Wuzzuf Data Exploration

Contents :

1. Preface

I'm going to explore the data from 3 sides.

- First I'm going to see the changes of (number of views, years of experience needed, number of job applications posted, number of open vacancies available and the average of minimum/maximum salaries) over time.
- Then I'll pay attention to cities, clean the city_name column to see the weight of each city, then see the changes for the above variables over time for the top cities.
- Finally I'll look at the two columns ["job description" and "job requirements"] applying NLP techniques to them, and get the most commonly used words.

Notes :

- Cleaning and exploration parts are done using ipython-notebooks and plot.ly. You can find the code used in this project in the github repository.
- You can jump for the last section for a quick summary
- For high quality graphs open the repository.
- All data-sets and dictionaries used can be found in the repository inside the folder "data"
- For more information open the 'readme' file in the repository, this report is meant for non-technicals.

2. Data cleaning

Each of the 3 parts mentioned above requires certain format of the data, and so this step is most important one.
I’m going to summary what I did to get the data ready without any technical details. For full details with code, you can find it in the repository.

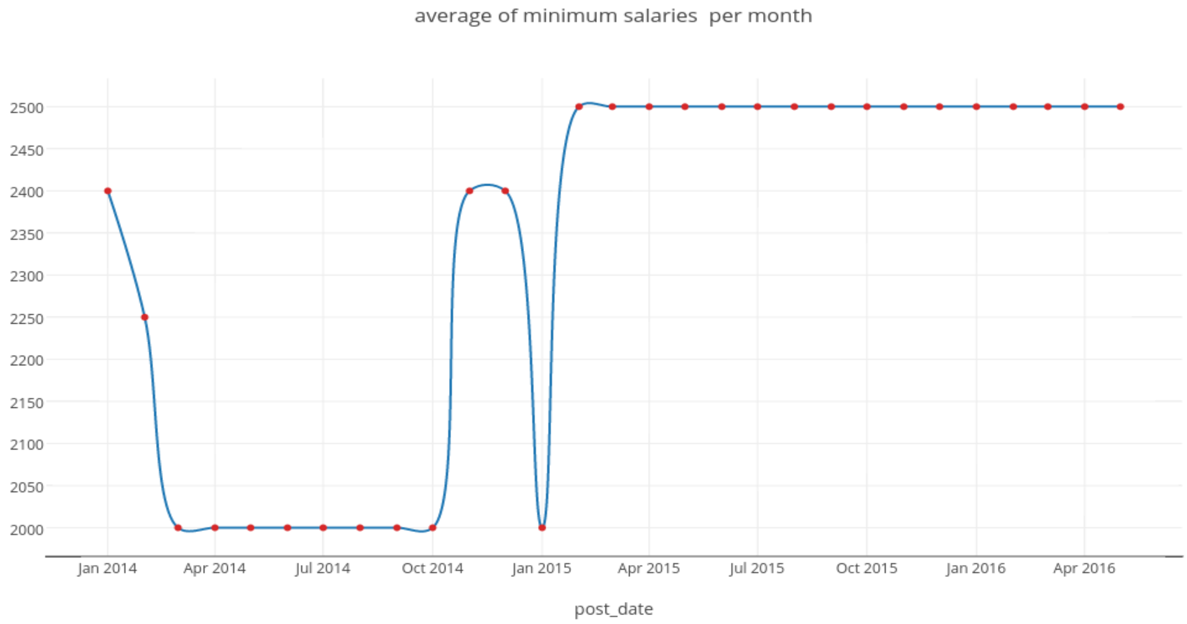
- Cleaning city_name :
    - My target here is categorize the column with fixed name for each city.
    - Problems I have countered here :
        - Cities are written with many different ways. Ex (alex, alexandria)
        - There are many typos !
        - Some fields holds more than one city, Ex (“alex and giza”) and the delimiter is different from field to the other ex(“,” , “and”, “or”, “.”)
    - So what I did is to get a list of all the unique names with number of occurrences of each one. And what I got :
        - There are 478 different names.
        - If we took only the names which occurred at least 3 times then we will have “104” names corresponds to more than 98% of the data.
        - They includes 6 names written in Arabic and 98 in English.
    - Then I made a dictionary with a the fixed city name as a key, and a list of the possible names written for this city as a value, and saved it in a json file.
    - After that I change All the city_names with names from the keys of the json file. And if the city_name is not included there, then I will change it to “other(`given name`)”.

