Saturday, April 25, 2015

Thumbstack Take Home

Introduction

The question to investigate was the following:
" Is there evidence that product changes over the last two months have caused site-wide shifts in quoting behavior?"

My answer:
After a fairly thorough investigation, I have not seen any evidence of product changes that caused site-wide shifts in quoting behavior

As you can find in the README, here was my general approach to the problem:

1. Data Exploration to get a better understanding of the problem and the data

2. Resolving the problem with an MVI : Data Analysis
a) Define in details the operational questions to answer
b) Answer each of these questions

3. Discussion about moving this further into production
a. What are the new data requirements from the analysis
b. Datamart design for an automated analysis
c. Prediction models: Machine Learning applications to this problem

1. Data Exploration

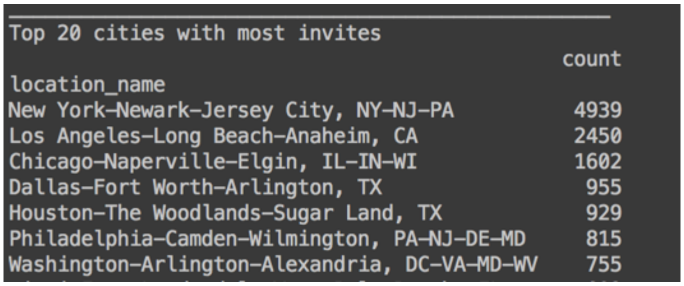
I wanted a rough understanding from the dataset in 3 angles

- An analysis of the requests
- An analysis of the service providers
- An analysis of the customers

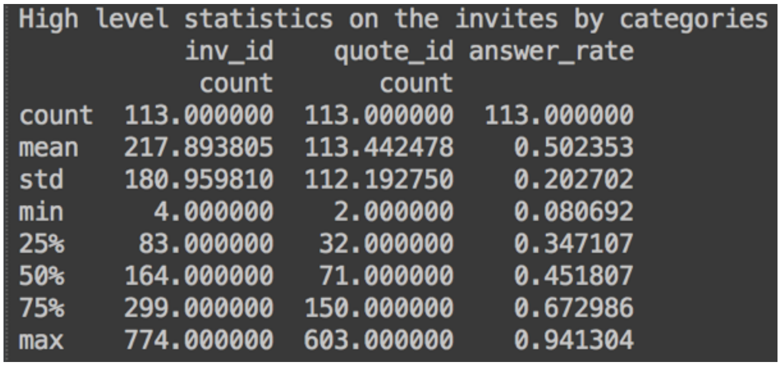
You can see this analysis in the files : requestAnalysis.py, serviceProviderAnalysis.py, customerAnalysis.py

Here are the key takeaways

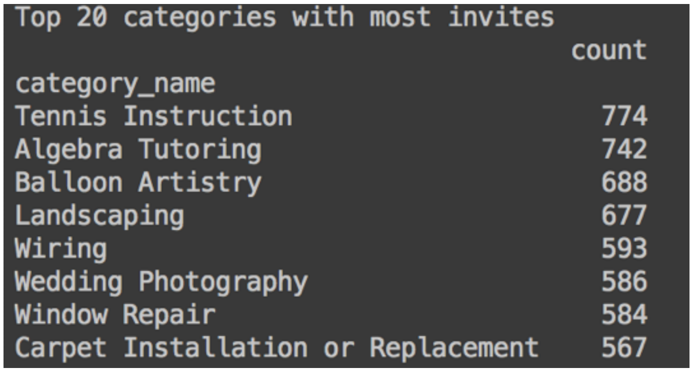
- 4961 requests, 24622 invites, 12819 quotes in the dataset
- Category with most invite: Tennis Instruction
- Location with most invites: New York



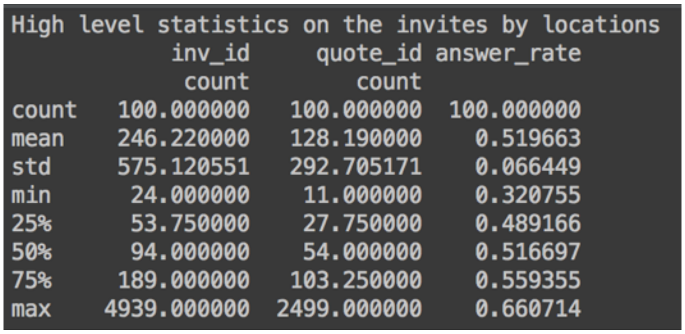
```
Top 20 cities with most invites
                                              count
location_name
New York-Newark-Jersey City, NY-NJ-PA          4939
Los Angeles-Long Beach-Anaheim, CA             2450
Chicago-Naperville-Elgin, IL-IN-WI             1602
Dallas-Fort Worth-Arlington, TX                 955
Houston-The Woodlands-Sugar Land, TX            929
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD     815
Washington-Arlington-Alexandria, DC-VA-MD-WV    755
```



```
High level statistics on the invites by categories
           inv_id    quote_id answer_rate
            count       count
count  113.000000  113.000000  113.000000
mean   217.893805  113.442478    0.502353
std    180.959810  112.192750    0.202702
min      4.000000    2.000000    0.080692
25%     83.000000   32.000000    0.347107
50%    164.000000   71.000000    0.451807
75%    299.000000  150.000000    0.672986
max    774.000000  603.000000    0.941304
```

