French Social Media Data – Exercise

Background

The data in the spreadsheet are a subset of a larger number of fields captured from Twitter, Facebook, Reddit and Digg as a test during the French Presidential Election of 2012 – see https://en.wikipedia.org/wiki/French_presidential_election,_2012 for the contestants.

What can you learn from the data with some simple manipulation in Excel?

Examine the spreadsheet

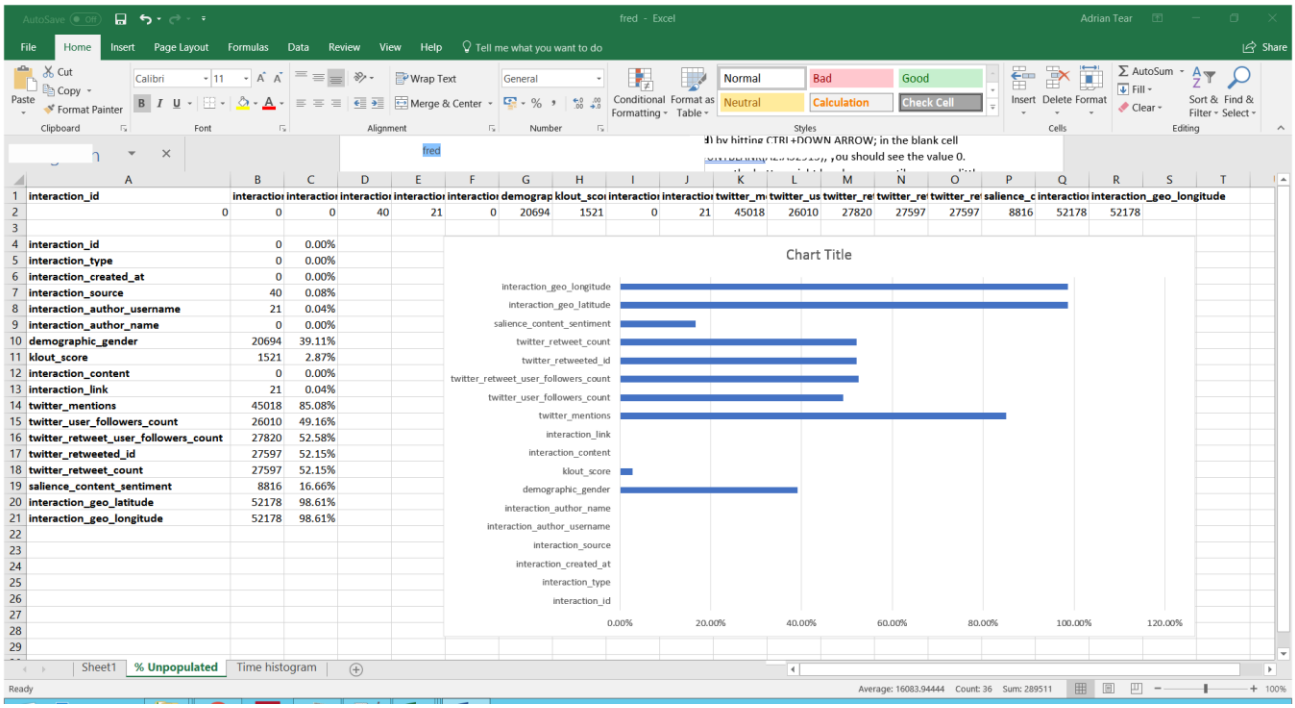
1. How many fields (columns) are there?
2. How many records (rows) are there?
3. What do you notice about the data (clues below)?

Do all the fields contain values?

Many of the fields are only partially populated (i.e. containing records) with plenty of blank, or NULL, values. How can you count these to get a feel for sparsity across the dataset?

Tip: Use the Excel function COUNTBLANK to calculate the number of NULL values in each field. Go to the bottom of the first column (**interaction_id**) by hitting CTRL+DOWN ARROW; in the blank cell beneath the last record type the formula =COUNTBLANK(A2:A52915), you should see the value 0. Highlight this cell and place the mouse cursor over the bottom right hand corner until you see a little cross. Click on this cross and pull it rightwards until you've copied the formula across all the columns in the spreadsheet.

Question: Can you calculate the number of blanks as a percentage of the total and produce a meaningful graph?



Tip: You'll need to know the total number of records, create another formula and – ideally – copy and transpose (Paste Special…) these values into a new sheet. Use the plus button inside a circle at the bottom of Excel to create the new sheet; select the first row and the COUNTBLANK row you created above. Hit CTRL+C to copy and paste these values into a new sheet. Select them then hit CTRL+C again then go to the Home ribbon, Paste, Paste Special, Transpose. Use a formula to create percentages and insert a graph. Your solution may look something like the above.

Question: Can you order the graph so the most populated columns are shown first? Does it make more sense to show the %populated or the %unpopulated?

How are the records distributed in time?

1. When was the first record created?
2. When was the last record created?
3. How much time has elapsed?

The field **interaction_created_at** contains record creation date/time (and timezone offset). Unfortunately, Excel cannot understand the human readable format…

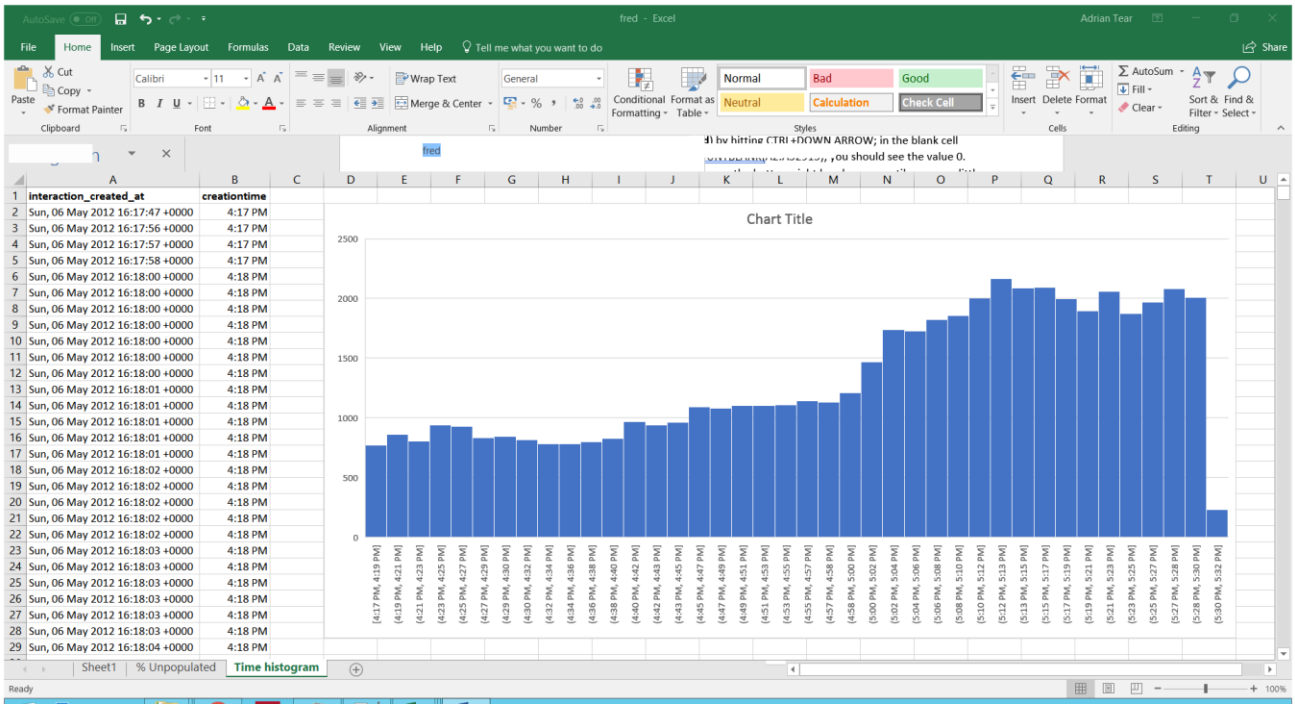
Tip: Click on the column heading to select all rows in this column, hit CTRL+C to copy these into a new sheet (click the cross within the circle at the bottom of Excel to create one). In cell B1 type a heading for your new column (e.g. **creationtime**). In B2 devise a formula that will turn the date/time literal into a time data type. Look at the first record:

```
Sun, 06 May 2012 16:17:47 +0000
12345678901234567890123456789012
```

The numbers below the date/time show character positions 1 through 10 (shown as 0); the field length is 31 overall. You can use the function MID to select parts of this field and pass days, months and years into Excels TIME function using the formula:

=TIME(MID(A2,18,2),MID(A2,21,2),MID(A2,24,2))

Work out what's going on! The TIME function expects hours, minutes and seconds. See how the MID function is used to select these (e.g. starting at character 18 reading 2 characters to get hours) from the long (human readable) date/time with timezone offset format and convert them into something Excel can understand.

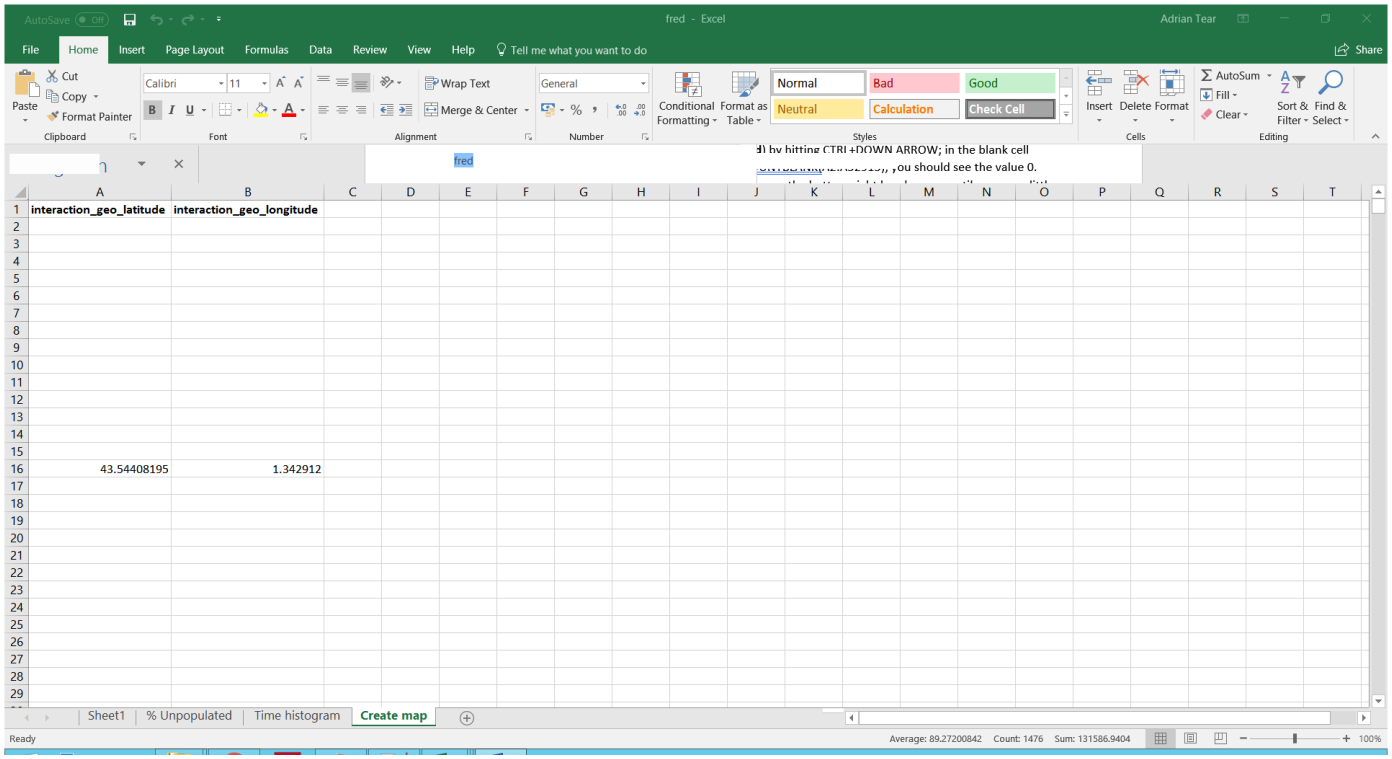


Once you've got your formula working select the cell you created it in and double click the little cross on the bottom right hand corner; this will copy the formula all the way down to the bottom of the sheet. Select your **creationtime** column and from the Insert ribbon insert a Histogram. It should look something like the above.

