

**तमसो मा ज्योतिर्गमय**

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

OpenRefine - Google Chrome

OpenRefine

127.0.0.1:3333

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

Create Project

Open Project

Import Project

Language Settings

Version 2.7 [TRUNK]

Help

About

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 added with OpenRefine extensions.

Get data from

This Computer

Web Addresses (URLs)

Clipboard

Google Data

Locate one or more files on your computer to upload:

Choose Files

auto-mpg.data-original

Next »

OpenRefine - Google Chrome

OpenRefine

127.0.0.1:3333

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

Create Project

Open Project

Import Project

Language Settings

Version 2.7 [TRUNK]

Help

About

« Start Over

Configure Parsing Options

Project name auto mpg data original

Create Project »

	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
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)

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

RDF/N3 files

XML files

Open Document Format spreadsheets (.ods)

Character encoding

Columns are separated by

☒ commas (CSV)

☐ tabs (TSV)

☐ custom ,

Escape special characters with \

☐ Ignore first 0 line(s) at beginning of file

☒ Parse next 1 line(s) as column headers

☐ Discard initial 0 row(s) of data

☐ Load at most 0 row(s) of data

☐ Parse cell text into numbers, dates, ...

☒ Quotation marks are used to enclose cells containing column separators

☒ Store blank rows

☒ Store blank cells as nulls

☐ Store file source (file names, URLs) in each row

Update Preview

Facet / Filter Undo / Redo 0

407 rows

Extensions:

Show as: rows records Show: 5 10 25 50 rows

« first &lt; previous 1 - 10 next &gt; 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	▼ 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

Facet / Filter Undo / Redo 0

407 rows

Extensions:

Show as: rows records Show: 5 10 25 50 rows

« first &lt; previous 1 - 10 next &gt; 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	▼ mpg	▼ cylinders	▼ displacement	▼ horsepower	▼ weight	▼ acceleration	▼ modelyear	▼ origin	▼ carname
☆ ↗ 1.	Facet ▸		307.0	130.0	3504	12.0	70	1	chevrolet chevelle malibu
☆ ↗ 2.	Text filter		350.0	165.0	3693	11.5	70	1	buick skylark 320
☆ ↗ 3.	Edit cells ▸		318.0	150.0	3436	11.0	70	1	plymouth satellite
☆ ↗ 4.	Edit column ▸		304.0	150.0	3433	12.0	70	1	amc rebel sst
☆ ↗ 5.	Transpose ▸		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.	Sort... ▸		454.0	220.0	4354	9.0	70	1	chevrolet impala
☆ ↗ 8.	View ▸		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.	Reconcile ▸		390.0	190.0	3850	8.5	70	1	amc ambassador dpl

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

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - ( x )

127.0.0.1:3333/project?project=2022532368492

Refine auto mpg data original Permalink Open... Export Help

Facet / Filter Undo / Redo 0

407 rows

Show as: rows records Show: 5 10 25 50 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](#)

Sort by mpg

Sort cell values as

☒ text ☐ case-sensitive

☐ numbers

☐ dates

☐ booleans

Position blanks and errors

Blanks

Valid values

Errors

Drag and drop to re-order

☒ a - z ☐ z - a

OK Cancel

		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					70	1	plymouth satellite
☆	4.	16.0	8					70	1	amc rebel sst
☆	5.	17.0	8					70	1	ford torino
☆	6.	15.0	8					70	1	ford galaxie 500
☆	7.	14.0	8					70	1	chevrolet impala
☆	8.	14.0	8					70	1	plymouth fury iii
☆	9.	14.0	8					70	1	pontiac catali
☆	10.	15.0	8					70	1	amc ambassador dpl

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - ( x )

127.0.0.1:3333/project?project=2022532368492

Refine auto mpg data original Permalink Flag row 368 Undo Open... Export Help

Facet / Filter Undo / Redo 8

407 rows

Show as: rows records Show: 5 10 25 50 rows Sort ▾ « 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](#)

