Cleaning Data with Refine

What you’ll need:

1. Refine - Download it from openrefine.org
2. The sample Dataset - Download it from Africa Open Data

Step 1: Creating a new Project¶

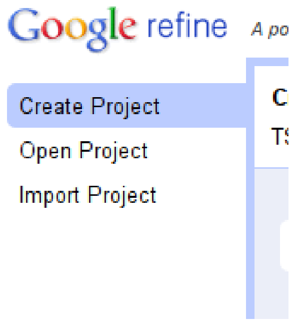
Open Refine (previously Google Refine) is a data cleaning software that uses your web browser as an interface. This means it will look like it runs on the internet but all your data remains on your machine and you do not need internet connection to work with it.

The main aim of Refine is to help you exploring and cleaning your data before you use it further. It is built for large datasets - so don’t worry as long as your spreadsheets can keep the information: Refine can as well.

To work with your data in Refine you need to start a new project:

Walkthrough: Creating a Refine project

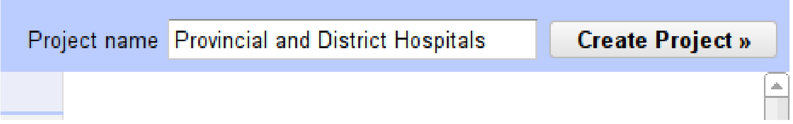
1. Start Refine - this will open a browser window pointing to http://127.0.0.1:3333 if this doesn’t happen open the link with your browser directly
2. Create a new project: On the left tab select the “Create Project” tab:



3. Click on “Choose Files” to choose your downloaded file and click on “next” - you can also use the URL to the CSV directly if your data is hosted on the web.
4. You will get a preview on how refine will interpret your data - if you have selected a well formatted CSV or other file: this should be pretty automatic.
5. Review the preview carefully to make sure the data looks right. Double check character encoding. Mu all data uses UTF-8 these days, but make sure you don’t see any funny characters in preview.

v: latest

6. You may want to turn of “guess data types”, particularly if you have data that contains leading zeros in numbers or identifiers which are significant.
7. Name your project in the box on the top right side and click on “Create Project”



Project name Provincial and District Hospitals

8. The project will open in the project view, this is the basic interface you are going to work with: by default refine shows only 10 rows of data, you can change this on the bar above the data rows. Also you can use the navigation on the right to see the next or previous rows.

You now have successfully created your first Refine project. **Remember:** although it runs in a web-browser, the Refine server is still on your machine - all the data is there (so no worries if you handle sensitive information)

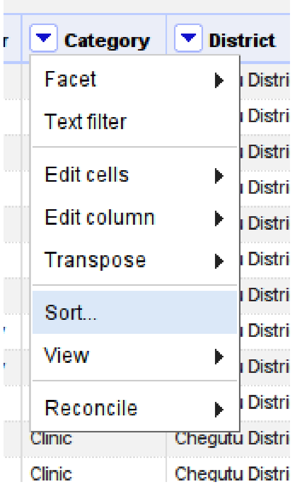
Step 2: Sorting and Facetting

Once we created our project, let’s go and explore the data and the Refine interface a bit. Using Refine might be intimidating at first, since it seems so different from spreadsheets, once you get used to it you will notice how easily you can do things with it.

