An analysis of Thumbtack user behavior over time: focus on product changes and quoting metrics

Dr. Rozmin Daya, Data Scientist Applicant

March 5, 2015

1 Introduction

Thumbtack is an online marketplace allowing consumers to obtain quotes, and eventually, services, from providers in a very simple way:

- A consumer user makes a **request** entering their location and their need,
- Thumbtack sends up to five matched service provider users an **invitation** to bid on the request,
- If the service provider user accepts this invitation, they send a **quote** to the consumer user.

A critical question is, are service providers becoming more or less inclined to quote over time? This can be assessed by measuring the invite-to-quote rate, R_{i2q} over time. Specifically, it is important to understand how changes to the product effect quoting metrics R_{i2q}, and the overall quote rate over time, R_q.

The analysis presented in this report focuses on this issue, and is performed using an artificially constructed example dataset provided by Thumbtack to the author, described at a basic level in Section 2.1 and in detail in Appendix A. First, the question of whether or not product changes have occurred will be addressed; next, quoting metrics will be measured; finally, the two will be checked for correlation.

2 Data

2.1 Raw data

The raw data used is stored in a sqlite database file containing six tables: *categories*, *invites*, *locations*, *quotes*, *requests* and *users*. The table parameters are listed below.

- **categories**: 113 tuples across 2 attributes, the category ID and the category name.
- **locations**: 100 tuples across 2 attributes, the location ID and the location name.
- **users**: 5,961 tuples across 2 attributes, the user ID and email address.
- **requests**: 4,961 tuples across 5 attributes, the request ID, user ID, category ID, location ID and datetime the request was generated. It was found that each consumer user in the dataset corresponds to exactly one request.
- **invites**: 24,622 tuples across 4 attributes, the invite ID, request ID, user ID, and datetime the invitation was sent.
- **quotes**: 12,819 tuples across 3 attributes, the quote ID, invite ID and time the quote was sent.

In the above, identically named attributes refer to the same parameter across tables. The only exception to this is 'sent_time', an attribute for both the *invites* and *quotes* tables: in the first case it refers to the time an invitation was sent from Thumbtack to a service provider user, in the second, the time when the service provider user sent a quote to the consumer user.

2.2 Transformations or Changes to the Dataset

A new table called *requests_new* was added. It is identical to the *requests* table, except that a new attribute, 'meta_category_id', has been added. This is a factor that takes five levels, as listed below. Each category belongs to a meta-category; for a complete listing of which categories belong to which meta-category, see Appendix A.1.

1. Photography (including Wedding and Event Photography)
2. Event planning
3. Home related (decorations, repairs, remodeling, etc.)
4. Lessons
5. Health (includes some lessons, such as yoga, personal training)

3 Exploratory Analysis

3.1 Product changes

Possible ways that product changes could effect R_{i2q} and R_q:

- Thumbtack could change the number of invites given per request. If low-quality invites (those less likely to result in quotes) are decreased, then one expects to see R_{i2q} increase and R_{invite} decrease, while R_q increases or remains constant over time. If high-quality invites are increased, then the expectation is that R_{i2q}, R_{invite} and R_q all increase.
- It is possible that new locations were added over time. A new pool of users may change the behavior of R_{i2q} and R_q, especially if those new users are unlike previous ones.
- It is possible that new categories were added over time. If users have more options for categorizing their request, invited service providers may be more accurately chosen, and more inclined to make quotes.

To check for changes in the way invites per requests were distributed, several plots were made. Firstly, the distribution of invites per request across all categories, locations and dates was made, shown in Figure 1a. The number of invites extended per request peaks at 5 and the distribution is slightly right-skewed. To study whether or not the pattern of inviting changed, this distribution was generated for each of the 9 weeks in the dataset, where the time parameter used was the request creation time. Two relevant summary statistics, the median and the standard deviation, were then plotted for each distribution. Figure 1b shows the median of the requests to invites distribution across the 9 weeks. The median is used as opposed to the mean, since the distribution of requests to invites is skewed. It is seen that the distribution of the median is flat over time. The standard deviation of the weekly distributions is plotted in the same way in Figure 1c. This decreases over time. Essentially, the number of invites is 'tightening up' around the median of the distribution–5 invites. The process of inviting seems to be becoming more uniform as time progresses.

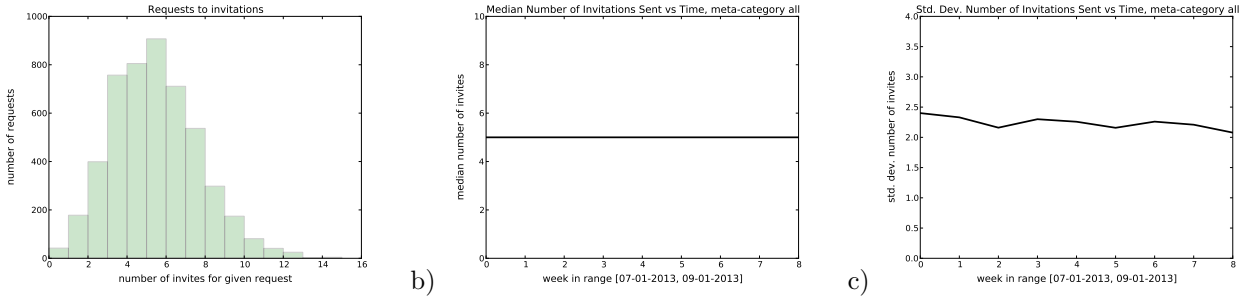


Figure 1: a) Histogram of number of invites made per request, across all categories, locations and dates. b) Median value for 9 distributions similar to the one shown in (a), with one distribution for each of the weeks in the dataset. c) As with (b), but showing the standard deviation of the 9 distributions.

To check if new locations were added over time, the number of unique location IDs was plotted for each week in the dataset, as shown in Figure 2a. As previously, the time used is the request creation time. It is possible that a given location ID was missing from a week due to an absence of requests from the corresponding metro area, rather than due to Thumbtack not having expanded to that area yet. However, such an effect should produce only minor fluctuations. Over time an obvious increase in number of unique location IDs should occur if Thumbtack expands to new areas. However, a notable increase is not seen, and this distribution is not investigated more rigorously.