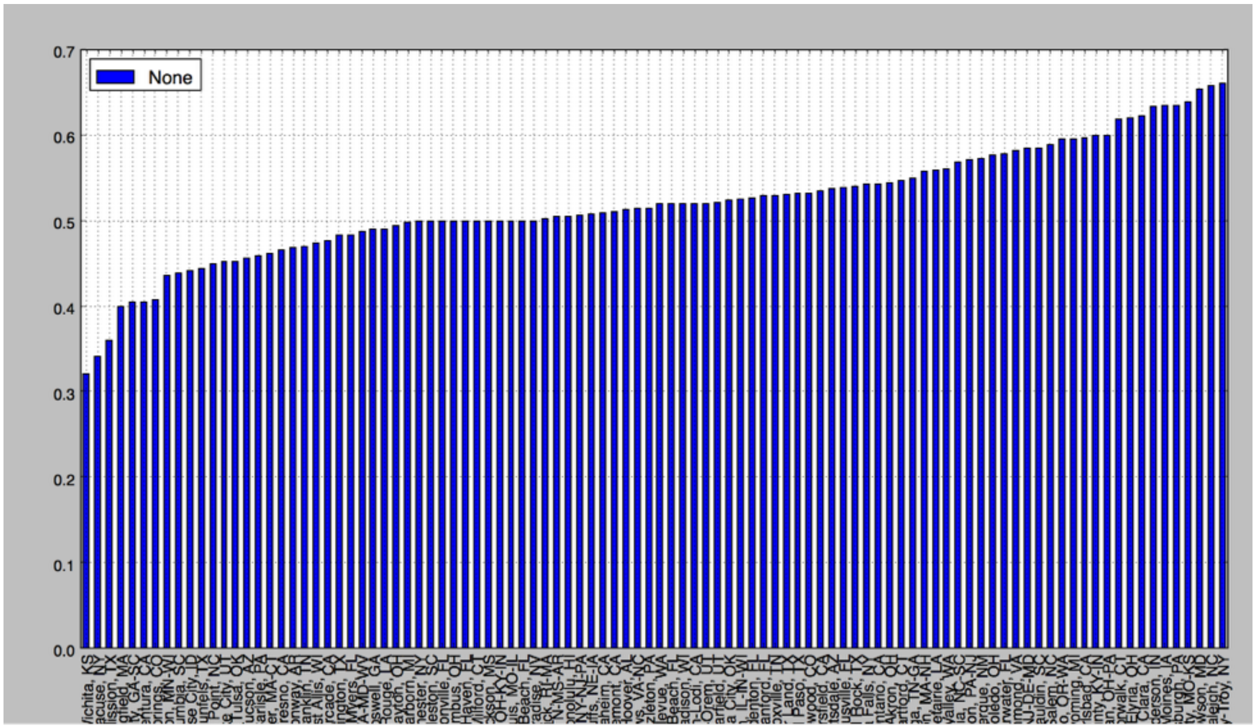


```
Top 20 categories with most invites
                                  count
category_name
Tennis Instruction                  774
Algebra Tutoring                    742
Balloon Artistry                    688
Landscaping                         677
Wiring                              593
Wedding Photography                 586
Window Repair                       584
Carpet Installation or Replacement  567
```

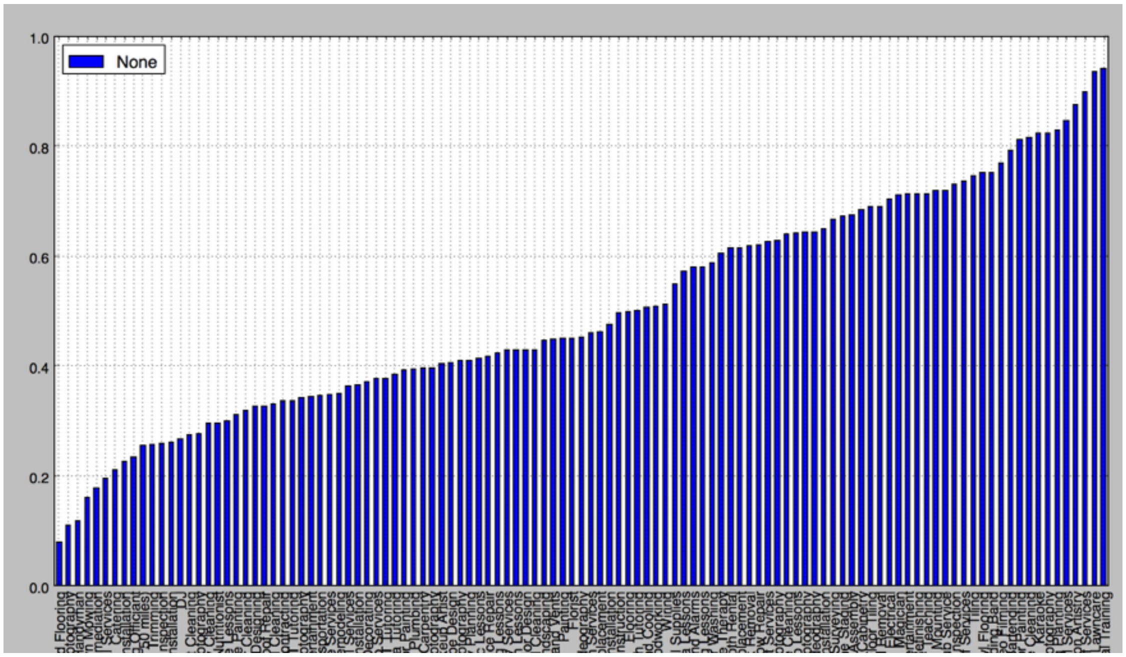


```
High level statistics on the invites by locations
            inv_id     quote_id answer_rate
             count        count
count   100.000000   100.000000  100.000000
mean    246.220000   128.190000    0.519663
std     575.120551   292.705171    0.066449
min      24.000000    11.000000    0.320755
25%      53.750000    27.750000    0.489166
50%      94.000000    54.000000    0.516697
75%     189.000000   103.250000    0.559355
max    4939.000000  2499.000000    0.660714
```

