Exploring enigma.io

enigma.io is a company that is aggregating open data from accross the web into one platform and providing powerful ways to search and filter the data to discover insight. In this example we shall look at employment trends in software companies over the last number of years. By keeping the data within the enigma.io platform allows the use of powerful filters wihtout having to download, store and filter all of the data locally.

Step 1

Browse to enigma.io and search for **google**, you will be prompted to register an account or login.

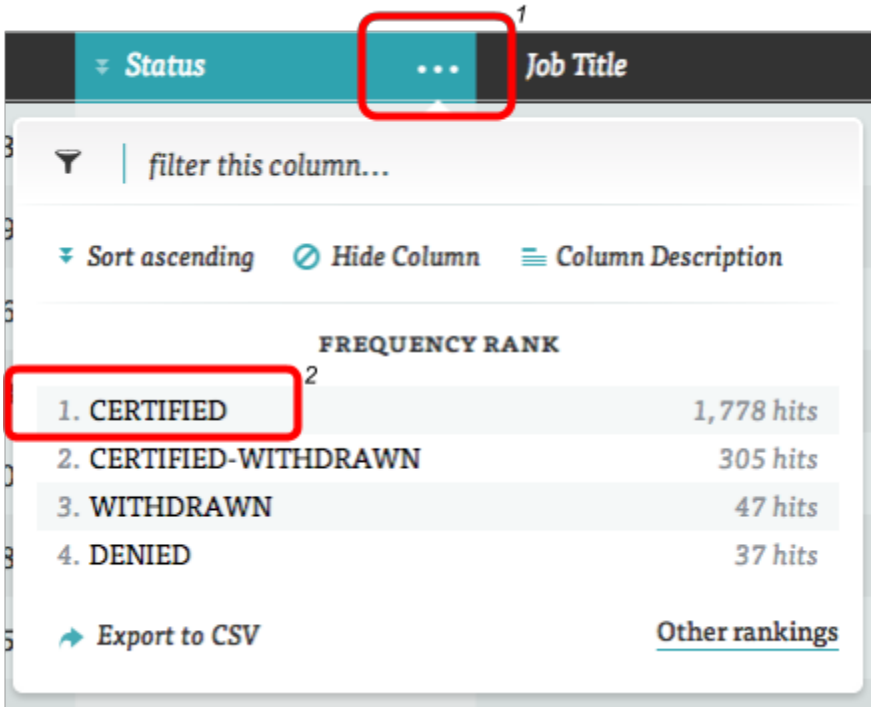
Step 2

The result screen will display all the datasets that contain our search term. For this exercise we are going to look at the **Office of Foreign Labor Certification** datasets.

This dataset lists the non-immigrant visa applications in the US allowing companies to temporarily employ foreign workers. Start by exploring the lastest dataset to see what sort of information it contains.