The same approach was used to check if new categories were added, as shown in Figure 2b. Note that there seems to be a sharp decrease for the last bin in this plot. It is possible that this is partially due to the fact that this 'week' is one day shorter than the others. While it spans the dates [08-26, 09-13], the last request is made on 08-31. As a check, the same plot is generated using daily binning, shown in Figure 2c. Upon examination of this plot, the decrease at the end of the time period seems to be a real effect. The plot of Figure 2b will be used for further analysis. Note: it is possible to fill the missing day in the last week with generated values, either randomly selected or predicted by a model. While that is not done for this report, it should be done for analysis of real data.

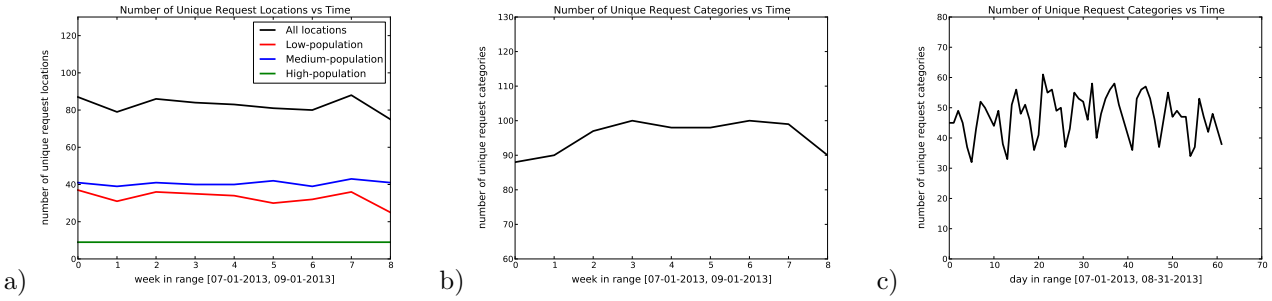


Figure 2: a) Unique location IDs over the 9 weeks of the dataset. Results are plotted for all areas, for areas with large population (greater than 7 million), with medium population (between 1 and 7 million) and for low population (less than 1 million). Population information is taken from Wikipedia. b) Unique category IDs over the 9 weeks of the dataset. c) Unique category IDs for the 62 days in the dataset on which requests were created, spanning the range [07-01, 08-31].

3.2 Quoting metrics

Two quoting metrics are studied in this analysis: the overall quoting rate over time, R_q, and the invite-to-quote rate over time, R_{i2q}. Figure 3 shows the overall quoting rate in daily and weekly bins. One of the analysis goals is to obtain a measurement of this parameter. This will be done in the following sections via a fit to the weekly distribution as a more generic model can be used to describe the weekly rate as opposed to the daily one.

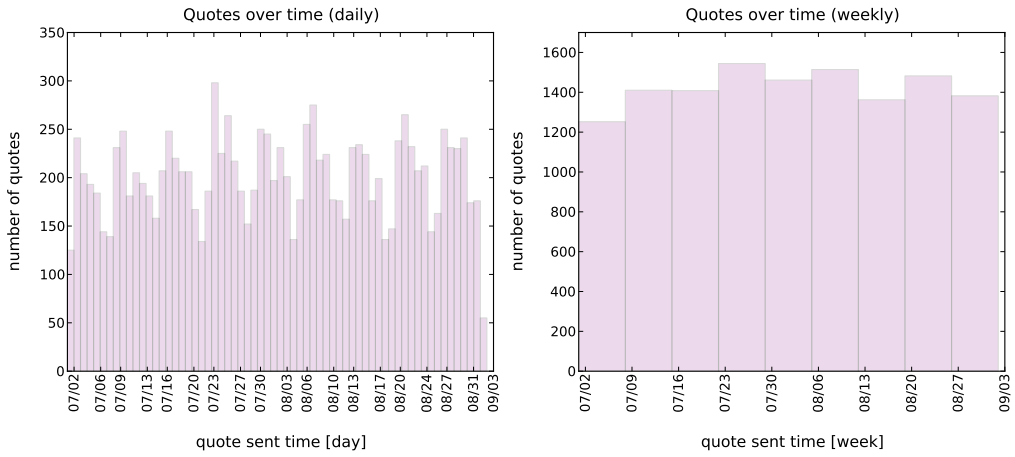


Figure 3: Overall quoting rate, R_q, in daily (left) and weekly (right) bins.

To study R_{i2q}, the ratio of the number of quotes to the number of invites is computed for each of the 9 weeks in the dataset. This is shown across all categories in Figure 4b, and in blocks of meta-category in Figure 4c. The overall distribution binned in days is shown in 4a.

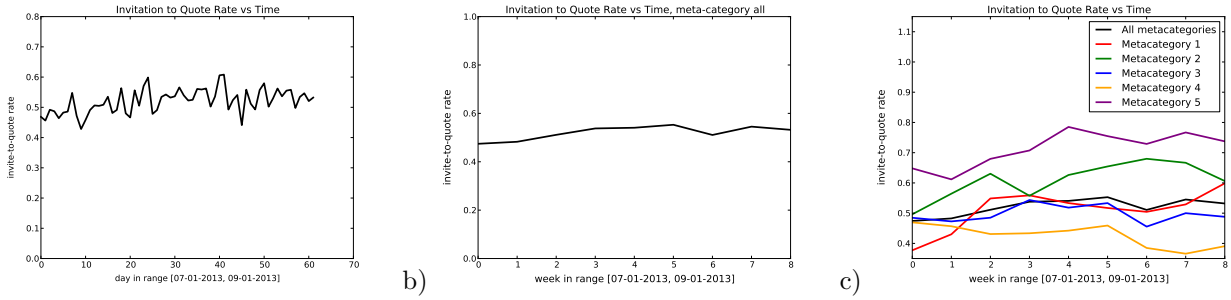


Figure 4: R_{i2q} for all categories (a) daily, (b) weekly) and in blocks of meta-category (c).

