**Web Scraper Issues and Pivot Tables**

For those of you who have scraped restaurant reviews and ratings from Yelp, you may have noticed that when a review is scraped the ratings column is blank and vice versa. This is a bug in web scraper (we have written to them about it). However, fortunately it does not affect your analysis in any way. Read on.

Your data file will look something like this – you can delete the other columns.

|  |  |  |  |
| --- | --- | --- | --- |
| Web scraper order **(after sorting from A to Z)** | Restaurant name | Review | Rating |
| 50249 | Azul | Great place. |  |
| 50250 | Azul | Good. |  |
| 50251 | Azul | Very bad. |  |
| 50252 | Azul |  | 5 |
| 50253 | Azul |  | 3 |
| 50254 | Azul |  | 1 |

Note that you get something similar to the above table after sorting the web scraper order column from A to Z. Either the Review or the Rating column is blank because the scraper is collecting the review and the ratings data in two separate passes. So the review “Great place” actually corresponds to rating 5, and the review “Bad” corresponds to rating 1. While that is kind of stupid, it poses no problem for us in this assignment. We can still do the assignment with the data file you have without even trying to align the reviews and the ratings.

For the similarity analysis the ratings don’t matter at all. So you do the similarity analysis with the Review column (make sure the label of the column matches what is in the script). Then sort the similarity score column from highest to lowest similarity, and choose the top 200 similarity scores. Now perform sentiment analysis on these 200 reviews and sort once more from high to low sentiment scores.

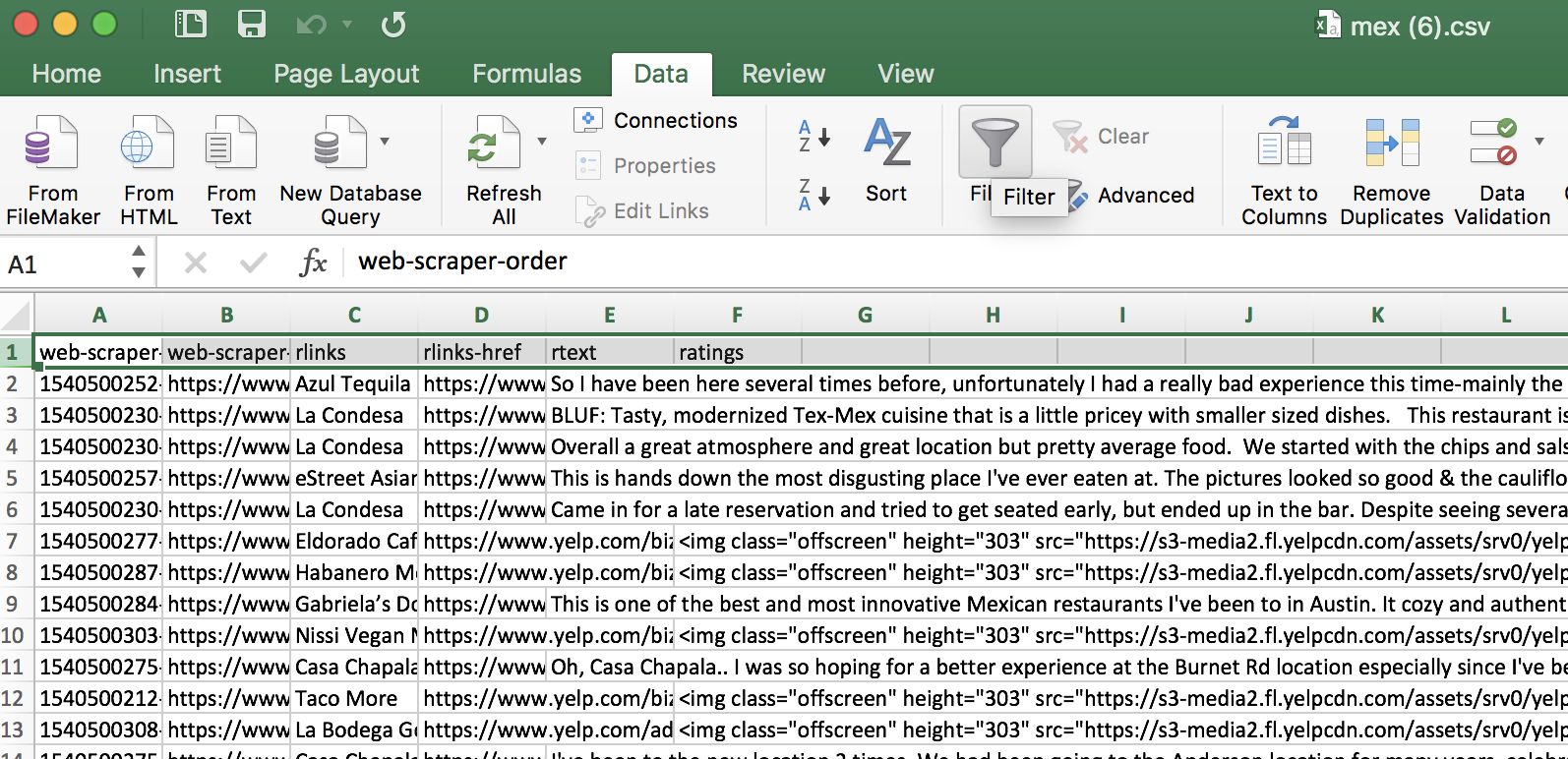
In order to do sentiment analysis, you must pass a single column file as input to sentiment\_new.py (remember, sentiment\_new.py was written to process the output of parserforsentiment.py, which we don’t need here). So copy and paste the reviews in new Excel file, and use it as the input for sentiment analysis. Paste the compound column sentiment scores back to the data file next to the reviews column.