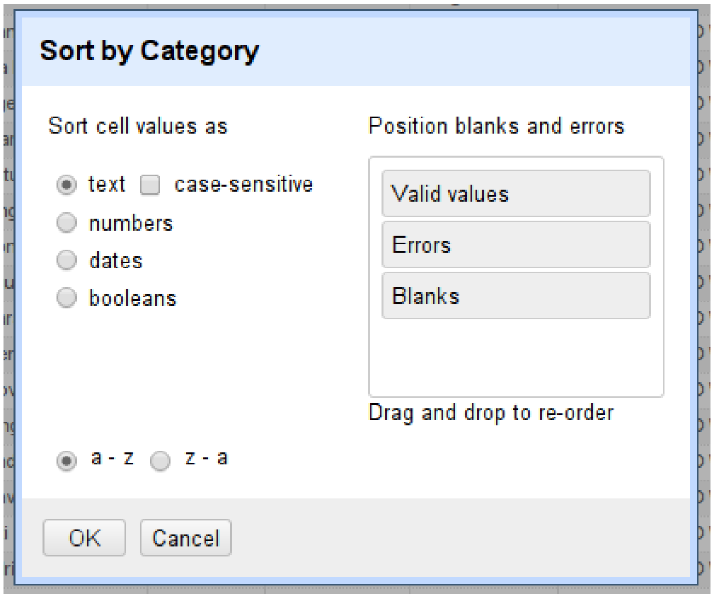
One of the commonly used functions in spreadsheets is sorting and filtering data - to figure out minima, maxima or things about certain categories. Refine can do the same thing.

Walkthrough: Sorting rows

1. Refines handles data similar to a spreadsheet: you have rows, columns and cells - a cell is a field defined by a row and a column.
2. To sort your rows based on a specific column click on the small downward triangle next to the column



3. Select “Sort...” to open the sorting dialog



Sort by Category

Sort cell values as

● text □ case-sensitive
○ numbers
○ dates
○ booleans

● a - z ○ z - a

Position blanks and errors

Valid values
Errors
Blanks

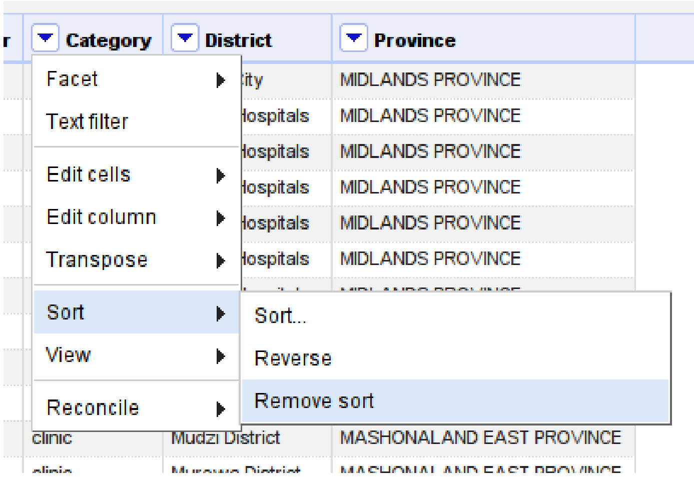
Drag and drop to re-order

OK Cancel

4. You can select what to sort the values as and then what order to sort in. (We'll sort in text, since for now we only have text columns)

5. Click "OK" and your rows will be sorted based on the column

6. To undo the sort, click on the column options again, select "sort" then "remove sort"

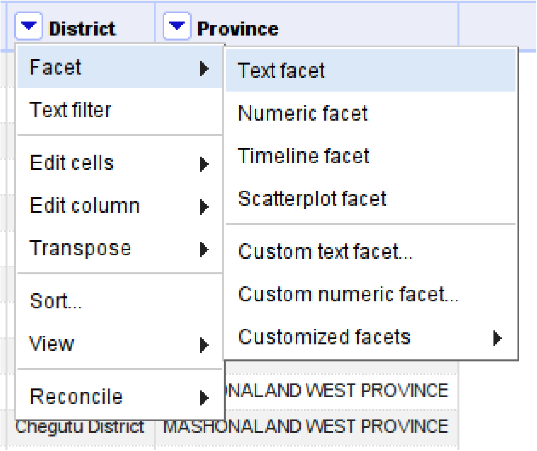


The other frequently used function in Spreadsheets is filtering - in Refine this is called facetting. Facetting in Refine is really powerful - you will see in most of the rest of the Recipe we'll use facets.