4 Modeling

4.1 Training, Validation and Test Data

The goal of this report is to obtain a descriptive model, and making accurate predictions is not explicitly important here. Therefore, in the interest of time, the data is not split into training, validation and test sets. This step is important for predictions, as one avoids overtraining the model, tailoring it too specifically to the dataset being used to derive the model parameters. If a predictive model was being researched, the split of the dataset would depend on it's size and on the computing power necessary for the model being studied.

4.2 Implementation

The distributions of the variables of interest–R_q, R_{i2q} and the number of unique category IDs (N_{cat})–over time display both periodicity and an overall increasing trend. These aspects are examined separately.

4.2.1 trend

A simple linear fit model was used to describe the trend of the distributions. This was done to the weekly distributions, in order to minimize the error of the fit. Figure 5 shows the results of these fits. According to the linear model, R_q increased by 18 ± 10 quotes for each successive week in the dataset, while R_{i2q} increased by $0.7 \pm 0.3\%$ and N_{cat} increased by 0.6 ± 0.7 categories.

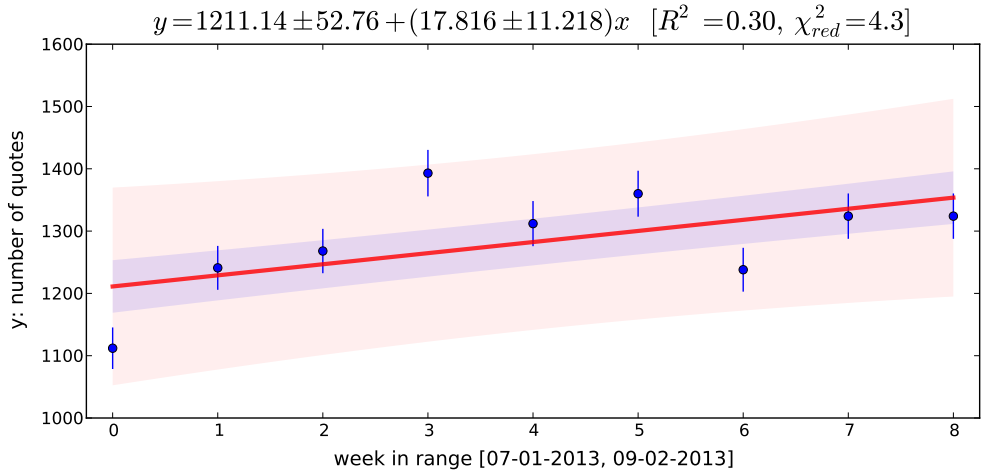
4.2.2 periodicity

Before further studying the periodicity of the daily distributions, the overall linear trend is removed. Next, Python matplotlib's psd() function is used to compute the power spectral density for each of the distributions, using the trend corrected values. The power spectral density function transforms time series to plot the power vs. the frequency. It is a good way of distinguishing noise from actual periodic signals–true aspects of the signal correspond to large peaks on the psd plot. Figure 6 shows the psd for N_{cat}, R_q and R_{i2q}.

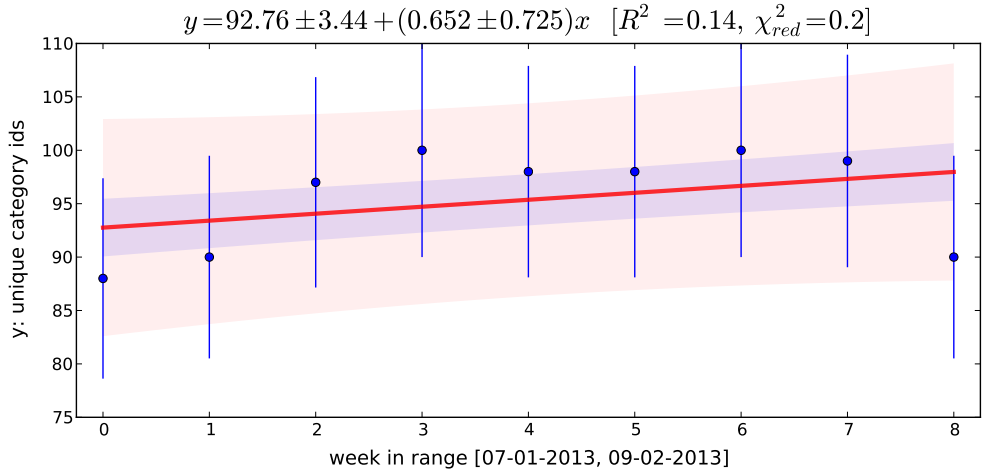
Next, Python matplotlib's cohere() function is used to study the coherence between the quoting metrics R_q and R_{i2q}, and a change that was made to the site over time, the number of categories added, represented by N_{cat}. The coherence essentially indicates whether two time-series datasets are correlated, and if so at which frequencies. Two perfectly correlated signals–for example, if one were to take the coherence of a dataset with itself–would have a straight-line coherence of 1. Figure 7 shows the coherence distributions for R_q with N_{cat} and for R_{i2q} with N_{cat}. In both cases, periodic coherence between N_{cat} and the quoting metrics is observed.

4.3 Reproducibility

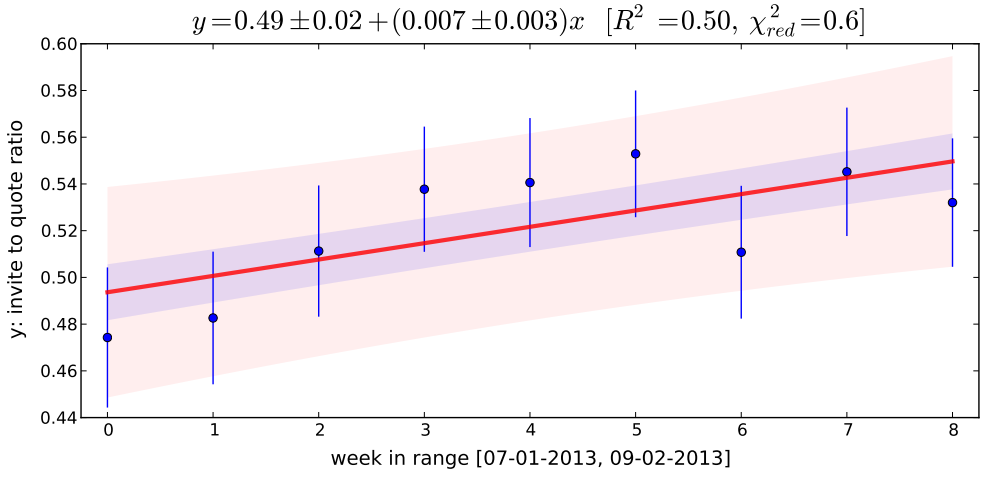
All code used in this analysis is in the 'code' file attached to this report. The README file explains which files are used for which functions.



a)



b)



c)

Figure 5: a) R_q, b) N_{cat} and c) R_{i2q} binned weekly, with linear fits. The uncertainty on data points is estimated using $\sqrt{N}$ (in the case of R_{i2q}, N is the number of quotes used in the ratio). The shaded bands show the 95% (red) and 68.5% (blue) prediction intervals for the fit.

