Google openrefine

Google refine

A powerful tool dealing with messy data

A GUIDE TO transform and deal with data

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; extending it with web services; and linking it to databases like Freebase.

Even journalists with little database expertise should be using Refine to organize and analyze data; it doesn't require much more technical skill than clicking through a webpage.

So normally, before attempting any calculations or analysis in Excel, Google Fusion Tables or visualization in Tableau, we often use Google Refine to clean the data.

Chapter 1

INTRODUCTION

What is “Messy data”?

“Messy data” refers to data that’s riddled with inconsistencies, either because of human error or poorly designed record systems. So, a column that contains a city’s name may hold values such as “New York”, “new york,” “New york city” and etc.

These inconsistencies can wreak havoc when trying to perform analysis on the data, so they have to be addressed before starting any analysis.

Hence, Google Refine, will help us do data cleaning before analyzing it.

Chapter 1

INTRODUCTION

Download and Installation

Step 1:

Please go to the website to install Google Refine on your computer.

http://openrefine.org/download.html

Remarks: If you are using Mac, please click "Mac Kit"; if you are using Windows, please click "Windows Kit".

OpenRefine Core

Google Refine 2.5 - Stable version

- **Windows kit**, Download, unzip, and double-click on *google-refine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it. **NOTE:** If you have issues installing Refine on Mac, please refer to issue 590 - OpenRefine 2.5 for mac support java 6 and 7 only
- **Linux kit**, Download, extract, then type *./refine* to start. **NOTE:** OpenRefine 2.5 for linux support java 6 and 7 only

Chapter 1

INTRODUCTION

Importing your data

Step 2:

Open the application on your computer, and it will be open in the browser, click “Create Project” and “choose file” from your computer. After uploading the file, click “Next”.

Download the data from this website: https://drive.google.com/file/d/0Bx8m5CZw-gtgcVZzLVBYQ0l5VTQ/view?usp=sharing

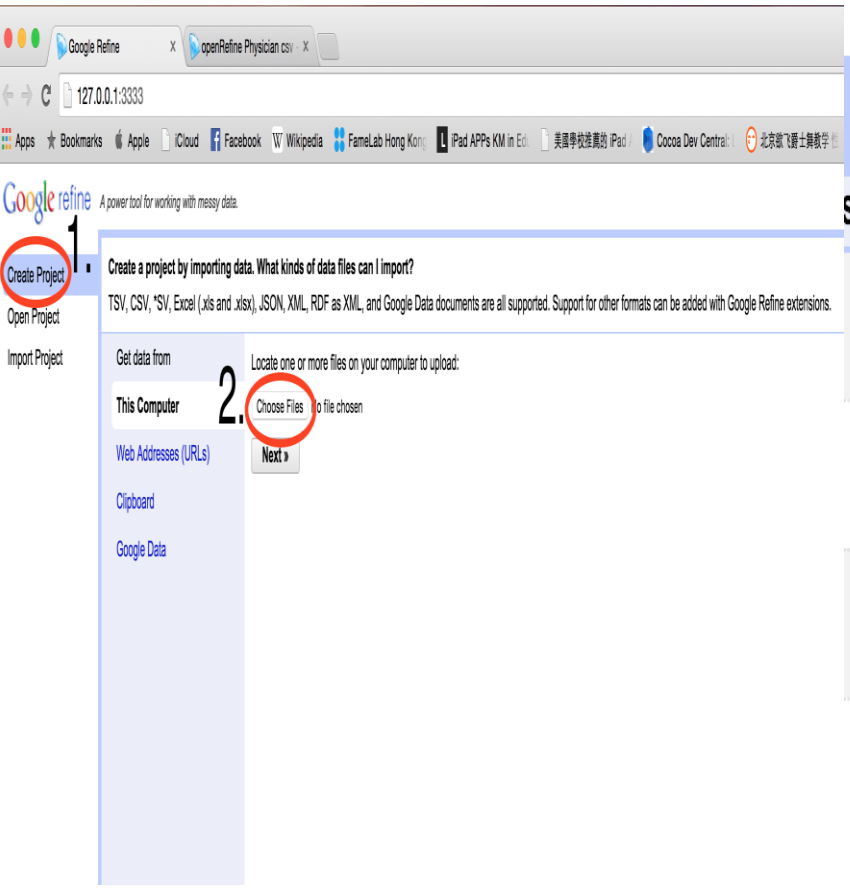
(There is a problem to view it, but no problem to download it)

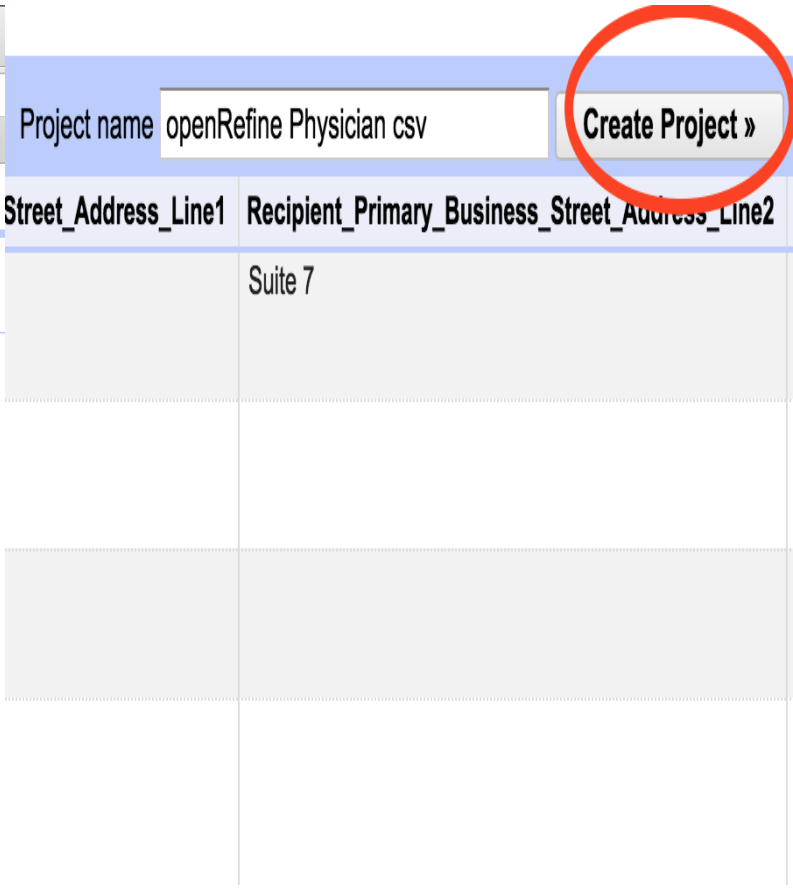
Chapter 1

INTRODUCTION

Importing your data

See the example below and then click “create project” on the right top corner.





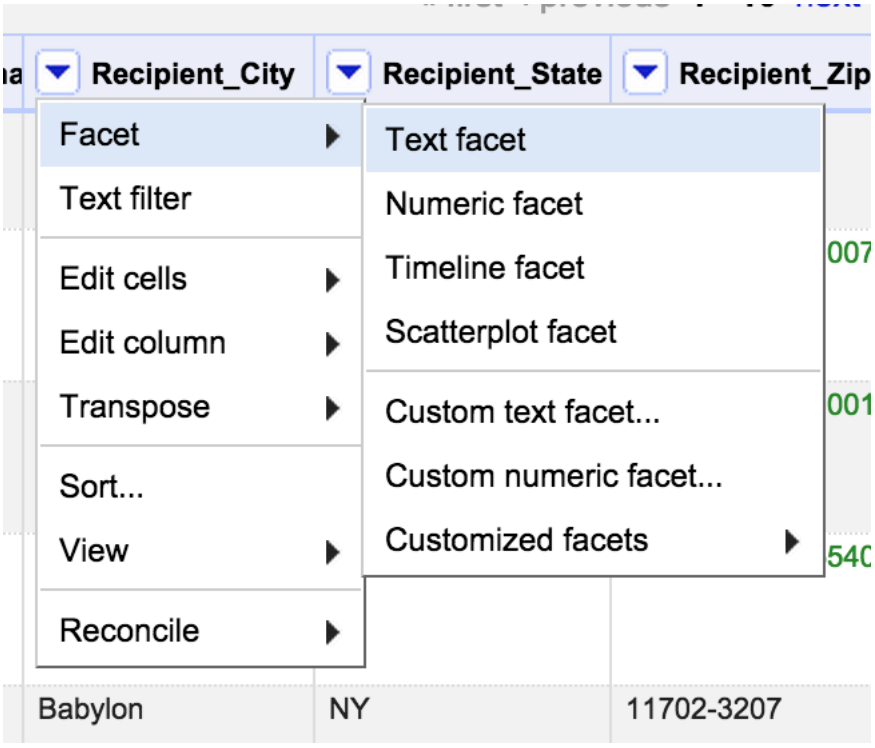
Chapter 1

INTRODUCTION

Faceting data-Text facet

Step 3:

The data includes the information of physicians and recipients and we would like to see the recipient's city. Find the column named "Recipient_City", click the arrow->facet->Text facet.



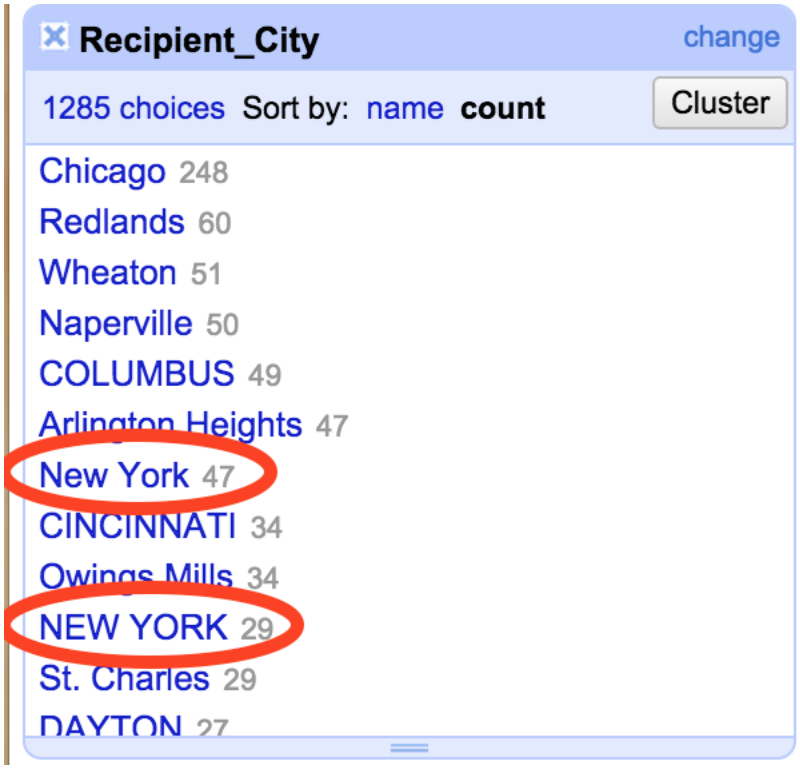
Chapter 1

INTRODUCTION

Faceting data-Text facet

Step 4:

On the left side of your webpage, click “sort by count”, you will see most of the recipients are from Chicago. And also 47 from “New York” and 29 from “NEW YORK”, which represents the same city.