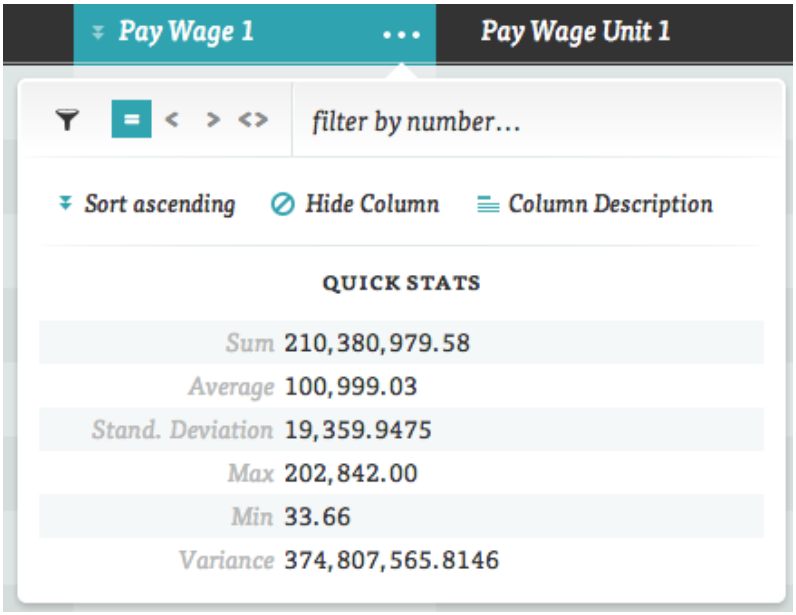
Step 3

For this exercise we would like to discover employment trends within some of the major software companies and chart the rise and fall of certain job titles. Starting with the latest dataset we first need to filter out those applications that were not certified. The data relating to this is contained in the **status** column. To view a summary of which types of status exist for Google, click the three **...** that appear when you hover over the status column.



We want to filter this column for only those applications that were certified. From the drop down simply click the **certified** option in the frequency rank list.

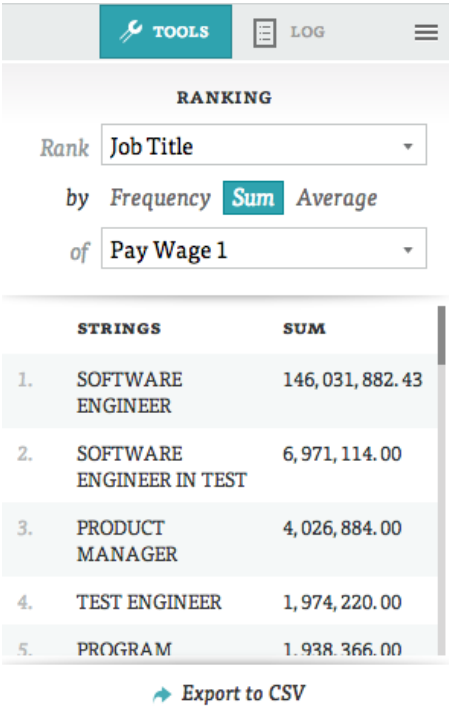
With the data filtered it is now possible to look at how much Google spent on employing temporary non-immigrant workers within the US. To view the stats, take a look at the drop down stats from the **Pay Wage 1** column.



Step 4

In order to make sense of the data we are going to group it by **Job Title**, to do this first pull in the **tools sidebar**.

We can then group the data by **Job Title** summed by the numerical values in the **Pay Wage 1** column.



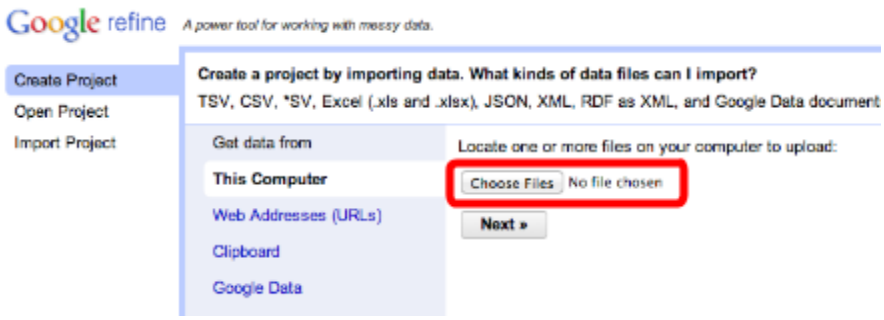
Finally we can export this data to csv in order to build a dataset for visual analysis.