For the last question involving Ratings, get the average rating for each restaurant from the Ratings column (you can use Pivot to do this). The blanks should not matter in your average rating calculation. Now sort from highest to lowest average ratings, and pick the top 3 restaurants. For these three restaurants you need the average similarity scores, which you can get with Pivot.

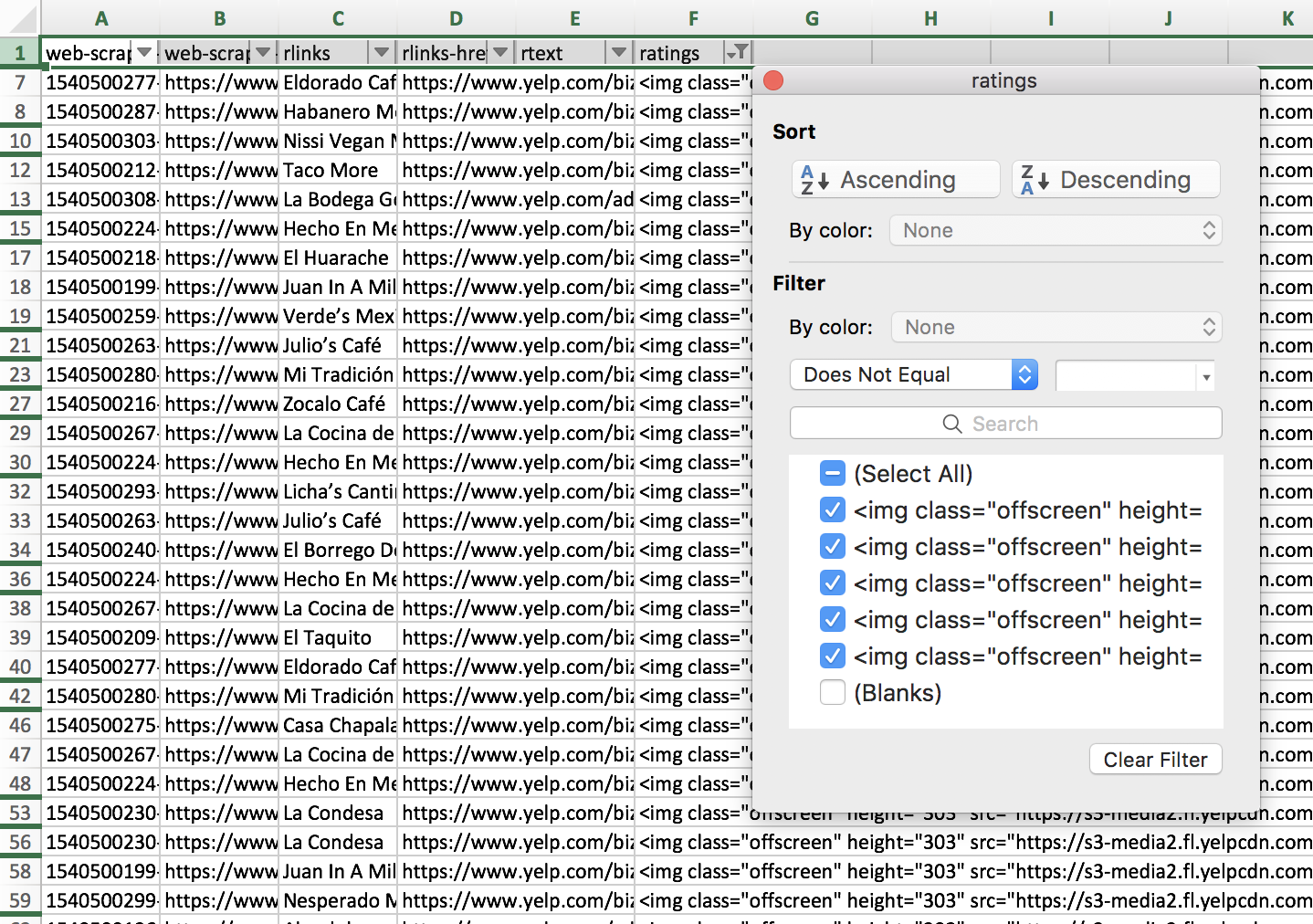
**If you are not familiar with pivot tables, here is a quick tutorial:**