Chapter 1

INTRODUCTION

Clustering data

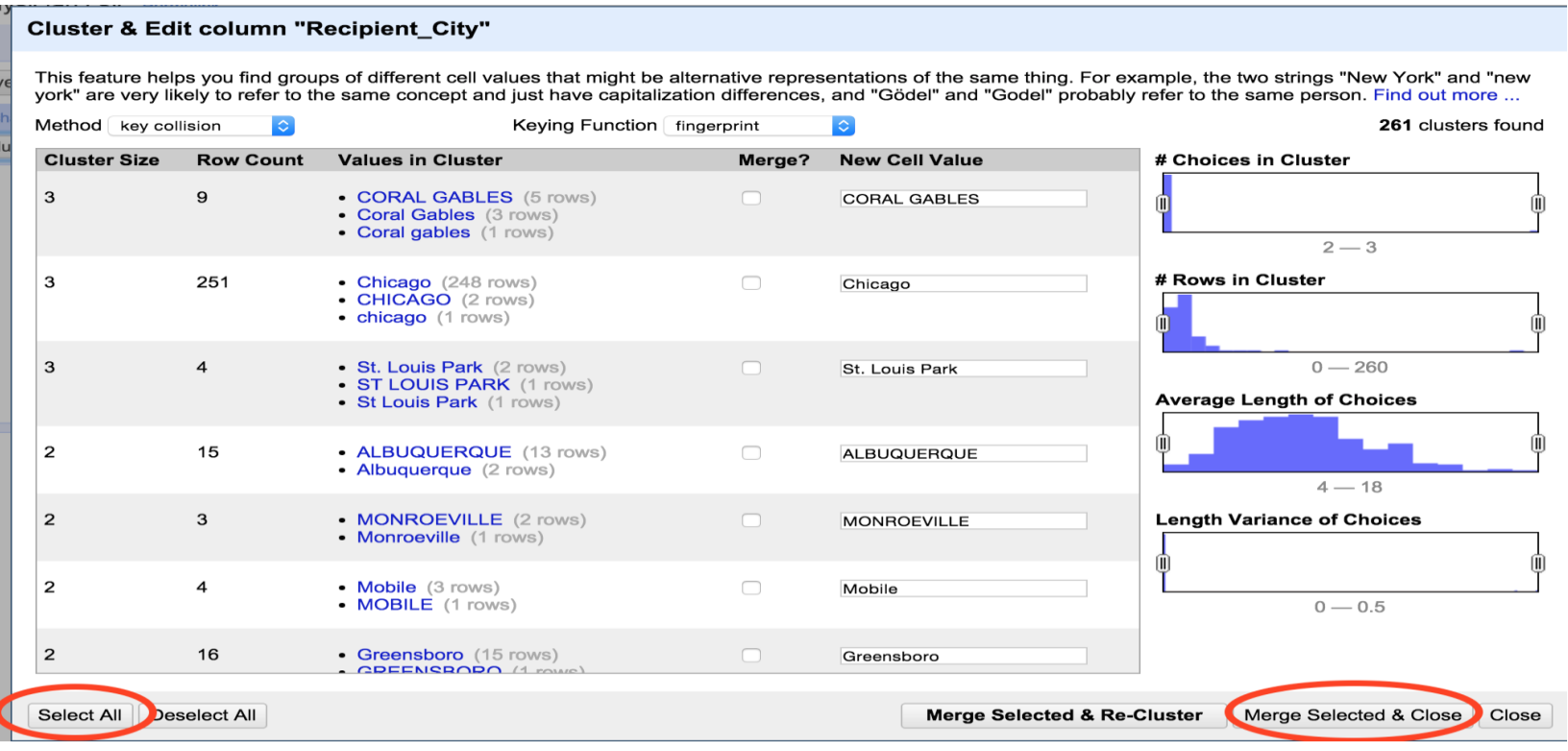
Step 4:

To solve this problem, we could use “cluster” function on the top right corner.

1.Click “cluster”, you will see the same city name will be clustered together;

2.Click “select all”;

3.Click “merge selected & close”.

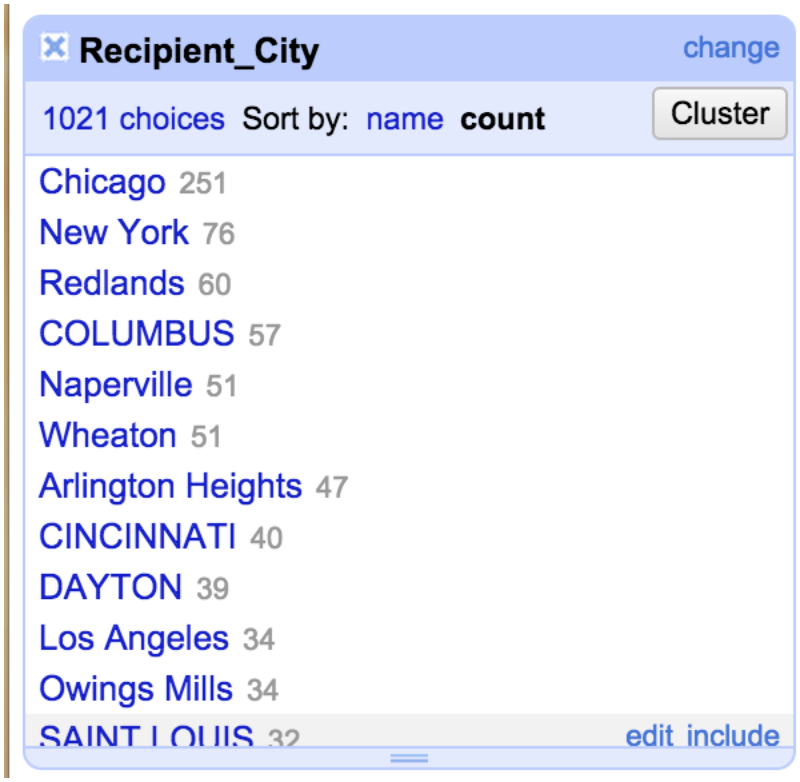


Chapter 1

INTRODUCTION

New Problem Occurs

Then you could see the new lists on the left side. But new problem occurs that the names of some cities are in upper case while some of them are in lower case. To overcome this, we use another method.



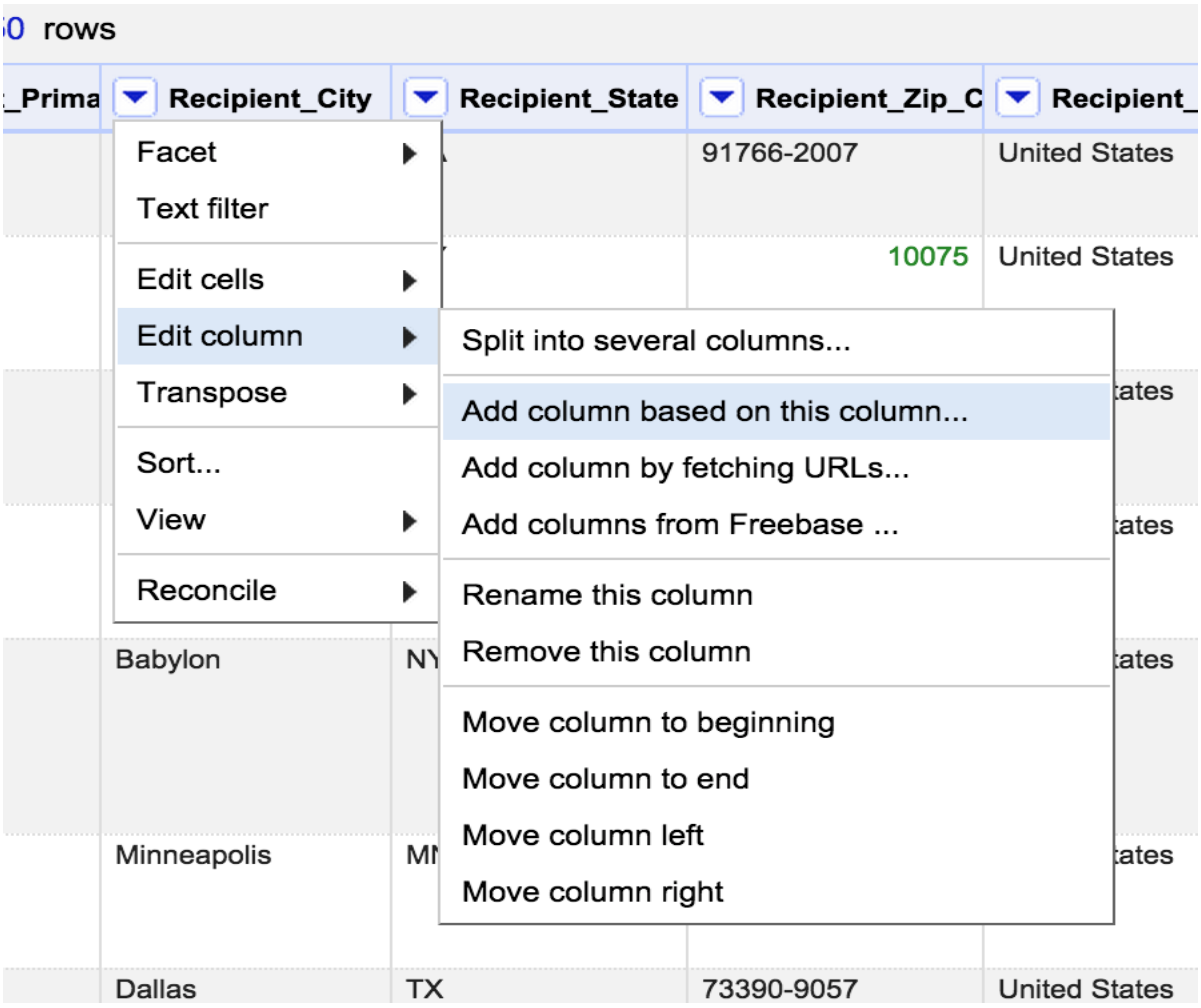
Chapter 1

INTRODUCTION

Duplicate the column

Step 5:

Back to the "Recipient_City" column, click the arrow->Edit column->Add column based on this column



Chapter 1

INTRODUCTION

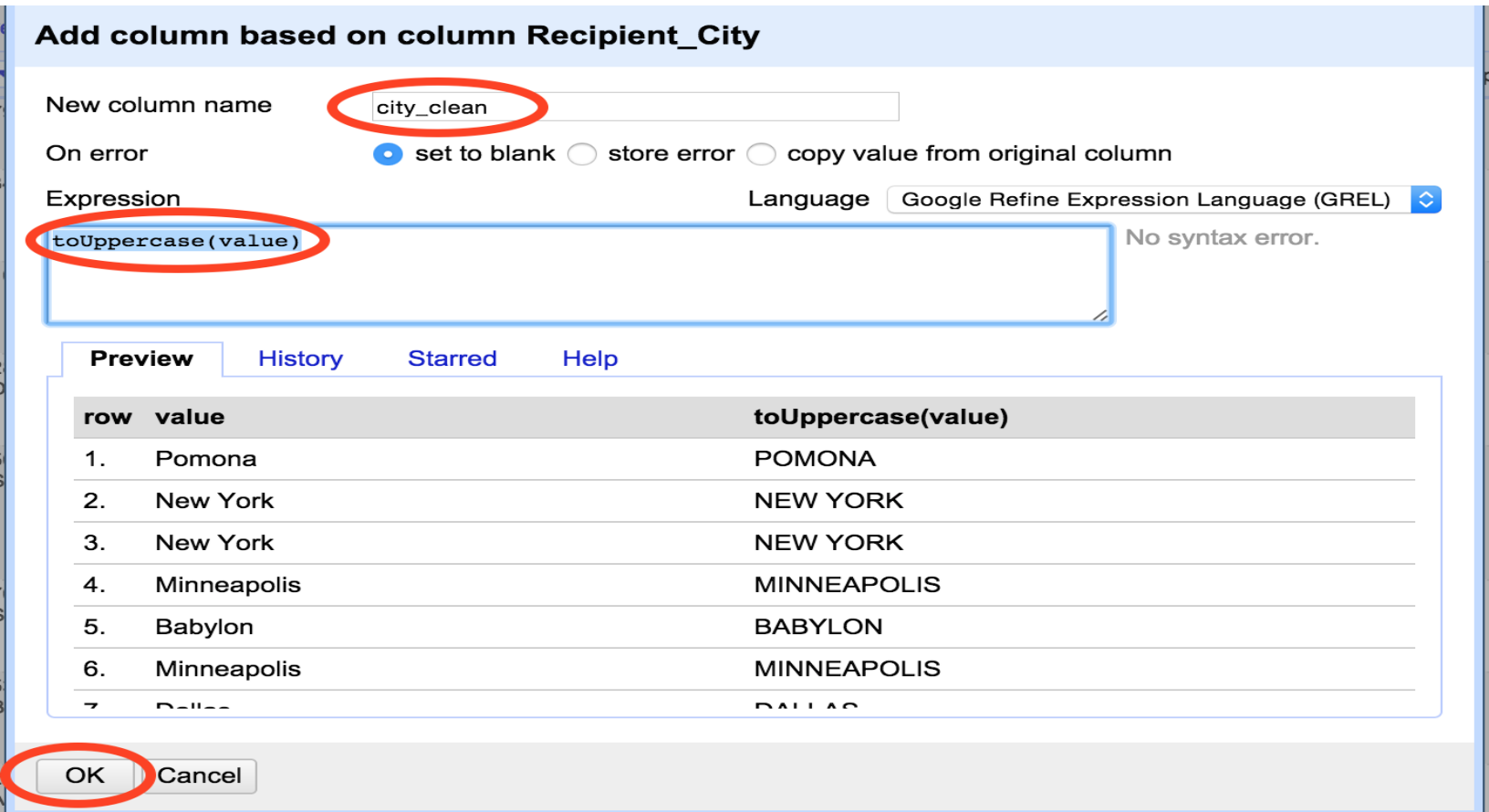
Duplicate the column

Step 5:

1.Enter new column name “city_clean”;

2.Change “value” to “toUppercase(value)”, you will see the value originally and see the new value on the right of it.

3.Then click “OK”.



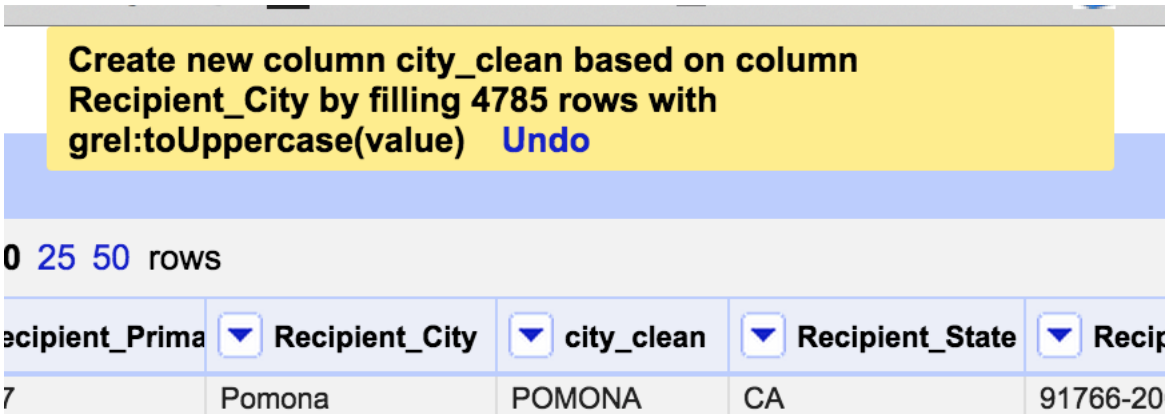
Chapter 1

INTRODUCTION

Duplicate the column

Then you will get a column of city names in uppercase. And you could delete the original column “Recipient_City”.

One thing good about the refine is you could “Undo” the operation you’ve done to the data if you’ve do something wrong.



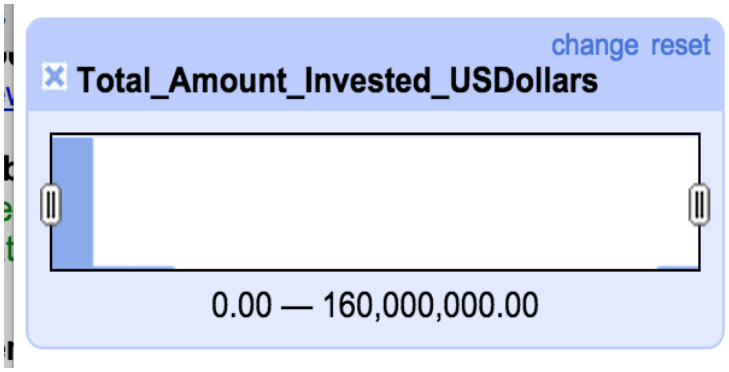
Chapter 1

INTRODUCTION

Facet data-Numeric facet

Step 6:

Find “Total_Amount_Invested_USDollars”, click the arrow->Facet->Numeric facet. Then you could see the result on the left side of the webpage like below.



Chapter 1

INTRODUCTION

Facet data-Numeric facet

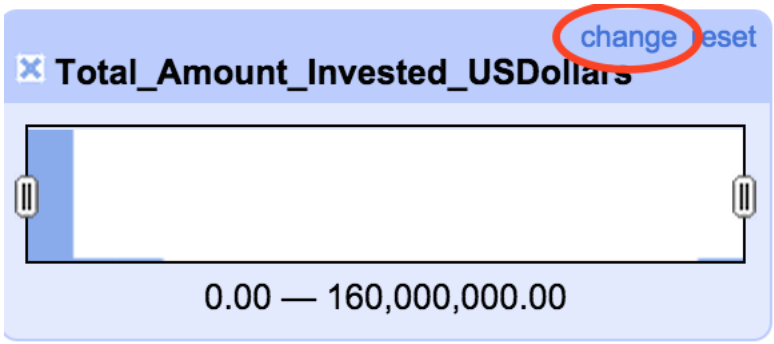
Step 6:

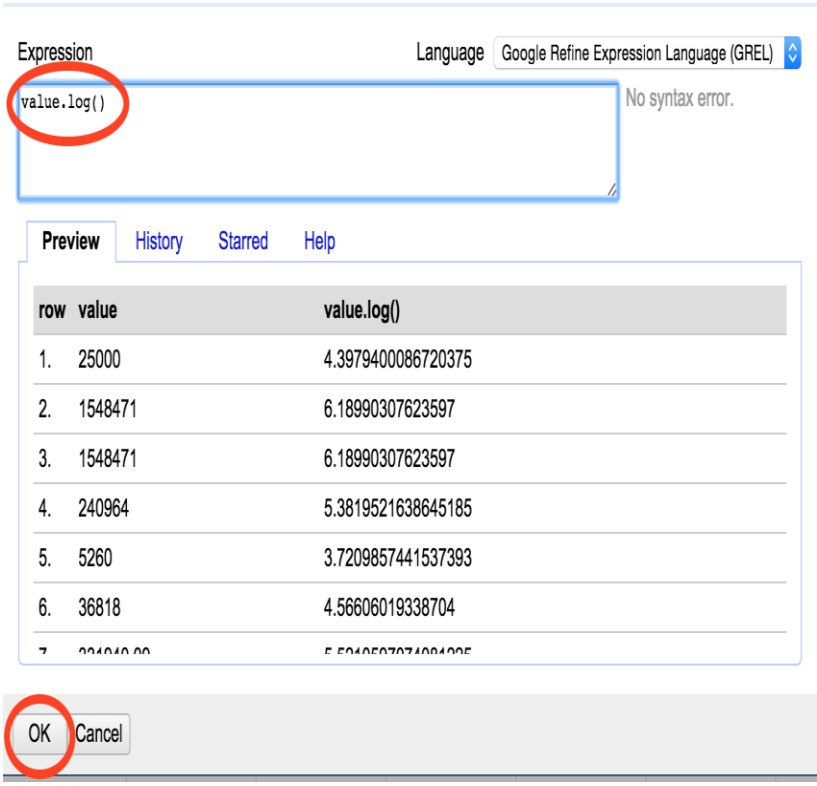
As the range is so large, we need to modify it for better understanding.

