on 'displacement')
mpg
cylinders
displacement
horsepower
weight

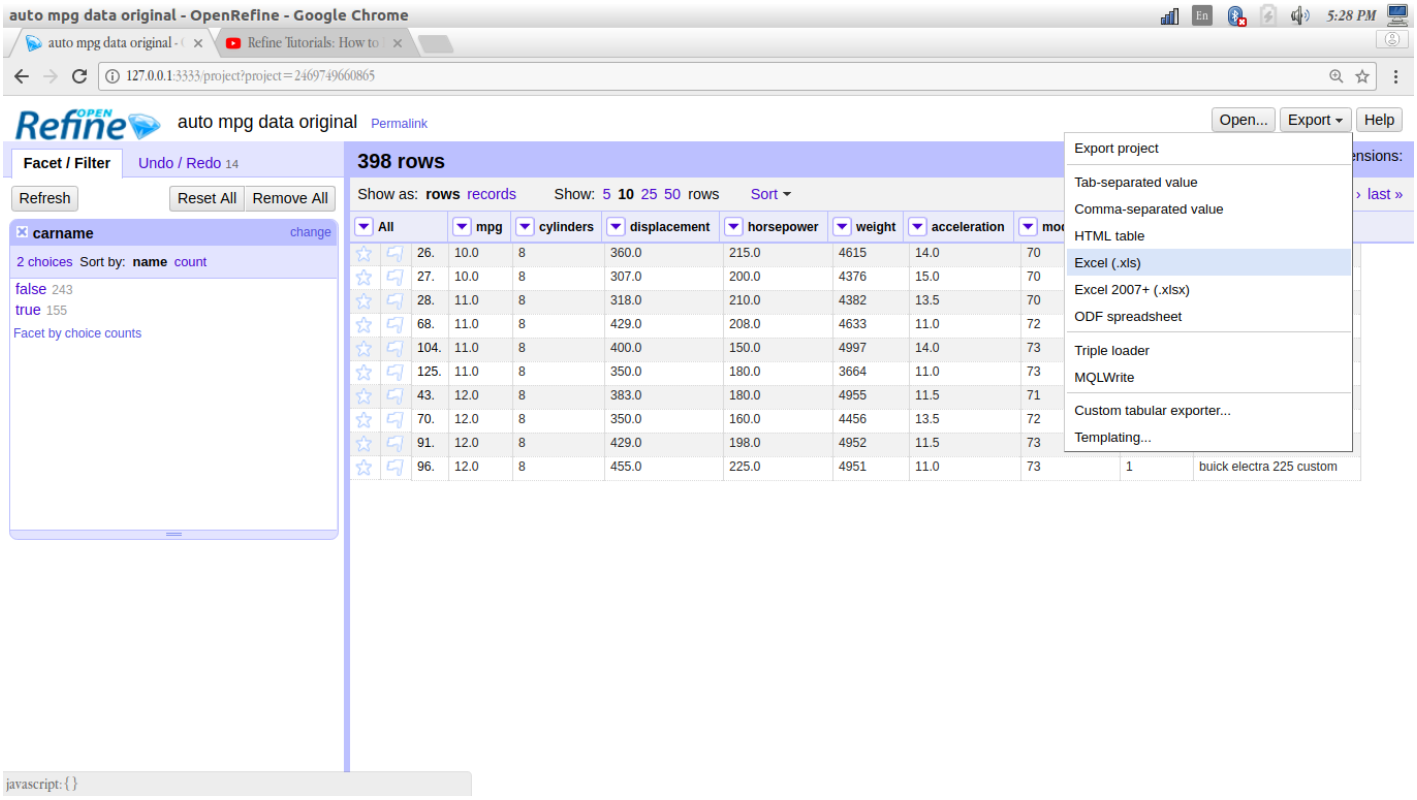
9.Removing duplicate Rows

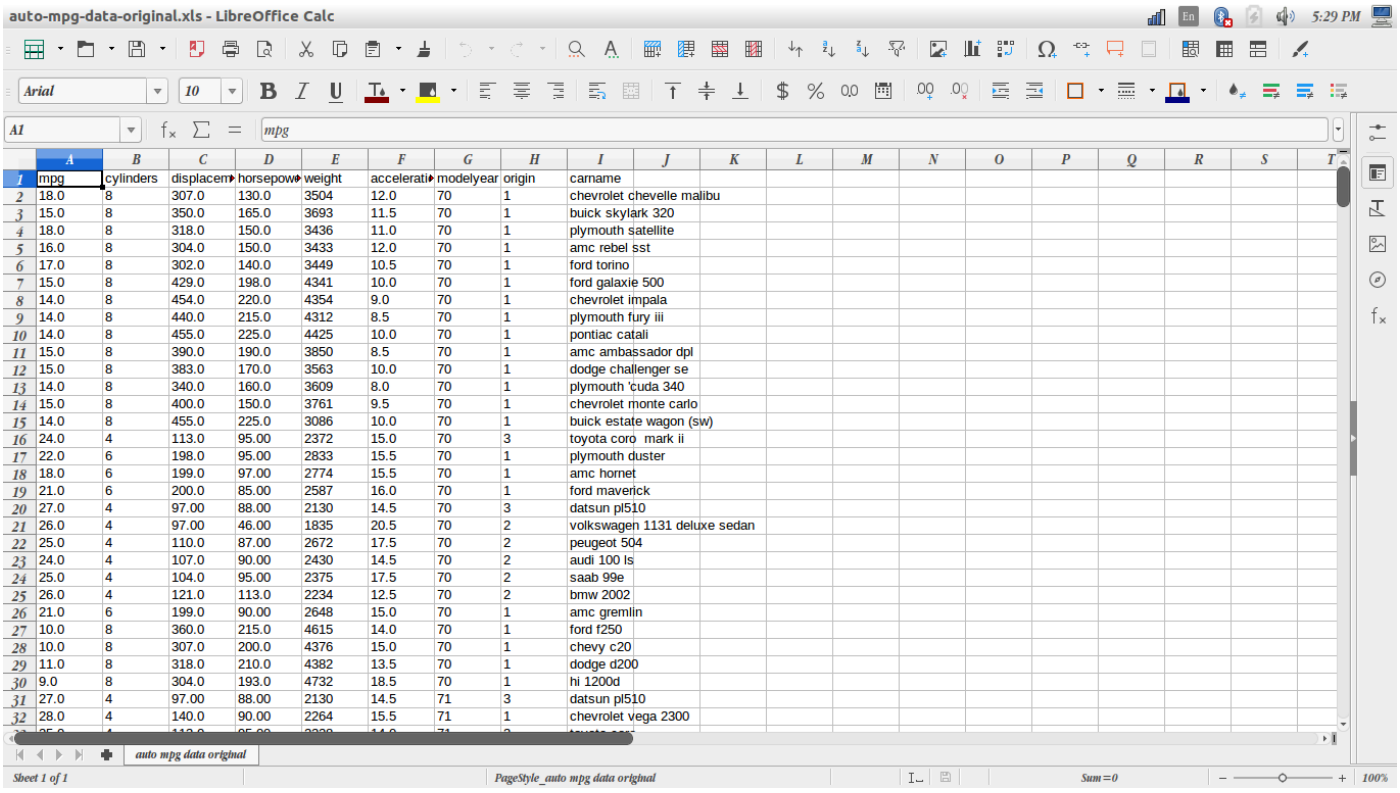


(
After clustering is over, we can select the cluster with higher no of rows, and search for duplicate values.
Flag those rows and remove .
)

No Duplicate is available in this data Set :(

10.Exporting cleaned data to Excel





	A	B	C	D	E	F	G	H	I
1	mpg	cylinders	displacem	horsepow	weight	acceleratio	modelyear	origin	carname
2	18.0	8	307.0	130.0	3504	12.0	70	1	chevrolet chevelle malibu
3	15.0	8	350.0	165.0	3693	11.5	70	1	buick skylark 320
4	18.0	8	318.0	150.0	3436	11.0	70	1	plymouth satellite
5	16.0	8	304.0	150.0	3433	12.0	70	1	amc rebel sst
6	17.0	8	302.0	140.0	3449	10.5	70	1	ford torino
7	15.0	8	429.0	198.0	4341	10.0	70	1	ford galaxie 500
8	14.0	8	454.0	220.0	4354	9.0	70	1	chevrolet impala
9	14.0	8	440.0	215.0	4312	8.5	70	1	plymouth fury iii
10	14.0	8	455.0	225.0	4425	10.0	70	1	pontiac catali
11	15.0	8	390.0	190.0	3850	8.5	70	1	amc ambassador dpl