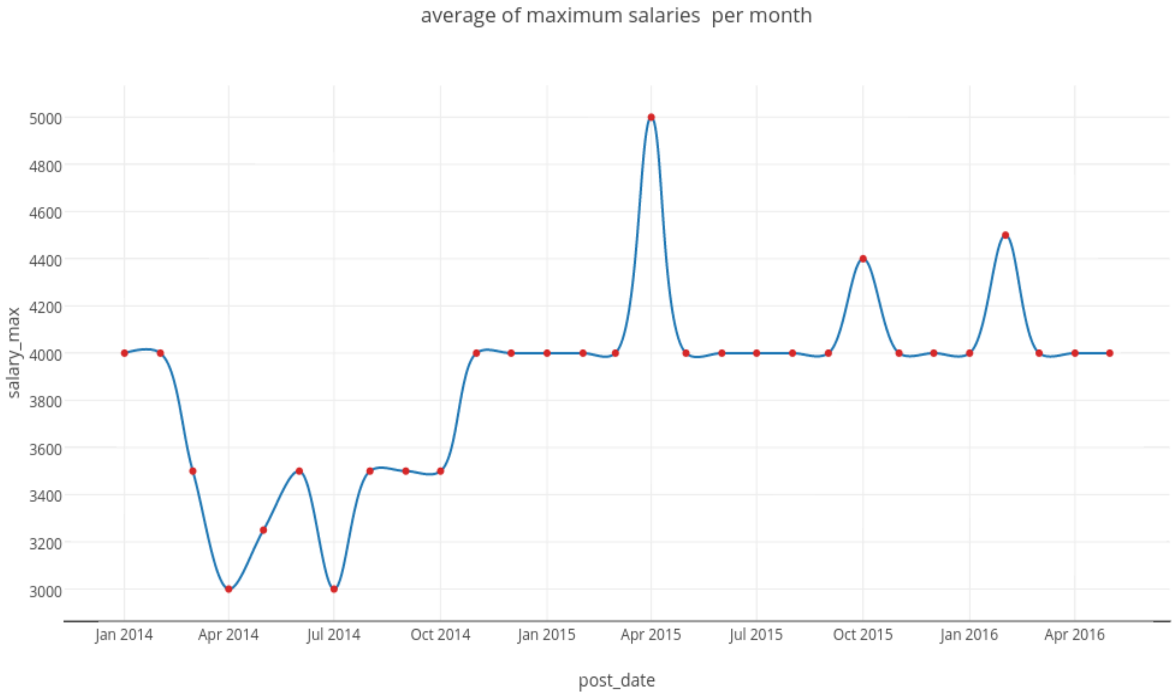
- The next thing I cleaned is the post_date, I formated it to “year-month”. ex(‘2015-06’).
- Also I extracted the “minimum experience years needed” from “experience_years” column using regular expressions.
- To explore the insights over times, I grouped the data by “post_date”, dropped the non-numerical columns. And used the median for (min/max salaries and min experience years) and the sum for (views, job applications and vacancies).
- Also to explore per city over time. I firstly choose cities with high number of job applications (== number of samples) so we can rely on its result. Then group the data by city, and for each city group it’s data by “post_date” like we did above.
- For the text processing step, I just needed a list of the two columns job descriptions/requirements.