Walkthrough: Facetting rows based on a column

1. Select the column options for the column you want to facet with

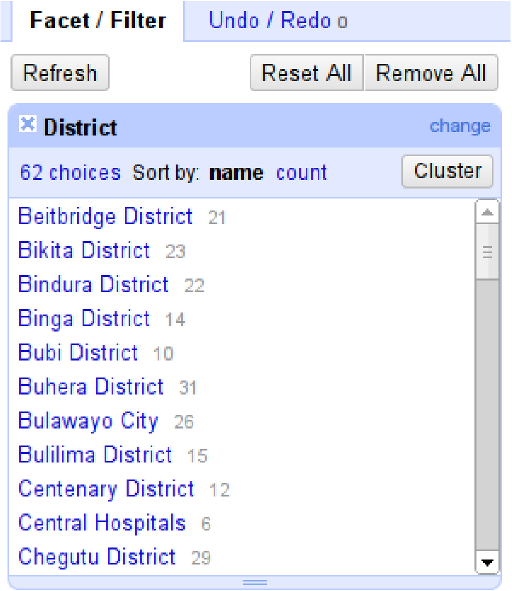
2. Select "Facet"



	NALAND WEST PROVINCE
Chegutu District	MASHONALAND WEST PROVINCE

3. You can facet differently for text, numbers or dates - let's facet as text - click on "Text facet"

4. This will open a facet in the left bar



Beitbridge District	21
Bikita District	23
Bindura District	22
Binga District	14
Bubi District	10
Buhera District	31
Bulawayo City	26
Bulilima District	15
Centenary District	12
Central Hospitals	6
Chegutu District	29

5. Now select one or more of the choices and you'll see how your data rows are limited to just those selected.

6. Of course you can add more than one facet and thus filter more than once.

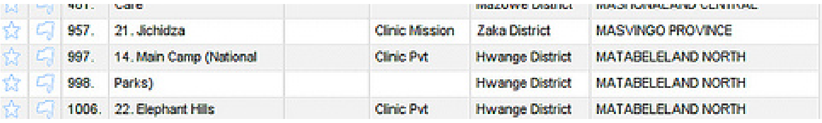
Step 3 Dealing with Blank Cells

If you look closely at your facets, you'll notice that on the bottom you have a selector saying "(blank)" for this - we'll need to deal with it.

Walkthrough: Filling in the (blank)s

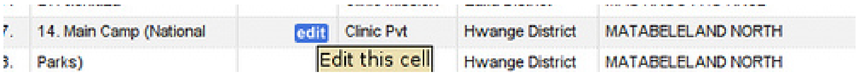
1. Choose the "(blank)" facet in your "Owner" column

2. If you look at some of the rows, you'll see that there was a mis-split of the columns and the owner actually ended up in the "Category" Column

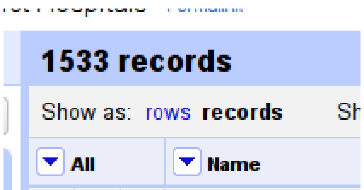


481.	Care			Mazowe District	MASHONALAND CENTRAL
957.	21. Jichidza		Clinic Mission	Zaka District	MASVINGO PROVINCE
997.	14. Main Camp (National		Clinic Pvt	Hwange District	MATABELELAND NORTH
998.	Parks)			Hwange District	MATABELELAND NORTH
1006.	22. Elephant Hills		Clinic Pvt	Hwange District	MATABELELAND NORTH

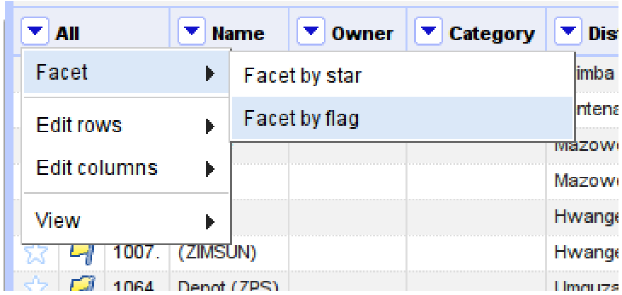
3. To fill this into the "Owner" Field hover over the cell you want to fill in and click the "edit" button.