5 Conclusion

A clear correlation was observed between a product change during the time of the dataset (N_{cat} vs. time) and the quoting metrics R_{i2q} and R_q. This was present at both the overall trend level, with all parameters increasing over

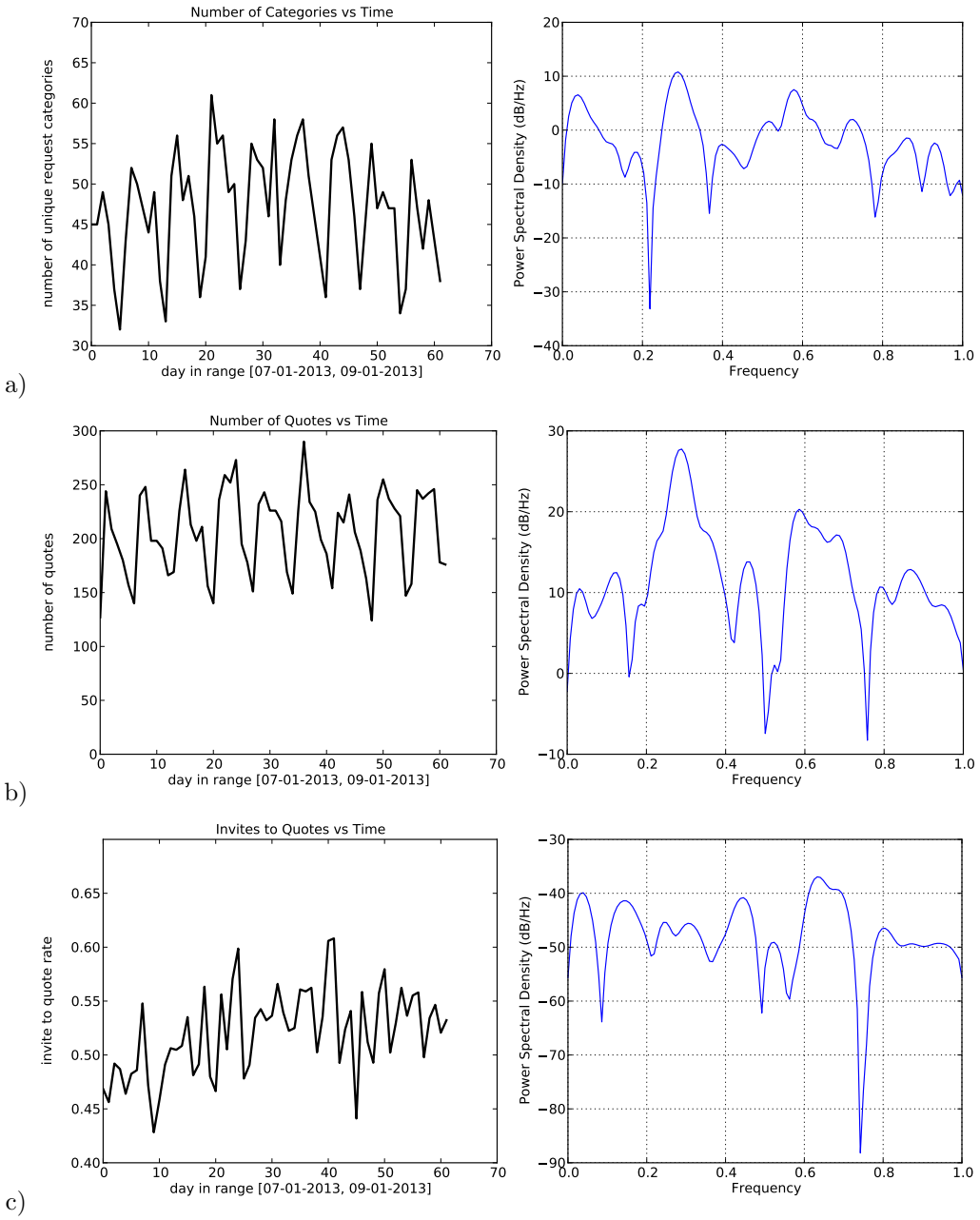


Figure 6: Power spectral density distribution for a) N_{cat} vs. time (daily), b) R_q vs. time (daily) and c) R_{i2q} vs. time (daily).

the 9 weeks described by this data, and at the periodic level. This analysis does not test for causality between the change in N_{cat} and the quoting metrics. However, **if** the change in N_{cat} does cause the increase in R_q and R_{i2q}, it can be said that increasing the number of categories available for users to make requests in improves the invite-to-quote ratio and increases the number of quotes that are given. A possible mechanisms for this as follows:

- If consumer users have a wider range of categories to choose from, they can more specifically identify their need, and Thumbtack can more accurately match them with a service provider user who potentially meets that need,
- If service provider users are better matched with requests that they are well-qualified to fulfill, they are more likely to extend a quote.