Here we show you how to get the average rating of the restaurants.

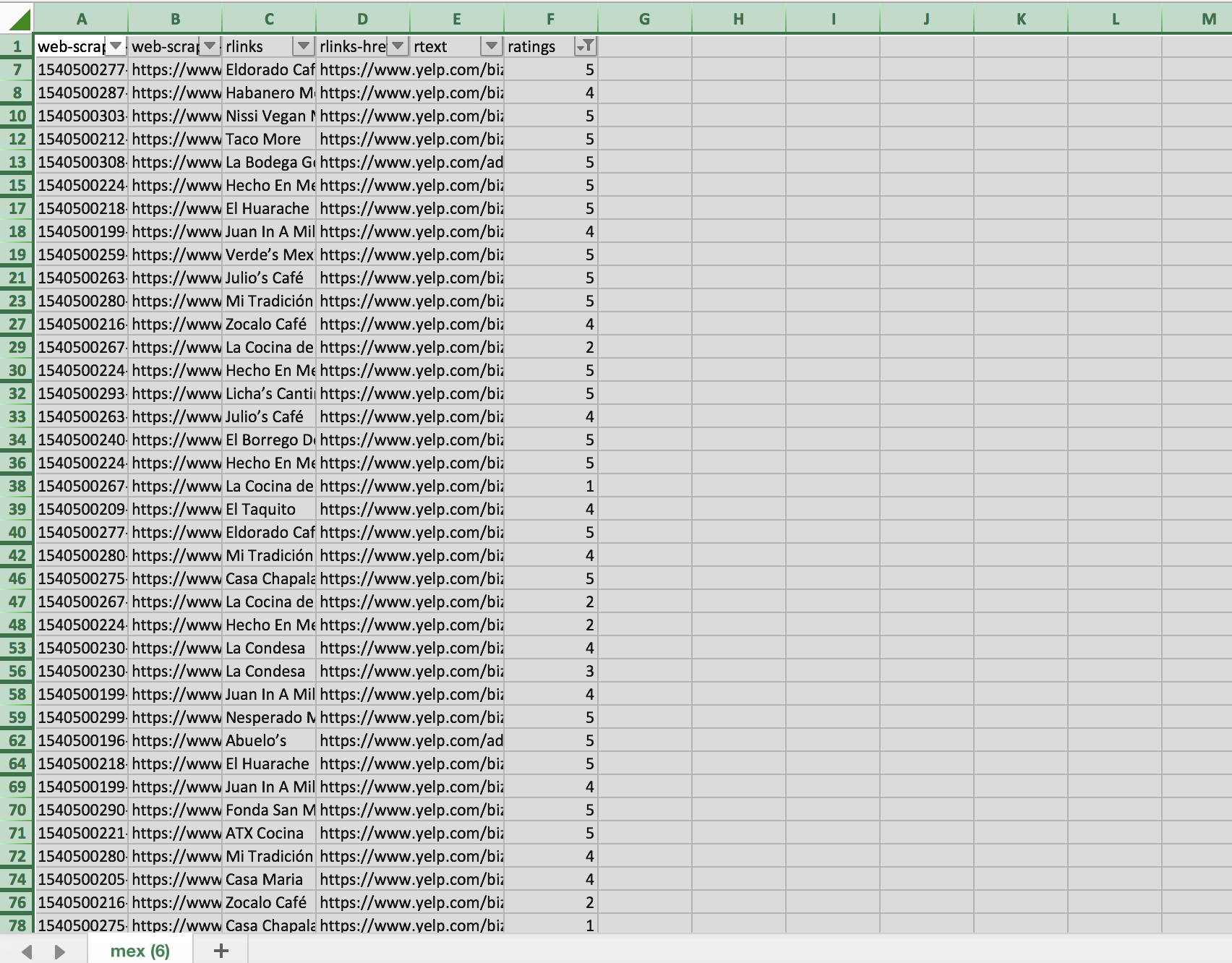
1. Select the topmost row ( clicking on 1) and click on Filter