- Overall invite to quote ratio does not really depend on the city (here is the ratio per locations over the 2 months) (many locations around 0.5 - the avergage)

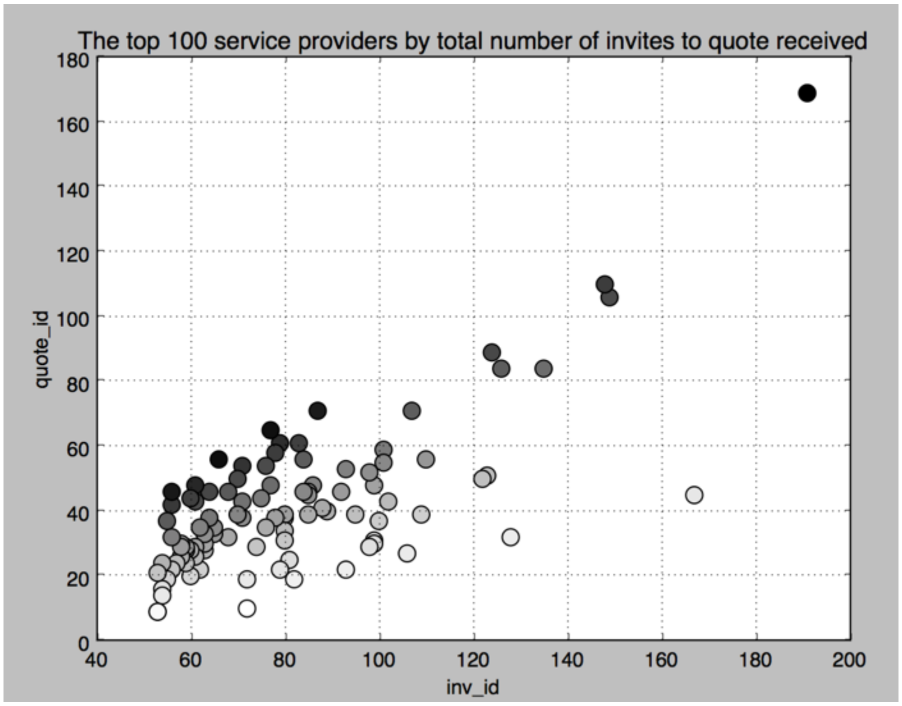


- Overall invite to quote ratio depends a lot on the category (can go from 0.1 to 0.8

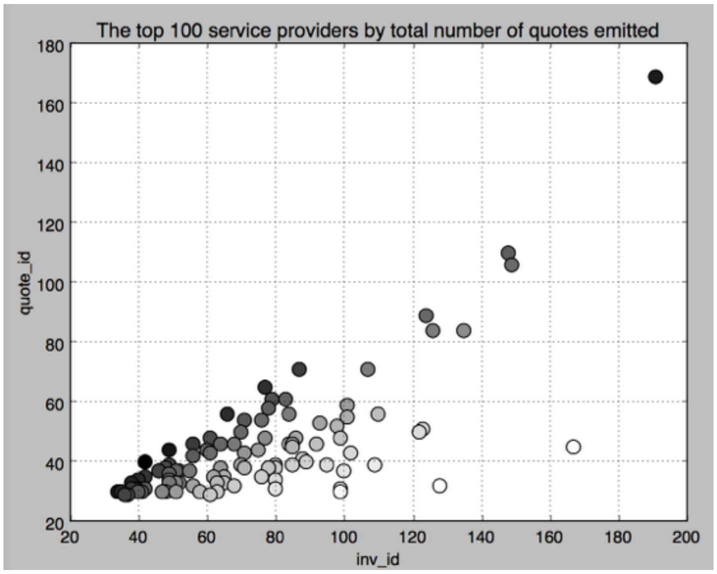


So we can expect that in some categories, some service providers are more “Professional” and answer to the invites with quotes

- Here (the color gradient is the answer ratio) we see among the top service providers receiving the most quotes that there is equally a number of them who answer a lot (dark points) and who does not (white points)



- And we can see from here that the “good service providers” who answer a lot do not necessarily receive a lot of invites



2. Question Analysis

Is there evidence that product changes over the last two months have caused site-wide shifts in quoting behavior?