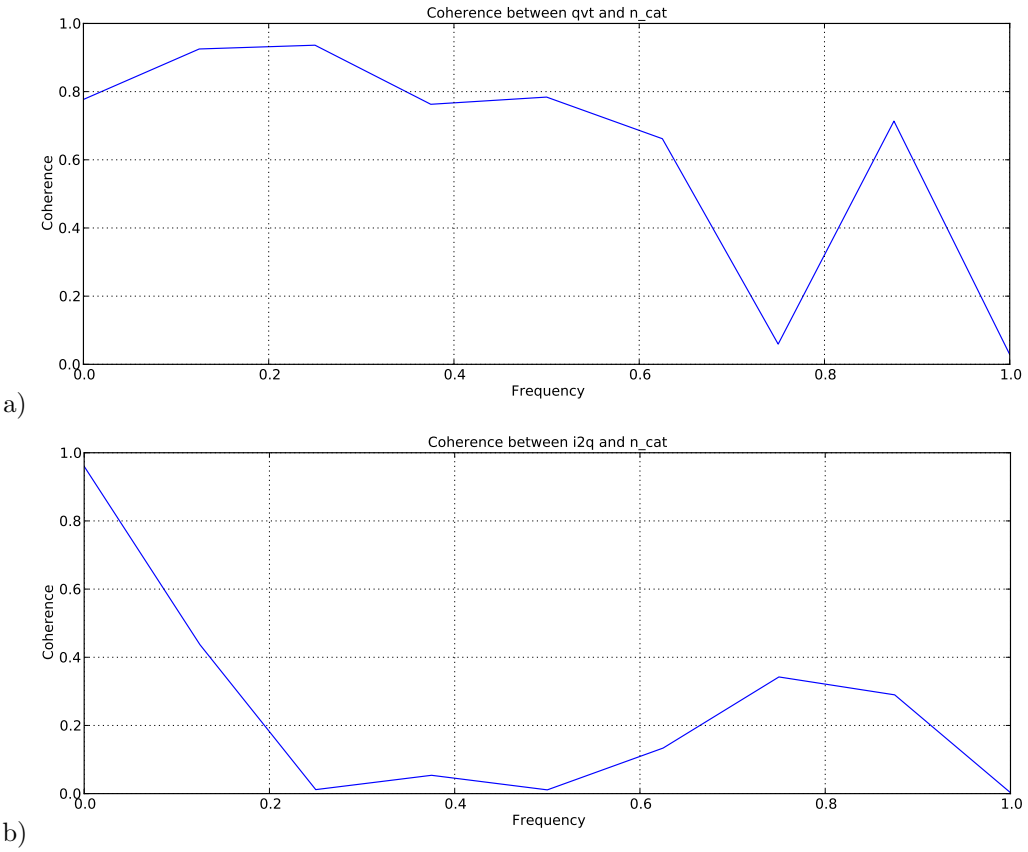


Figure 7: Coherence for a) N_{cat} with R_q, b) N_{cat} with R_{i2q}.

The validity of this hypothesis can be tested in future experiments, using a control group where the number of available categories is held constant over time, and an experimental group where a number of new categories is introduced.

In addition to changes to N_{cat} over time, it was observed that the standard deviation of the number of invites distributed per request was decreasing over time. While this was not studied further, it reflects more uniform product behavior, which could possibly be correlated with improvements in quoting metrics. This would be an interesting area of future study.

A Raw data details

This appendix contains details of the dataset that the author found interesting, but which are not directly relevant to the report.

A.1 categories

The categories table contains 113 tuples across 2 attributes, the category ID and the category name. For the purpose of this report, a new variable, the *meta-category* is identified by levels 1-5, as described in Section 2.2:

1. Photography: Contains 10 categories (1, 3, 19, 24, 47, 52, 64, 81, 82 and 89)
2. Event planning: Contains 26 categories (4, 23, 28, 32, 33, 40, 42, 43, 44, 46, 51, 67, 72, 76, 83, 93, 94, 96, 99, 100, 103, 106, 108 and 110)
3. Home related: Contains 61 categories (2, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 20, 21, 25, 27, 29, 30, 31, 35, 36, 37, 39, 45, 48, 49, 55, 57, 58, 59, 60, 62, 63, 65, 66, 69, 70, 71, 74, 75, 77, 78, 79, 80, 84, 85, 86, 87, 88, 90, 91, 92, 98, 101, 102, 105, 107, 109 and 111)
4. Lessons: Contains 12 categories (17, 22, 34, 38, 50, 53, 54, 56, 61, 68, 95 and 113)
5. Health: Contains 4 categories (26, 97, 104 and 112)

For information on the number of users/requests per category and meta-category, see Section A.3.

A.2 locations

The locations table contains 100 tuples across 2 attributes, the location id and the location name. The locations correspond to Metropolitan Statistical Areas (MSAs) in the USA, and appear to be ordered by decreasing MSA population.

For information on the number of users/requests per location, see Section A.3.

A.3 users

The users table contains 5,961 tuples across 2 attributes, the user ID and the user email address. The user email addresses appear to be randomly generated, with the domain portion of the address consisting of a random string of 6 letters. The name portion of the email address seems to be sampled from the 10 most popular names for boys and girls born in the USA in 2013. The *users* table was studied using the file `users_basic_info.py`, which is attached to this report.

A.3.1 Gender

Using the above stated assumption regarding how the name portion of email addresses was derived, the number of male and female users in the dataset was counted: it was found that there are 2,931 female users and 3,030 male users. This is consistent with a 50% probability of a user being male or female, so it is likely that this selection is random and that no information can be drawn from the users' gender.

A.3.2 Customer or Provider

In order to ascertain the number of users that are consumers vs. service providers, the number of users with no associated requests was counted. These were taken to be service providers. There are 1,000 service providers, having user IDs 1-1000, and 4,961 consumers, having user IDs 1001-5961.

A.3.3 Location

The number of consumer users was counted for each location ID and plotted in Figure 8. In general, the number of consumer users for a location ID is positively correlated with population size for the corresponding MSA. In Figure 8, those bars above the black line correspond to locations where the number of users is more than would be expected based on normalization to the median value, and bars below the line correspond to locations where the number of users is less than this value. The statistical significance of such deviations has not been assessed, as it is not the focus of this report. However, the author suggests that this is an interesting area for future investigation. The median number of consumers across the 100 locations is 20.5.