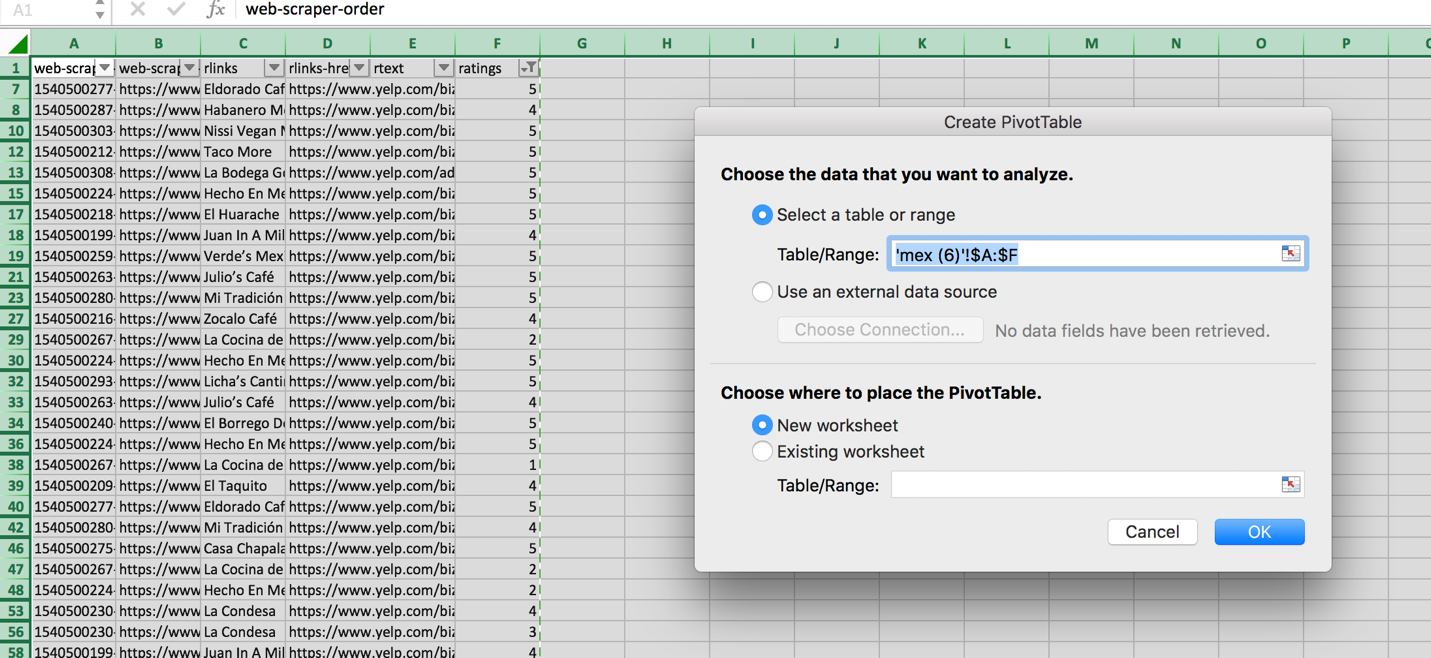
1. Click on the filter sign next to ratings and then click on the box before Blank to remove all the blank ratings (note that you should first remove the text from the ratings column as shown in the assignment document so that this column only contains a number like 4 or 3)