12	15.0	8	383.0	170.0	3563	10.0	70	1	dodge challenger se
13	14.0	8	340.0	160.0	3609	8.0	70	1	plymouth 'cuda 340
14	15.0	8	400.0	150.0	3761	9.5	70	1	chevrolet monte carlo
15	14.0	8	455.0	225.0	3086	10.0	70	1	buick estate wagon (sw)
16	24.0	4	113.0	95.00	2372	15.0	70	3	toyota coro mark ii
17	22.0	6	198.0	95.00	2833	15.5	70	1	plymouth duster
18	18.0	6	199.0	97.00	2774	15.5	70	1	amc hornet
19	21.0	6	200.0	85.00	2587	16.0	70	1	ford maverick
20	27.0	4	97.00	88.00	2130	14.5	70	3	datsun pl510
21	26.0	4	97.00	46.00	1835	20.5	70	2	volkswagen 1131 deluxe sedan
22	25.0	4	110.0	87.00	2672	17.5	70	2	peugeot 504
23	24.0	4	107.0	90.00	2430	14.5	70	2	audi 100 ls
24	25.0	4	104.0	95.00	2375	17.5	70	2	saab 99e
25	26.0	4	121.0	113.0	2234	12.5	70	2	bmw 2002
26	21.0	6	199.0	90.00	2648	15.0	70	1	amc gremlin
27	10.0	8	360.0	215.0	4615	14.0	70	1	ford f250
28	10.0	8	307.0	200.0	4376	15.0	70	1	chevy c20
29	11.0	8	318.0	210.0	4382	13.5	70	1	dodge d200
30	9.0	8	304.0	193.0	4732	18.5	70	1	hi 1200d
31	27.0	4	97.00	88.00	2130	14.5	71	3	datsun pl510
32	28.0	4	140.0	90.00	2264	15.5	71	1	chevrolet vega 2300