---

3. Changes over time

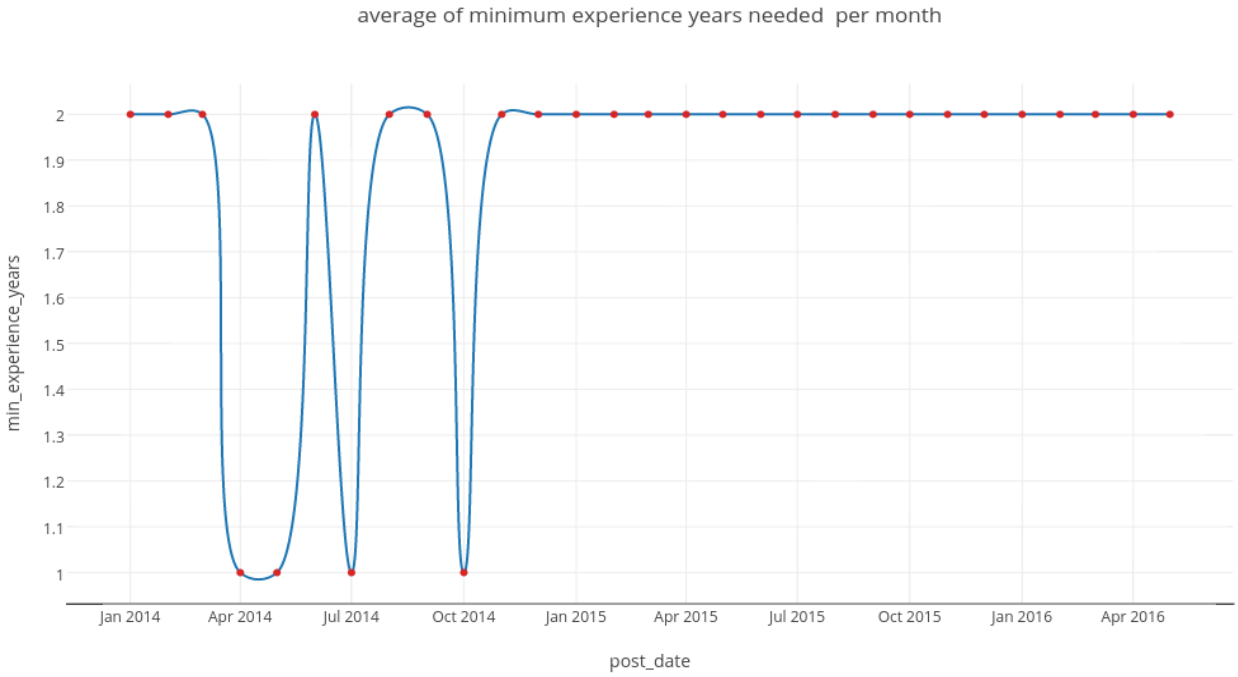
➔ Average minimum/maximum salaries :





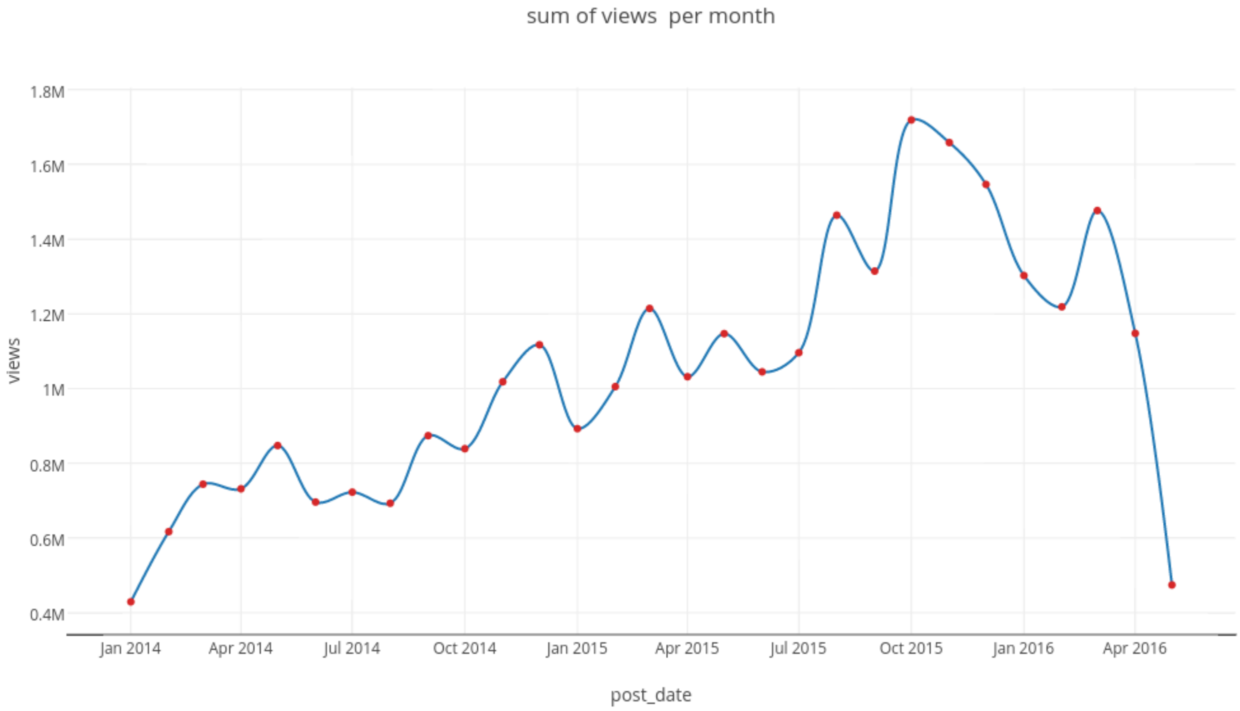
Minimum salaries we around 2K before jan 2015 and increased to 2.5K after that.
Maximum salaries after jan 2015 increased to range [4K to 4.5K].