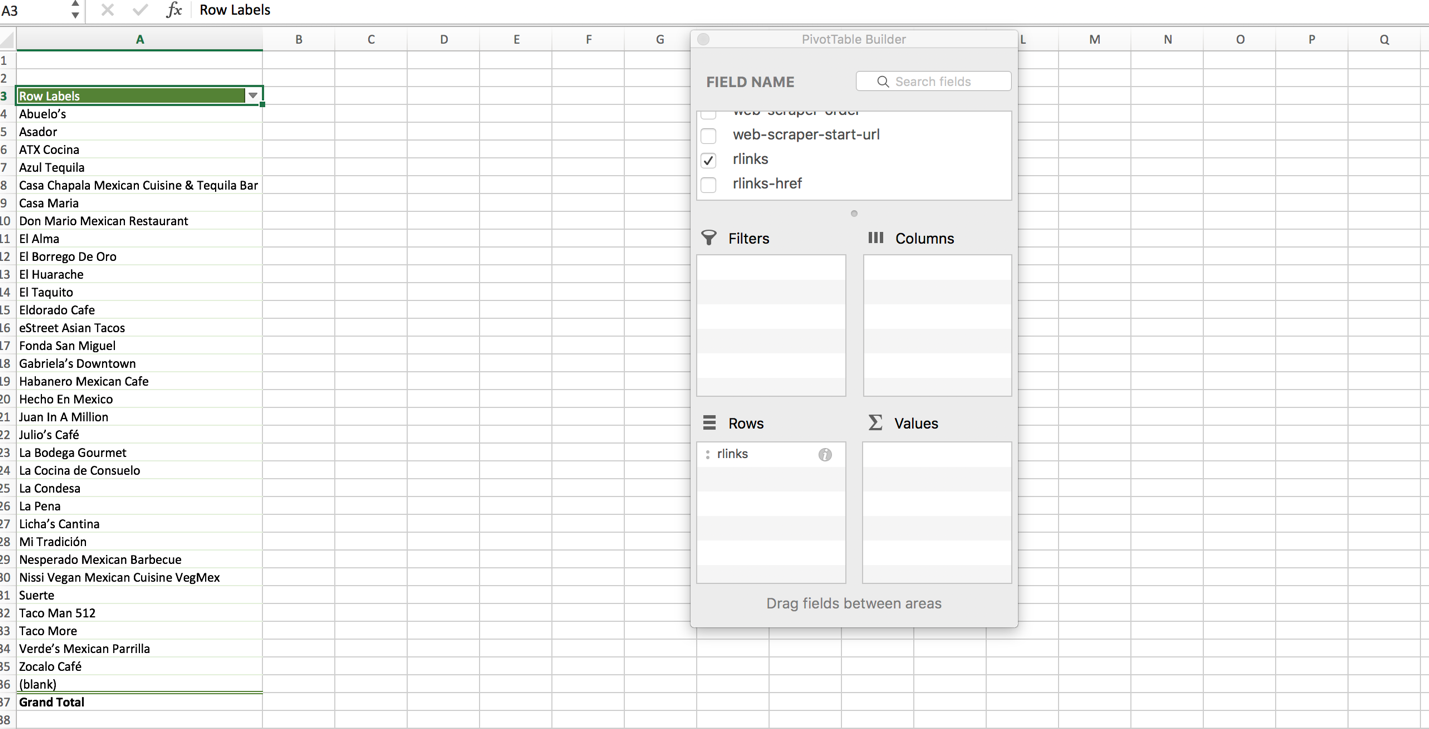
1. Now you have filtered all the non-blank ratings