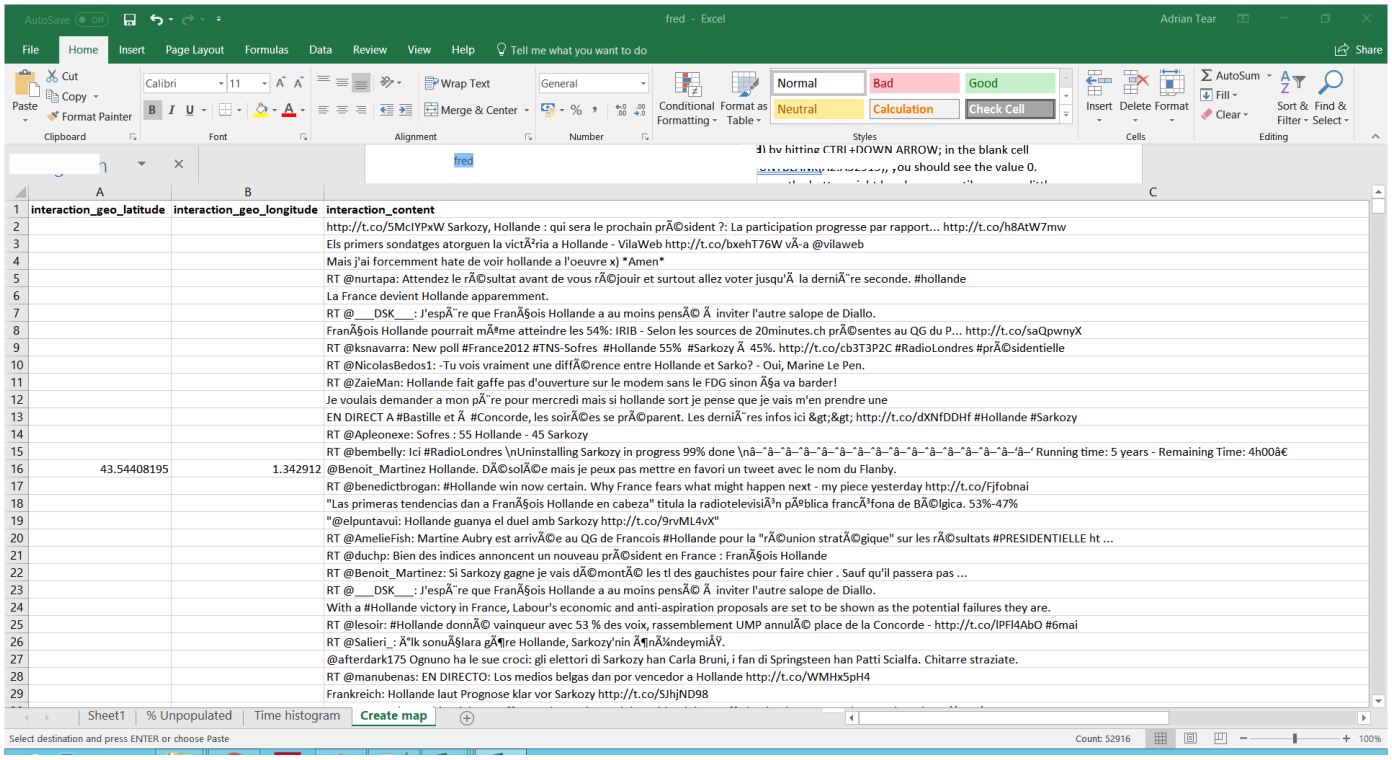
Can any of the Tweets be mapped?

You will notice the spreadsheet contains two fields/columns named **interaction_geo_latitude** and **interaction_geo_longitude** – some users, when they Tweet, agree to post their geographic coordinates alongside their message. These coordinates are saved from the Global Positioning System (GPS) chip in their phones etc.

Excel does not have great capabilities for mapping from Latitude and Longitude, but this can be overcome! Copy the two columns named above into another new sheet. It should look something like this:

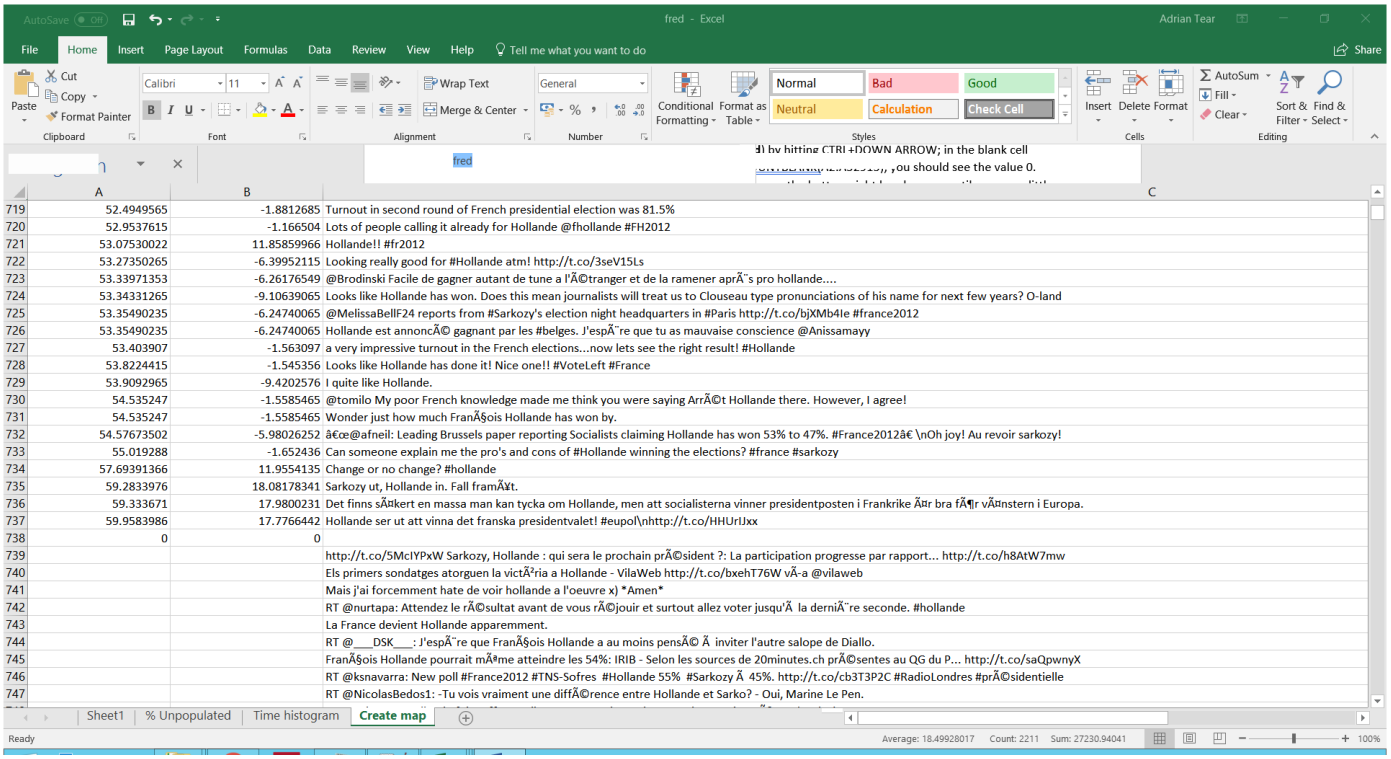


Then copy/paste the **interaction_content** column into this sheet as well, it should look like this:



From the first exercise you'll know that very few records (how many?) are geocoded. To map them select the whole sheet (by clicking on the little arrow top right hand side near A1) then hold down CTRL while unselecting the first row with the column names. Now, in the Home ribbon, click Sort & Filter on the right hand side and select Smallest to Largest. This will put the non-NULL records together at the top of the sheet.

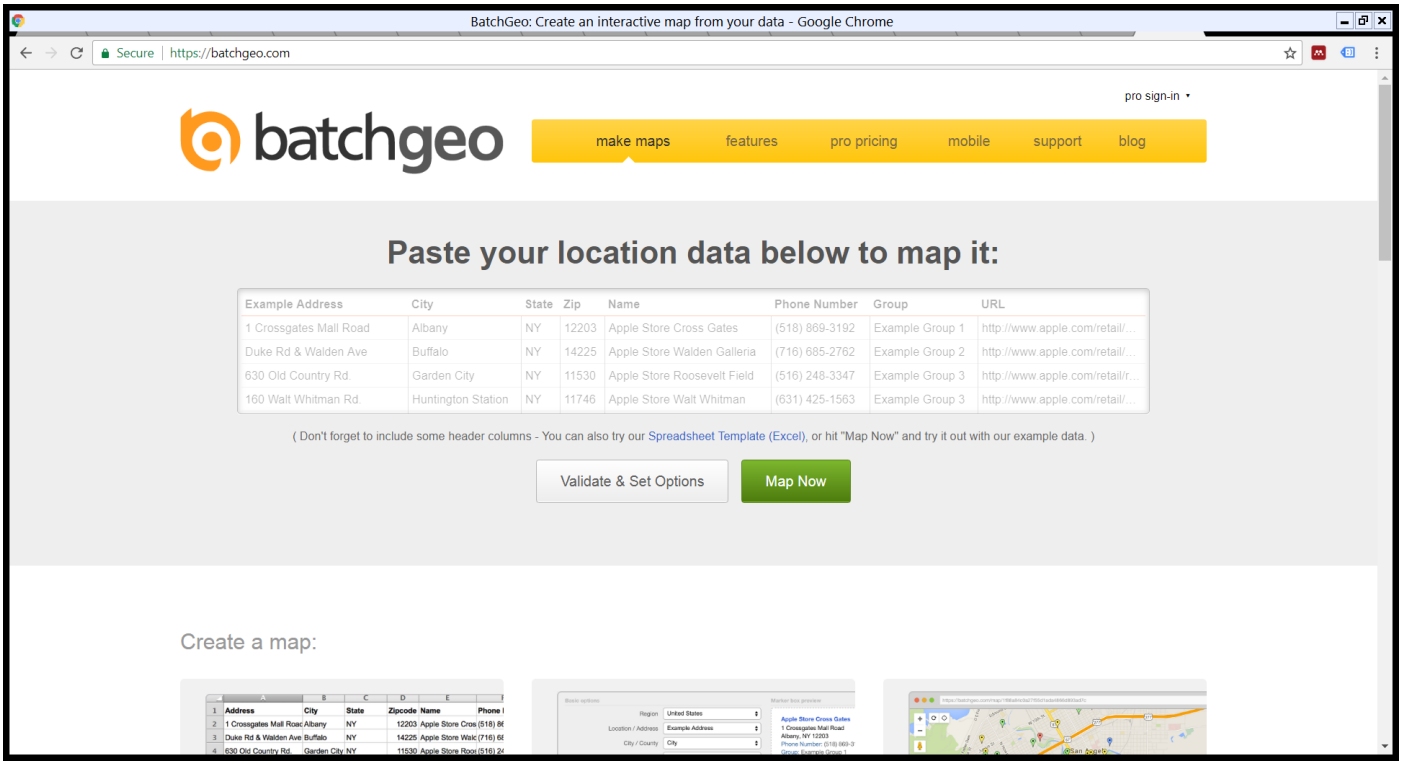
Use the mouse to select the cells that have Latitude and Longitudes along with the text of the Tweet and hit CTRL+C to copy them. You should go from row 1 to row 737 as below:



	A	B	C
719	52.4949565	-1.8812685	Turnout in second round of French presidential election was 81.5%
720	52.9537615	-1.166504	Lots of people calling it already for Hollande @fhollande #FH2012
721	53.07530022	11.85859966	Hollande!! #fr2012
722	53.27350265	-6.39952115	Looking really good for #Hollande atm! http://t.co/3seV15Ls
723	53.33971353	-6.26176549	@Brodinski Facile de gagner autant de tune a l'Ã©tranger et de la ramener aprÃ¨s pro hollande....
724	53.34331265	-9.10639065	Looks like Hollande has won. Does this mean journalists will treat us to Clouseau type pronunciations of his name for next few years? O-land
725	53.35490235	-6.24740065	@MelissaBellF24 reports from #Sarkozy's election night headquarters in #Paris http://t.co/bjXMb4Ie #france2012
726	53.35490235	-6.24740065	Hollande est annoncÃ© gagnant par les #belges. J'espÃ¨re que tu as mauvaise conscience @Anissamayy
727	53.403907	-1.563097	a very impressive turnout in the French elections...now lets see the right result! #Hollande
728	53.8224415	-1.545356	Looks like Hollande has done it! Nice one!! #VoteLeft #France
729	53.9092965	-9.4202576	I quite like Hollande.
730	54.535247	-1.5585465	@tomilo My poor French knowledge made me think you were saying ArrÃ©t Hollande there. However, I agree!
731	54.535247	-1.5585465	Wonder just how much FranÃ§ois Hollande has won by.
732	54.57673502	-5.98026252	â€œ@afneil: Leading Brussels paper reporting Socialists claiming Hollande has won 53% to 47%. #France2012â€ \nOh joy! Au revoir sarkozy!
733	55.019288	-1.652436	Can someone explain me the pro's and cons of #Hollande winning the elections? #france #sarkozy
734	57.69391366	11.9554135	Change or no change? #hollande
735	59.2833976	18.08178341	Sarkozy ut, Hollande in. Fall framÃ¥t.
736	59.333671	17.9800231	Det finns sÃ¤kert en massa man kan tycka om Hollande, men att socialisterna vinner presidentposten i Frankrike Ã¤r bra fÃ¶r vÃ¤rlden i Europa.
737	59.9583986	17.7766442	Hollande ser ut att vinna det franska presidentvalet! #eupol\nhttp://t.co/HHUrIJxx
738	0	0	
739			http://t.co/5MclYPxW Sarkozy, Hollande : qui sera le prochain prÃ©sident ?: La participation progresse par rapport... http://t.co/h8AtW7mw
740			Els primers sondatges atorguen la victÃ²ria a Hollande - VilaWeb http://t.co/bxehT76W vÃ-a @vilaweb
741			Mais j'ai forcemment hate de voir hollande a l'oeuvre x) *Amen*
742			RT @nurtapa: Attendez le rÃ©sultat avant de vous rÃ©jouir et surtout allez voter jusqu'Ã la derniÃ¨re seconde. #hollande
743			La France devient Hollande apparemment.
744			RT @___DSK___: J'espÃ¨re que FranÃ§ois Hollande a au moins pensÃ© Ã inviter l'autre salope de Diallo.
745			FranÃ§ois Hollande pourrait mÃªme atteindre les 54%: IRIB - Selon les sources de 20minutes.ch prÃ©sentes au QG du P... http://t.co/saQpwnyX
746			RT @ksnavarra: New poll #France2012 #TNS-Sofres #Hollande 55% #Sarkozy Ã 45%. http://t.co/cb3T3P2C #RadioLondres #prÃ©sidentielle
747			RT @NicolasBedos1: -Tu vois vraiment une diffÃ©rence entre Hollande et Sarko? - Oui, Marine Le Pen.

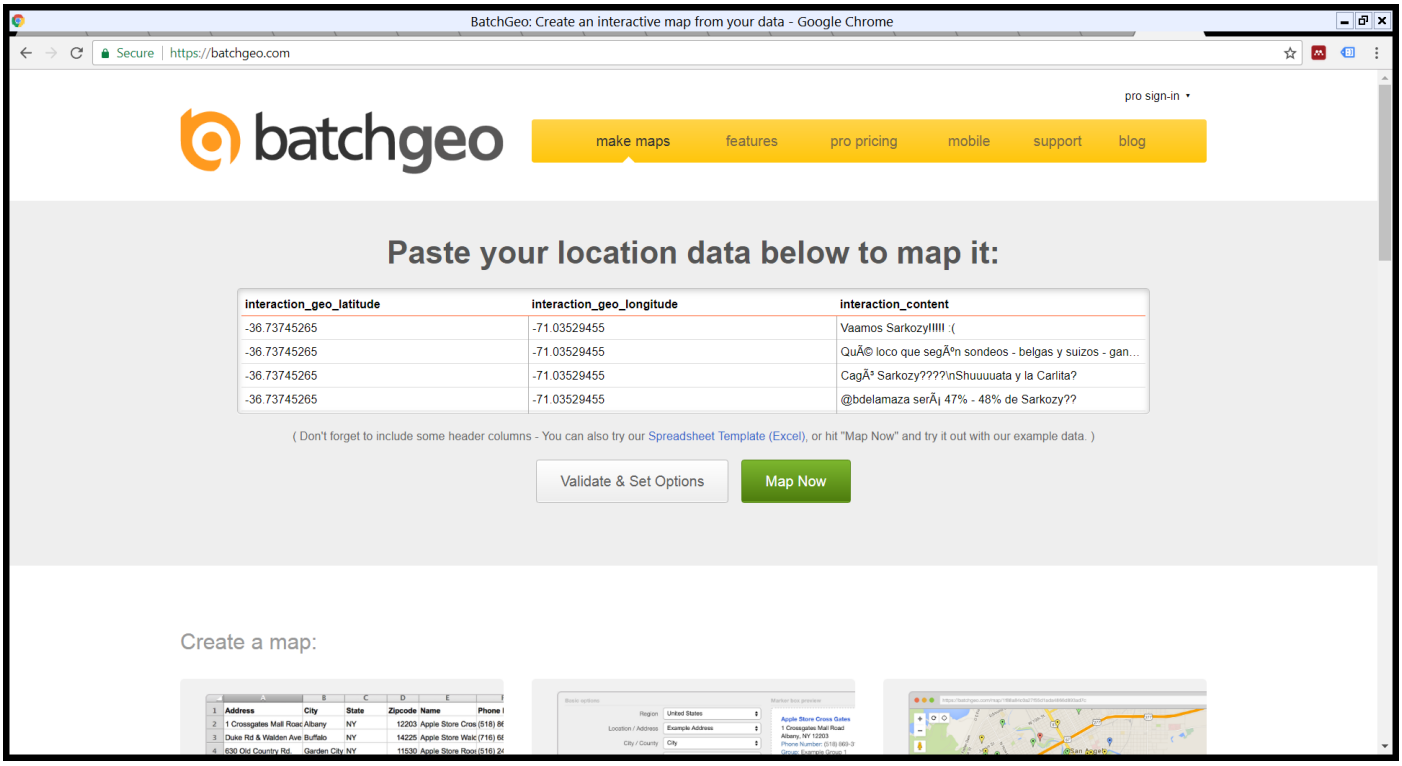
In the Chrome web browser go to:

https://batchgeo.com/

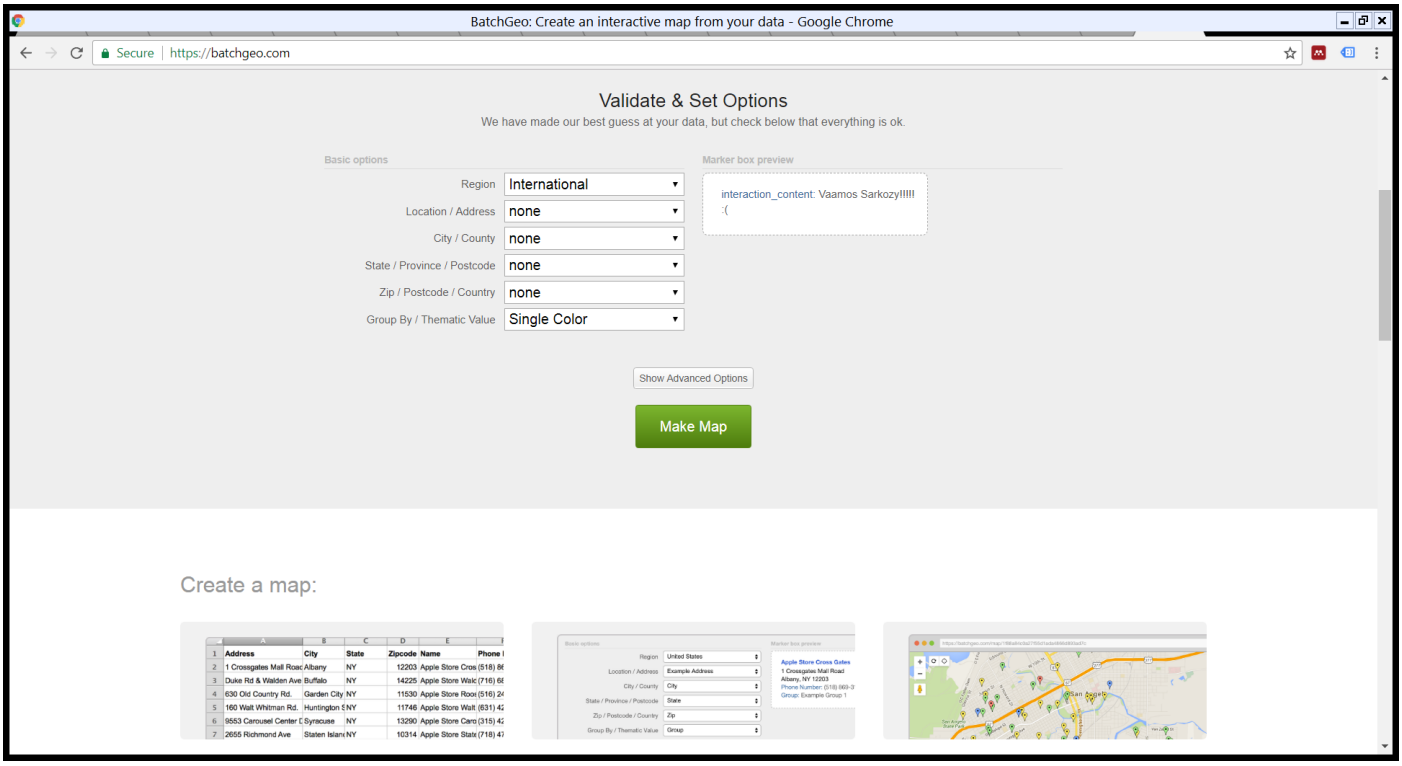


Paste the cells you have copied from Excel into the obvious box (it goes green outlined when you click on it)…