- ➔ Average of minimum experience years needed :



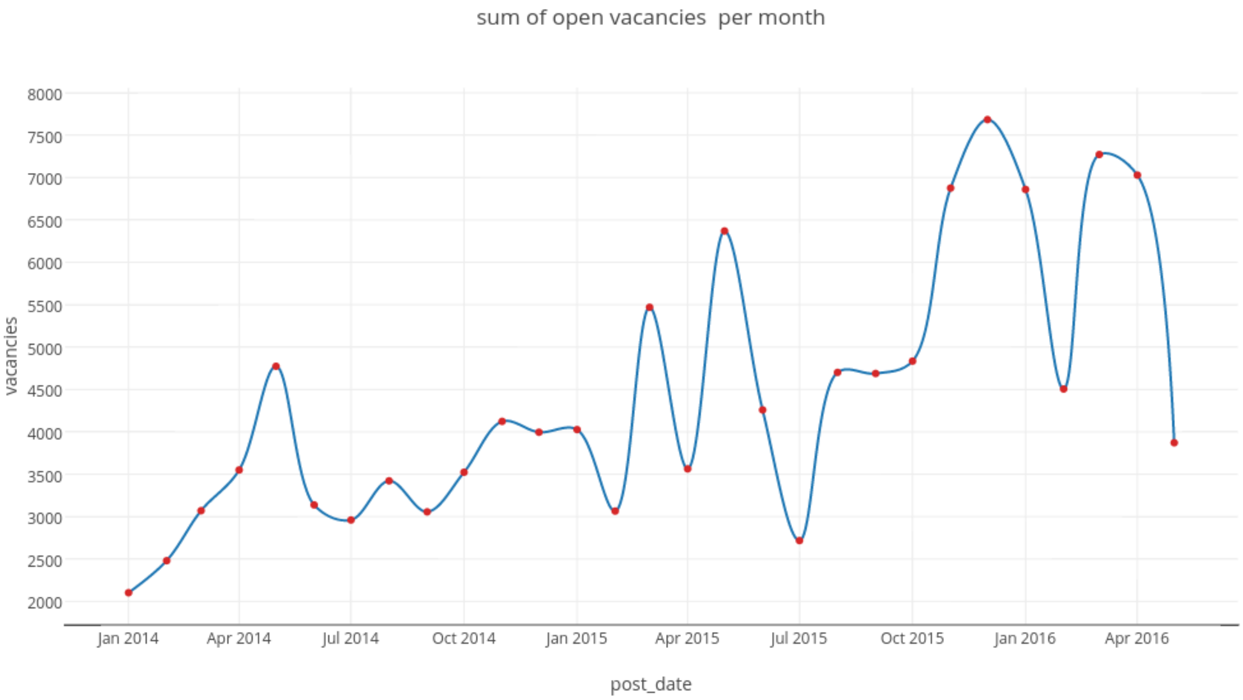
As it's clear the minimum experience years needed in most jobs are 2 years.

- ➔ Sum of views :