1.Click “change” on the top right corner;

2.Change “value” to “value.log()”, you will see the result on the right.

3.Click “OK”.



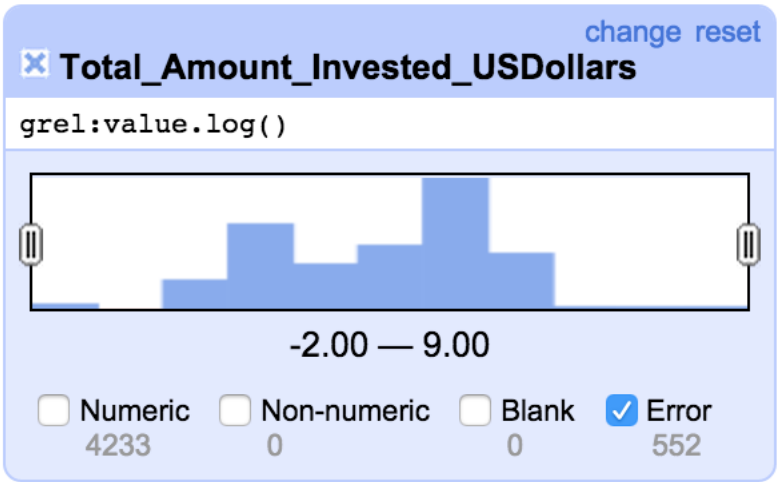


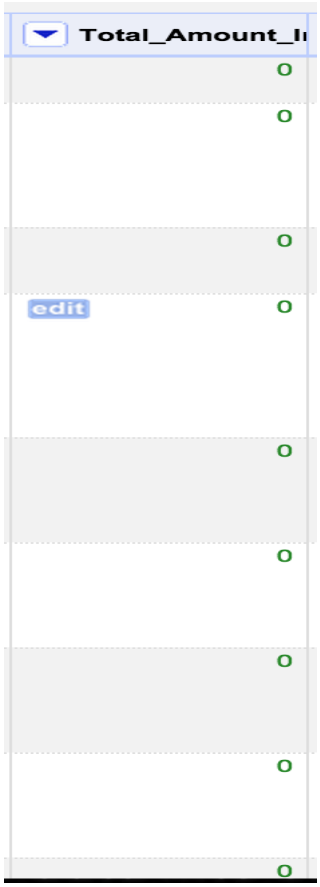
Chapter 1

INTRODUCTION

Facet data-Numeric facet

Then you will see the result like below: the range will change to from -2 to 9. We could see that there are errors, if we uncheck the numeric, we will see like below. The values are 0, which are correct.



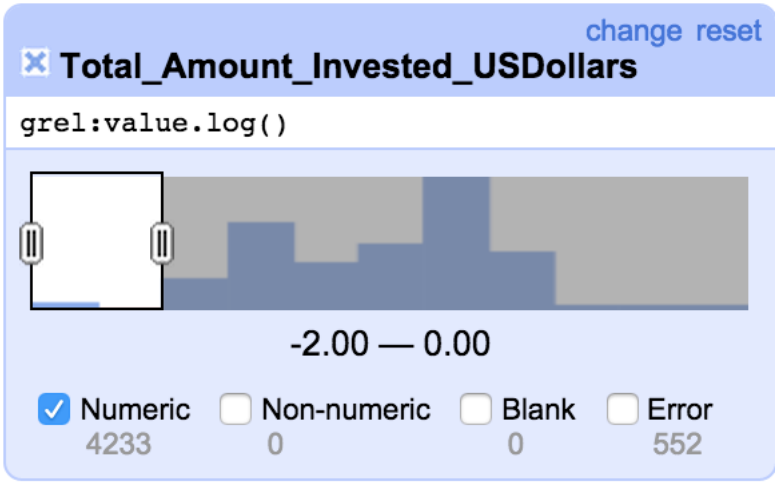


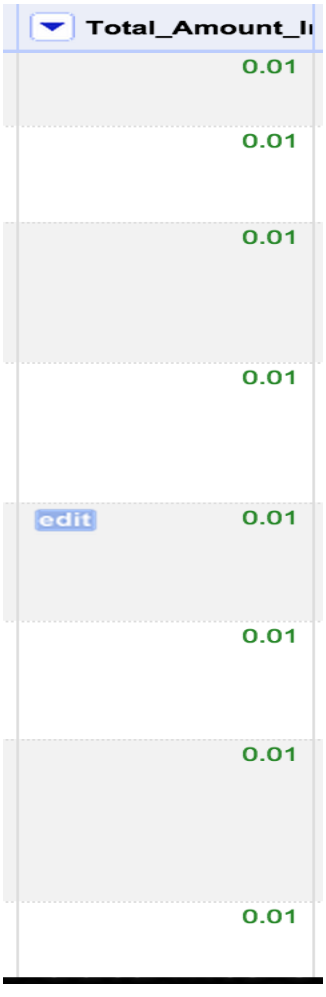
Chapter 1

INTRODUCTION

Facet data-Numeric facet

And the range includes negative numbers. If we change the range from -2 to 0, and uncheck the error, we will see the value are 0.1, which is less than 1, we know that log (0~1) is negative, and they're correct, too.





Chapter 2

Data Transformation

Download original webpage data

Besides the messy data, the data might just not be in the form that you want. And using Google Refine, you can quickly transform it into the form that you do want.

Go to the webpage to get the data!
https://en.wikipedia.org/wiki/Filmfare_Award_for_Best_Actress#List_of_winners_and_nominees

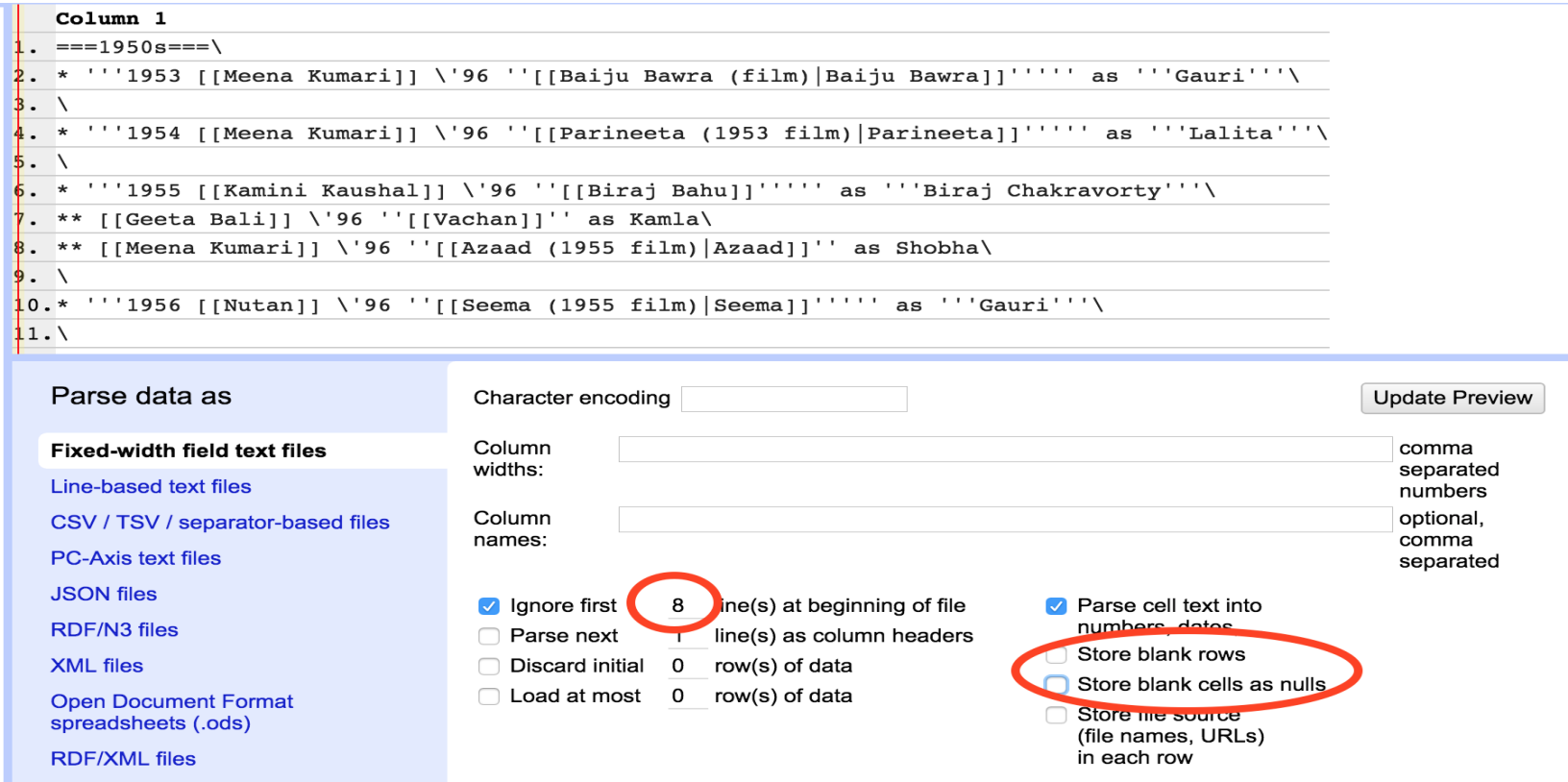
1.You need to scroll down to find “List of winners and nominees”;
2.Click the “edit” button beside it;
3.Copy the codes and paste them in any text editor in your computer;
4.Save the file and name it “List of winners”.

Chapter 2

Data Transformation

Import the data

Import the data like the chapter 1 said. You may see the data like this after uploading it. Before cleaning it, we need to do something. If your data does not start with the "===1950s===", you need to ignore the first "a number" rows to ignore them. Uncheck "store blank rows" and "store blank cells as nulls".