7. 14. Main Camp (National edit Clinic Pvt Hwange District MATABELELAND NORTH
3. Parks) Edit this cell Hwange District MATABELELAND NORTH

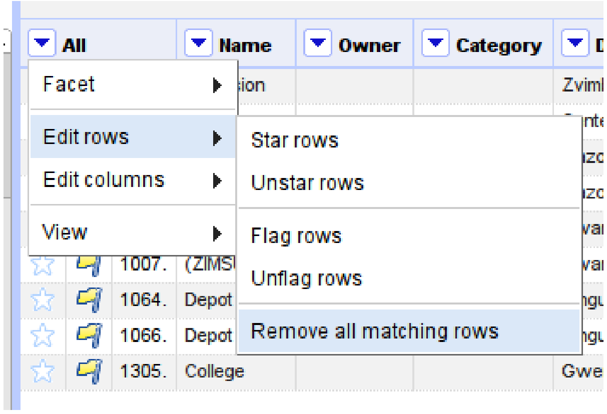
4. If you click the “Edit” button you can add the Owner there - don’t forget to also correct the “Category” cell.
5. You’ll notice some rows seem to be erratic - they don’t have a name that makes sense and no further information - you can flag these for deletion by clicking on the little flag.
6. Do the same with the “Category” Column - the Category is sometimes joined with the “Name” column
7. Now let’s delete the flagged rows - make sure you are in row mode for this: for this click on “row” in the top left corner above the data.



8. Open the column options for “All” and select “Facet” - “Facet by Flag”



9. Now you can select “true” in your flag facet on the left.
10. Now let’s delete the flagged rows: in the Column options for “All” select “Edit rows” - “Remove all matching rows”



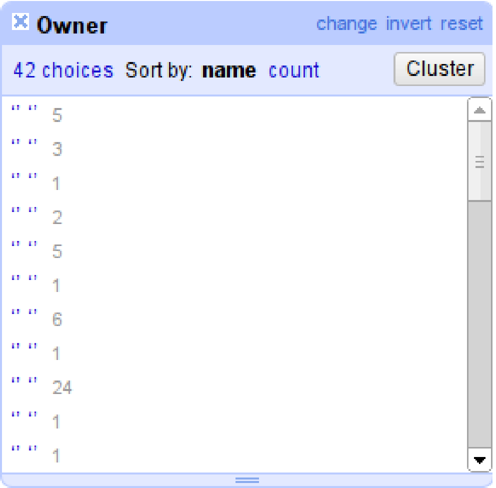
Step 4: Fighting the Invisible Man

As illustrated in The Invisible Man is in your Spreadsheets having spaces or newlines in your datafields is a problem. Since this is a very common problem, Refine has specific functions to remove whitespaces that shouldn’t be there.

Walkthrough: Removing hidden whitespaces

1. Let’s start cleaning our Dataset with the Owner Column
2. Create a Text Facet for the Owner Column as described above

3. You will notice that there are several things odd in the column: It starts with a long list of similar looking entries - we'll deal with it later.