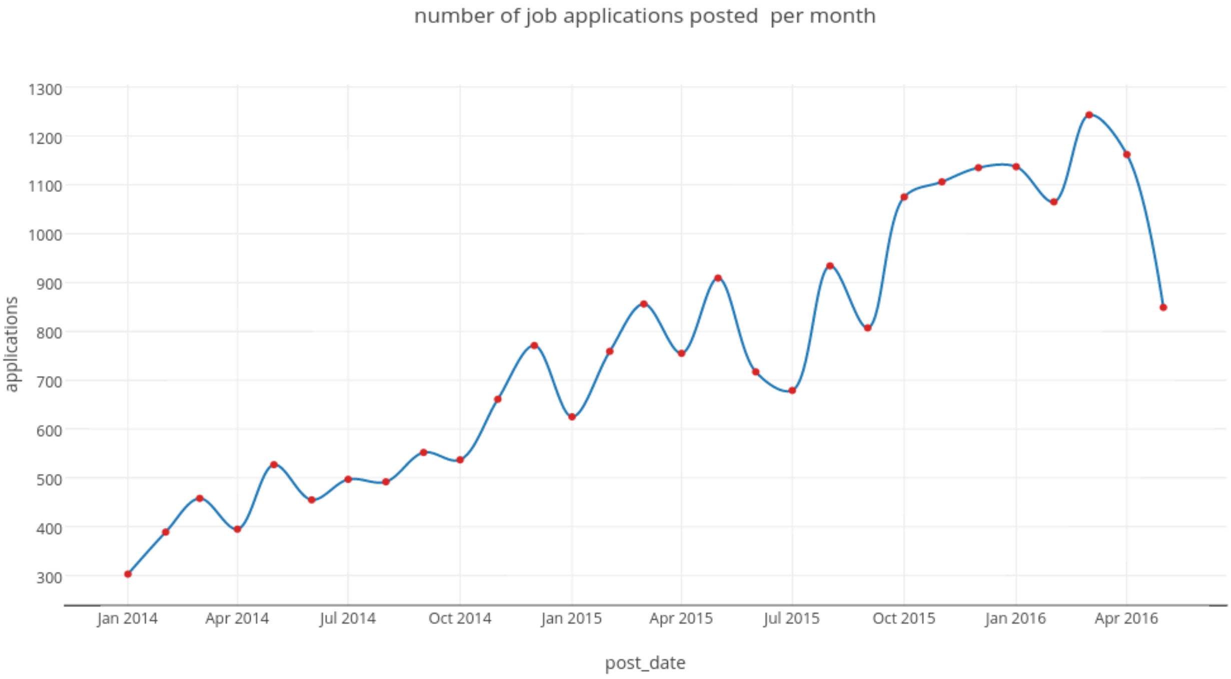
Views increases over time, highest value is around 1.7M views in Oct 2015. Also for some reason we will investigate later views are decreased after Apr 2016.

➔ Sum of open vacancies :



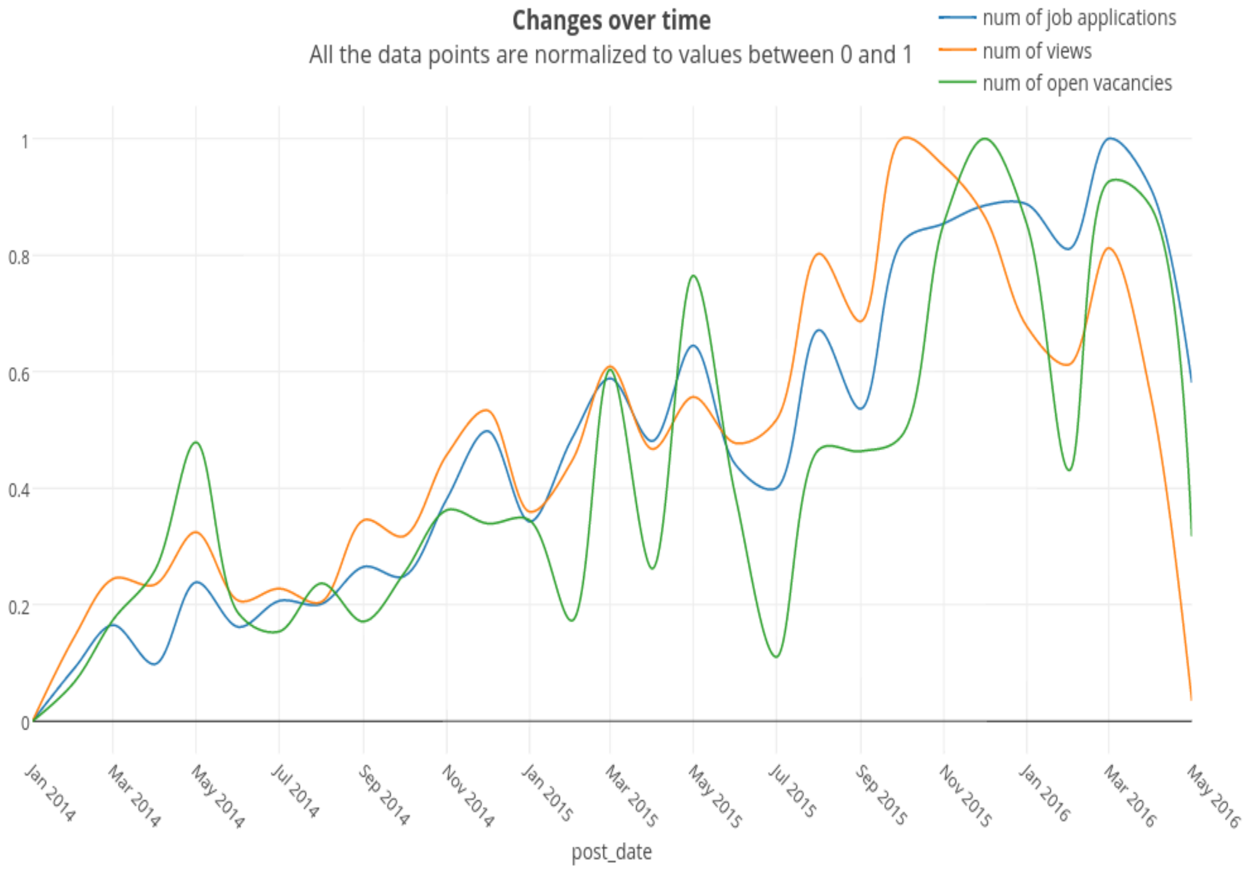
Also the number of vacancies increases over time and as the previous one it dropped down after Apr 2016.