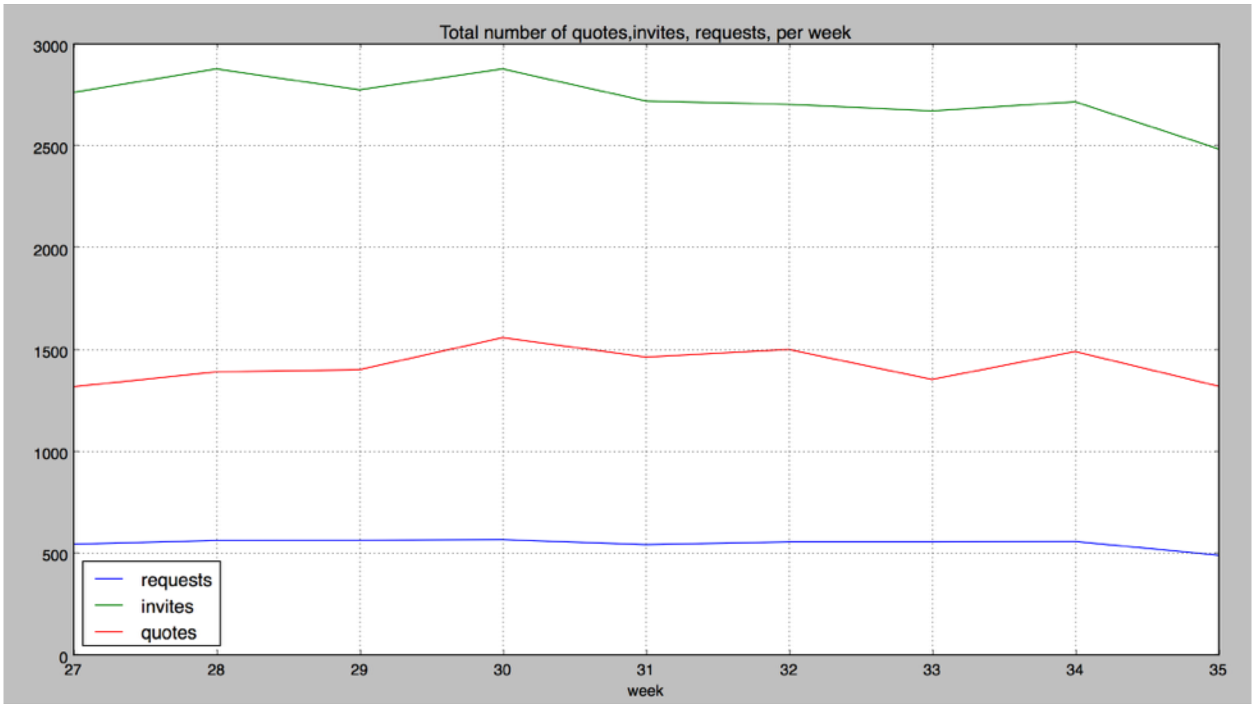
This will be in the file question.py

a. Questions to investigate

- Is there an evolution of the invite quote to rate over time site wide?
- Is there an evolution of the time at which the invites were sent to service providers?
- Is there an evolution of the delay between sending a quote after receiving a quote for service providers?
- What is the detailed evolution of the invite to quote rate over time for
  - the top categories
  - the top service providers

b. Results

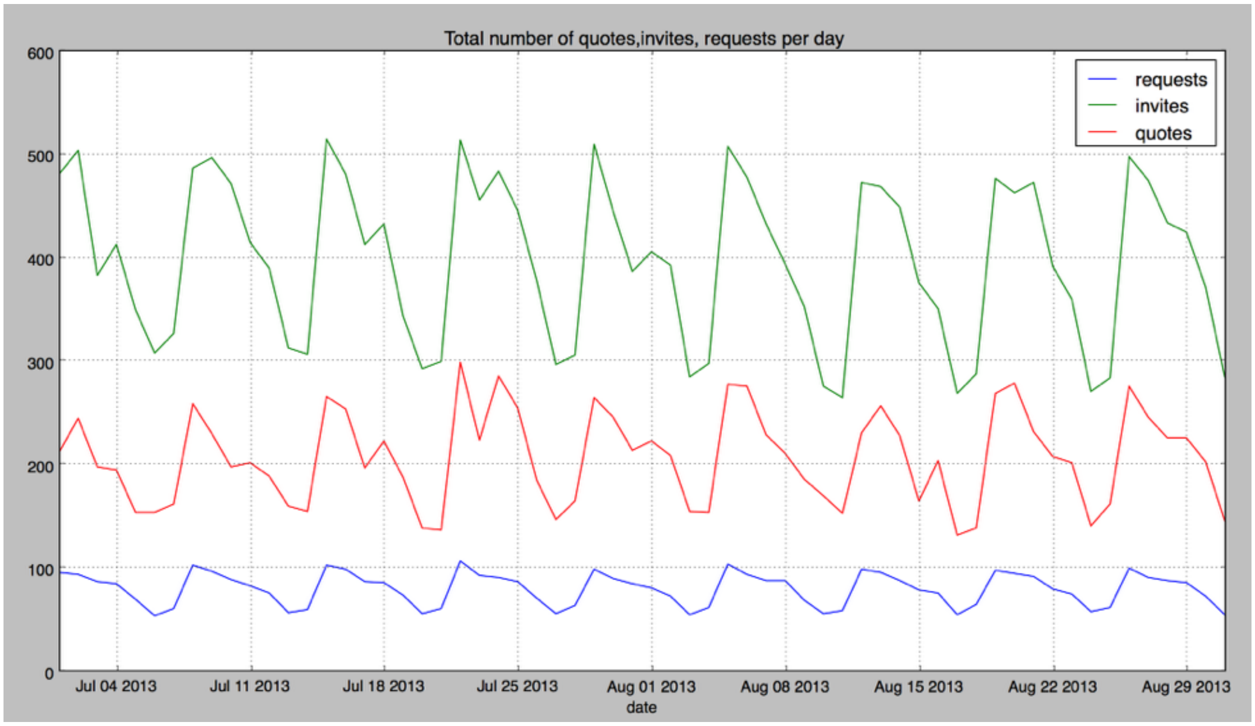
At a high level, the ratio of invite to quote is constant over the 2 month period
The number of invites and quotes are very proportional to the number of requests