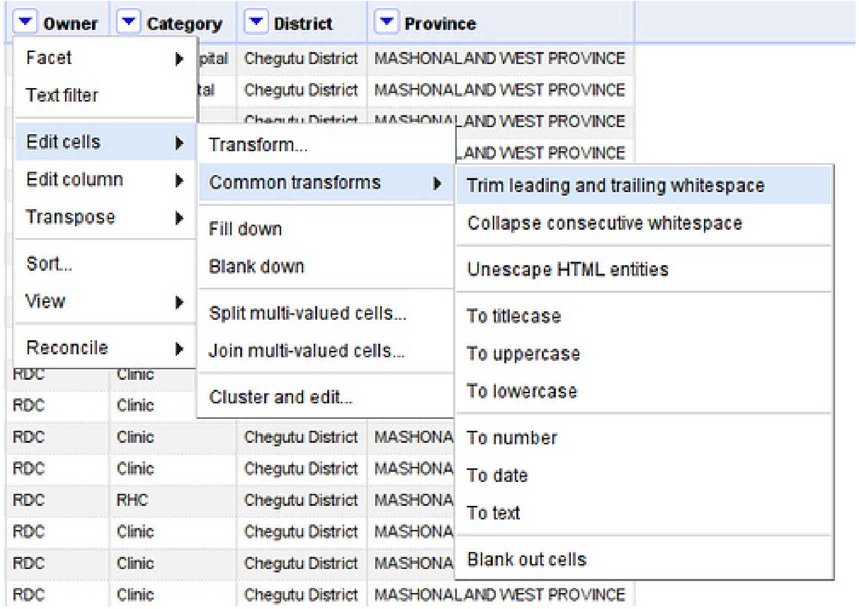


Although they look similar to you, they are different for the computer - there is a different number of spaces between the quotes.

4. Scroll down and you'll notice that some entries will be there twice - although they look similar. There are two entries for Municipality that look exactly the same. This is because they have whitespaces at the end.

5. Refine can help you clean this up in an instant - open the column options for the "Owner" column

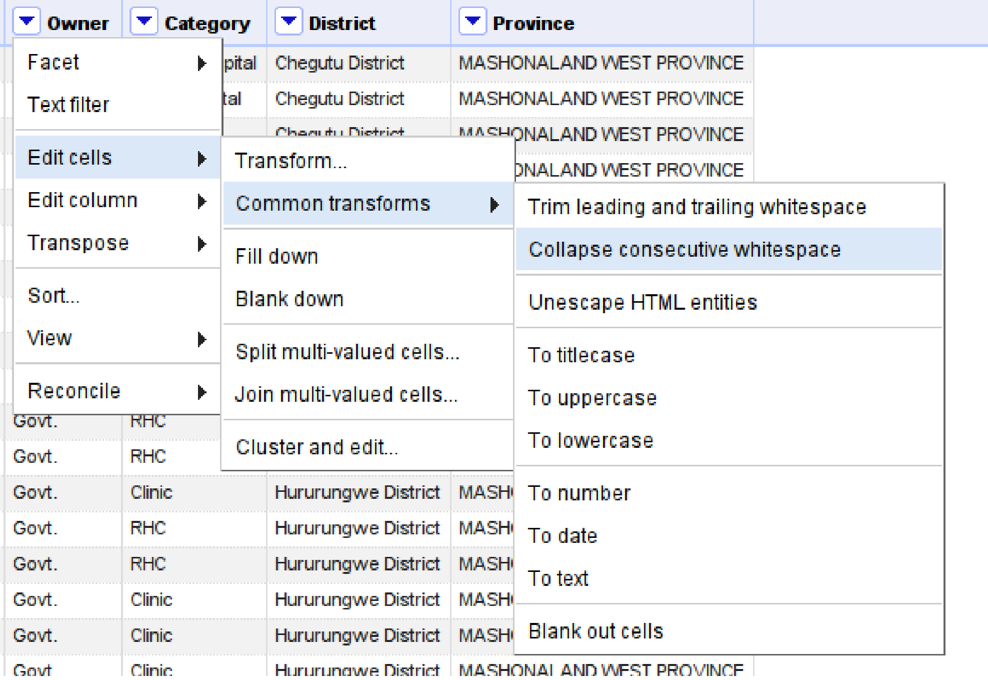
6. Select "Edit Cells" - "Common Transforms" - "Trim leading and trailing whitespaces"