➔ Sum of job applications :



Number of job applications posted on wuzzuf increases over time and it slightly dropped down after Apr 2016.

What we notice here is for “views, vacancies and job applications” the graph dropped after Apr 2016, although the three of them increases over time. So lets draw the three of them together to notice any correlations.



And as we can see the three of them of strongly correlated with each other.
And the statistical correlation between them is as following :

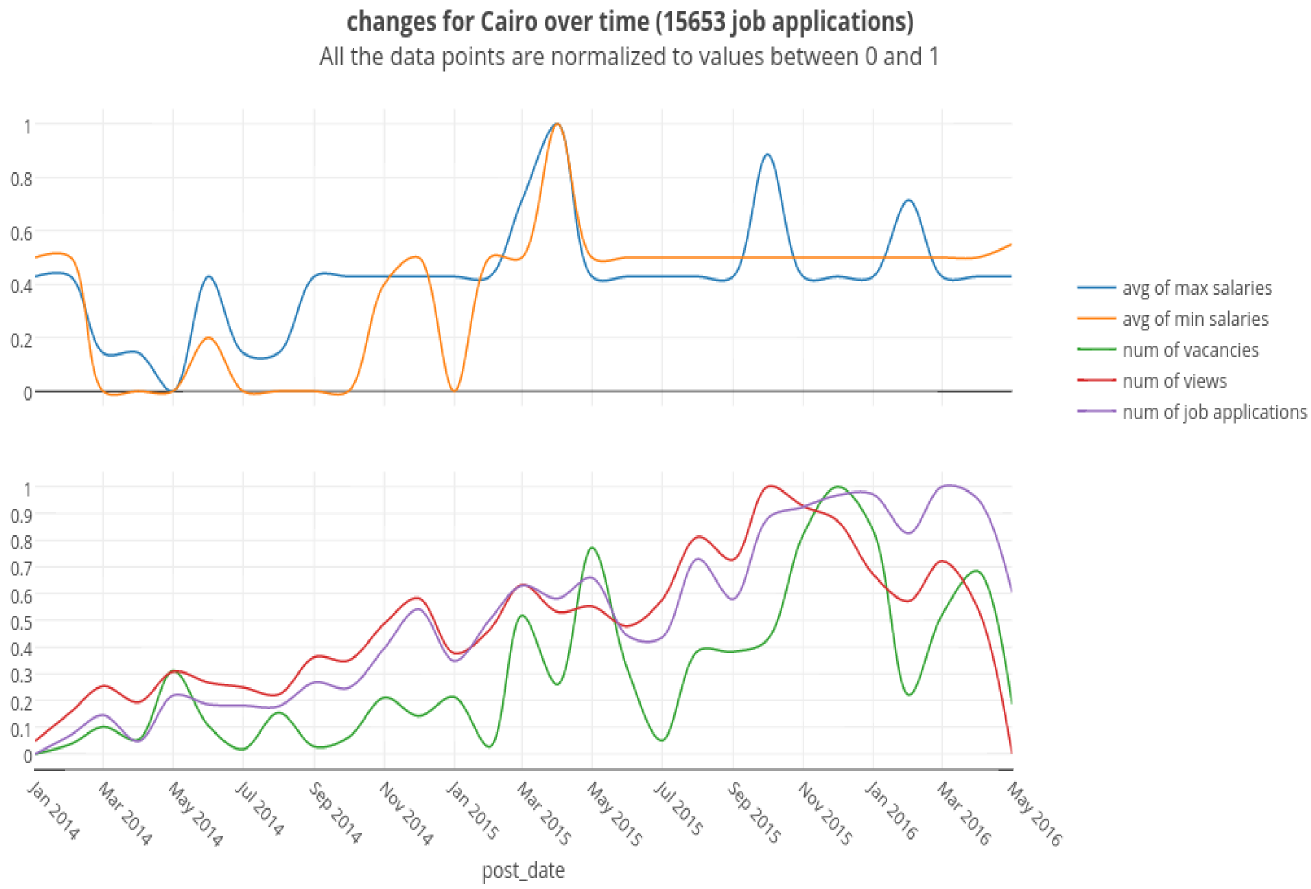
	views	Open vacancies	Job applications
views	1	0.838	0.741
Open vacancies	0.838	1	0.858
Job applications	0.741	0.858	1