If we now look at this metric at the day level, we cannot see any change either

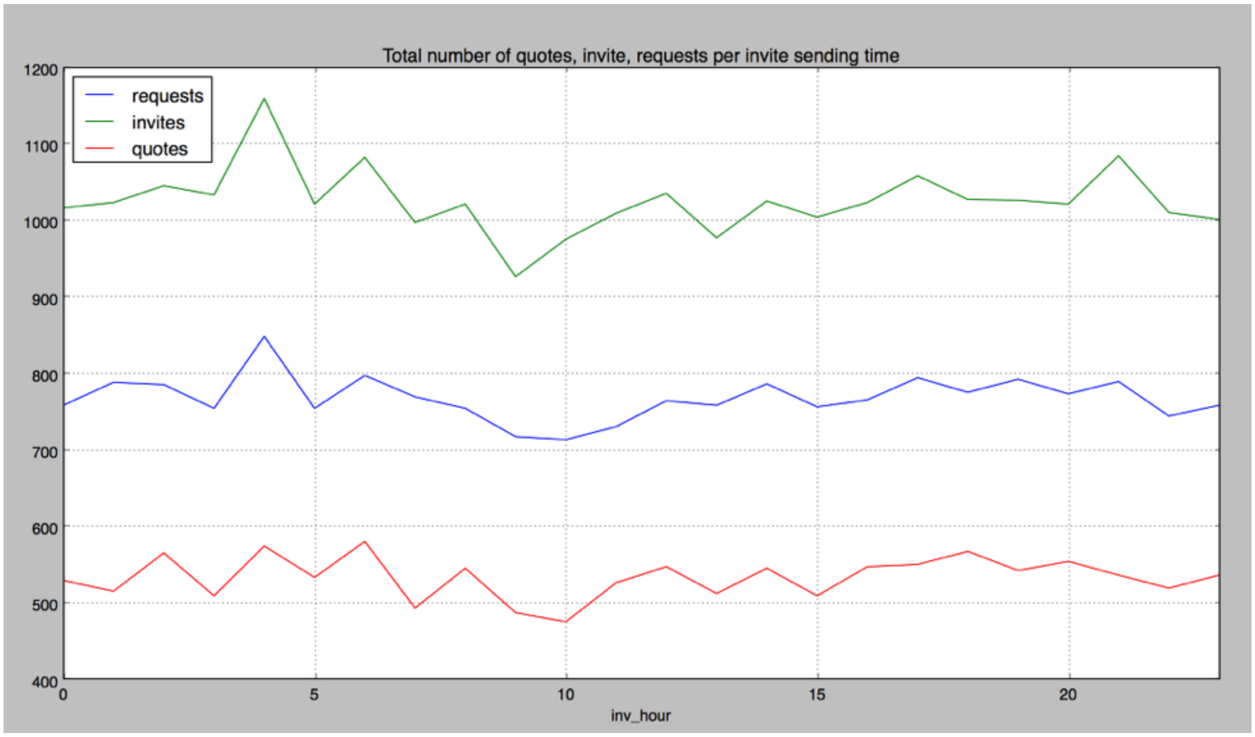
Saturday, April 25, 2015

It is only interesting to note that (the 4th of july 2013 being a thursday), we see a trend that people tend to send less requests on the week ends



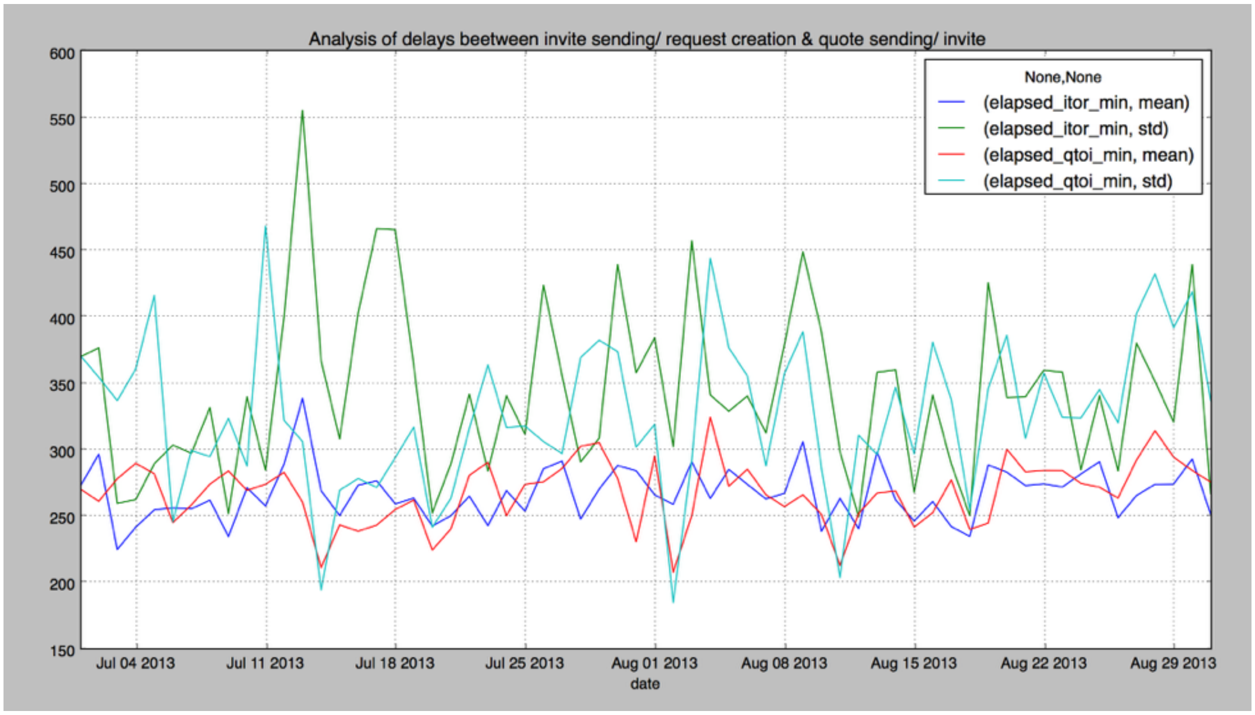
The only “abnormality that we can see on this graph is on August 16th, the only day where quote rate is higher than usual