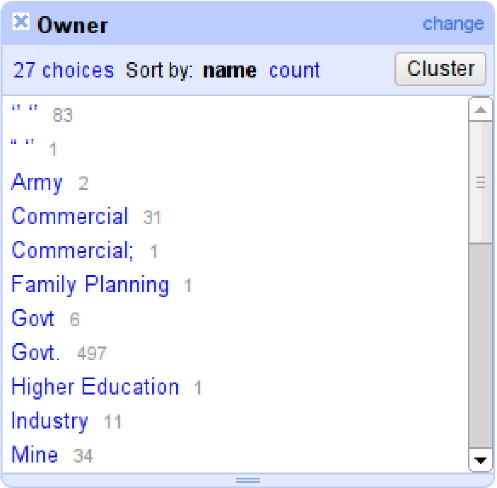
7. This will remove whitespaces in the beginning and at the end of your column

8. Check Municipality and you'll note that there's only one choice now - perfect. Now let's deal with the list at the beginning.

9. Select "Edit Cells" - "Common Transforms" - "Collapse consecutive Whitespaces"



10. You'll see the multiple choices have been reduced to two choices in an instant

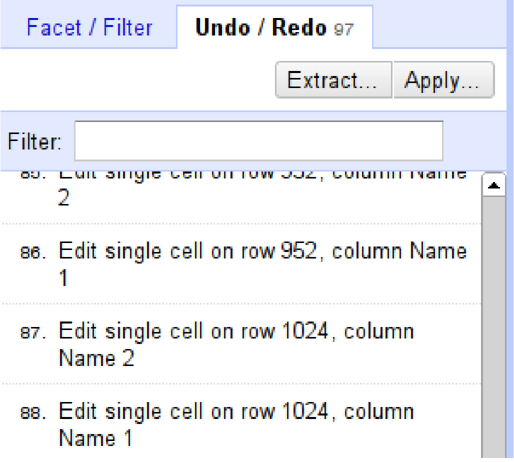


" "	83
" "	1
Army	2
Commercial	31
Commercial;	1
Family Planning	1
Govt	6
Govt.	497
Higher Education	1
Industry	11
Mine	34

11. Now our list already looks a lot cleaner!

12. Go ahead and apply the two transforms to all your columns.

Once you made your transforms you might wonder: What if I made a mistake? Also if you work with data you generally want to keep track of what you did to the data. Since Refine was build with data processing in mind it keeps track of what you're doing with your data and allows you to go back and forth in time. To see your history of changes click on the "Undo/Redo" tab on the left.



85. Edit single cell on row 952, column Name 2

86. Edit single cell on row 952, column Name 1

87. Edit single cell on row 1024, column Name 2

88. Edit single cell on row 1024, column Name 1

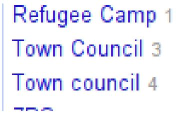
You see all the changes you made - by simply clicking on one of the steps you'll be undoing all the changes after the step (don't worry you can redo pretty much the same way). Play with this system until you are comfortable.

Step 5. Reconciling categories

A quick look at our categories and you'll notice that not everything is well in Owner land - still some categories that should be the same are not. The same for the "Category" column - let's reconcile them.

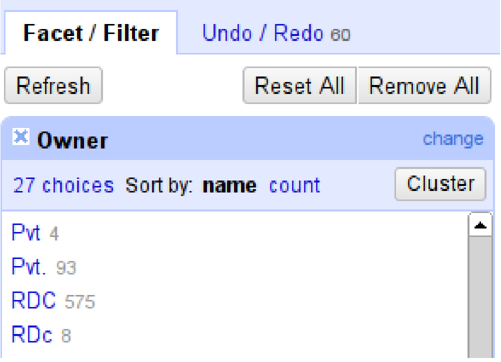
Walkthrough: Reconciling Categories

1. Create a Facet for the column you want to reconcile (in our case this is "Owner")
2. The first step is to bring the categories to the same case - see for example "Town Council" and "Town council" - the difference is just one letter.