4. Changes for cities over time

After cleaning the city-name to categorized data. Here are the top cities with number of job applications.

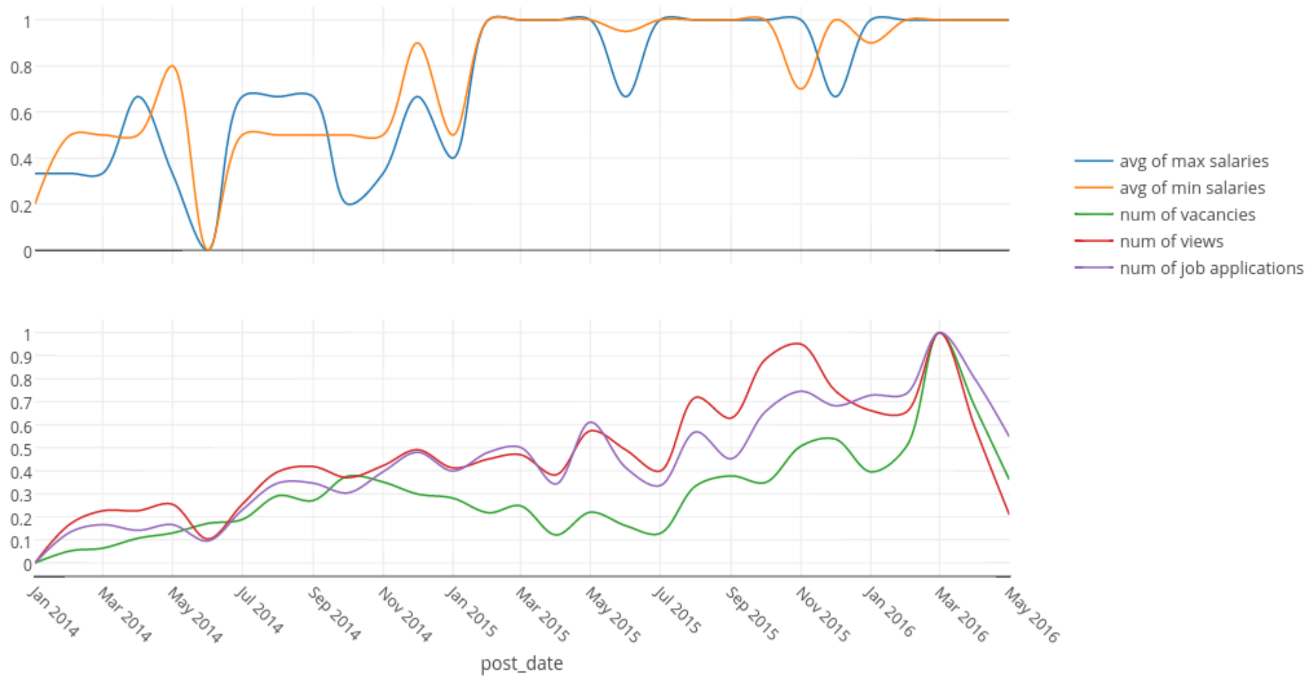
city	Job applications
cairo	15653
giza	4117
alexandria	1162
6th of october	225
10th of ramadan	115
mansoura	140

As I'm following the change over 2.5 years, less than 1000 samples(job applications) won't be enough. So I'll follow the change over the first three cities only (cairo, giza and alexandria).