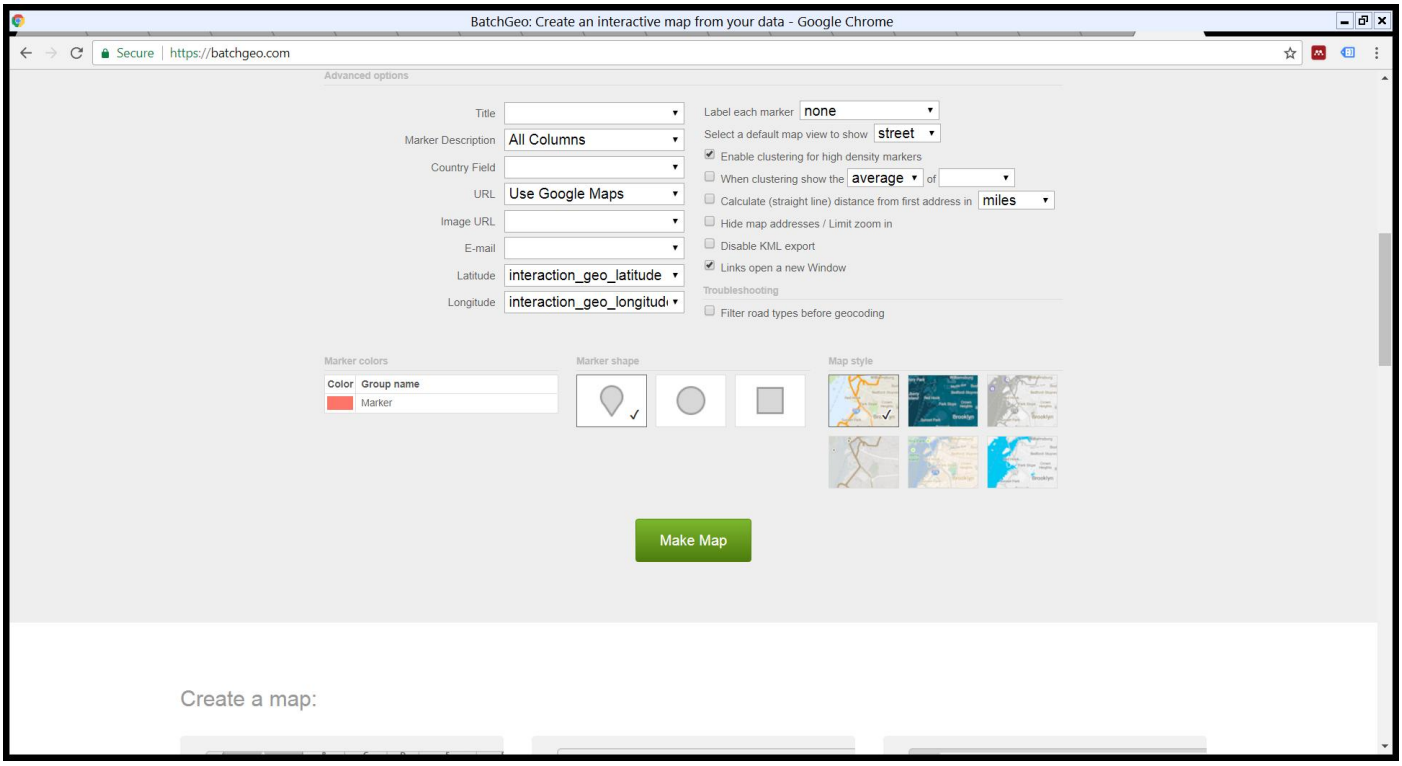
Once it's pasted in you should see something like the below:



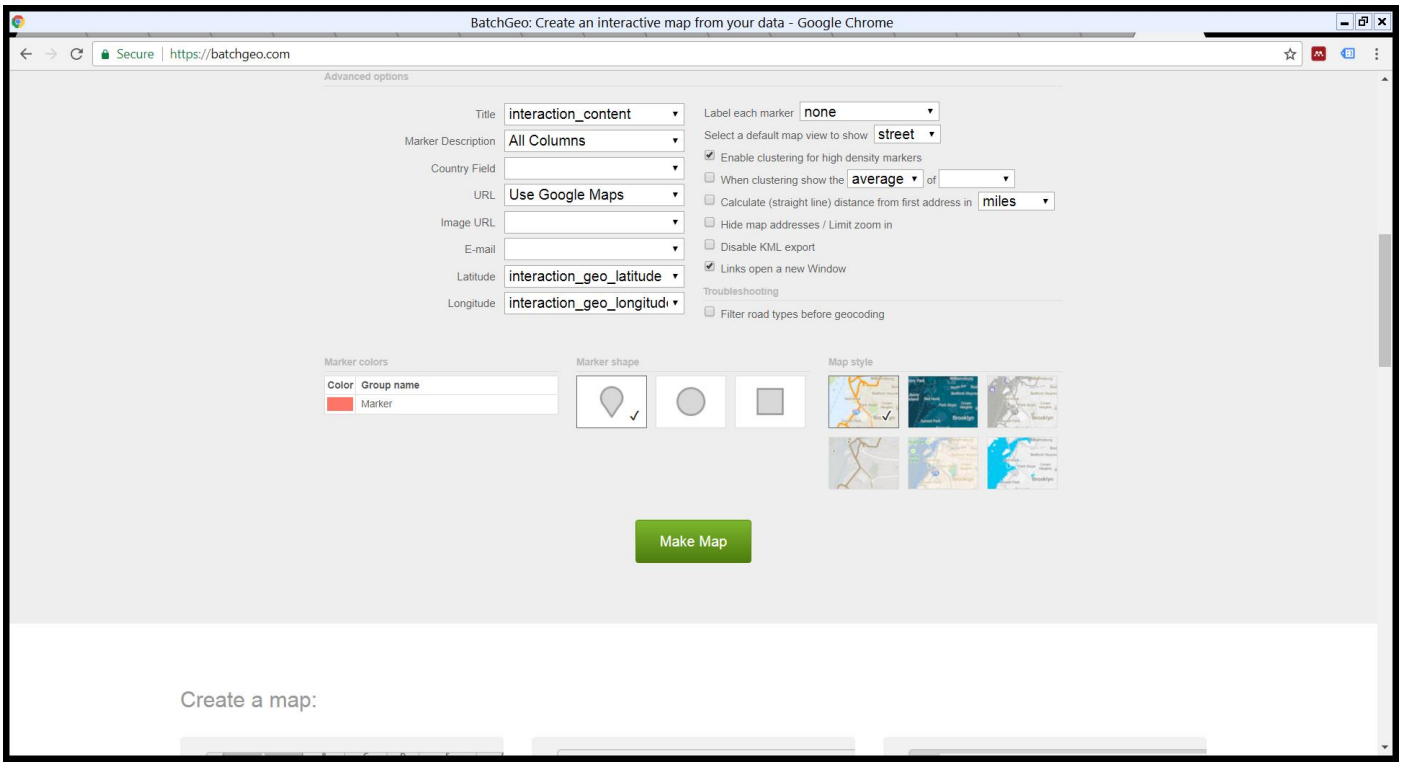
Next click on the button Validate & Set Options…



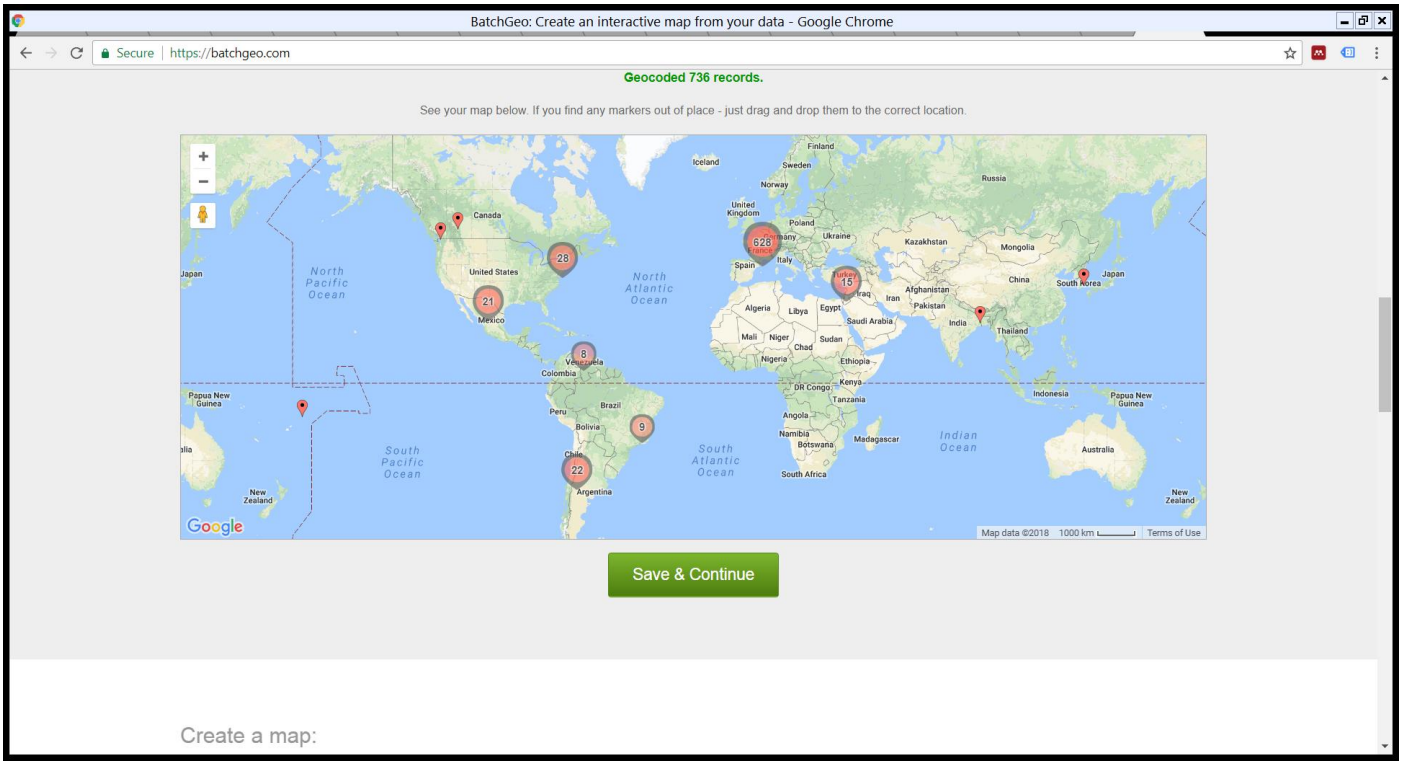
Then click the further button Show Advanced Options…