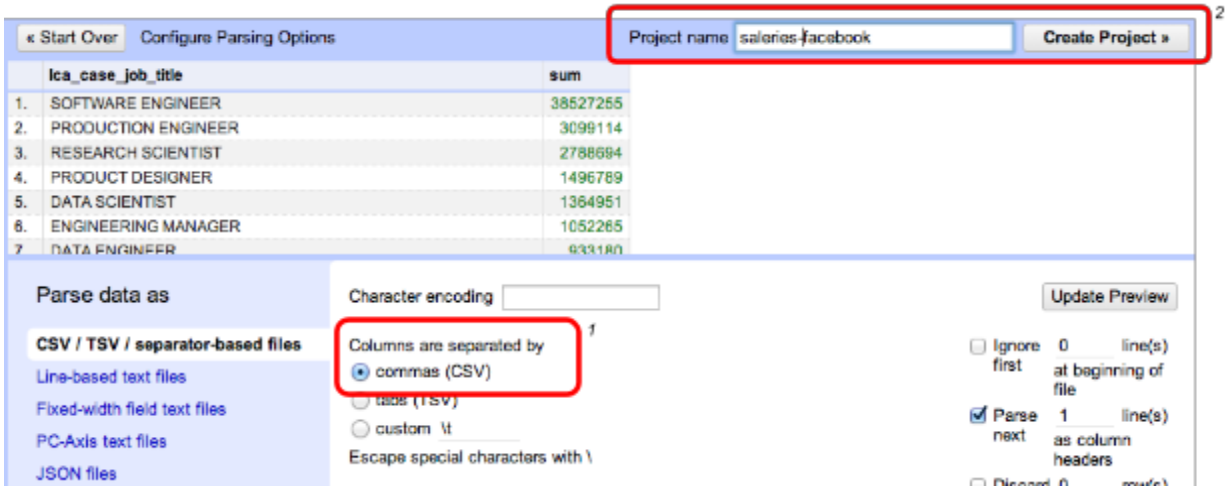
Step 5

Now repeat all the steps for **facebook** in order to end up with two CSV files.

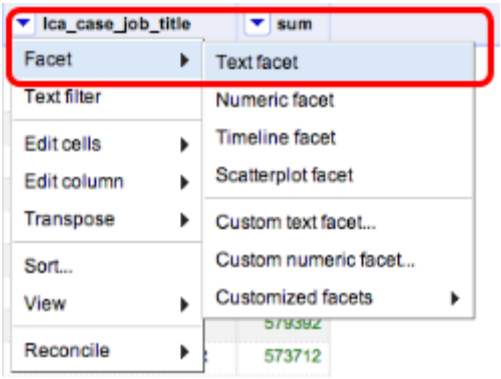
Step 6

The next step involves cleaning the data. All data is dirty, so we need a copy of Google/Open Refine. Upload both datasets into refine, creating a separate project for each one.

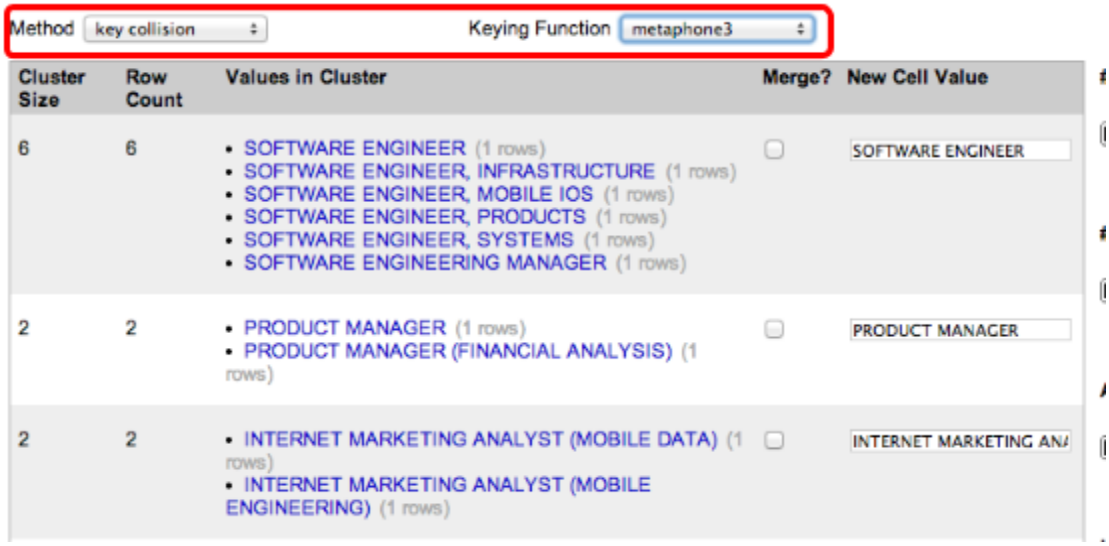




With the data loaded, apply a **text facet** to the job title column. From the facet box, click the **cluster** button.



