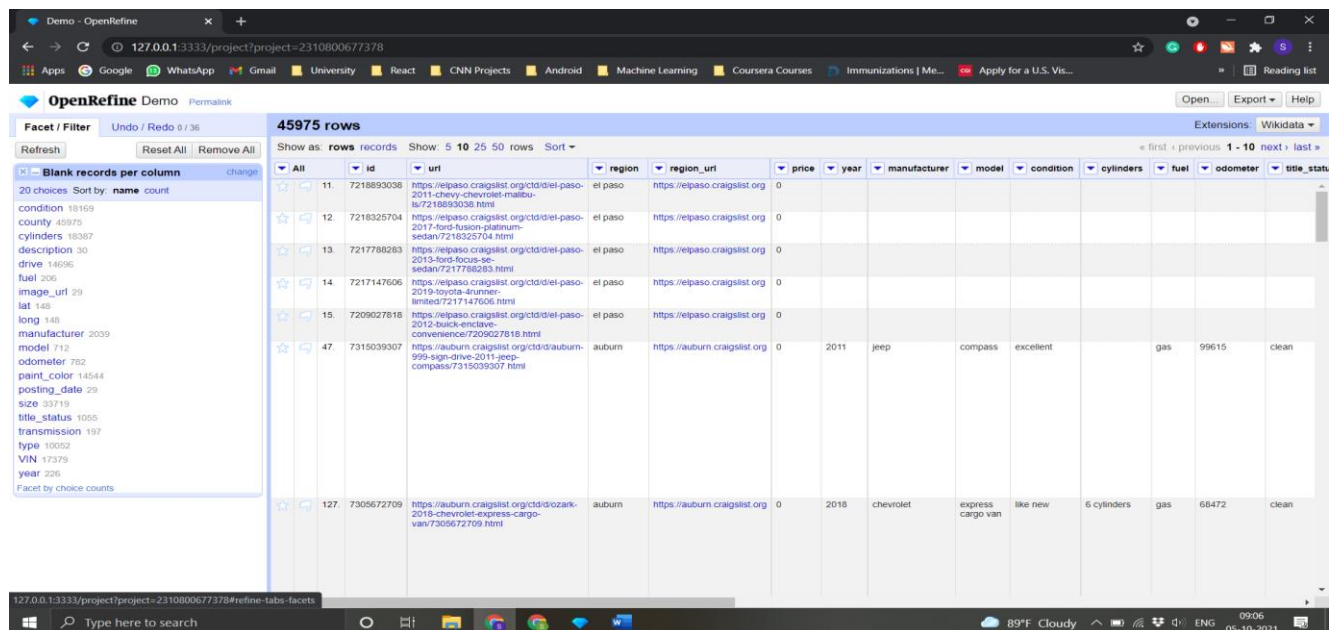


Data Cleaning with Google Open Refine Tool:

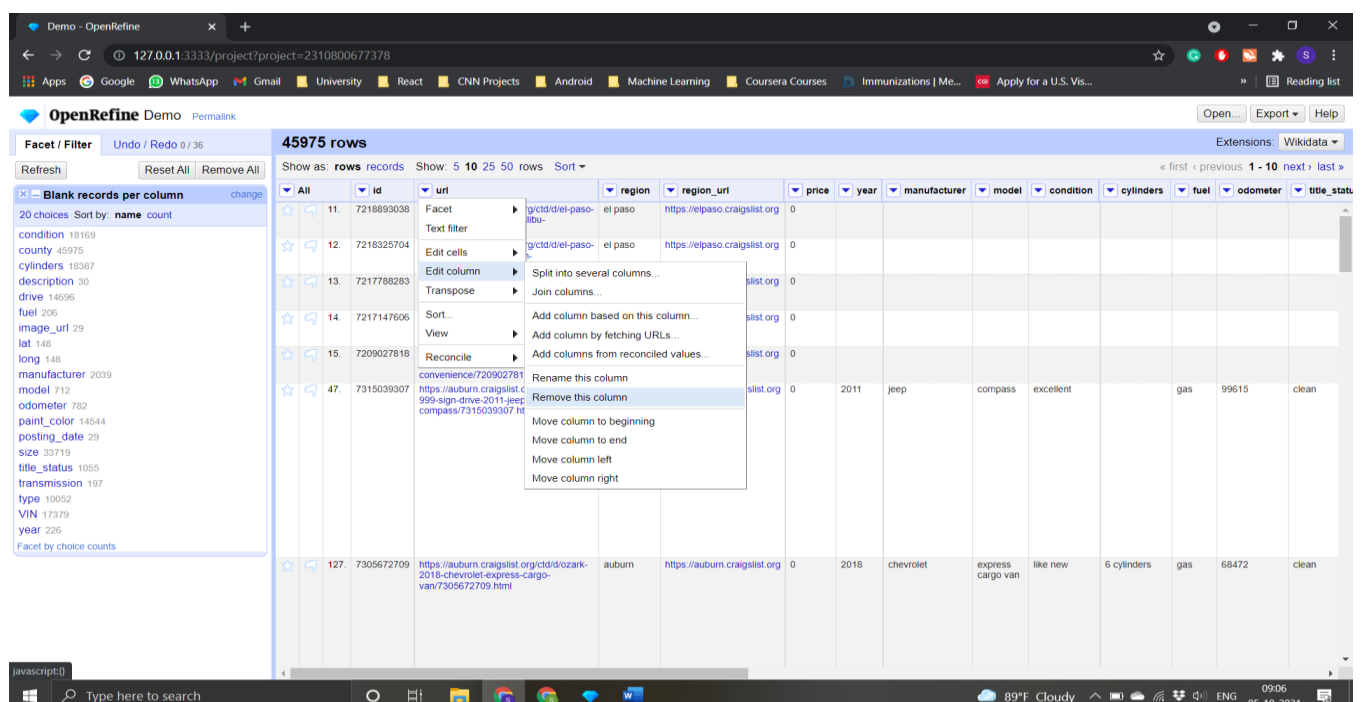
Because of the memory, I have just uploaded a sample of the original dataset in openrefine.

After uploading the sample of the dataset, in the below screenshot you can see all the null values present in each column.



The screenshot shows the OpenRefine interface with a dataset of 45975 rows. The 'url' column is highlighted, and a context menu is open over it, showing options like 'Facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', 'Rename this column', 'Remove this column', 'Move column to beginning', 'Move column to end', 'Move column left', and 'Move column right'.

url column is not useful to predict the price of the car, So I just removed that column. Below, I have attached the screenshot of how to remove the column in the openrefine.



The screenshot shows the OpenRefine interface with the 'url' column being removed. The context menu is open over the 'url' column, and the 'Remove this column' option is selected.

Similarly, size, image_url, description, county, posting_date, region_url, id, VIN which are also not useful in the predicting the car price, So I just removed these columns from the dataset.

OpenRefine Demo interface showing a dataset of 45975 rows. The interface includes a sidebar with a list of actions (0-9) and a main table view. The table has columns: region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, state. The 'cylinders' column is highlighted, and a context menu is open over it, showing options like 'Facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', 'Rename this column', 'Remove this column', 'Move column to beginning', 'Move column to end', 'Move column left', and 'Move column right'.

Moreover, I've transformed the column Cylinders and put just number e.g. from "6 Cylinders" to "6" only. So that, we can use those numerical values in predicting the car price.

Below, I have attached the screenshot of that process.

OpenRefine Demo interface showing a dataset of 45827 rows. The interface includes a sidebar with a list of actions (0-9) and a main table view. The table has columns: region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, state. The 'cylinders' column is highlighted, and a context menu is open over it, showing options like 'Facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', 'Rename this column', 'Remove this column', 'Move column to beginning', 'Move column to end', 'Move column left', and 'Move column right'.

Here you can see, two new columns were made, Cylinders 1 and Cylinders 2.