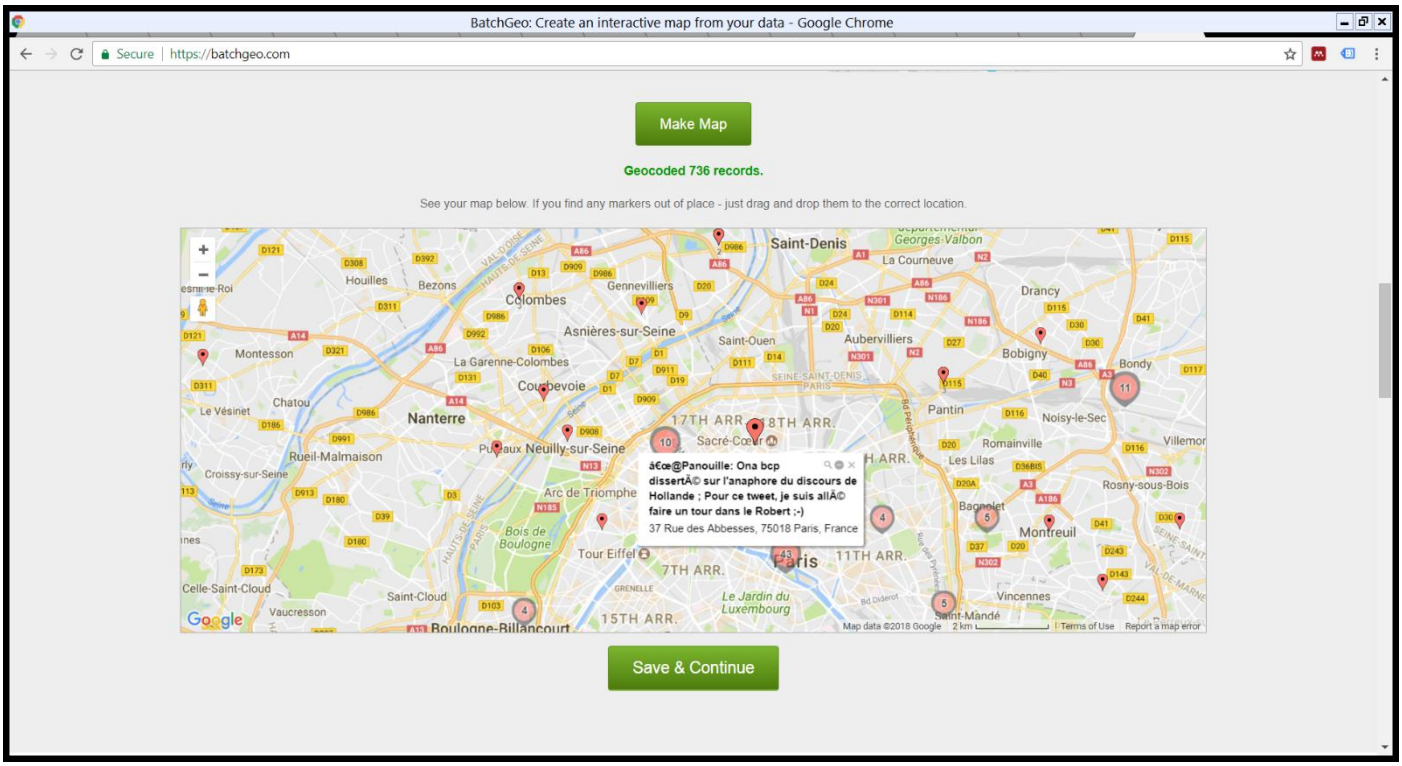
It should have correctly picked out the Latitude and Longitude columns. Adjust the Title column to use the field **interaction_content** as below:



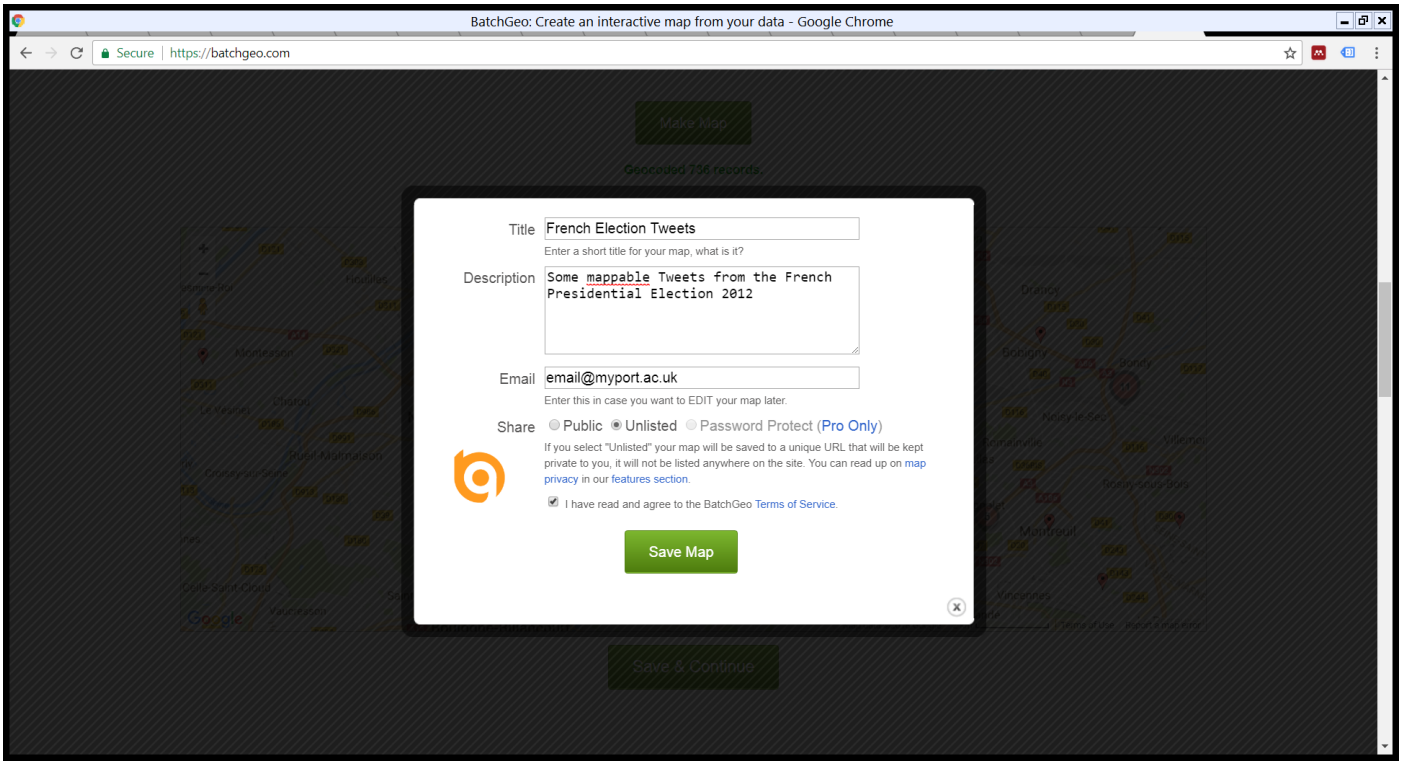
Finally, click the Make Map button to plot your geolocated Tweets…



You will see where most geolocated Tweets are concentrated world-wide. Zoom in on Paris (or wherever you like) to click on individual Tweets and see what the user is saying…



This processing has used Excel and BatchGeo to produce the map. You can save this using the Save & Continue button if you wish; please **do not Share** the map, **keep it Unlisted** and email yourself the link…

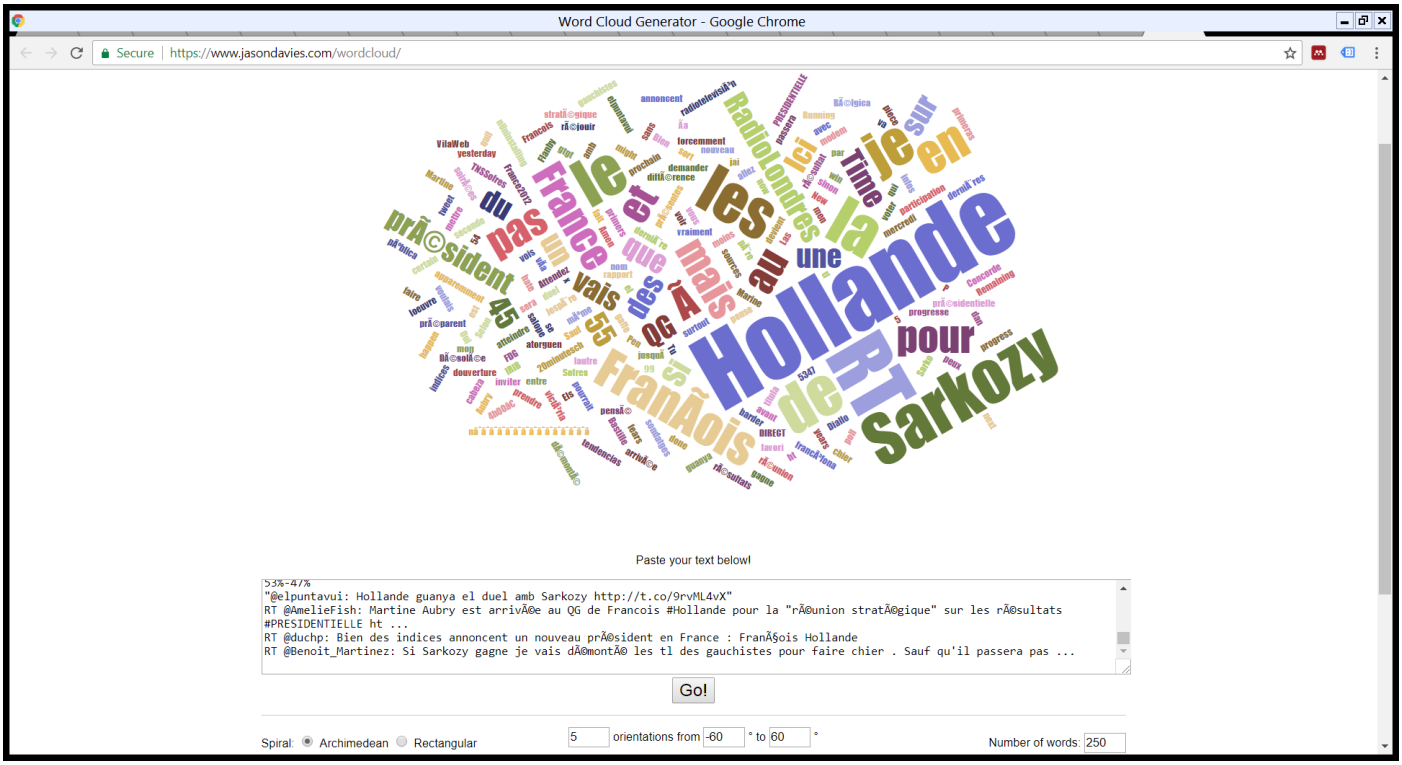


The Department uses Geographic Information System (GIS) software to perform these sorts of operations on, potentially, much larger data sets. Data may also be saved in the Relational Database Management System (RDBMS) called Oracle – on in other relational databases – for query and analysis. Ask one of the supervisors for more detail if you're interested or consider the GIS module as an option for your Degree programme…

What else?