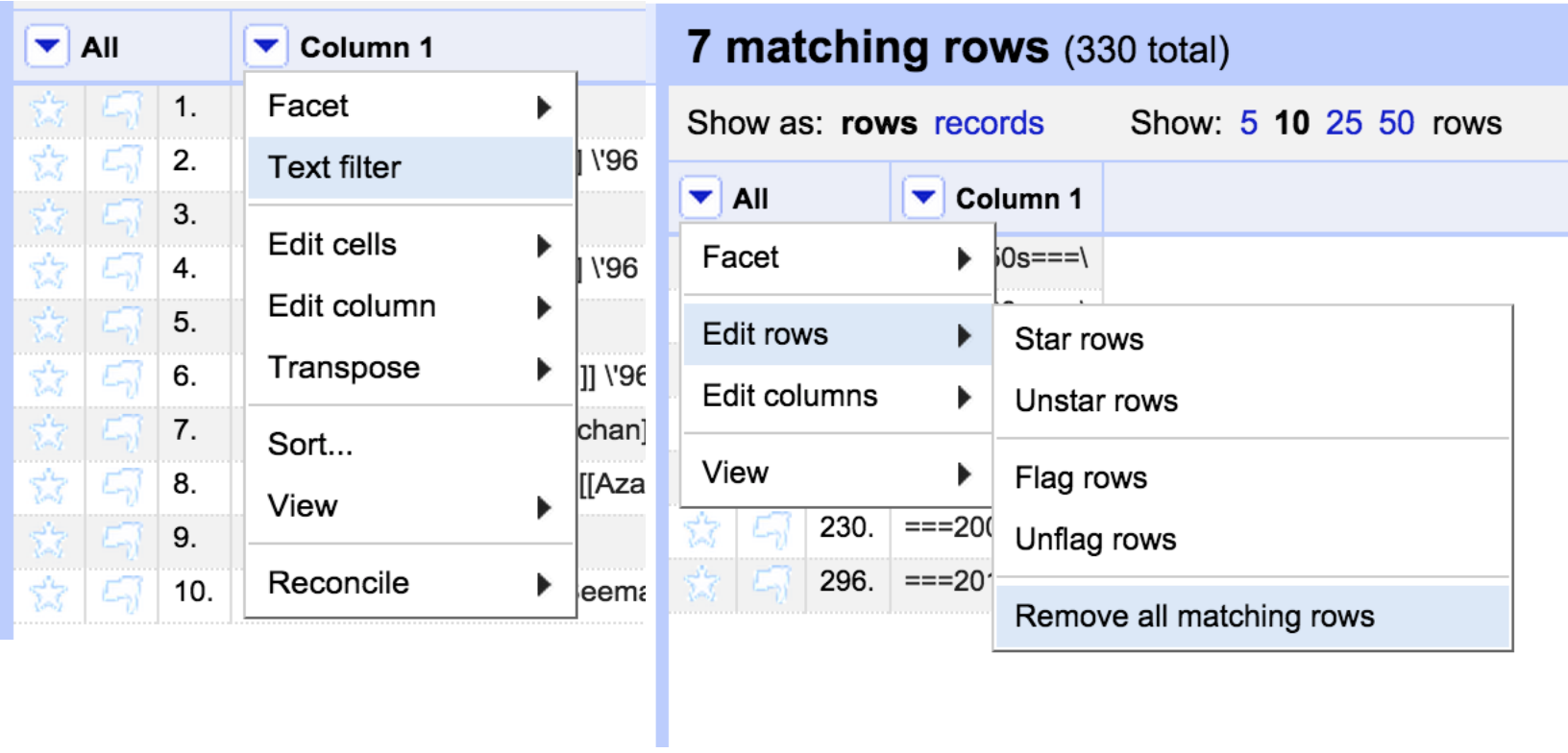
	Column 1
1.	===1950s===\
2.	* '''1953 [[Meena Kumari]] \'96 ''[[Baiju Bawra (film)|Baiju Bawra]]''''' as '''Gauri'''\
3.	\
4.	* '''1954 [[Meena Kumari]] \'96 ''[[Parineeta (1953 film)|Parineeta]]''''' as '''Lalita'''\
5.	\
6.	* '''1955 [[Kamini Kaushal]] \'96 ''[[Biraj Bahu]]''''' as '''Biraj Chakravorty'''\
7.	** [[Geeta Bali]] \'96 ''[[Vachan]]'' as Kamla\
8.	** [[Meena Kumari]] \'96 ''[[Azaad (1955 film)|Azaad]]'' as Shobha\
9.	\
10.	* '''1956 [[Nutan]] \'96 ''[[Seema (1955 film)|Seema]]''''' as '''Gauri'''\
11.	\

Chapter 2

Data Transformation

Delete unnecessary rows

1. Choose Column 1, then click “Text filter”;
2. Type “===“ to find all rows with “===“, you will see 7 matching rows;
3. Choose “All”, click “Edit rows” then “Remove all matching rows” to delete them aall.

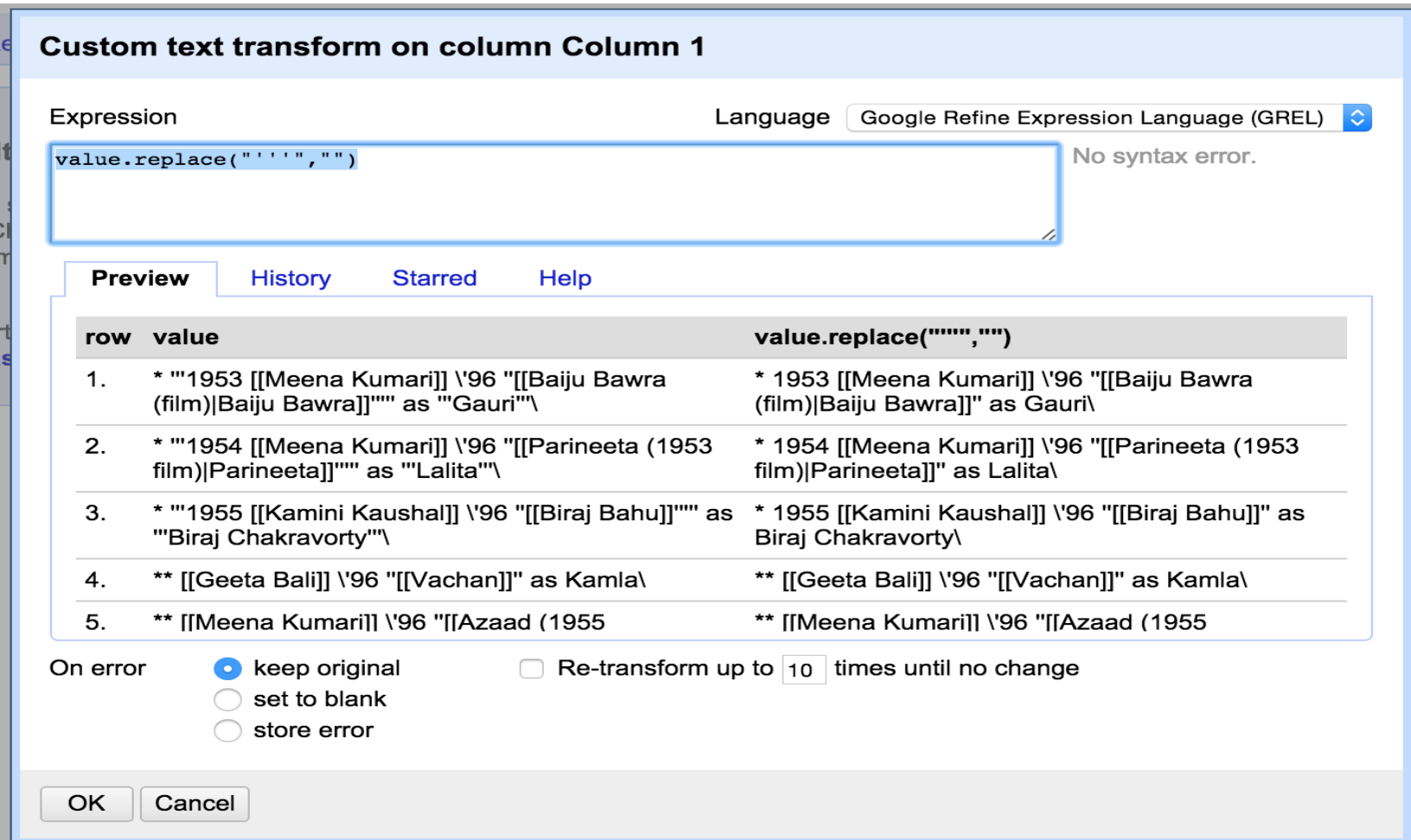


Chapter 2

Data Transformation

Transform Function

1. Choose “Column 1”, click “Edit cells” and then “Transform”;
2. Modify the Expression like below, you will see the different appearance of the data from below, then click “OK”.