There is no indication that there is a specific time where more invites are sent to customers

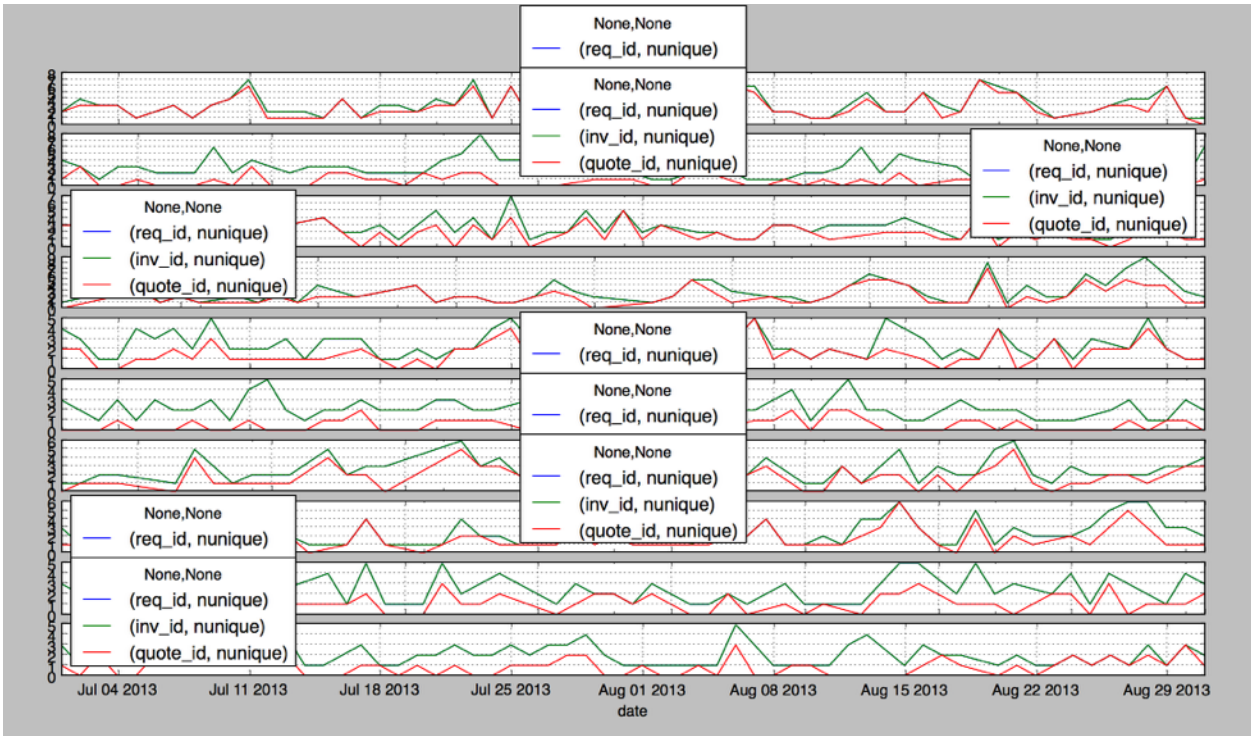


4 AM seems to be a peek but I imagine that there is a server time in west coast and that it corresponds to the time in East Coast (to be verified). That would be the numerous NY requests

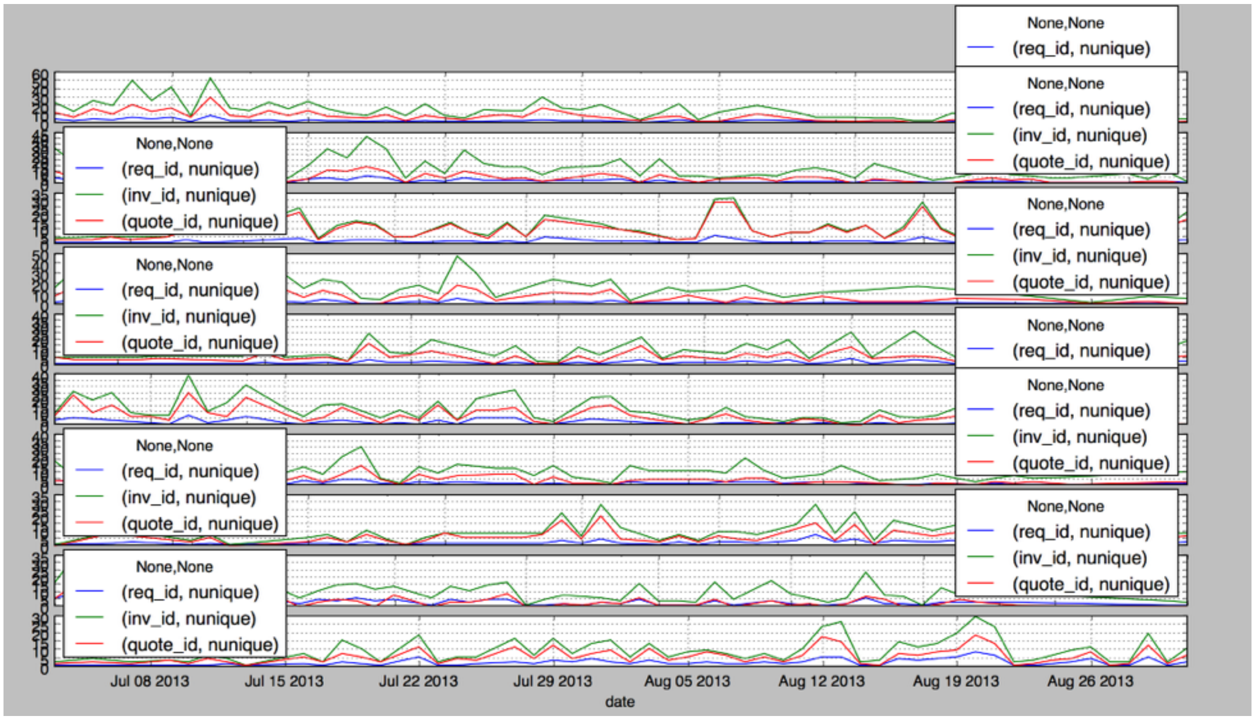
There hasn't been any trend in the changes of the time elapsed between invite to quote either