Even though this spreadsheet only contains a subset of the many fields/columns saved when downloading data from Twitter (see the file **French Social Media Data - All Data (for those interested).xlsx** if you're interested) there are still a few other things you can easily achieve with some copying/pasting and free online tools…

Wordclouds



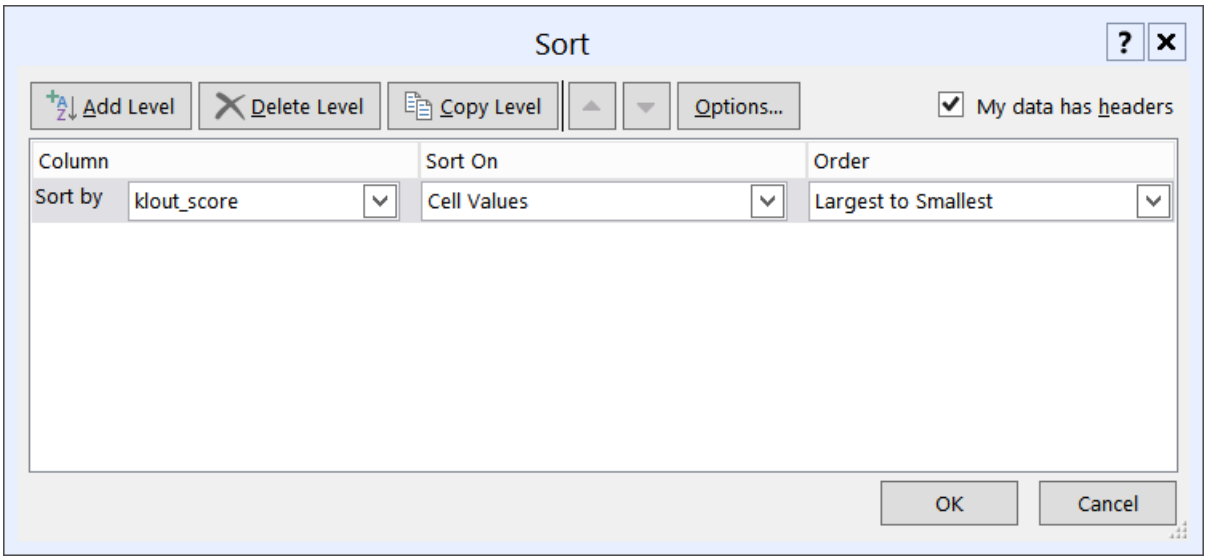
Copy *some* of the **interaction_content** cells and paste them into https://www.jasondavies.com/wordcloud/

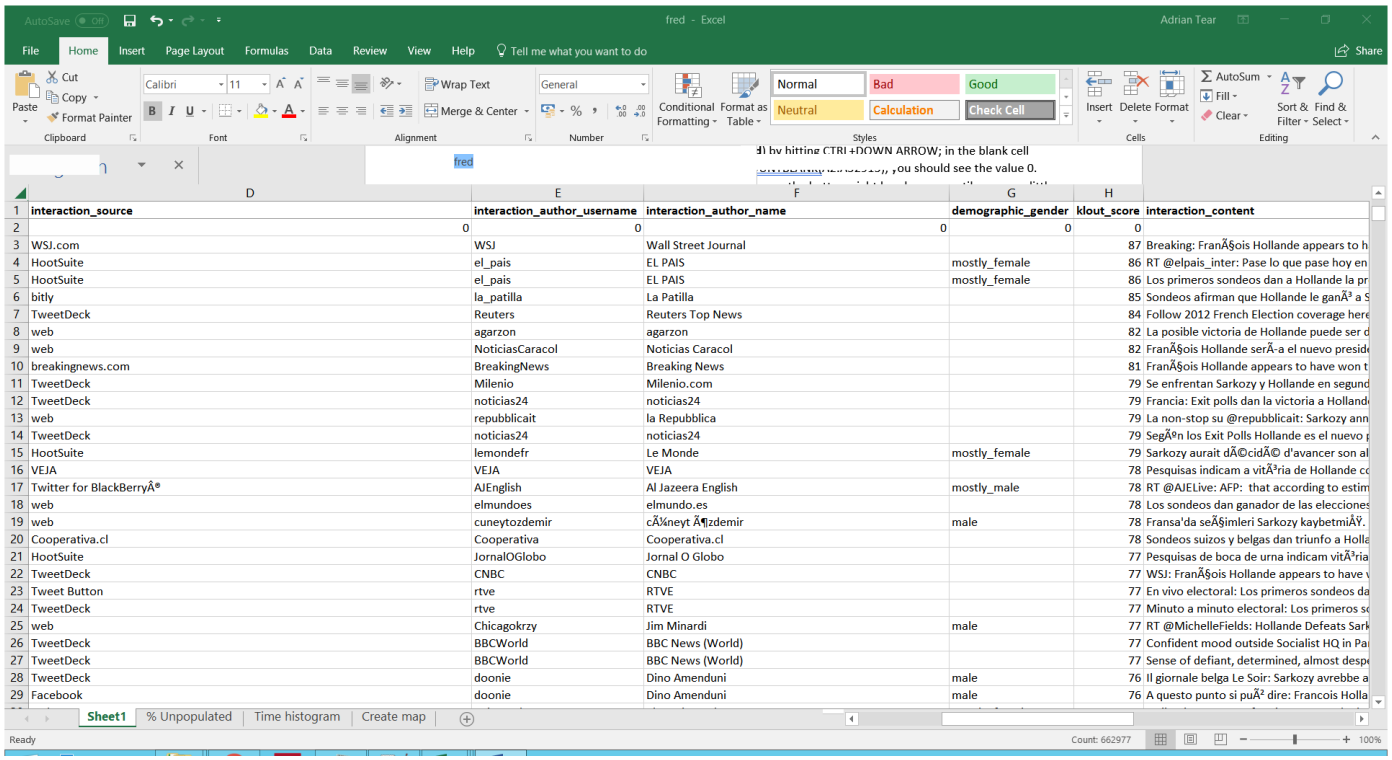
Other word cloud generators online let you upload files. You can try finding another one (e.g. http://www.wordle.net/create or https://tagcrowd.com/) and copy/paste more cells from the **interaction_content** column into a text file (use Notepad), saving this as WORDS.TXT and try uploading this to analyse all of the words. Note that some sites (e.g. Tagcrowd) have a 5MB limit of file size so you may not be able to create a word cloud visualisation of *all* the records in the spreadsheet.

Question: What do word clouds help you achieve? Does the relevant visual weight of the words identified in the text help you draw any conclusions about what people were Tweeting at the time the results came in during the 2012 French Presidential Election? Who, somewhat surprisingly, won that election?

Ranking

There are several fields in the Excel spreadsheet (e.g. **klout_score**, **twitter_user_followers_count**, **twitter_retweet_user_followers_count** and **twitter_retweet_count**) that can be used to rank (or sort) data. Highlight all cells in the spreadsheet (little arrow near cell A1) then from Home ribbon hit Sort & Filter and Custom Sort… Select one or more levels of sortation (e.g. on **klout_score** largest to smallest as above) and look at the Tweeters at the top of this list:





	D	E	F	G	H	
1	interaction_source	interaction_author_username	interaction_author_name	demographic_gender	klout_score	interaction_content
2	0	0	0	0	0	
3	WSJ.com	WSJ	Wall Street Journal		87	Breaking: FranÃ§ois Hollande appears to h
4	HootSuite	el_pais	EL PAIS	mostly_female	86	RT @elpais_inter: Pase lo que pase hoy en
5	HootSuite	el_pais	EL PAIS	mostly_female	86	Los primeros sondeos dan a Hollande la pr
6	bitly	la_patilla	La Patilla		85	Sondeos afirman que Hollande le ganÃ³ a S
7	TweetDeck	Reuters	Reuters Top News		84	Follow 2012 French Election coverage here
8	web	agarzon	agarzon		82	La posible victoria de Hollande puede ser d
9	web	NoticiasCaracol	Noticias Caracol		82	FranÃ§ois Hollande serÃ-a el nuevo presid
10	breakingnews.com	BreakingNews	Breaking News		81	FranÃ§ois Hollande appears to have won t
11	TweetDeck	Milenio	Milenio.com		79	Se enfrentan Sarkozy y Hollande en segund
12	TweetDeck	noticias24	noticias24		79	Francia: Exit polls dan la victoria a Holland
13	web	repubblicait	la Repubblica		79	La non-stop su @repubblicait: Sarkozy ann
14	TweetDeck	noticias24	noticias24		79	SegÃºn los Exit Polls Hollande es el nuevo p
15	HootSuite	lemondefr	Le Monde	mostly_female	79	Sarkozy aurait dÃ©cidÃ© d'avancer son al
16	VEJA	VEJA	VEJA		78	Pesquisas indicam a vitÃ³ria de Hollande c
17	Twitter for BlackBerryÂ®	AJEnglish	Al Jazeera English	mostly_male	78	RT @AJELive: AFP: that according to estim
18	web	elmundoes	elmundo.es		78	Los sondeos dan ganador de las elecciones
19	web	cuneytozdemir	cÃ¼neyt Ã¶zdemir	male	78	Fransa'da seÃ§imleri Sarkozy kaybetmiÅŸ.
20	Cooperativa.cl	Cooperativa	Cooperativa.cl		78	Sondeos suizos y belgas dan triunfo a Holla
21	HootSuite	JornalOGlobo	Jornal O Globo		77	Pesquisas de boca de urna indicam vitÃ³ria
22	TweetDeck	CNBC	CNBC		77	WSJ: FranÃ§ois Hollande appears to have w
23	Tweet Button	rtve	RTVE		77	En vivo electoral: Los primeros sondeos da
24	TweetDeck	rtve	RTVE		77	Minuto a minuto electoral: Los primeros s
25	web	Chicagokrzy	Jim Minardi	male	77	RT @MichelleFields: Hollande Defeats Sark
26	TweetDeck	BBCWorld	BBC News (World)		77	Confident mood outside Socialist HQ in Pa
27	TweetDeck	BBCWorld	BBC News (World)		77	Sense of defiant, determined, almost despe
28	TweetDeck	doonie	Dino Amenduni	male	76	Il giornale belga Le Soir: Sarkozy avrebbe a
29	Facebook	doonie	Dino Amenduni	male	76	A questo punto si puÃ² dire: Francois Holla

You should see that the top Tweeters, in terms of **klout_score** (see https://klout.com/corp/score for further details), are large news organisations like the *Wall Street Journal* and *El Pais*.