The screenshot shows the OpenRefine Demo interface. At the top, a notification bar states: "Split 27523 cell(s) in column cylinders into several columns by separator Undo". The main table displays 45827 rows. The columns include: All, region, price, year, manufacturer, model, condition, cylinders, cylinders 1, cylinders 2, fuel, odometer, title_status, transmission, and drive. The 'cylinders' column is highlighted, and the 'cylinders 1' and 'cylinders 2' columns are visible. The left sidebar shows a facet filter for 'Blank records per column' with a count of 12 choices. The bottom status bar shows the system time as 09:15 on 05-10-2021.

We don't need a column which contain only "Cylinder" Value, So I just removed that column.

The screenshot shows the OpenRefine Demo interface with the 'cylinders' column selected for removal. A context menu is open over the 'cylinders' column header, showing options: Facet, Text filter, Edit cells, Edit column, Transpose, Sort..., View, Reconcile, Rename this column, Remove this column, Move column to beginning, Move column to end, Move column left, and Move column right. The 'Remove this column' option is highlighted. The main table displays 45827 rows. The left sidebar shows a facet filter for 'Blank records per column' with a count of 12 choices. The bottom status bar shows the system time as 09:15 on 05-10-2021.

After that, Click the dropdown menu for fuel column and select text Facet which helps us to show the unique values in that column.

The screenshot shows the OpenRefine web interface. The top bar indicates 45827 rows. The left sidebar shows a list of columns with their respective choice counts. The main table displays columns: All, region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, and state. A dropdown menu is open for the 'fuel' column, showing options: Facet, Text filter, Edit cells, Edit column, Transpose, Sort..., View, and Reconcile. The 'Facet' option is selected, and a sub-menu is visible with options: Text facet, Numeric facet, Timeline facet, Scatterplot facet, Custom text facet..., Custom Numeric Facet..., Customized facets, clean, automatic, 4wd, truck, silver, al.

After clicking the Text facet on fuel, we can get the result like below. As per the screenshot, diesel, electric, gas, hybrid, other and blank values were present in fuel column.

After clicking on blank, openrefine only shows the record in which fuel is blank. So, we need to remove these rows.

The screenshot shows the OpenRefine web interface after filtering by fuel = blank. The top bar indicates 177 matching rows (45827 total). The left sidebar shows the 'fuel' column with 5 choices: diesel (3119), electric (273), gas (37567), hybrid (770), and other (3921). The main table displays columns: All, region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, and state. The 'fuel' column is filtered to show only blank values.

So, to remove those blank fuel rows, I have converted those blank values to the one which are most frequent fuel choice (Mode value) which is **gas** in our case.

The screenshot shows the OpenRefine web interface. On the left, the 'Facet / Filter' panel shows a facet for 'fuel' with 5 choices: diesel (3119), electric (273), gas (37567), hybrid (770), and other (3921). The '(blank)' category has 177 records. The main table displays 177 matching rows for the 'fuel' facet. A context menu is open over the 'fuel' column, showing options like 'Transform...', 'Common transforms', 'Fill down', 'Blank down', 'Split multi-valued cells...', 'Join multi-valued cells...', 'Cluster and edit...', and 'Replace'.

row	region	price	year	manufacturer	model	condition	cylinders	fuel	odometer	title_status	transmission	drive	type	paint_color	state
507	birmingham	0	2012	chevrolet	silverado 2500hd		8 cylinders			clean	automatic	nwd		white	al
9384	mohave county	0	2017	chevrolet	silverado 1500		8 cylinders				automatic	nwd	pickup	red	az
12249	phoenix	0			Revero GT							nwd		black	az
1160	birmingham	1950	1999	dodge	caravan		6 cylinders					nwd	van		al
338	birmingham	3350	2010	chevrolet	hhr		4 cylinders					nwd	SUV	red	al
626	birmingham	3450	2007	toyota	prius		4 cylinders					nwd		grey	al
976	birmingham	3450	2007	toyota	prius		4 cylinders					nwd		grey	al
1702	birmingham	3450	2003	ford	explorer		6 cylinders					nwd	SUV	red	al
1423	birmingham	3650	2008	chevrolet	impala		6 cylinders					nwd	sedan	grey	al
1754	birmingham	4980	2004	gmc	yukon xl		8 cylinders					nwd	SUV	silver	al