And if we look at the top 10's over time, no real clues either..
Below, when looking at the top 10 services provides in Number of quotes sent, we don't see a pattern.
There are service providers which have a higher answer ratio, but it does not evolve over time.
If we take the service provider one for example, his ratio is always great, so is number 3
But number 4 for example, always stay at a fairly low rate.



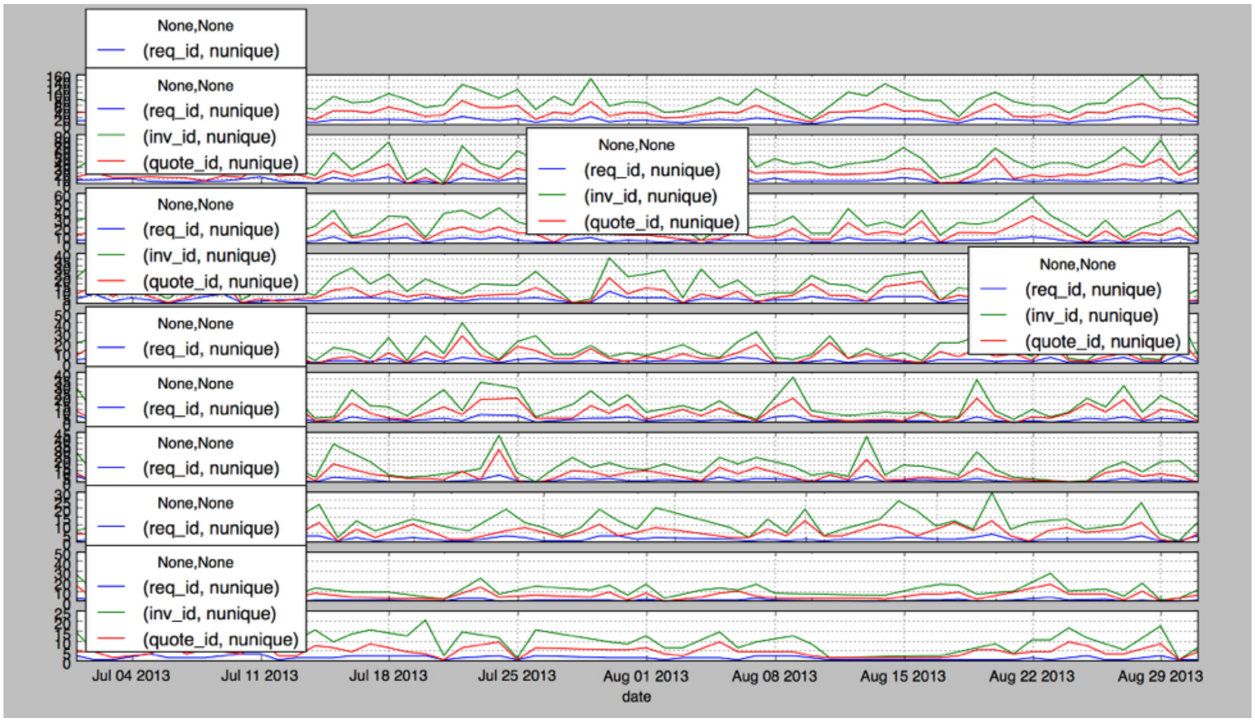
And the breakdown per top 10 categories does not bring much more information