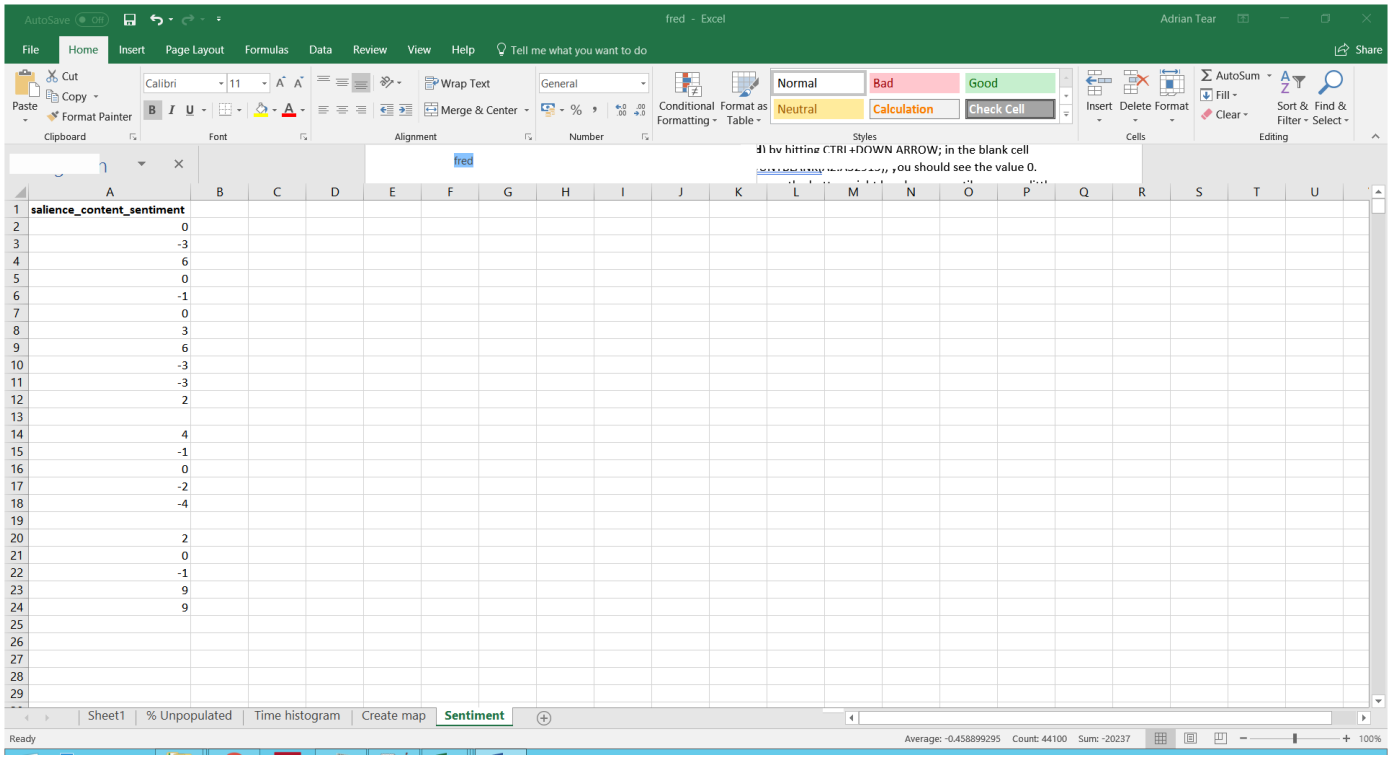
Question: Who are the least important Tweeters by **klout_score**? Is the average **klout_score** for those Tweeting with geographic coordinates higher or lower than for those who Tweet without coordinates?

Tip: Use Excel to calculate averages and make these comparisons using what you have learnt above.

Sentiment analysis

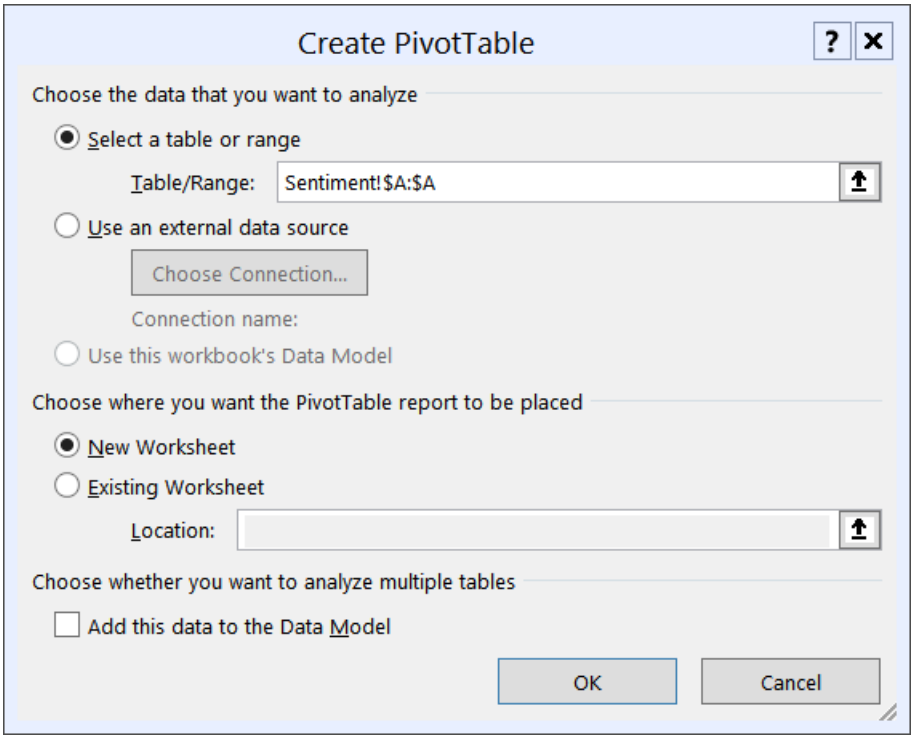
The spreadsheet also contains a column **salience_content_sentiment** which gives an indication of negative, neutral or positive sentiment for the Tweet.

Select this column and paste it into a new sheet…

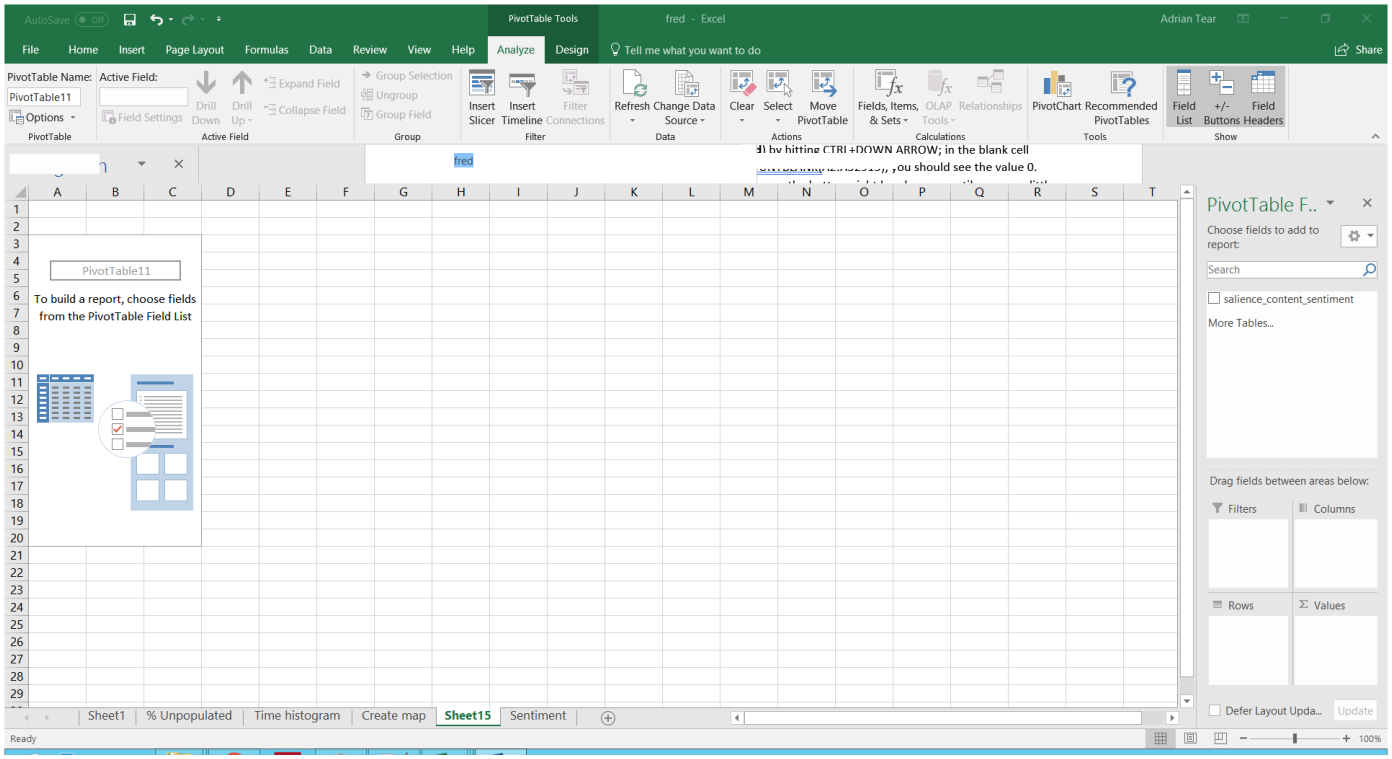


	A
1	salience_content_sentiment
2	0
3	-3
4	6
5	0
6	-1
7	0
8	3
9	6
10	-3
11	-3
12	2
13	
14	4
15	-1
16	0
17	-2
18	-4
19	
20	2
21	0
22	-1
23	9
24	9

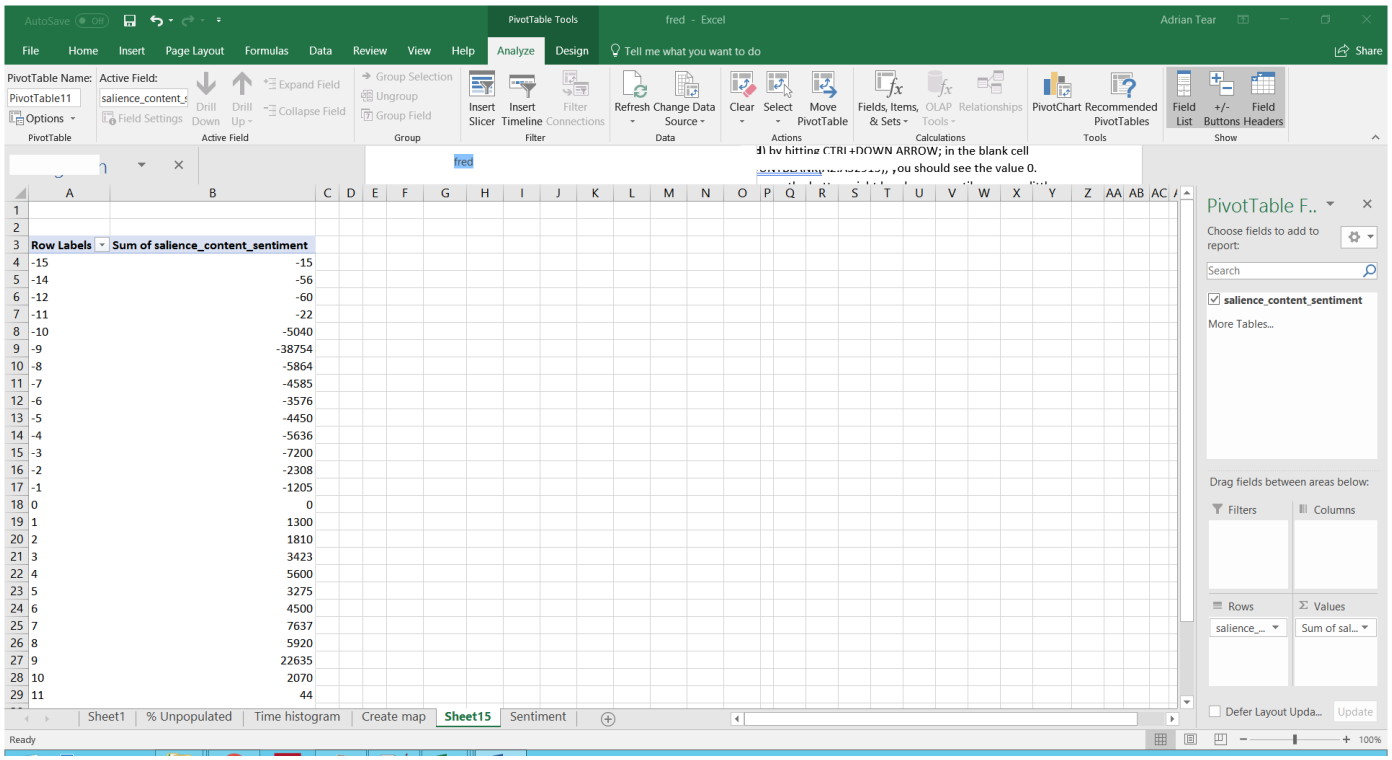
From the Insert ribbon choose Pivot Table and follow the defaults to create another new sheet:



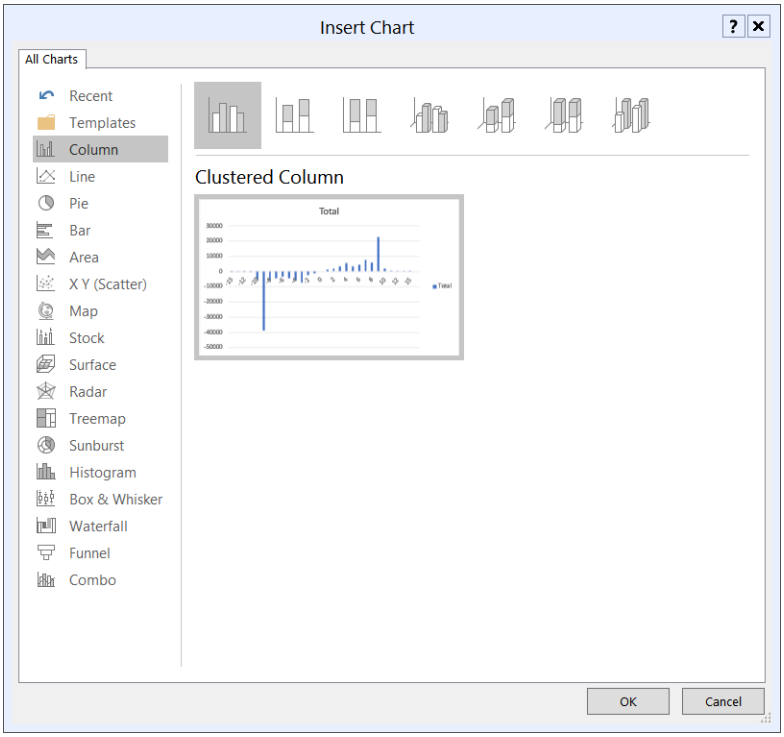
This will create a blank definition:



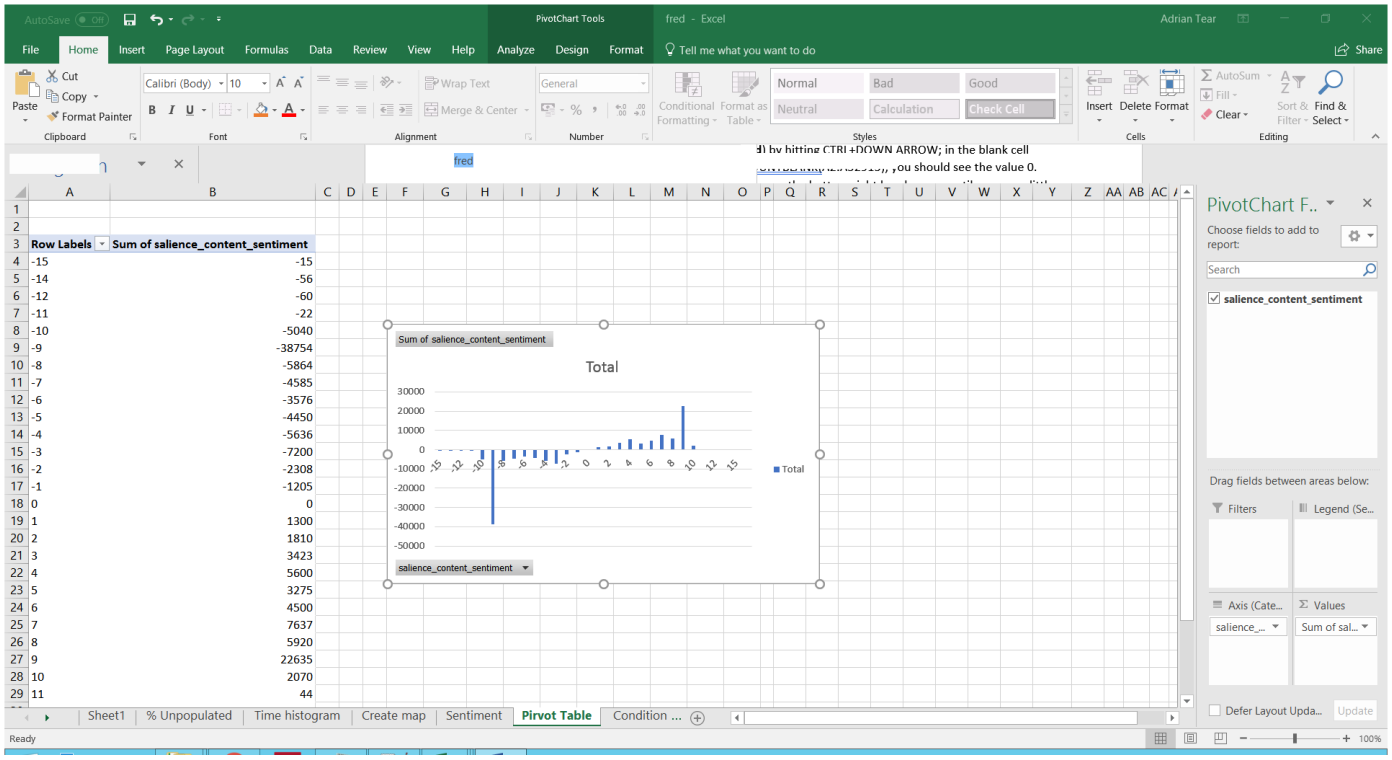
On the right hand side drag **salience_content_sentiment** into the Rows and Sum Values boxes:



Then from the Insert ribbon choose Insert Recommended Chart:

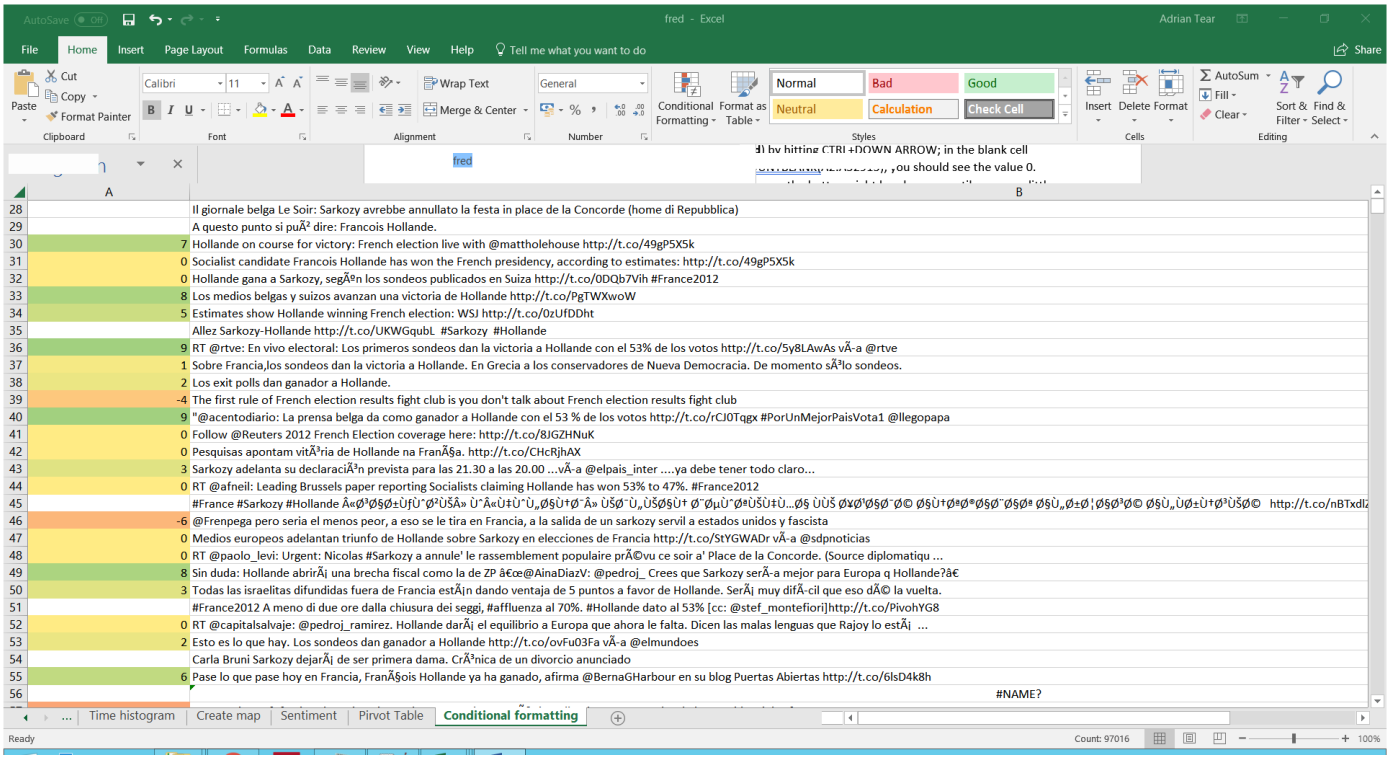


This should suggest a Clustered Column chart as below:



Question: What does this data tell us? It appears that large numbers of Tweets are Negative (-9) but that quite a few (+9) are Positive. Is this meaningful?

Tip: Use Conditional Formatting / Colour Scale to colour the spreadsheet according to **salience_content_sentiment** by copying this column, and **interaction_content** into another new sheet:



This should show you how the software has scored each Tweet. Is it accurate? You might need to get your dictionary out if your GCSE French is a bit rusty!!

Try some more advanced software from IBM at https://natural-language-understanding-demo.ng.bluemix.net and type in, or copy/paste in, some sentences of your own or, e.g., some copy from the BBC News website. Check the Keywords and Entities detected in the text. Is this software any better? How could you use it?

Summary

This Exercise has guided you through what can be achieved to analyse and map some Online Social Network (OSN) data, primarily sourced from Twitter.

Quite a lot can be achieved using Excel and some free, online resources. More systematic analysis requires a database for data storage, many more records, and more sophisticated processing using specialised software.

Think about the quality, and use, of this data. Also, are there any ethical problems with using data of this type?

Follow the references for more material covering these areas.

References

Fuchs, C. (2017). *Social Media: A Critical Introduction* (2nd ed.). SAGE Publications. Retrieved from https://uk.sagepub.com/en-gb/eur/social-media/book250870

Scott, J. (2017). *Social Network Analysis*. SAGE Publications. Retrieved from https://uk.sagepub.com/en-gb/eur/social-network-analysis/book249668

Tear, A. (2014). SQL or NoSQL? Contrasting Approaches to the Storage, Manipulation and Analysis of Spatio-temporal Online Social Network Data. In B. Murgante, S. Misra, A. M. A. C. Rocha, C. Torre, J. G. Rocha, M. I. Falcão, … O. Gervasi (Eds.), *Computational Science and Its Applications -- ICCSA 2014: 14th International Conference, Guimarães, Portugal, June 30 -- July 3, 2014, Proceedings, Part I* (Vol. 8579 LNCS, pp. 221–236). Springer International Publishing. http://doi.org/10.1007/978-3-319-09144-0_16