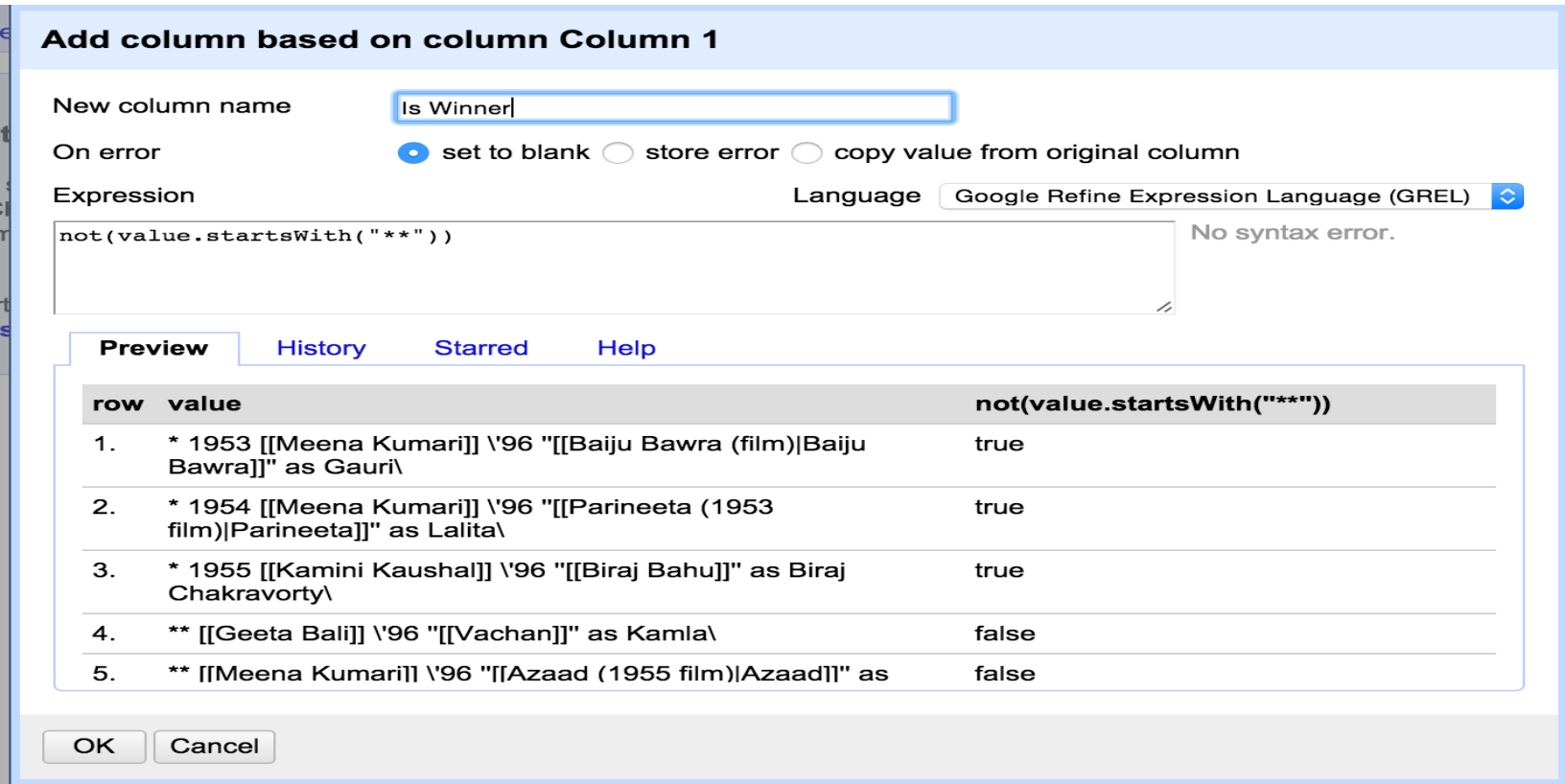


Chapter 2

Data Transformation

Add Column to see the winner

1. Do you still remember how to add column based on this column we described in chapter 1? You need to click “Edit column” then choose “add column based on this column”.
2. Modify like the screenshot below, pay attention to the name and the expression.

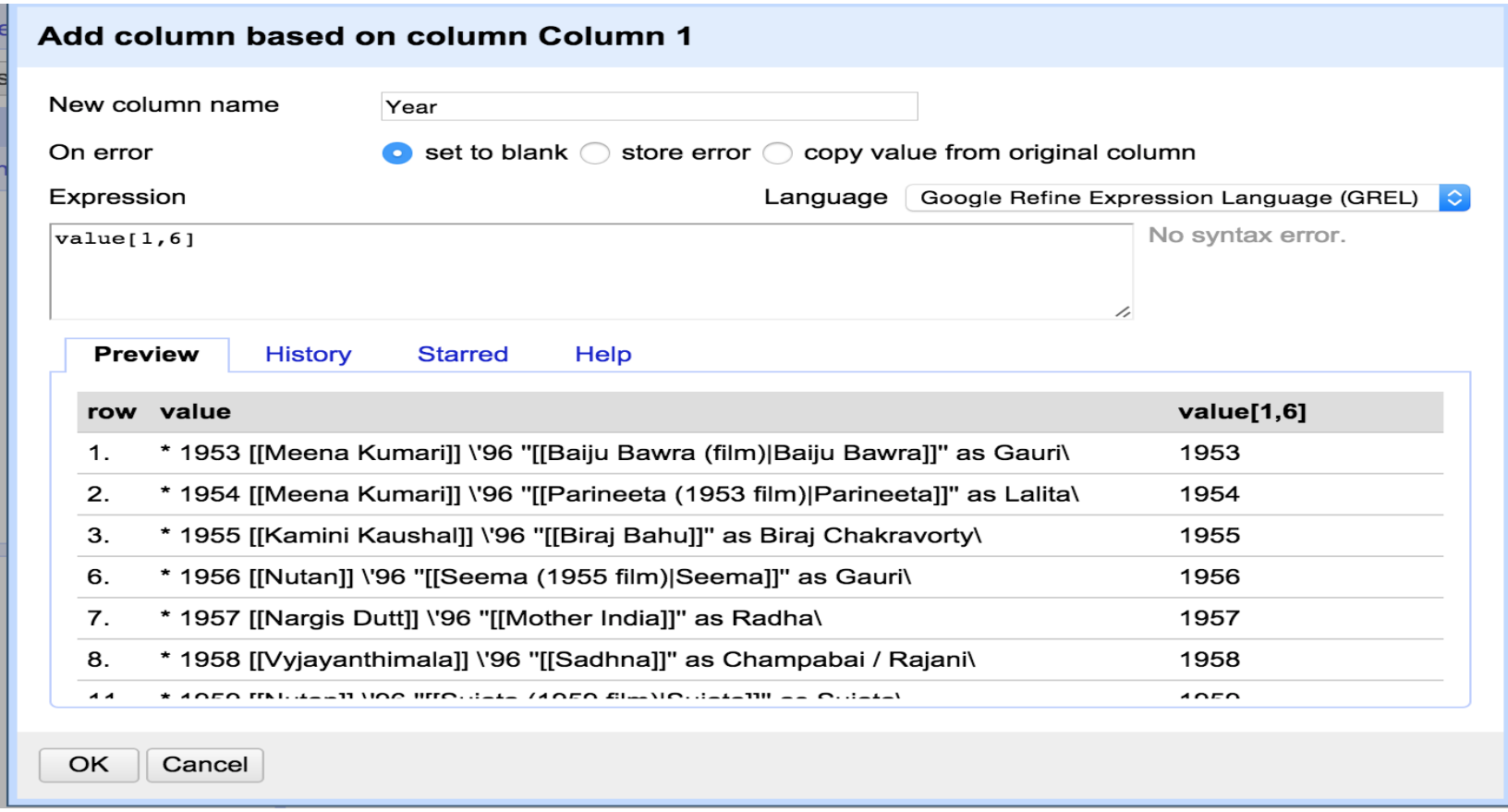


Chapter 2

Data Transformation

Add the “Year” Column

1. Then you will see a new column with values “true” or “false” occurs, facet it with text facet and click “true” to isolate the winners.
2. Click “add column based on this column” again and modify like below.



Chapter 2

Data Transformation

Delete year from the original column

1. Choose “Column 1”, click “Edit cells” then “Transform”;
2. Modify the expression like below. You will see the result.

Expression Language Google Refine Expre

```
value.substring(7)
```

All		Column 1	Year	Is Winner
1.		[[Meena Kumari]] \'96 "[[Baiju Bawra (film)|Baiju Bawra]]" as Gauri\	1953	true
2.		[[Meena Kumari]] \'96 "[[Parineeta (1953 film)|Parineeta]]" as Lalita\	1954	true
3.		[[Kamini Kaushal]] \'96 "[[Biraj Bahu]]" as Biraj Chakravorty\	1955	true
6.		[[Nutan]] \'96 "[[Seema (1955 film)|Seema]]" as Gauri\	1956	true
7.		[[Nargis Dutt]] \'96 "[[Mother India]]" as Radha\	1957	true
8.		[[Vyjayanthimala]] \'96 "[[Sadhna]]" as Champabai / Rajani\	1958	true
11.		[[Nutan]] \'96 "[[Sujata (1959 film)|Sujata]]" as Sujata\	1959	true
14.		[[Bina Rai]] \'96 "[[Ghunghat (1960 film)|Ghunghat]]" as Parvati / Jamuna\	1960	true
17.		[[Vyjayanthimala]] \'96 "[[Gunga Jumna]]" as Dhanno\	1961	true
20.		[[Meena Kumari]] \'96 "[[Sahib Bibi Aur Ghulam]]" as Chhoti Bahu\	1962	true

Chapter 2

Data Transformation

Deal with nominees (not winners)

1. Choose “false” in the “Is Winner” facet;
2. Choose “Column 1”, “Edit cells” then “Transform”;
3. Modify like below and you will see the result.

193 matching rows (256 total)

Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ pre

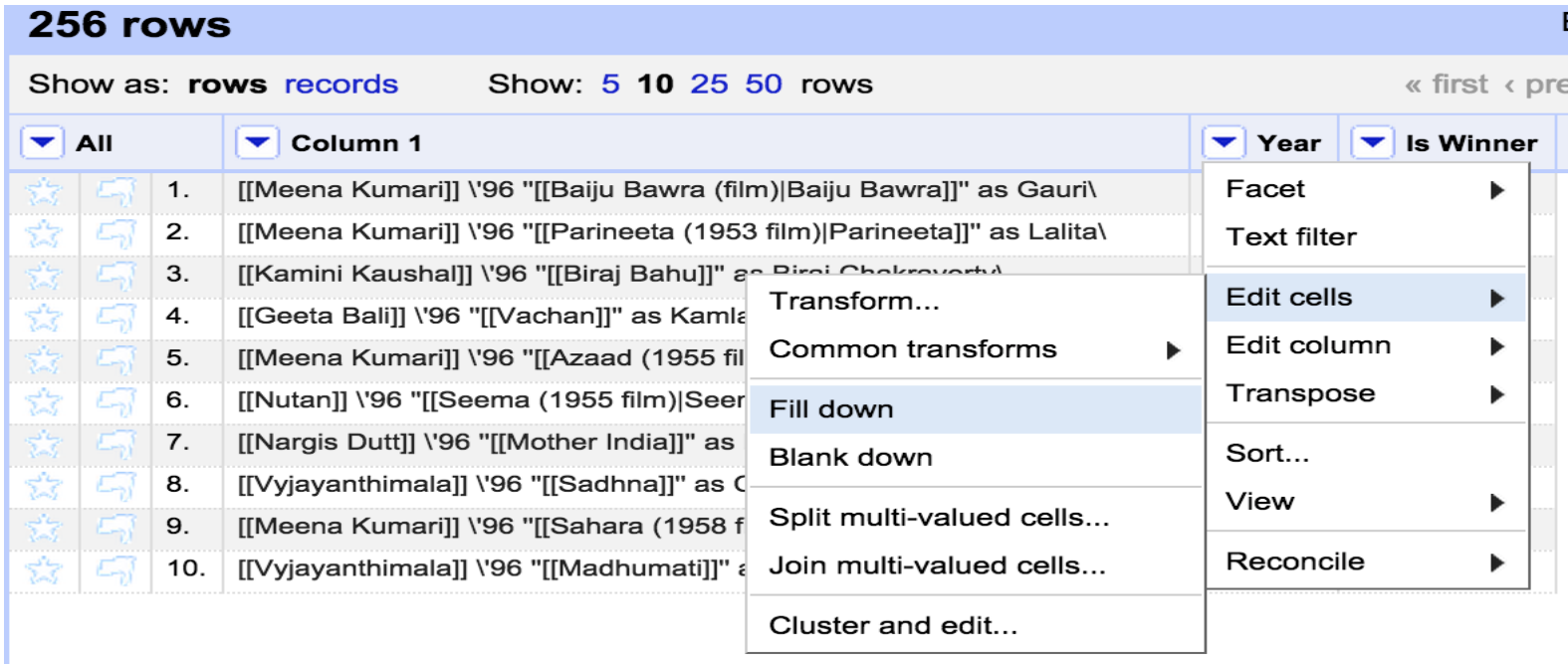
All		Column 1	Year	Is Winner
4.		[[Geeta Bali]] \'96 "[[Vachan]]" as Kamla\		false
5.		[[Meena Kumari]] \'96 "[[Azaad (1955 film)|Azaad]]" as Shobha\		false
9.		[[Meena Kumari]] \'96 "[[Sahara (1958 film)|Sahara]]" as Leela\		false
10.		[[Vyjayanthimala]] \'96 "[[Madhumati]]" as Madhumati / Madhavi / Radha\		false
12.		[[Mala Sinha]] \'96 "[[Dhool Ka Phool]]" as Meena Khosla\		false
13.		[[Meena Kumari]] \'96 "[[Chirag Kahan Roshni Kahan]]" as Ratna\		false
15.		[[Madhubala]] \'96 "[[Mughal-e-Azam]]" as [[Anarkali]]\		false
16.		[[Nutan]] \'96 "[[Chhalia]]" as Shanti\		false
18.		[[Padmini]] \'96 "[[Jis Desh Men Ganga Behti Hai]]" as Kammo\		false
19.		[[Saira Banu]] \'96 "[[Junglee]]" as Rajkumari 'Raj'\		false