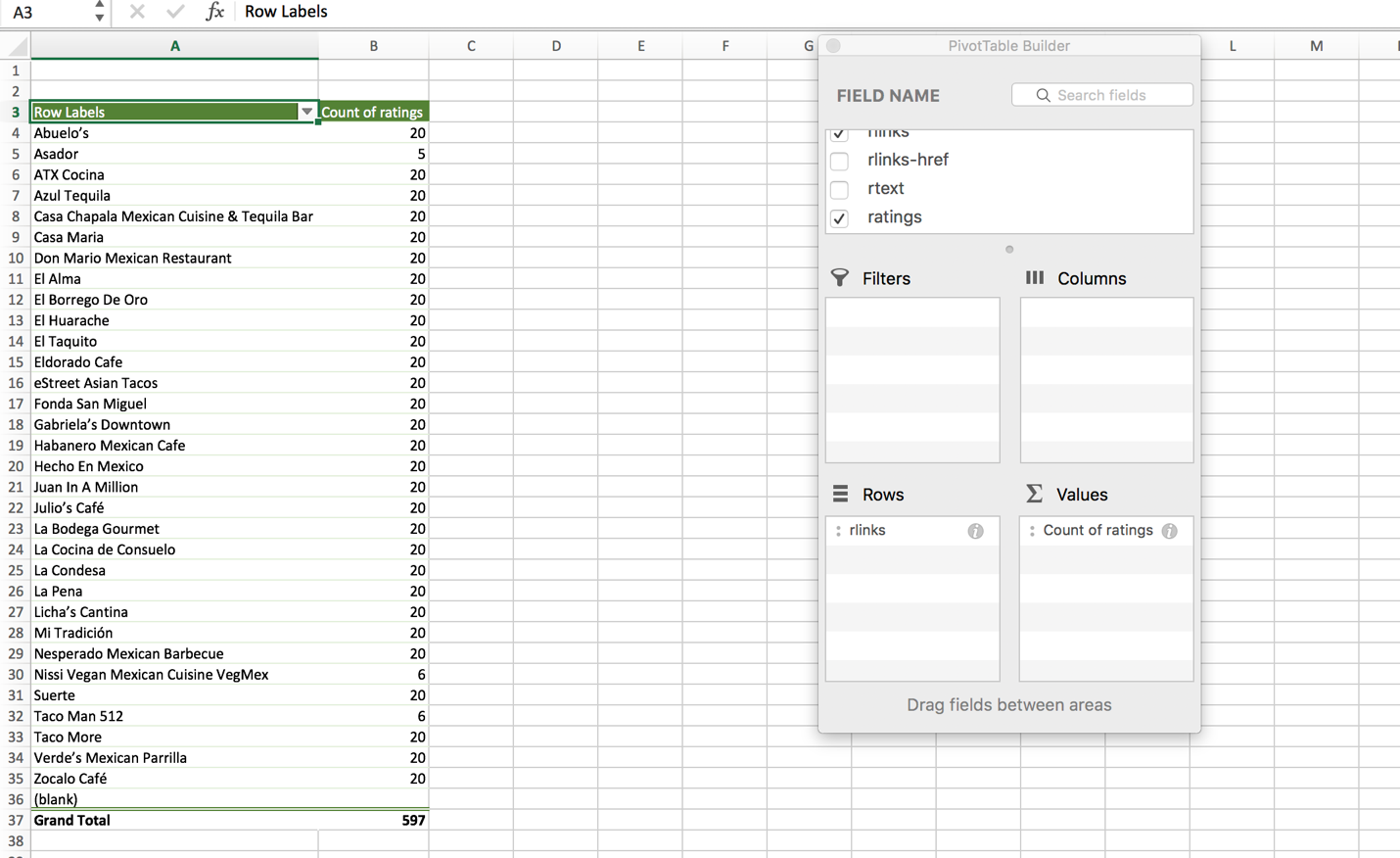
1. Now select all the data by clicking on the (left topmost cell) click on Insert tab -> Pivot table and you will see the screen below – click on OK