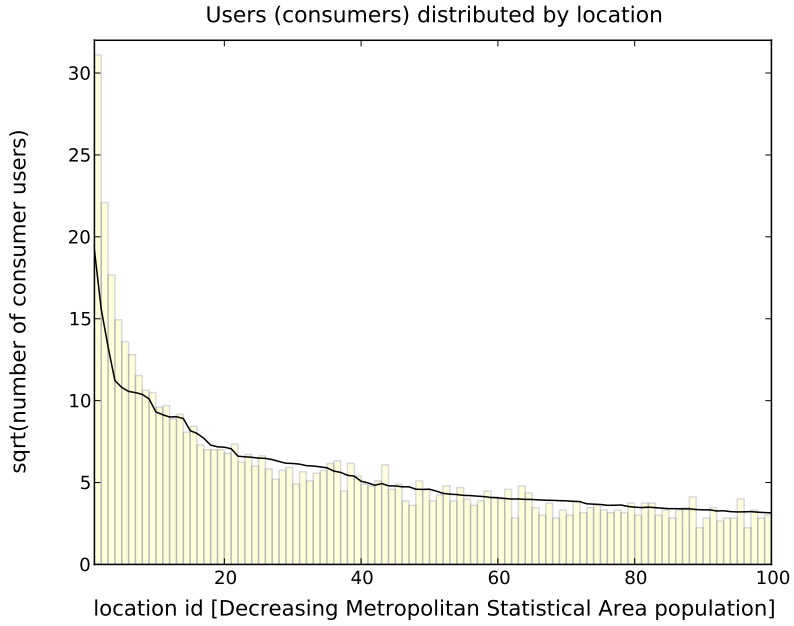


Figure 8: Number of consumer users for each location. The black line shows, for each location ID corresponding to a given MSA, $\sqrt{P_{MSA}/54083.73}$. Here, P_{MSA} is the population of the given MSA, and 54083.73 is a normalization factor obtained by dividing the median population value by the median number of consumers.

A.3.4 Category

The number of consumer users was counted for each category ID and for the meta-categories described in Section A.1, and plotted in Figure 9. The median value of consumers across categories is 34, and across meta-categories is 632, which is indicated by the black line in the plot.

A.4 requests

The requests table contains 4,961 tuples across 5 attributes, the request ID, user ID, category ID, location ID and time the request was generated. Since each consumer user in the dataset corresponds to one request, the distribution of requests by location and by category are essentially already shown in Section A.3. The *requests* table was studied using the file `requests_basic_info.py`, which is attached to this report. All requests were generated in the 2 month time period from 07/01/2013 to 08/31/2013. Generally, the number of requests is highest on Mondays, and decreases over the course of the week.

A.5 invites

The invites table contains 24,622 tuples across 4 attributes, the invite id, request id, user id, and time (and date) the invitation was sent.

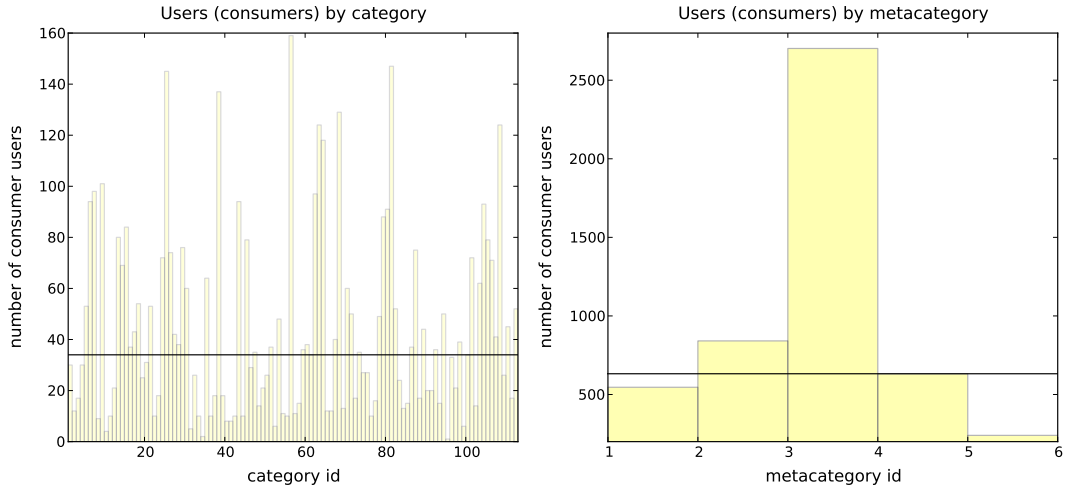


Figure 9: Number of consumer users for each category (left) and meta-category (right). In both cases the black line shows the median value, 34 and 632, respectively.