Chapter 2

Data Transformation

Deal with nominees (not winners)

1. Move the “Is Winner” facet on the left side, you will see that the nominees’ rows are missing years.
2. To deal with this, we can just choose the “Year” column, click “Edit cells” then “Fill down” because the nominees’ year are the same with the winners’.



Chapter 2

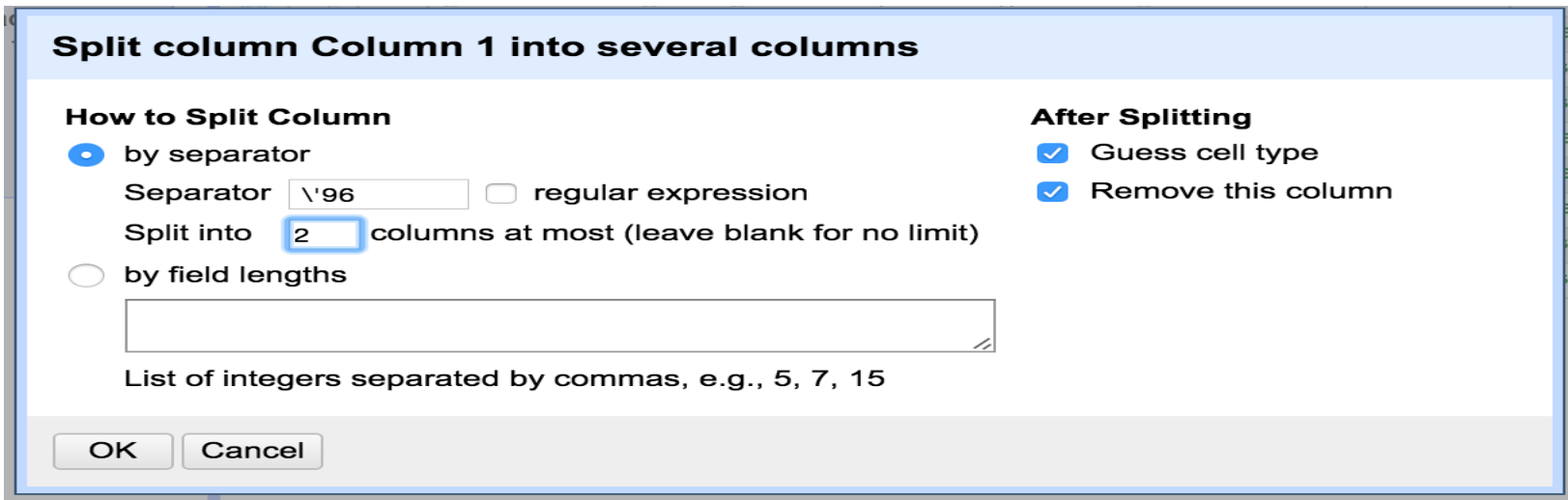
Data Transformation

Separate columns

In this part, we need to separate the first column into 3 separate columns: actress, film and character.

1.Choose “Column 1”, “Edit column” then “split into several columns”.

2.Type into the separator like below, pay attention to the space at the beginning and the end. If the separator are not these, just copy the separator between the actress and the film. Then type in “2” columns, and click “OK”.

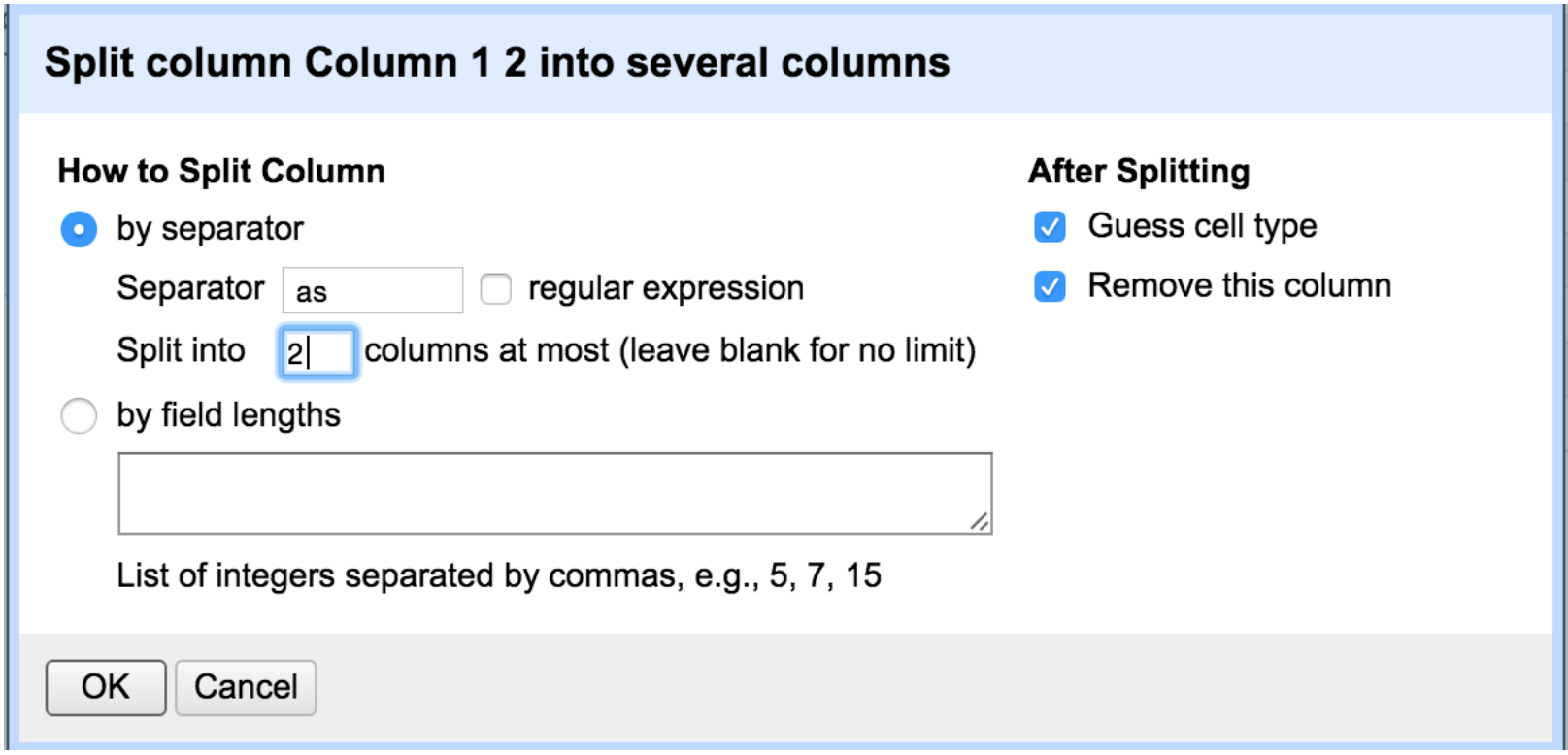


Chapter 2

Data Transformation

Separate columns

1. Choose the new column you added before;
2. Still choose “Edit column”-> “split into several columns”;
3. For this time, you need to type in “space->as->space”, and still 2 columns to separate.