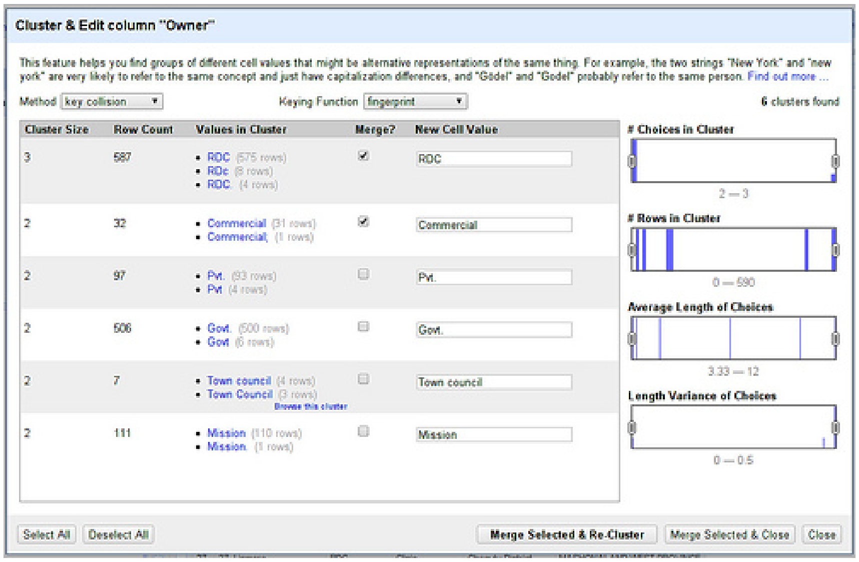
Refugee Camp	1
Town Council	3
Town council	4

3. Refine can help you to automatically find the categories that belong together - a feature it calls "Clustering". To activate clustering click on the "Cluster" button in your facet.



Pvt	4
Pvt.	93
RDC	575
RDc	8

4. You will end up in the clustering menu - as you can see Refine is pretty smart about which things should belong together



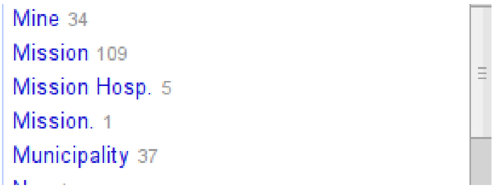
Cluster & Edit column "Owner"

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 ▾ 6 clusters found

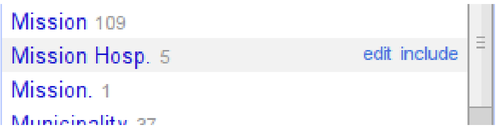
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	587	RDC (575 rows) RDc (8 rows) RDC. (4 rows)	☑	RDC
2	32	Commercial (31 rows) Commercial, (1 rows)	☑	Commercial
2	97	Pvt. (93 rows) Pvt (4 rows)	☐	Pvt.
2	506	Govt. (500 rows) Govt (6 rows)	☐	Govt.
2	7	Town council (4 rows) Town Council (3 rows) Browse this cluster	☐	Town council
2	111	Mission (110 rows) Mission. (1 rows)	☐	Mission

5. Check the "merge" checkbox if you want the two categories to be the same. Once you marked all the categories you want to merge click on "Merge selected & Re-Cluster"
6. If Refine doesn't find more values to be similar change the "Keying Function" and see whether you can find more similar categories - if not: simply click close to continue.
7. This reconciled some of your values - let's go on.
8. Look at "Mission" for example we have three different categories for what should be one - Refine did not automatically find them.