		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

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

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - x

127.0.0.1:3333/project?project=2469749660865

Refine auto mpg data original Permalink

Open... Export Help

Facet / Filter Undo / Redo 9

Refresh Reset All Remove All

407 rows

Show as: rows records Show: 5 10 25 50 rows Sort

« first < previous 1 - 10 next > last »

Extensions:

carname change

309 choices Sort by: name count Cluster

chrysler new yorker brougham 1  
chrysler newport royal 1  
citroen ds-21 pallas 1  
datsun 1200 1  
datsun 200-sx 2  
datsun 210 2  
datsun 210 mpg 1  
datsun 280-zx 1  
datsun 310 1

	mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	origin	carname
407.									
11.	?	4	133.0	115.0	3090	17.5	70	2	
12.	?	8	350.0	165.0	4142	11.5	70	1	
13.	?	8	351.0	153.0	4034	11.0	70	1	
14.	?	8	383.0	175.0	4166	10.5	70		
15.	?	8	360.0	175.0	3850	11.0	70		
18.	?	8	302.0	140.0	3353	8.0	70		
40.	?	4	97.00	48.00	1978	20.0	71		
368.	?	4	121.0	110.0	2800	15.4	81		
32.	10.0	8	360.0	215.0	4615	14.0	70		

Facet

Text filter

Edit cells

Transform...

Trim leading and trailing whitespace

Collapse consecutive whitespace

Unescape HTML entities

To titlecase

To uppercase

To lowercase

To number

To date

To text

Blank out cells

Common transforms

Fill down

Blank down

Split multi-valued cells...

Join multi-valued cells...

Cluster and edit...

javascript: {}

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - x

127.0.0.1:3333/project?project=2022532368492

Refine auto mpg data original Permalink

Open... Export Help

Facet / Filter Undo / Redo 8

Refresh Reset All Remove All

407 rows

Show as: rows records Show: 5 10 25 50 rows Sort

« first < previous 1 - 10 next > last »

Extensions:

carname change

312 choices Sort by: name count Cluster

amc ambassador brougham 1  
amc ambassador dpl 1  
amc ambassador sst 1  
amc concord 2  
amc concord dl 1  
amc concord dl 1  
amc concord dl 6 1  
amc gremlin 4  
amc hornet 4  
amc hornet sportabout (sw) 1  
amc matador 5  
amc matador (sw) 2

	mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	origin	carname
368.	?	4	121.0	110.0	2800	15.4	81	2	
11.	?	4	133.0	115.0	3090	17.5	70	2	
40.	?	4	97.00	48.00	1978	20.0	71	2	
18.	?	8	302.0	140.0	3353	8.0	70	1	
13.	?	8	351.0	153.0	4034	11.0	70	1	
12.	?	8	350.0	165.0	4142	11.5	70	1	
15.	?	8	360.0	175.0	3850	11.0	70	1	
14.	?	8	383.0	175.0	4166	10.5	70	1	
33.	10.0	8	307.0	200.0	4376	15.0	70	1	
32.	10.0	8	360.0	215.0	4615	14.0	70	1	

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Text facet

Numeric facet

Timeline facet

Scatterplot facet

Custom text facet...

Custom Numeric Facet...

Customized facets

javascript: {}

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - 127.0.0.1:3333/project?project=2022532368492