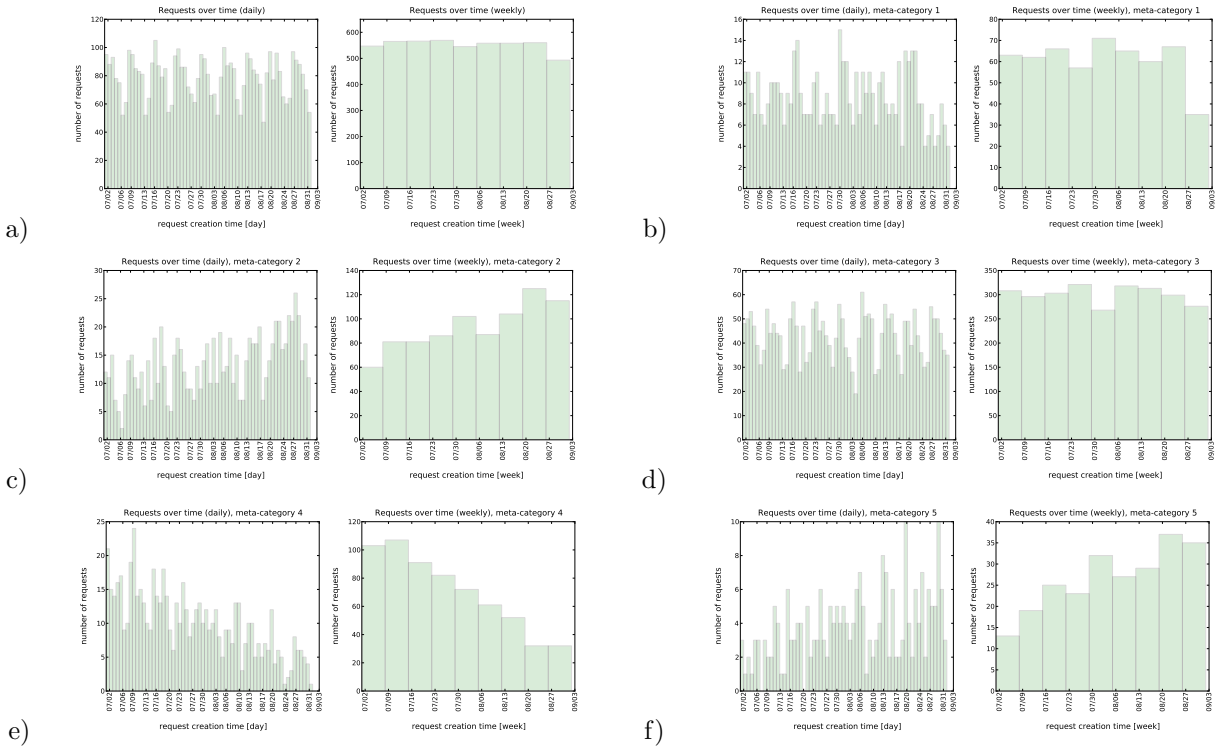


Figure 10: Number of requests made vs. time request was created, shown in one day and one week bins, a) across all meta-categories, b) for meta-category 1 (photography), c) for meta-category 2 (event planning), d) for meta-category 3 (home related), e) for meta-category 4 (lessons) and f) for meta-category 5 (health).

A.6 quotes

The quotes table contains 12,819 tuples across 3 attributes, the quote id, invite id and time the quote was sent.

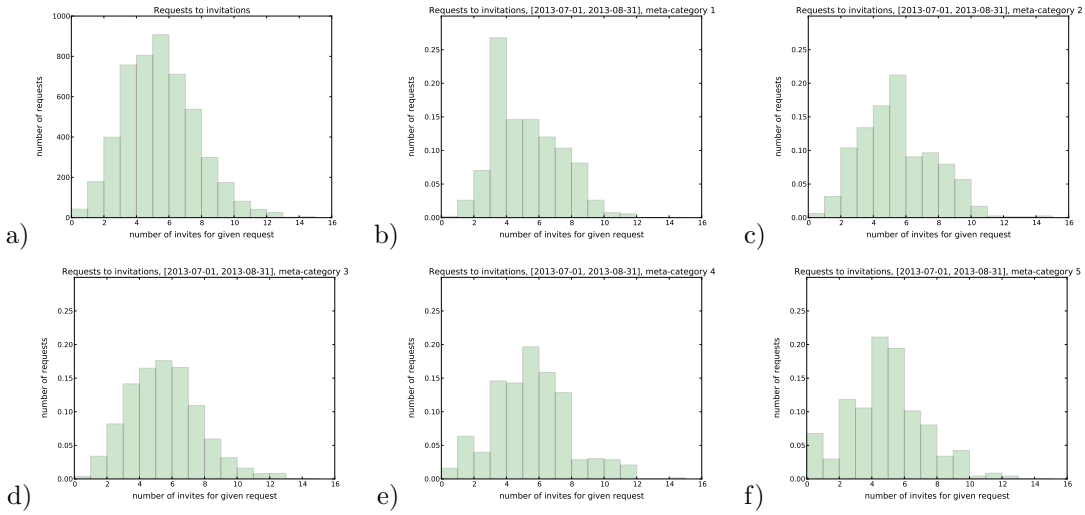


Figure 11: Number of invites given per request, a) across all meta-categories, b) for meta-category 1 (photography), c) for meta-category 2 (event planning), d) for meta-category 3 (home related), e) for meta-category 4 (lessons) and f) for meta-category 5 (health).

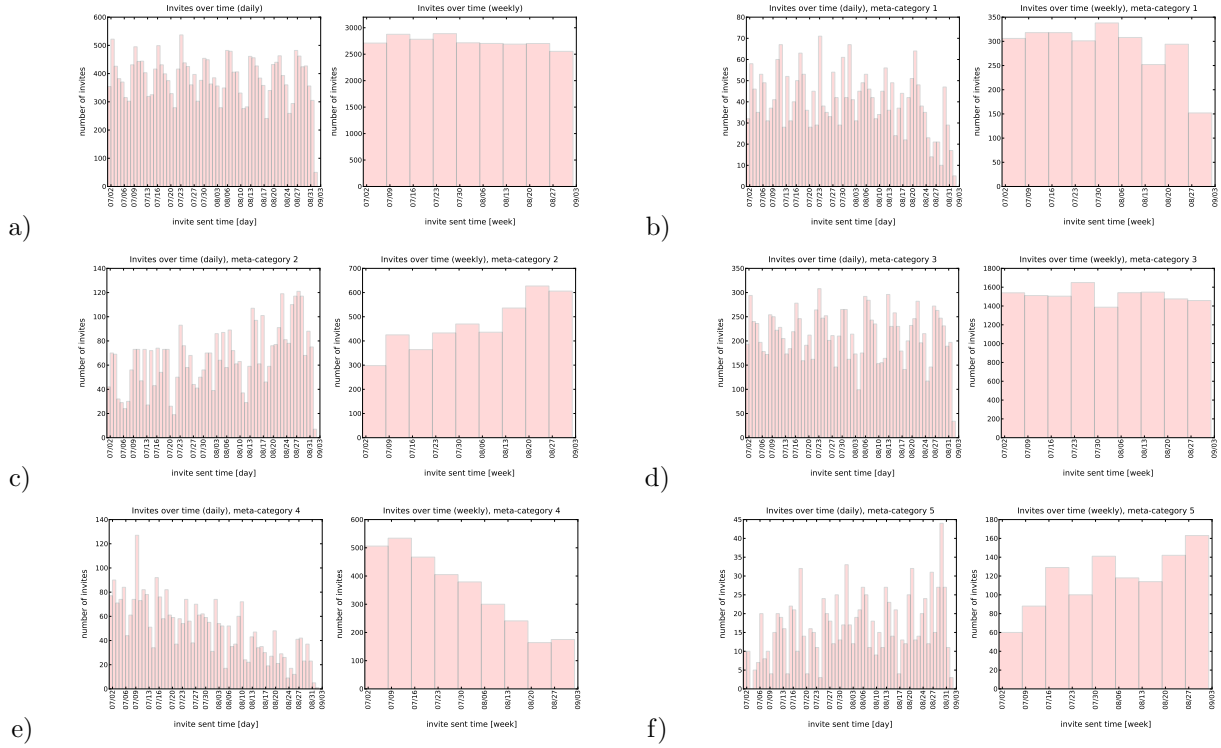


Figure 12: Number of invitations extended vs. time invite was sent, shown in one day and one week bins, a) across all meta-categories, b) for meta-category 1 (photography), c) for meta-category 2 (event planning), d) for meta-category 3 (home related), e) for meta-category 4 (lessons) and f) for meta-category 5 (health).