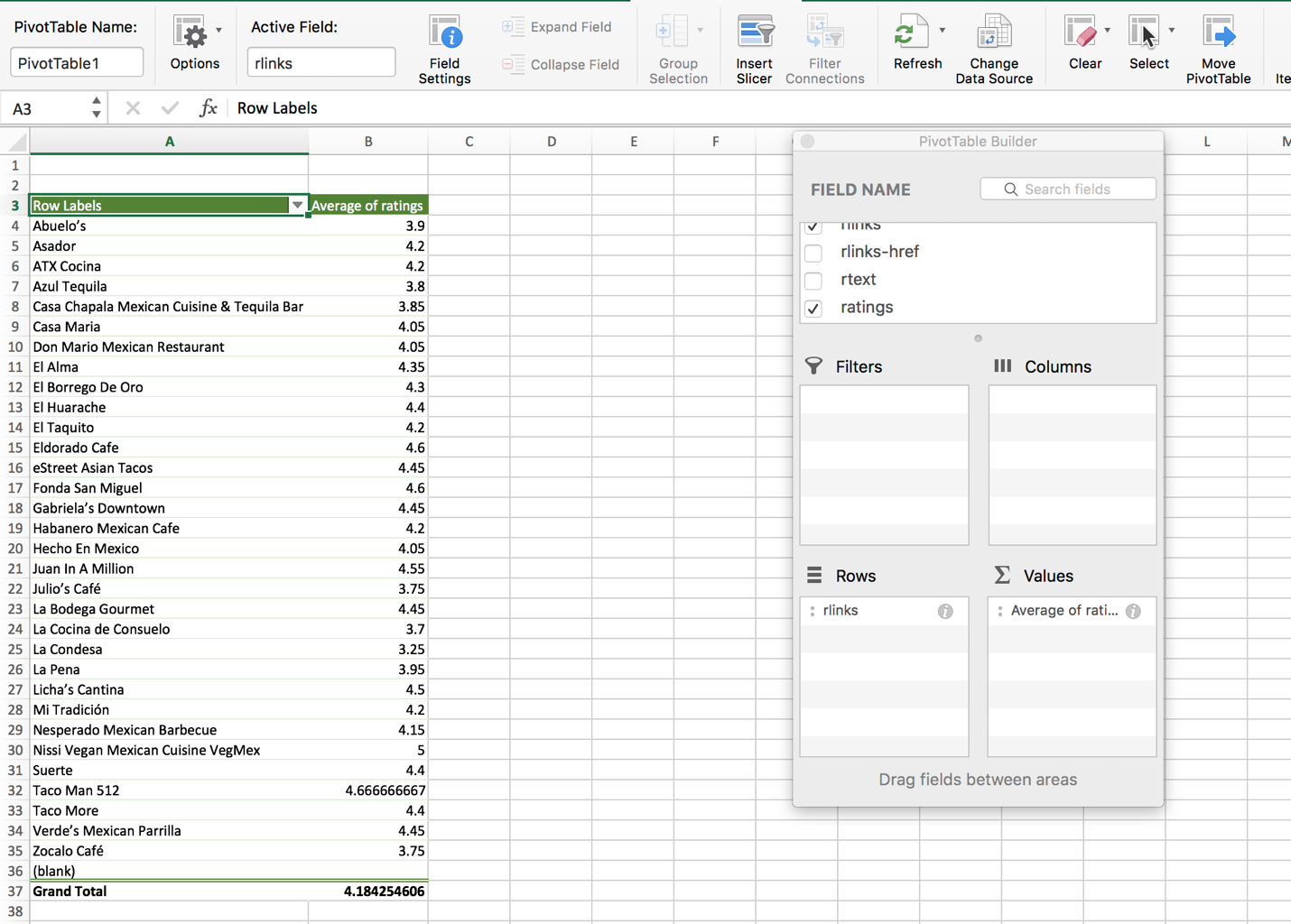
1. In the pivot sheet click on rlinks (restaurant links) and drag it to Rows box



1. Now click on rating and drag it to Values box and then click on the i symbol next to Count and select Average



1. You will see something like this. Now select the two columns and paste by value in a new Excel file, and sort from high to low to get the top rated restaurants.



**You will follow similar steps to get the average similarity score of each restaurant.**

If you still have doubts about pivot analysis with Excel, please take a look at the videos below:

<https://www.youtube.com/watch?v=Vx-Fuw46VbY>

<https://support.office.com/en-us/article/video-create-a-pivottable-74ce8afc-2446-4816-80ee-20ca7fb71793>

<https://www.youtube.com/watch?v=9NUjHBNWe9M>