Refine auto mpg data original Permalink Open... Export Help

Facet / Filter Undo / Redo 8

Refresh Reset All Remove All

407 rows

Show as: rows records Show: 5 10 25 50 rows Sort

« first < previous 1 - 10 next > last »

	mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	origin	carname
368. ?	4	121.0	110.0	2800	15.4	81	2	saab 900s	
11. ?	4	133.0	115.0	3090	17.5	70	2	citroen ds-21 pallas	
40. ?	4	97.00	48.00	1978	20.0	71	2	volkswagen super beetle 117	
18. ?	8	302.0	140.0	3353	8.0	70	1	ford mustang boss 302	
13. ?	8	351.0	153.0	4034	11.0	70	1	ford torino (sw)	
12. ?	8	350.0	165.0	4142	11.5	70	1	chevrolet chevelle concours (sw)	
15. ?	8	360.0	175.0	3850	11.0	70	1	amc rebel sst (sw)	
14. ?	8	383.0	175.0	4166	10.5	70	1	plymouth satellite (sw)	
33. 10.0	8	307.0	200.0	4376	15.0	70	1	chevy c20	
32. 10.0	8	360.0	215.0	4615	14.0	70	1	ford t250	

Facet: carname

312 choices Sort by: name count Cluster

amc ambassador brougham 1  
amc ambassador dpi 1  
amc ambassador sst 1  
amc concord 2  
amc concord d/i 1  
amc concord dl 1  
amc concord dl 6 1  
amc gremlin 4  
amc hornet 4  
amc hornet sportabout (sw) 1  
amc matador 5  
amc matador (sw) 2

Facet: carname

case sensitive regular expression

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

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - 127.0.0.1:3333/project?project=2469749660865

Refine auto mpg data original Permalink Open... Export Help

Facet / Filter Undo / Redo 8

Refresh Reset All Remove All

Cluster & Edit column "carname"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: key collision Keying Function: fingerprint 3 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none"> <li>amc concord d/i (1 rows)</li> <li>amc concord dl (1 rows)</li> </ul>	<input type="checkbox"/>	amc concord d/i
2	2	<ul style="list-style-type: none"> <li>datsum b-210 (1 rows)</li> <li>datsum b210 (1 rows)</li> </ul>	<input type="checkbox"/>	datsum b-210
2	2	<ul style="list-style-type: none"> <li>datsum 200-sx (1 rows)</li> <li>datsum 200sx (1 rows)</li> </ul>	<input type="checkbox"/>	datsum 200-sx

Average Length of Choices: 11.5 — 14.5

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

## 5) Extracting and cleaning values for dates

The screenshot shows the OpenRefine web interface in Google Chrome. The browser address bar shows the URL: 127.0.0.1:3333/project?project=2469749660865. The OpenRefine interface has a top bar with 'Refine' logo, 'auto mpg data original', and 'Permalink'. Below this is a 'Facet / Filter' section with 'Undo / Redo 10' and buttons for 'Refresh', 'Reset All', and 'Remove All'. The main table displays 407 rows. The 'carmake' column is selected, and the 'Edit cells' menu is open, showing options like 'Facet', 'Text filter', 'Edit cells', 'Transform...', 'Common transforms', 'Fill down', 'Blank down', 'Split multi-valued cells...', 'Join multi-valued cells...', 'Cluster and edit...', 'To titlecase', 'To uppercase', 'To lowercase', 'To number', 'To date', 'To text', and 'Blank out cells'. The 'To date' option is highlighted.

	mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	origin	carmake
11.	?	4	133.0	115.0	3090	17.5			citroen ds-21 pallas
12.	?	8	350.0	165.0	4142	11.5			chevrolet chevelle concours (sw)
13.	?	8	351.0	153.0	4034	11.0			ford torino (sw)
14.	?	8	383.0	175.0	4166	10.5			plymouth satellite (sw)
15.	?	8	360.0	175.0	3850	11.0			amc rebel sst (sw)
18.	?	8	302.0	140.0	3353	8.0			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

(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

The screenshot shows the OpenRefine web interface in Google Chrome. The browser address bar shows the URL: 127.0.0.1:3333/project?project=2469749660865. The OpenRefine interface has a top bar with 'Refine' logo, 'auto mpg data original', and 'Permalink'. Below this is a 'Facet / Filter' section with 'Undo / Redo 11' and buttons for 'Refresh', 'Reset All', and 'Remove All'. The main table displays 407 rows. The 'carmake' column is selected, and the 'Transform' menu is open, showing options like 'Transform', 'Facet', 'Edit rows', 'Edit columns', and 'View'. The 'Facet by star' option is highlighted.

	mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	origin	carmake
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

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - Introduction to OpenRefine

127.0.0.1:3333/project?project=2469749660865

### Refine

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 nomet sportabout (sw) 1
- amc matador 5
- amc matador (sw) 2
- amc pacer 1
- amc pacer d/i 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

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

auto mpg data original - Introduction to OpenRefine

127.0.0.1:3333/project?project=2469749660865

### Refine

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)

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



auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - Introduction to OpenRefine

127.0.0.1:3333/project?project=2469749660865

Refine auto mpg data original Permalink

Open... Export Help

Facet / Filter Undo / Redo 11

Refresh Reset All Remove All

**Facet by 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)

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
Transform	4	133.0	115.0	3090	17.5	70	2	citroen ds-21 pallas	
Facet	8	350.0	165.0	4142	11.5	70	1	chevrolet chevelle concours (sw)	
	8	351.0	153.0	4034	11.0	70	1	ford torino (sw)	
Edit rows	Star rows	175.0	4166	10.5	70	1	plymouth satellite (sw)		
Edit columns	Unstar rows	175.0	3850	11.0	70	1	amc rebel sst (sw)		
View	Flag rows	140.0	3353	8.0	70	1	ford mustang boss 302		
	Unflag rows	48.00	1978	20.0	71	2	volkswagen super beetle 117		
		110.0	2800	15.4	81	2	saab 900s		

368. ?

Remove all matching rows

javascript: {}

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

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - Introduction to OpenRefine

127.0.0.1:3333/project?project=2469749660865

Refine auto mpg data original Permalink

Open... Export Help

Facet / Filter Undo / Redo 14

Refresh Reset All Remove All

**Facet by carname** change

302 choices Sort by: name count Cluster

amc ambassador brougham 1  
amc ambassador dpl 1  
amc ambassador sst 1  
amc concord 2  
amc concord dl 2  
amc concord dl 6 1  
amc gremlin 4  
amc hornet 4  
amc hornet sportabout (sw) 1  
amc matador 5  
amc matador (sw) 2  
amc nacer 1

**Flagged Rows** change invert reset

1 choices Sort by: name count

false 398 exclude

Facet by choice counts

**Flagged Rows** change

1 choices Sort by: name count

false 398

Facet by choice counts

398 rows

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
26	10.0	8	360.0	215.0	4615	14.0	70	1	ford f250
27	10.0	8	307.0	200.0	4376	15.0	70	1	chevy c20
28	11.0	8	318.0	210.0	4382	13.5	70	1	dodge d200
68	11.0	8	429.0	208.0	4633	11.0	72	1	mercury marquis
104	11.0	8	400.0	150.0	4997	14.0	73	1	chevrolet impala
125	11.0	8	350.0	180.0	3664	11.0	73	1	oldsmobile omega
43	12.0	8	383.0	180.0	4955	11.5	71	1	dodge mo co (sw)
70	12.0	8	350.0	160.0	4456	13.5	72	1	oldsmobile delta 88 royale
91	12.0	8	429.0	198.0	4952	11.5	73	1	mercury marquis brougham
96	12.0	8	455.0	225.0	4951	11.0	73	1	buick electra 225 custom