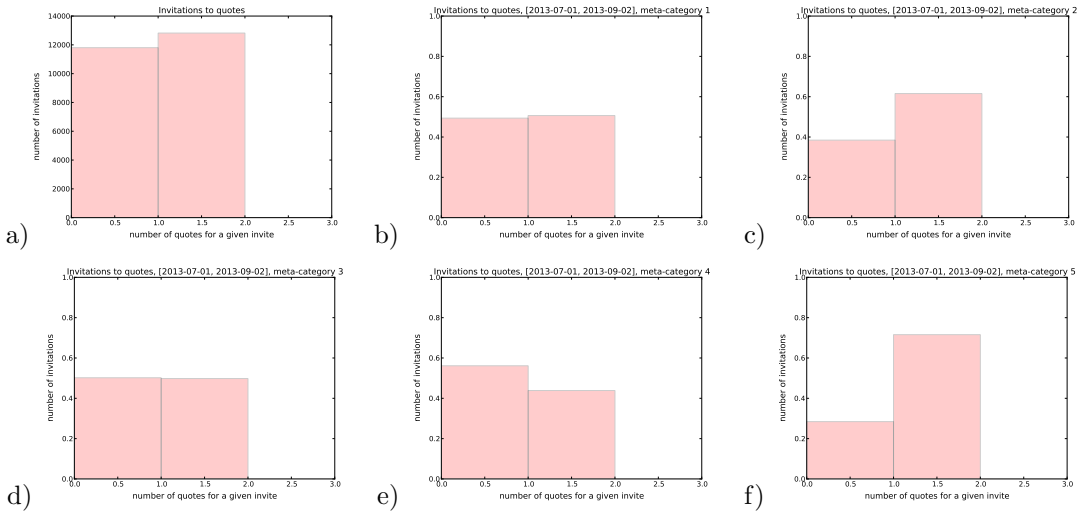


Figure 13: Number of quotes given per invite, a) across all meta-categories, b) for meta-category 1 (photography), c) for meta-category 2 (event planning), d) for meta-category 3 (home related), e) for meta-category 4 (lessons) and f) for meta-category 5 (health).

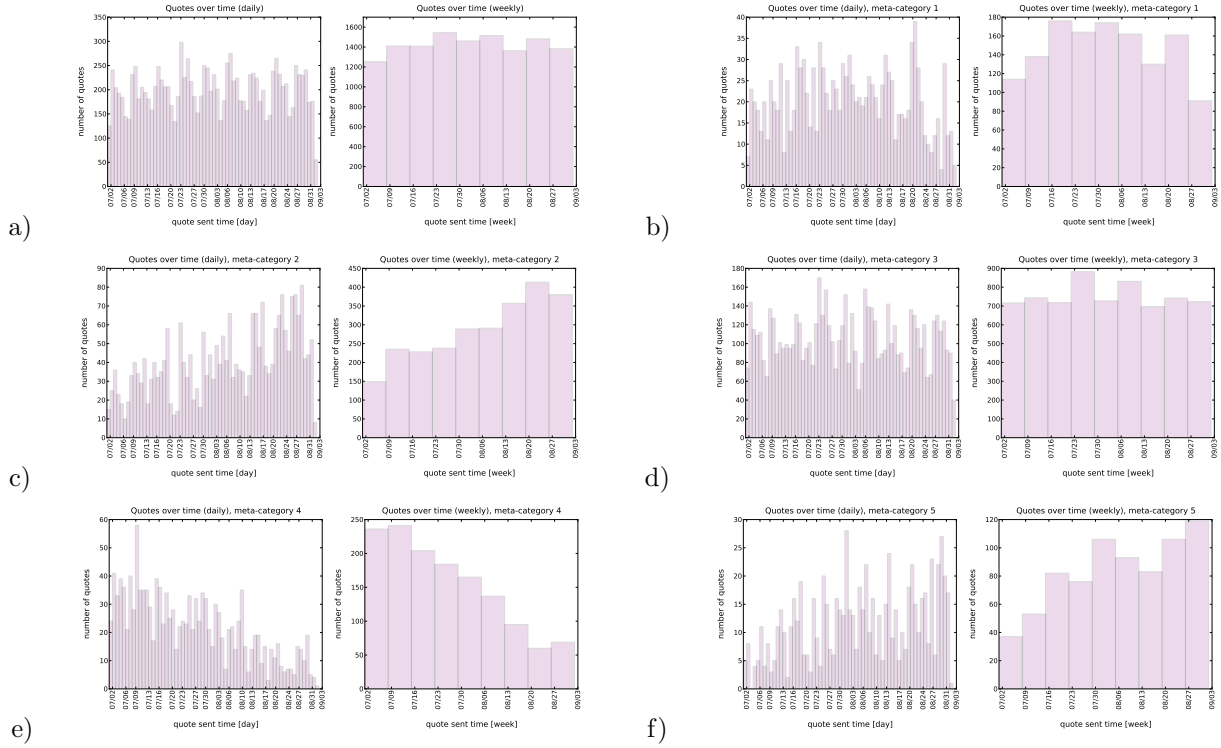


Figure 14: Number of quotes given vs. time quote was sent, shown in one day and one week bins, a) across all meta-categories, b) for meta-category 1 (photography), c) for meta-category 2 (event planning), d) for meta-category 3 (home related), e) for meta-category 4 (lessons) and f) for meta-category 5 (health).