Mine 34
Mission 109
Mission Hosp. 5
Mission. 1
Municipality 37

9. Let's change them all to mission
10. Hover over "Mission Hosp." notice the "edit" button at the end?



Mission 109
Mission Hosp. 5
Mission. 1
Municipality 37

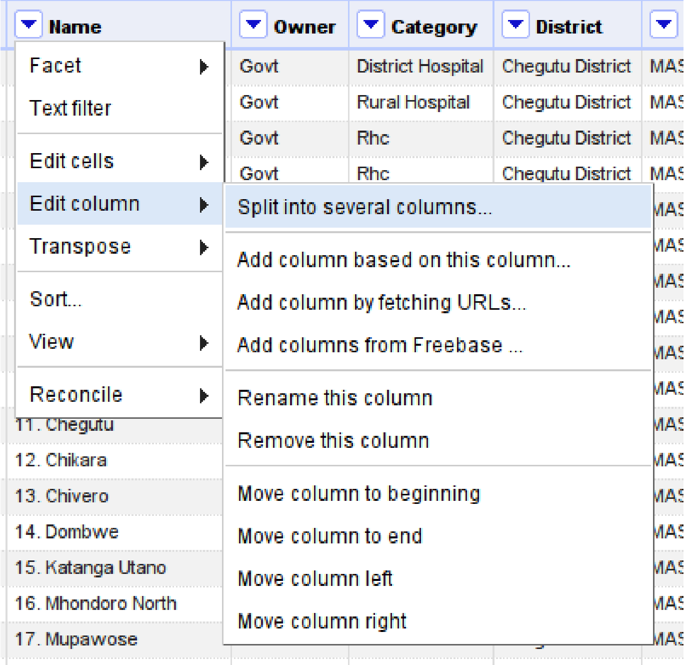
11. Click on Edit - this will open the field for editing. Change the name to "Mission" - this will change "Mission Hosp." to "Mission" in all cells where it appears - continue on to change all the fields you can find.
12. Repeat reconciling for "Category"

Step 6: Splitting Columns

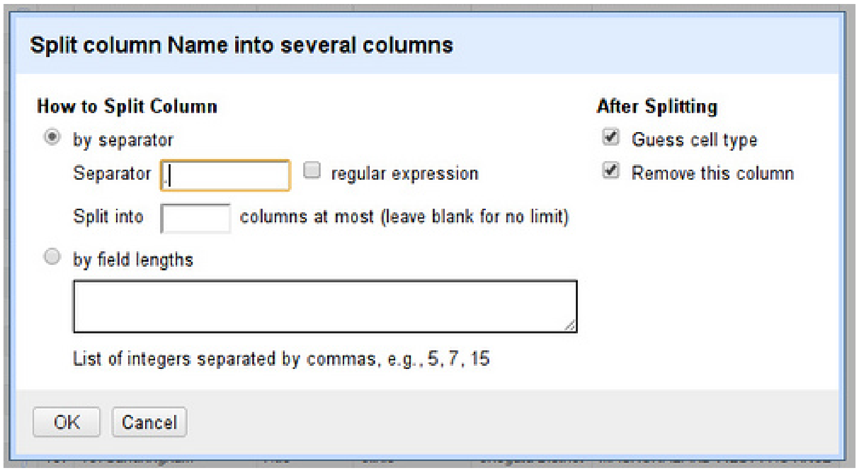
If you look at the "Name" column in our dataset you'll notice that the names commonly start with a number (this is an enumeration of hospitals in a district - and is an artifact from extracting the data). Let's clean this up and split the number and the name.

Walkthrough: Splitting Columns

1. To Split a column select "Edit Column" "Split into several columns"



2. We want to split at a "." since the number generally ends with a "."
3. Enter "." into the Seperator for in the split menu - since we only want to have two new columns enter 2 into the field below so the sentence reads "Split into 2 columns at most"



Split column Name into several columns

How to Split Column

● by separator

Separator . □ regular expression

Split into columns at most (leave blank for no limit)

○ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting

☑ Guess cell type

☑ Remove this column

4. Click on "OK" and you'll end up with two columns.
5. On some of the rows the split will fail - to fix those, create a facet on the second column and select "(blank)"
6. You can now manually fix the cells.

Congratulations - You successfully cleaned up a dataset using Refine!

However there is even more you can do with Refine: For example did you notice how there is always a number next to the categories in the facet - telling you how many rows are in that category? By combining two facets, can you find out how many clinics the government owns? And who owns the Provincial Hospitals?

Any questions? Got stuck? Ask School of Data!

v: latest

Last updated on Dec 09, 2013.

v: latest