## 7) Removing duplicate columns(in some cases)

The screenshot shows the OpenRefine web interface in Google Chrome. The browser tab is 'auto mpg data original - OpenRefine'. The address bar shows the URL '127.0.0.1:3333/project?project=2469749660865'. The interface has a top bar with 'Refine' logo, 'auto mpg data original', and 'Permalink'. Below this is a 'Facet / Filter' tab and 'Undo / Redo 14'. The main area shows '398 rows' and a table with columns: 'All', 'mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'modelyear', 'origin', and 'carmake'. A context menu is open over the 'carmake' column header, showing options like 'Facet', 'Text facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'Word facet', 'Duplicates facet', 'Numeric log facet', '1-bounded numeric log facet', 'Text length facet', 'Log of text length facet', 'Unicode char-code facet', 'Facet by error', and 'Facet by blank'. The 'Facet' option is selected, and a sub-menu is visible on the right showing 'Text facet', 'Numeric facet', 'Timeline facet', 'Scatterplot facet', 'Custom text facet...', 'Custom Numeric Facet...', and 'Customized facets'.

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - x Refine Tutorials: How to x

127.0.0.1:3333/project?project=2469749660865

Refine auto mpg data original Permalink

Open... Export Help

Facet / Filter Undo / Redo 14

398 rows

Show as: rows records Show: 5 10 25 50 rows Sort

« 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

Facet

Text facet

Numeric facet

Timeline facet

Scatterplot facet

Custom text facet...

Custom Numeric Facet...

Customized facets

Word facet

Duplicates facet

Numeric log facet

1-bounded numeric log facet

Text length facet

Log of text length facet

Unicode char-code facet

Facet by error

Facet by blank

javascript: {}

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - x Refine Tutorials: How to x

127.0.0.1:3333/project?project=2469749660865

Refine auto mpg data original Permalink

Open... Export Help

Facet / Filter Undo / Redo 14

398 rows

Show as: rows records Show: 5 10 25 50 rows Sort

« first < previous 1 - 10 next > last »

Refresh Reset All Remove All

carname change

2 choices Sort by: name count

false 243

true 155

Facet by choice counts

All mpg cylinders displacement horsepower weight

☆	26.	10.0	8	360.0	215.0	4615
☆	27.	10.0	8	307.0	200.0	4376
☆	28.	11.0	8	318.0	210.0	4382
☆	68.	11.0	8	429.0	208.0	4633
☆	104.	11.0	8	400.0	150.0	4997
☆	125.	11.0	8	350.0	180.0	3664
☆	43.	12.0	8	383.0	180.0	4955
☆	70.	12.0	8	350.0	160.0	4456
☆	91.	12.0	8	429.0	198.0	4952
☆	96.	12.0	8	455.0	225.0	4951

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

## 8. Scatter plot

Google Chrome

auto mpg data original - x

127.0.0.1:3333/project?project=2469749660865

Refine OPEN auto mpg data original Permalink

Facet / Filter Undo / Redo 16

398 rows

Show as: rows records Show: 5 10 25 50 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](#)

Facet / Filter

mpg cylinders displacement horsepower weight acceleration modelyear origin carname

1. 18.0 8 130 3504 12.0 70 1 chevrolet chevelle malibu

2. 15.0 8 165 3693 11.5 70 1 buick skylark 320

3. 18.0 8 150 2496 11.0 70 1 plymouth satellite

4. 16.0 8 150 2496 11.0 70 1 amc rebel sst

5. 17.0 8 150 2496 11.0 70 1 amc rebel sst

6. 15.0 8 150 2496 11.0 70 1 amc rebel sst

7. 14.0 8 150 2496 11.0 70 1 amc rebel sst

8. 14.0 8 150 2496 11.0 70 1 amc rebel sst

9. 14.0 8 150 2496 11.0 70 1 amc rebel sst

10. 15.0 8 150 2496 11.0 70 1 amc rebel sst

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Transform...