The screenshot shows the OpenRefine web interface with a 'Custom text transform on column fuel' dialog box open. The 'Expression' field contains the code `return 'gas'`. The 'Language' is set to 'Python / Jython'. A 'Preview' tab shows the result of the transform, where all 'null' values in the 'fuel' column are replaced with 'gas'. The 'On error' options are 'keep original' (selected), 'set to blank', and 'store error'. The 'Re-transform' checkbox is unchecked, and the 'times until no change' is set to 10.

row	value	return 'gas'
507	null	gas
9384	null	gas
12249	null	gas
1160	null	gas
338	null	gas
626	null	gas
...

After clicking the 'ok' button, openrefine fill all the blank values of the fuel column with the value **gas**.

The screenshot shows the OpenRefine interface with a text transform operation applied to the 'fuel' column. The operation is 'jython:return \'gas\''. The facet for 'fuel' shows 5 choices: diesel (3119), electric (273), gas (37744), hybrid (770), and (blank) (0). The 'fuel' column is selected in the main table view.

Similarly, I also replaced the blank values of transmission column with the value “automatic”.

The screenshot shows the OpenRefine interface with a text transform operation applied to the 'transmission' column. The operation is 'jython:return \'automatic\''. The facet for 'transmission' shows 3 choices: automatic (36078), manual (2903), and other (6678). The 'transmission' column is selected in the main table view. The main table view shows 45827 rows with columns: region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, and state.

The screenshot shows the OpenRefine web interface. A dialog box titled "Custom text transform on column transmission" is open. The "Expression" field contains the code `return 'automatic'`. The "Language" dropdown is set to "Python / Jython". A "No syntax error" message is displayed. Below the expression field, there is a "Preview" tab showing a table with columns "row", "value", and "return 'automatic'". The "value" column contains null values for rows 1660, 3477, 6217, 14495, 15995, and 21096. The "return 'automatic'" column shows the result of the transformation. At the bottom of the dialog, there are options for "On error": "keep original" (selected), "set to blank", and "store error". There is also a checkbox for "Re-transform up to 10 times until no change".

As you can see there are no blank records available for transmission column.

The screenshot shows the main OpenRefine interface. The top bar indicates "45827 rows". The left sidebar shows the "Facet / Filter" panel with various facets. The "transmission" facet is selected, showing 3 choices: "automatic" (36246), "manual" (2903), and "other" (6678). The main table displays 10 rows of data. The columns are: "All", "region", "price", "year", "manufacturer", "model", "condition", "cylinders", "fuel", "odometer", "title_status", "transmission", "drive", "type", "paint_color", and "state".

	region	price	year	manufacturer	model	condition	cylinders	fuel	odometer	title_status	transmission	drive	type	paint_color	state
20.	auburn	0	2011	jeep	compass	excellent	6 cylinders	gas	99615	clean	automatic	rwd	SUV	white	al
100.	auburn	0	2018	chevrolet	express cargo van	like new	6 cylinders	gas	68472	clean	automatic	rwd	van	white	al
101.	auburn	0	2019	chevrolet	express cargo van	like new	6 cylinders	gas	69125	clean	automatic	rwd	van	white	al
102.	auburn	0	2018	chevrolet	express cargo van	like new	6 cylinders	gas	66555	clean	automatic	rwd	van	white	al
165.	birmingham	0	2015	nissan	sentra	excellent	4 cylinders	gas	99505	clean	automatic	fwd	sedan	silver	al
213.	birmingham	0	2019	chevrolet	silverado 1500	excellent	8 cylinders	gas	25127	clean	automatic	4wd	truck	red	al
216.	birmingham	0	2014	Freightliner	Cascadia		100	diesel	100	clean	automatic				al
218.	birmingham	0	2017	chevrolet	silverado 2500hd 4x4	like new	8 cylinders	diesel	102000	clean	automatic	4wd	truck	silver	al
332.	birmingham	0	2017	chevrolet	silverado 2500 271 4x4	like new	8 cylinders	diesel	102000	clean	automatic	4wd	truck	silver	al
333.	birmingham	0	2016	ram	3500 laramie 4x4	like new	6 cylinders	diesel	120000	clean	automatic	4wd	truck	black	al