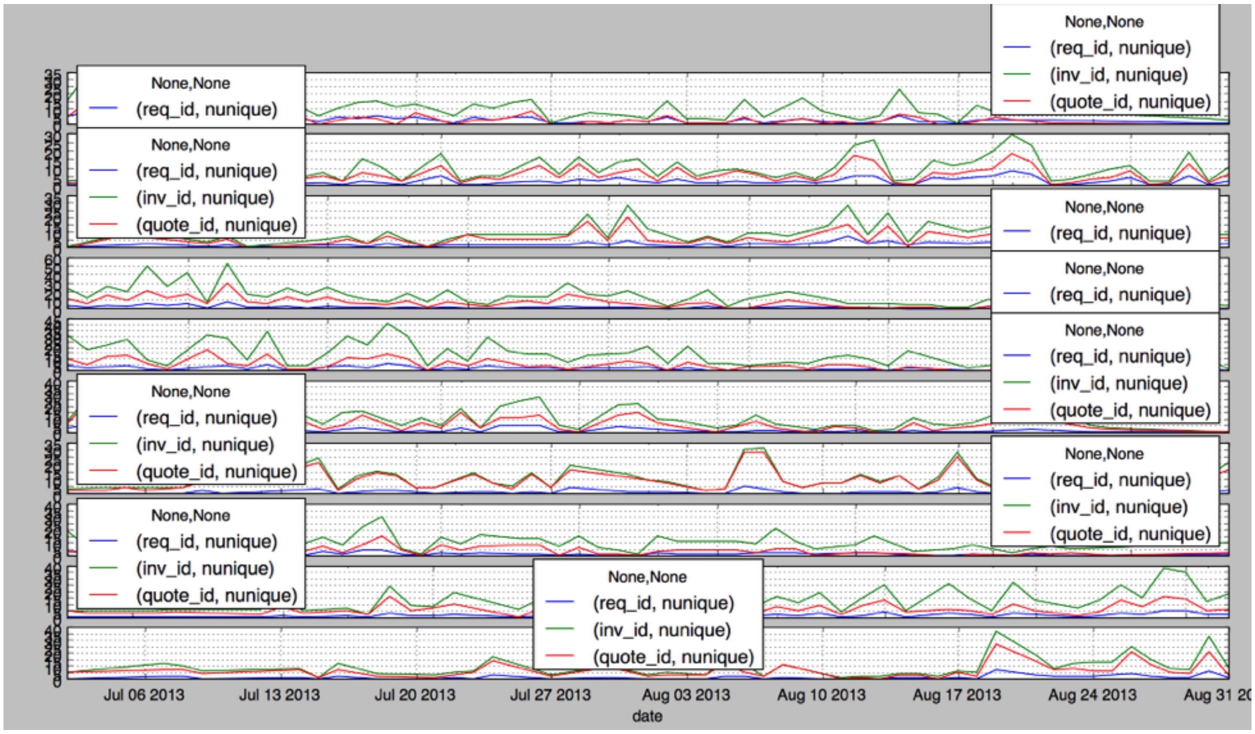
Saturday, April 25, 2015

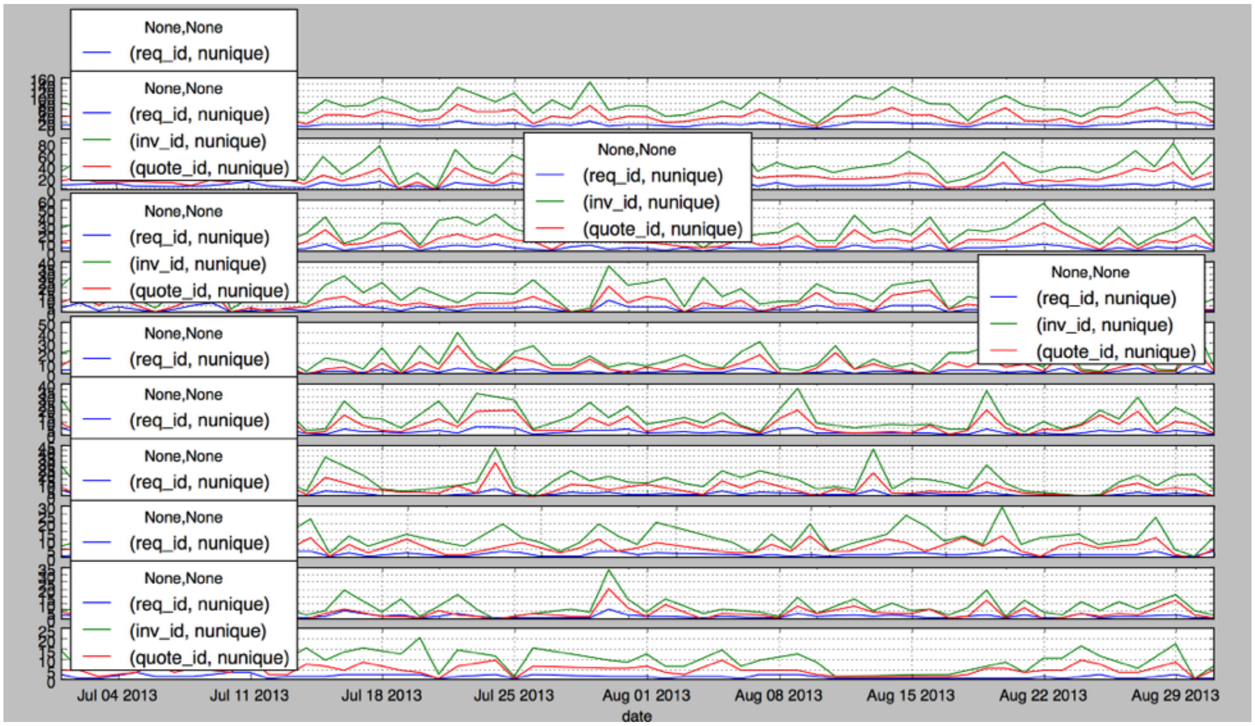
On thing to notice is that the rate for category 3 is always very high. a “professional” category. but not overall increase or decrease for several categories at the same time.

The last thing to consider is the breakdown per location.
Here again, nothing obvious



And considering the top 10 categories and location per requests receive does not show anything either





At this point, a blind investigation seems rather unproductive and I would ask for what the product changes have been in order to verify hypothesis , assumptions that were made on the expected impact of product changes .

3. Discussion about going further with investigation

Some technical considerations:

The analysis was done in python using pandas library.
I have imported the data quickly with a SQL join across all table.

This only works with such a small dataset. In production, we would have to do it differently and use a data mart with a push on event (service completion)

a. What are the new data requirements from the analysis

Check if timestamps are from servers or from locations.

b. Datamart design for an automated analysis

Creation of measure tables for quotes, requests invites and dimensions table of time, location and categories

c. Prediction models: Machine Learning applications to this problem

Since we have seen such a variations of ratio for users and categories, we could probably learn what will likely lead to a conversion and send an appropriate number of invites per requests (to not discourage service providers)