Common transforms

Trim leading and trailing whitespace

Collapse consecutive whitespace

Unescape HTML entities

To titlecase

To uppercase

To lowercase

To number

To date

To text

Blank out cells

javascript: {}

Google Chrome

auto mpg data original - x

127.0.0.1:3333/project?project=2469749660865

Refine OPEN auto mpg data original Permalink

Facet / Filter Undo / Redo 20

398 rows

Show as: rows records Show: 5 10 25 50 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](#)

Facet / Filter

mpg cylinders displacement horsepower weight acceleration modelyear origin carname

1. 18 8 12 70 1 chevrolet chevelle malibu

2. 15 8 11.5 70 1 buick skylark 320

3. 18 8 11 70 1 plymouth satellite

4. 16 8 12 70 1 amc rebel sst

5. 17 8 10.5 70 1 ford torino

6. 15 8 10 70 1 ford galaxie 500

7. 14 8 9 70 1 chevrolet impala

8. 14 8 8.5 70 1 plymouth fury iii

9. 14 8 10 70 1 pontiac catali

10. 15 8 8.5 70 1 amc ambassador dpl

Facet

Text facet

Numeric facet

Timeline facet

Scatterplot facet

Custom text facet...

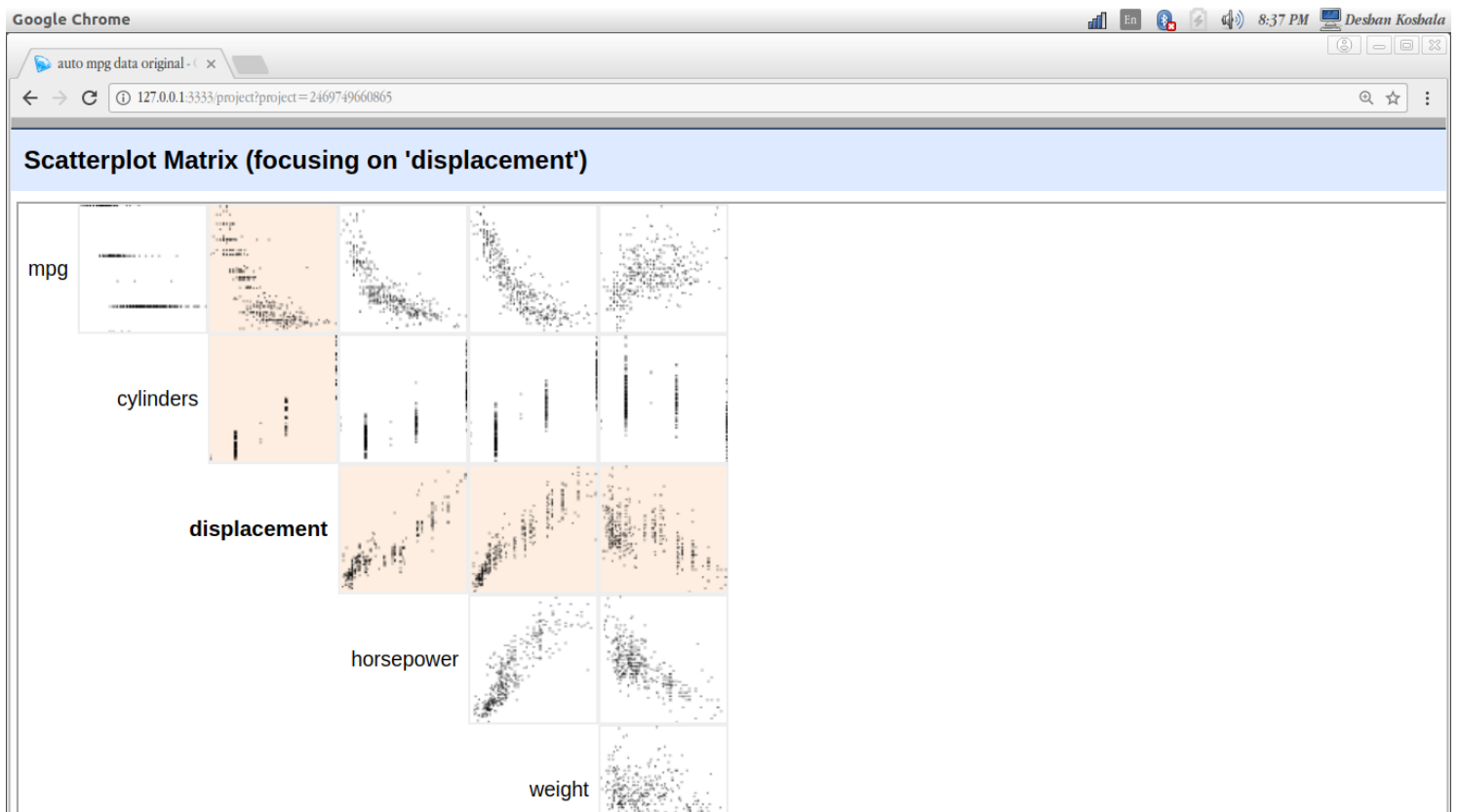
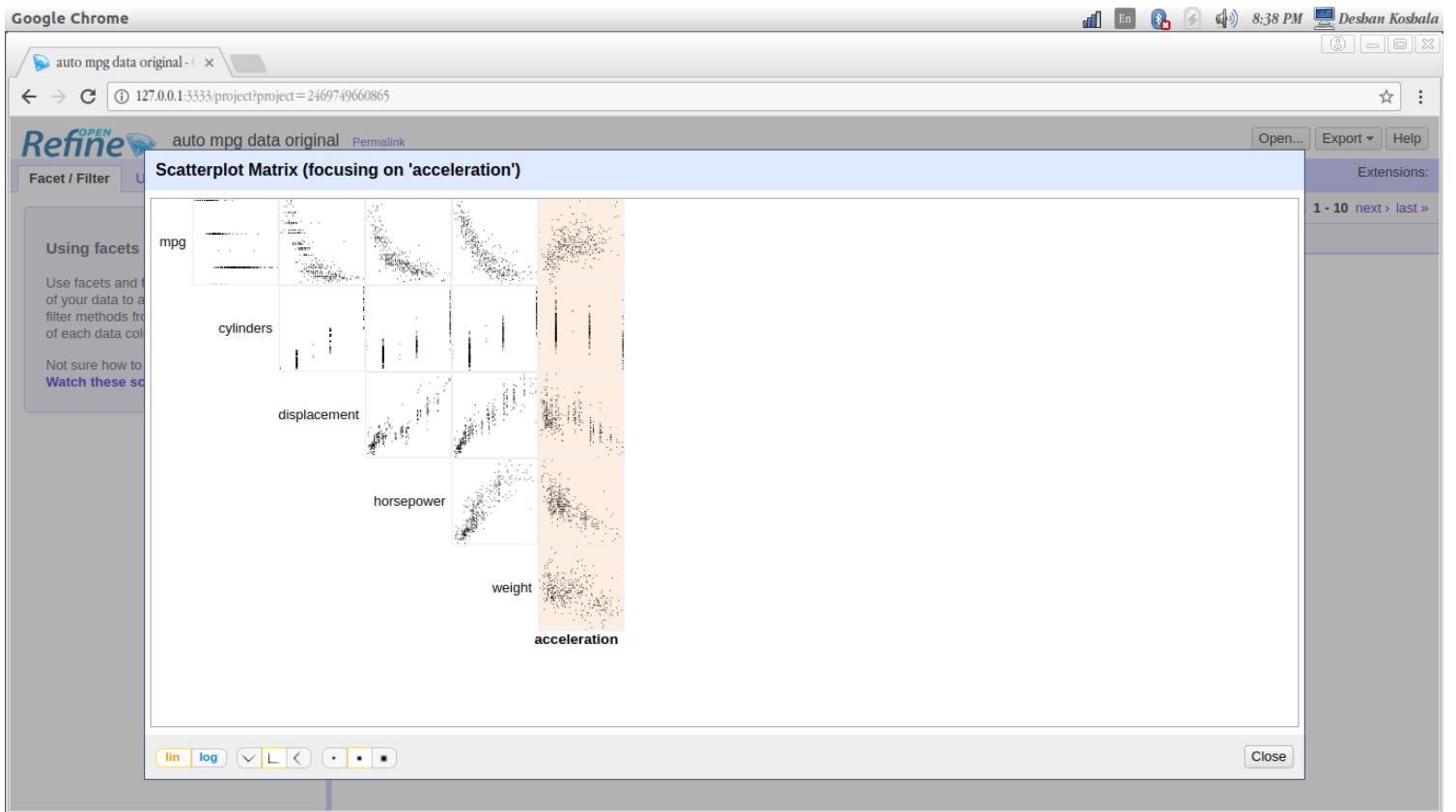
Custom Numeric Facet...