The clustering function allows you to fix errors in the dataset by grouping groups of cells. Select a number of different methods and keying functions in order to group together rows where the jobs are very similar in nature.



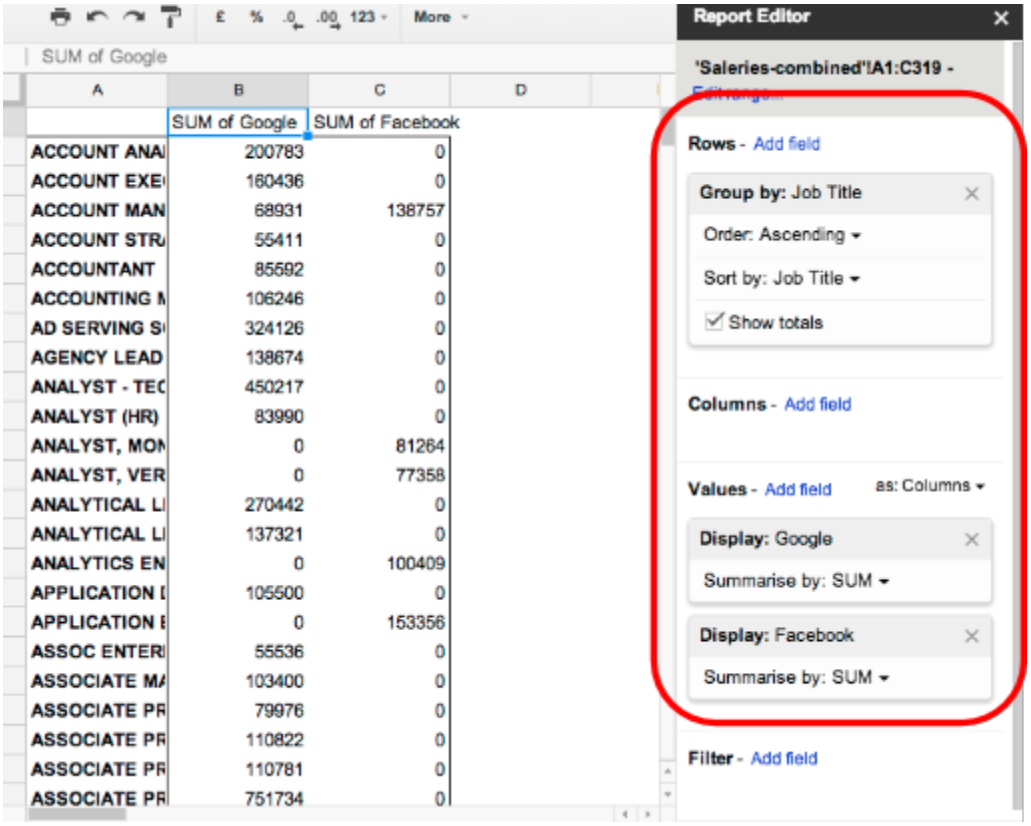
Step 7

With both datasets cleaned it is time to align them into one single dataset. To do this upload them both to Google sheets at docs.google.com. Once done open the one relating to google and create a new column to the right of the google salary data (you might also want to rename the columns for clarity). Title this column **facebook salary**. Finally scroll to the bottom of this dataset.

The next stage is to copy **just** the job title data from the facebook dataset and paste it under the job titles in the google spreadsheet. Now do the same with the salary data **remembering to put it in column 3 and not column 2!**.

At this point we should have a several hundred row spreadsheet with 3 columns of data. Now we can group the data again using a pivot table.

To create a pivot table in google sheets select the **pivot table report** option from the **data** menu.



By selecting the correct **rows** and **values** from the report editor (see above) we can group the data together, making our final report of facebook vs google job title and spending.

Extension exercises

Can you think of a way to visualise the data?

What are the problems with the data and how might you go about resolving them?

Can you add data from other companies and years to tell a more interesting story?