Moreover, for latitude column there were also null values present, but we can't recover null latitude values from any other column. So, I just dropped those null values. Below I have attached the screenshot of doing that.

OpenRefine Demo interface showing 148 matching rows (45975 total). The 'lat' facet is selected, showing a list of latitude values. The main table displays columns: region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, state, lat, and long. The 'lat' column shows values like 'tx', 'ai', 'al', 'ga', 'nc', 'sc', 'va', 'wv', 'md', 'de', 'pa', 'ny', 'nj', 'ct', 'ri', 'ma', 'nh', 'vt', 'me', 'ny', 'nj', 'ct', 'ri', 'ma', 'nh', 'vt', 'me'.

OpenRefine Demo interface showing 148 matching rows (45975 total). The 'lat' facet is selected, showing a list of latitude values. The main table displays columns: region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, state, lat, and long. The 'lat' column shows values like 'tx', 'ai', 'al', 'ga', 'nc', 'sc', 'va', 'wv', 'md', 'de', 'pa', 'ny', 'nj', 'ct', 'ri', 'ma', 'nh', 'vt', 'me', 'ny', 'nj', 'ct', 'ri', 'ma', 'nh', 'vt', 'me'. A context menu is open over the 'lat' column, showing options: Transform, Facet, Edit rows, Edit columns, View, Flag rows, Unflag rows, and Remove matching rows.

The screenshot shows the OpenRefine web interface. The top bar indicates '148 matching rows (45975 total)'. The left sidebar shows a facet for 'lat' with 6118 choices. The main table displays columns: region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, and state. A context menu is open over the table, with 'Remove matching rows' selected. The bottom status bar shows '60°F Mostly cloudy' and the date '05-10-2021'.

As you can see from the below screenshot, there are no null values present in the latitude column.

The screenshot shows the OpenRefine web interface after removing the 148 rows. The top bar indicates '0 matching rows (45827 total)'. The left sidebar shows the 'lat' facet with 6118 choices. The main table is empty. The bottom status bar shows '60°F Mostly cloudy' and the date '05-10-2021'.

To handle the missing values for title_status, paint_colour, model, odometer, drive, type, year, manufacturer, cylinders columns, I have used **fill down** method.

The screenshot shows the OpenRefine web application interface. The top bar indicates '1026 matching rows (45827 total)'. The left sidebar shows a facet for 'Blank records per column' with 10 choices. The main table displays columns: All, region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, paint_color, and st. A context menu is open over the 'title_status' column, showing options like 'Facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', 'Cluster and edit...', and 'Replace'. The 'Fill down' option is highlighted under 'Common transforms'.

After doing that, as you can see from the below screenshot there are no null values present in each column.

The screenshot shows the OpenRefine web application interface after data cleaning. The top bar indicates '39835 rows'. The left sidebar shows the 'Blank records per column' facet with 0 choices. The main table displays the same columns as the previous screenshot, but now all values are present, indicating successful data cleaning.

References:

- <https://openrefine.org/>
- https://www.youtube.com/watch?v=-JW67U_rK1M&t=135s&ab_channel=susanemcg
- https://www.youtube.com/watch?v=jyUIT8ohlG4&ab_channel=BiodiversityDataScience
- https://www.youtube.com/watch?v=nORS7STbLyk&t=237s&ab_channel=WebScraper