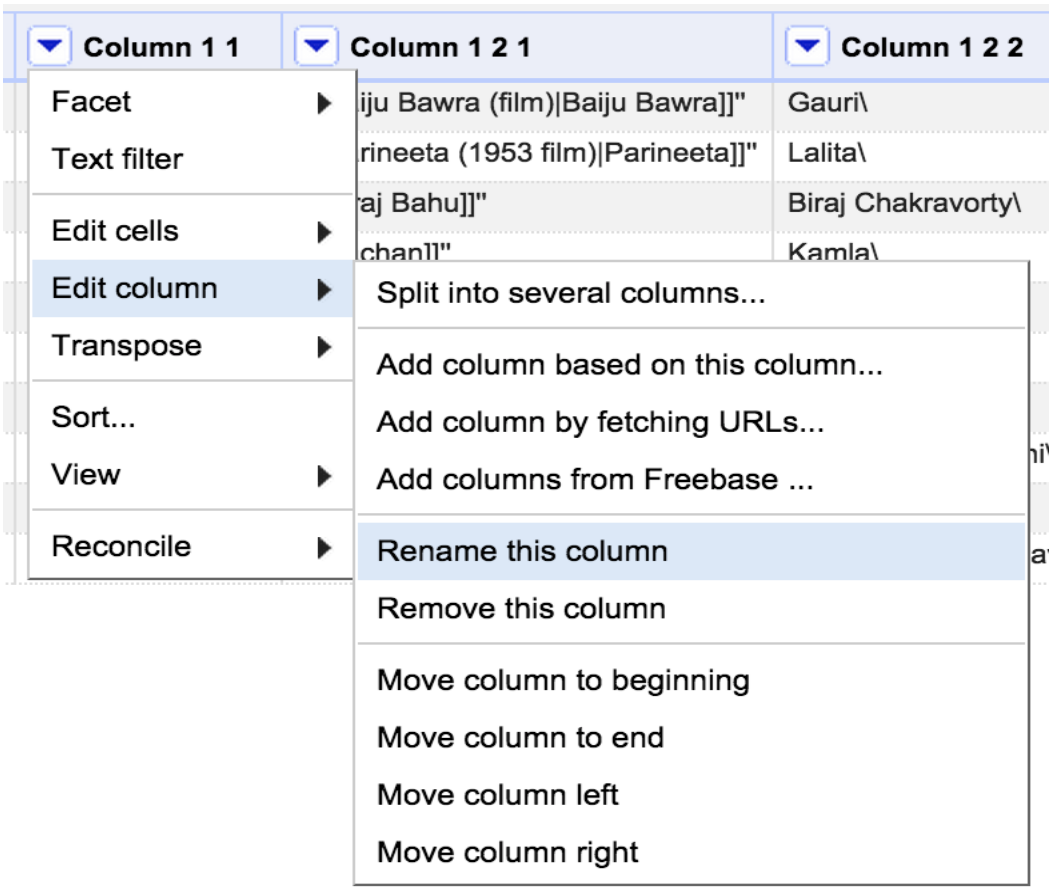


Chapter 2

Data Transformation

Rename columns

1. Choose “Edit column”-> “rename this column” to modify each column to “actress”, “film” and “character”.

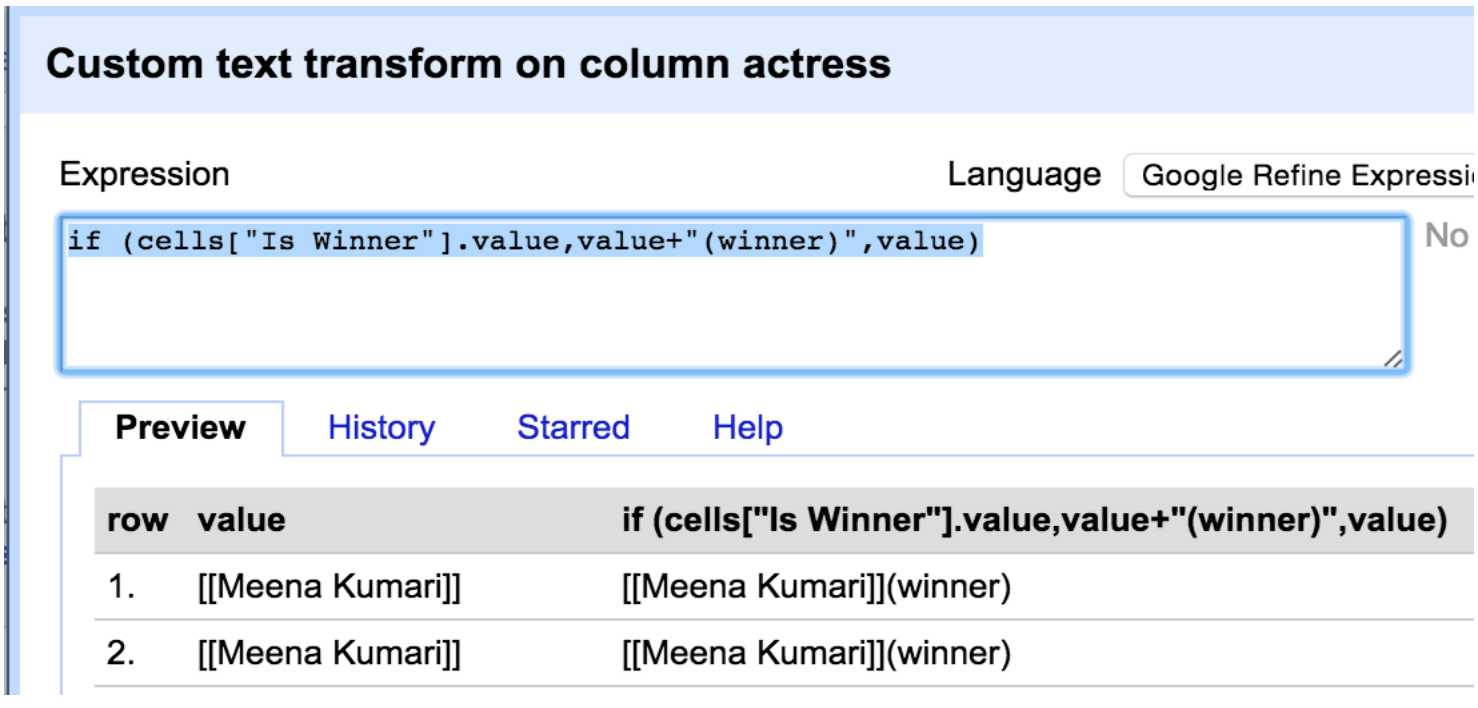


Chapter 2

Data Transformation

Do something different

We would like to know if the actress is the winner in the actress column. So we do like this:
1.In the actress column, choose “edit cell” and then “transform”;
2.Modify like below, you will see the result. (If there are some errors, just ignore this step due to the data errors).

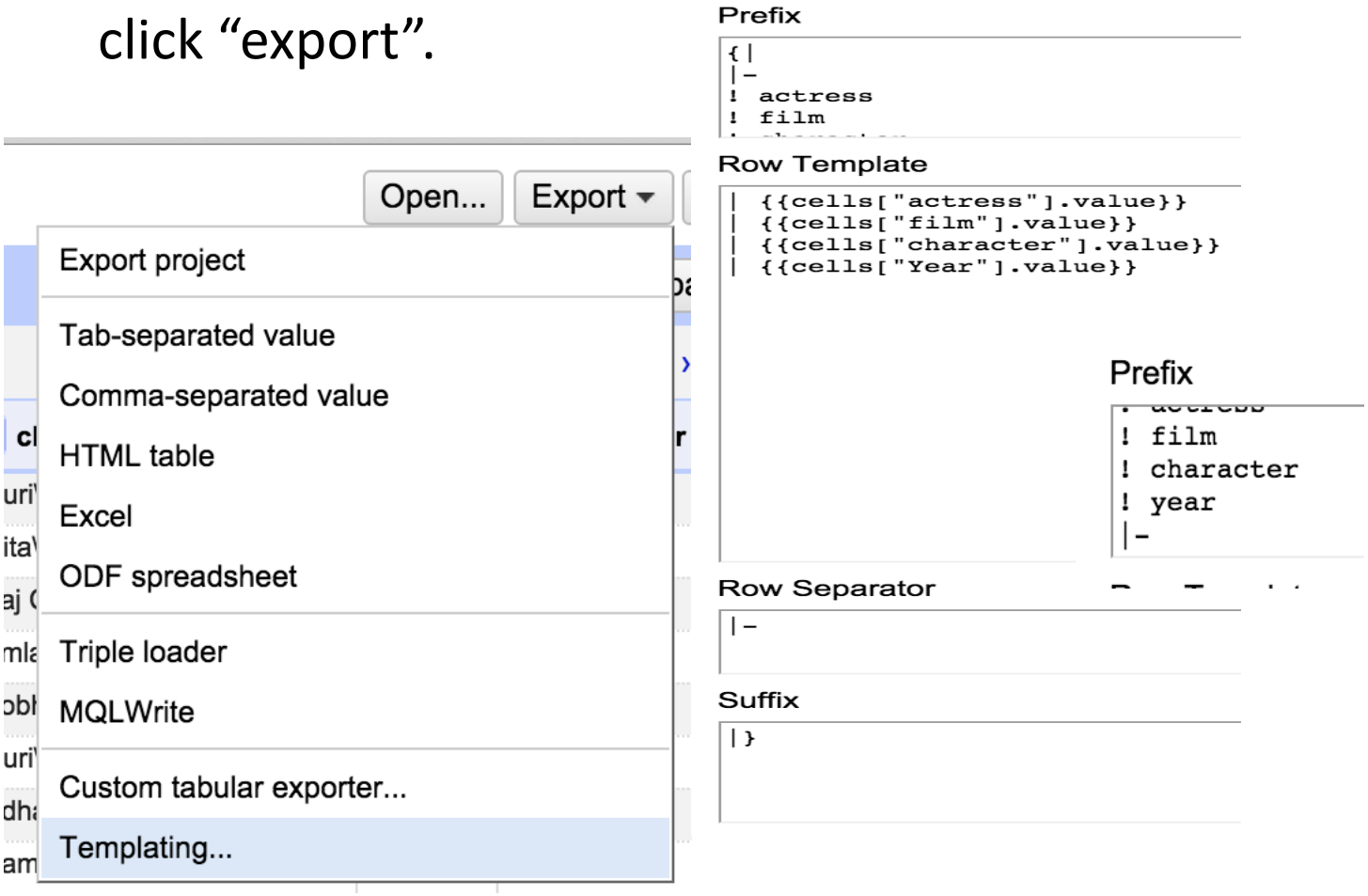


Chapter 2

Data Transformation

Export the data

After modifying the data, we need to export it. Click the “export” button then choose “Templating”. Modify the code like below, then click “export”.



Chapter 2

Data Transformation

Export the data

The file will be downloaded automatically, open the file, copy the codes and paste them in the original wiki editor, click “show preview”, you will see a table like this. (The layout may be a little different on your computer due to the different settings).

actress	film	character	year
Meena Kumari	*Parineeta*	Lalita	1954
Kamini Kaushal	*Biraj Bahu*	Biraj Chakravorty	1955
Geeta Bali	*Vachan*	Kamla	1955
Meena Kumari	*Azaad*	Shobha	1955
Nutan	*Seema*	Gauri	1956
Nargis Dutt	*Mother India*	Radha	1957
Vyjayanthimala	*Sadhna*	Champabai / Rajani	1958
Meena Kumari	*Sahara*	Leela	1958
Vyjayanthimala	*Madhumati*	Madhumati / Madhavi / Radha	1958
Nutan	*Sujata*	Sujata	1959
Mala Sinha	*Dhool Ka Phool*	Meena Khosla	1959
Meena Kumari	*Chirag Kahan Roshni Kahan*	Ratna	1959
Bina Rai	*Ghunghat*	Parvati / Jamuna	1960
Madhubala	*Mughal-e-Azam*	Anarkali	1960
Nutan	*Chhalia*	Shanti	1960
Vyjayanthimala	*Gunga Jumna*	Dhanno	1961