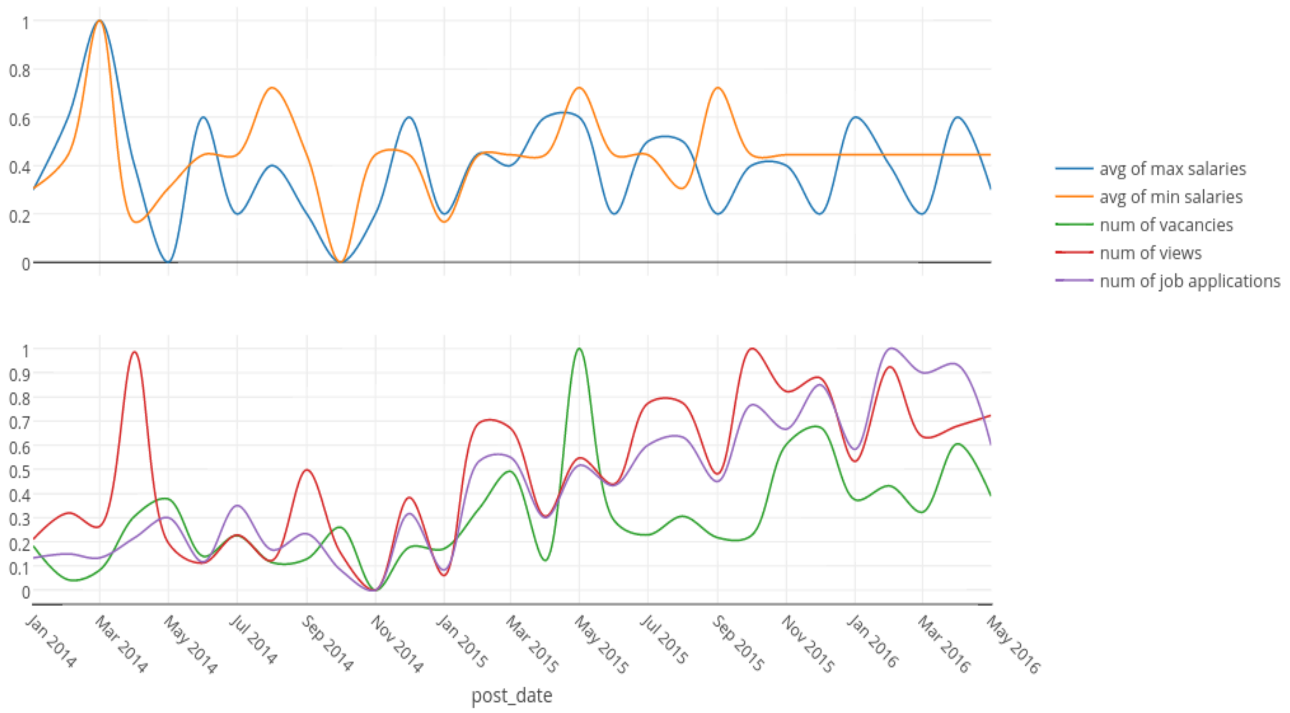
changes for Giza over time (4117 job applications)

All the data points are normalized to values between 0 and 1



changes for Alexandria over time (1162 job applications)

All the data points are normalized to values between 0 and 1



5. Analyzing job descriptions and requirements

The last thing I did , is that I took all the job descriptions and requirements each alone. Then for each of them, I firstly deleted all stopwords then stripped the rest, and finally I calculated the number of occurrences for each word, so I could know the most repeated words.
And here are the top of the list [word : frequency] :

Job requirements :

experience	15245
work	10546
skill	10194
year	10149
communication	8520
knowledge	8507
good	6170
english	5843

Job descriptions :

customer	17826
develop	16069
product	13392
design	12541
manage	12209
work	11886
market	10328
project	10240

From job requirements we can see that experience is the most important thing to companies, also skills, communication, knowledge and english are the top words written in job requirements.

6. Summary

To summary up what I gathered from the data in points.

- After Jan,2015 for most jobs, the range of the salaries is raised by 500 LE.
- The minimum number of experience years needed by most jobs are 2 years.
- Number of job views , open vacancies and job applications posted on wuzzuf are increasing over time, correlated with each other, have drops and rises, and all of them are strongly dropped in the last two months (April and May 2016).
- Cairo, Giza and Alexandria are the top cities with jobs with 15653, 4117, and 1162 job applications in order.
- Min and max salaries are correlated with each other, whenever one of them rise/drop the other goes with it.
- The word “experience” is the most frequent word in the requirements. After it comes “skills”, “communication”, “knowledge” and “english” which are the top words used in the job requirements field.