Customized facets

225 4425

190 3850

javascript: {}



## 9.Removing duplicate Rows

Google Chrome

auto mpg data original - x

127.0.0.1:3333/project?project=2469749660865

Refine OPEN auto mpg data original Permalink

Open... Export Help

Facet / Filter Undo / Redo 16

Refresh Reset All Remove All

398 rows

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

Extensions:

carname change

302 choices Sort by: name count Cluster

cadillac seville 1

capri ii 1

chevrolet chevelle malibu 1

chevrolet bel air 1

chevrolet camaro 1

chevrolet caprice classic 3

chevrolet cavalier 1

chevrolet cavalier 2-door 1

chevrolet cavalier wagon 1

chevrolet chevelle concours (sw) 1

chevrolet chevelle malibu 2

chevrolet chevelle malibu classic 2

		mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	origin	carname
1.	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu	
2.	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320	
3.	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite	
4.	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst	
5.	17.0	8	302.0	140	3449	10.5	70	1	ford torino	
6.	15.0	8	429.0	198	4341	10.0	70	1	ford galaxie 500	
7.	14.0	8	454.0	220	4354	9.0	70	1	chevrolet impala	
8.	14.0	8	440.0	215	4312	8.5	70	1	plymouth fury iii	
9.	14.0	8	455.0	225	4425	10.0	70	1	pontiac catali	
10.	15.0	8	390.0	190	3850	8.5	70	1	amc ambassador dpl	

(  
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

auto mpg data original - OpenRefine - Google Chrome

auto mpg data original - Refine Tutorials: How to

127.0.0.1:3333/project?project=2469749660865

Refine auto mpg data original Permalink

Facet / Filter Undo / Redo 14

Refresh Reset All Remove All

2 choices Sort by: name count

false 243

true 155

Facet by choice counts

398 rows

Show as: rows records Show: 5 10 25 50 rows Sort

Export project

- Tab-separated value
- Comma-separated value
- HTML table
- Excel (.xls)
- Excel 2007+ (.xlsx)
- ODF spreadsheet
- Triple loader
- MQLWrite
- Custom tabular exporter...
- Templating...

	mpg	cylinders	displacement	horsepower	weight	acceleration	model
26.	10.0	8	360.0	215.0	4615	14.0	70
27.	10.0	8	307.0	200.0	4376	15.0	70
28.	11.0	8	318.0	210.0	4382	13.5	70
68.	11.0	8	429.0	208.0	4633	11.0	72
104.	11.0	8	400.0	150.0	4997	14.0	73
125.	11.0	8	350.0	180.0	3664	11.0	73
43.	12.0	8	383.0	180.0	4955	11.5	71
70.	12.0	8	350.0	160.0	4456	13.5	72
91.	12.0	8	429.0	198.0	4952	11.5	73
96.	12.0	8	455.0	225.0	4951	11.0	73

1 buick electra 225 custom

javascript: {}

auto-mpg-data-original.xls - LibreOffice Calc

Arial 10 B I U T, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T

AI f x = mpg

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	mpg	cylinders	displacement	horsepower	weight	acceleration	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											

auto mpg data original

Sheet 1 of 1 PageStyle\_auto mpg data original Sum=0 100%



