Yelp Project 10 Capstone - An Analysis Report of the Yelp Data on Ambience and Users

Phil Burdi

Thursday, October 15, 2015

Title

Yelp Ambience Marketing Study

Introduction

In there a correlation between ambience in lodging establishments and attributes to the users who post these reviews? My Planned investigation and report work centered around the Yelp data of the businesses “Ambience” data frame (located in the “yelp_academic_dataset_business.json” data, but after determining there was insufficient data for testing the correlation with lodging (motels/hotels). The testing data **was** available using restaurants and the decision was made to switch the category. The first focus was the business that provided ambience data and this lead to the user that created a review to determine a targeted audience for a specific ambience in that type of food establishment. The analysis will look at other data files for user tips, checkins, reviews, and profile to determine if a correlation exists with attribute of the user.

The analysis will evolve in discovery to new areas uncovered in the analysis.

Methods & Data

The data-file yelp_academic_checkin.json was not used for this analysis. Graphs will be done using plot and ggplot2. Due to the heavy (RAM) memory requirements and processing time, all pre-processing was alo done earlier and saved in the global environment using the **pre-load_env.RData** file. Details on the pre-processed code can be found in the **pre-load_env.R file**. The data loaded into the pre-processed environment includes:

1. The yelp_academic_dataset_user.json - pre-read and entered into the environment as **juser**.

2. The yelp_academic_dataset_tip.json - pre-read and pre-process in the environment as **njtips**.

3. The yelp_academic_dataset_review.json - pre-read and pre-process in the environment as **nreviews**.

4. The yelp_academic_dataset_business.json - pre-read and entered into the environment as **jbusiness**.

To begin this analysis the methods using **R** and **Rstudio** are included in this work. Using *grepl* for example, to identify all the categories in the hotel industry and then later for restaurants. Matching this with the those in the ambience data frame.

Once the restaurant category was extracted and after processing only those businesses that have data in the ambience section. We’ll build another data-frame just for this group. Once examined, we continue to drill down into the data use a selected ambience called *trendy*. It doesn’t contain the most data of the least data but the number of observations is significant and can be used to map (first in charts) for relations in density to tips and reviews, but also to examine against the user data.

Our next stop is the **bus_map** data-frame, this includes the *business ID*, the number of *tips* and the number of *reviews*. Although our focus is not on the business, this produces a wealth of information that could indicate the amount of traffic from clients. However, for our purposes we are looking for patterns in *tips vs. reviews*, *tips vs. business*, and *reviews vs. business*. Overlapping patterns are evident in the *tips vs. reviews*, we’ll explore that later in this results analysis.

Our focus now turns to the users, we’ll examine users by cross referencing the *business_ids* into the nreviews and ntips data file to get the *user_id*’s for all tip posters and reviewers. We’ll then look and compare the top 100 trendy users (reviewing and providing tips) to the overall user population in the Yelp user’s data-set.

One area the Yelp data covers very well is the ability of the user to connect with others in their reviews, profile, tips, etc. These fields include Fans, Friends, and a multitude of compliment fields. To further dive into popularity of the user, we’ll focus on the *Fans* attribute in the **juser** data. Once the data is graphed, we begin to look at the average number of fans for the “typical” user and those who are reviewing and providing tips to the trendy restaurant.

Finally we perform a linear model for correlation testing between the user *review_counts* and the number of *Fans* using “lm” and “with” on the tip_posters & reviewers, graphing the data help to show the significance.

Results

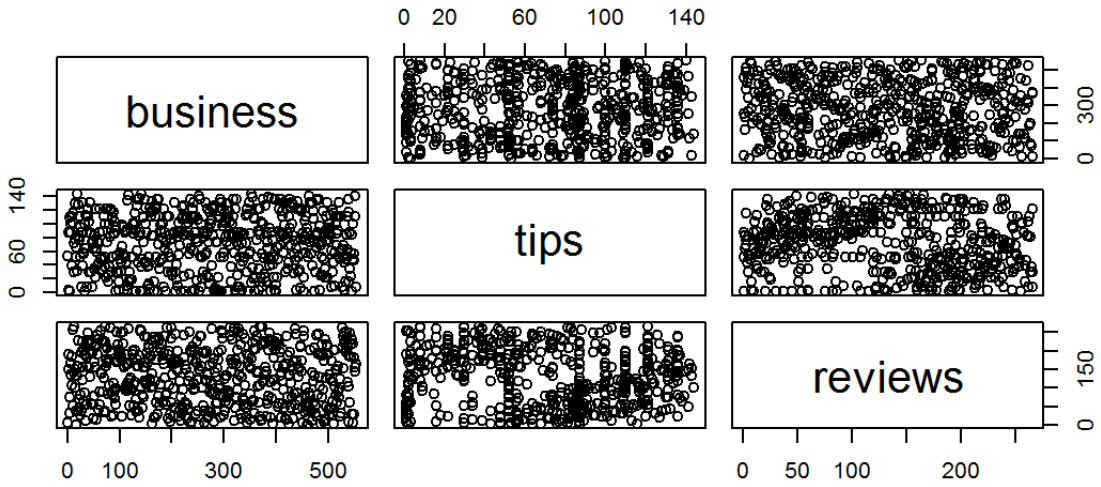
In reviewing the lodging data for ambience, even the highest number of observations is not enough to proceed to a full analysis. A total of 105 observervations in all attributes of the ambience data frame. Moving on to examination of ambience in the restaurant reviews.

Significantly more data is available for ambience under restaurant categories. For example the romantic *attribute* has 226 observations and the *upscale* has 146, all categories have at least 146 observations with the exception of *touristy* with only 83. The most data is in *causal* attribute with 7737 observations.

For this exercise, if we look at the dimensions of the *trendy* attribute it contains 553 observation. We'll make that attribute the focal point of the study and create a business map data file around the business, tips, and reviews for trendy food establishments.

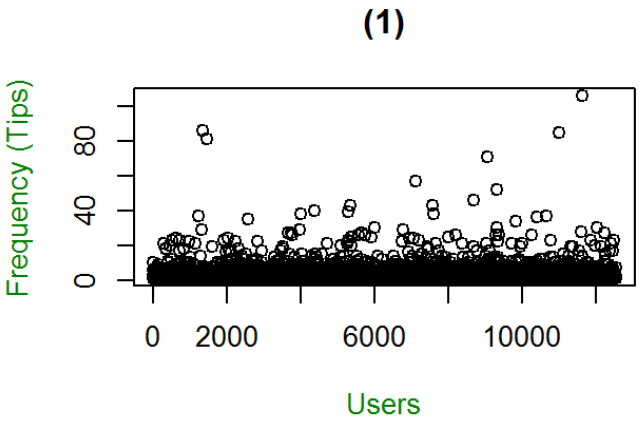
```
summary(bus_map)
```

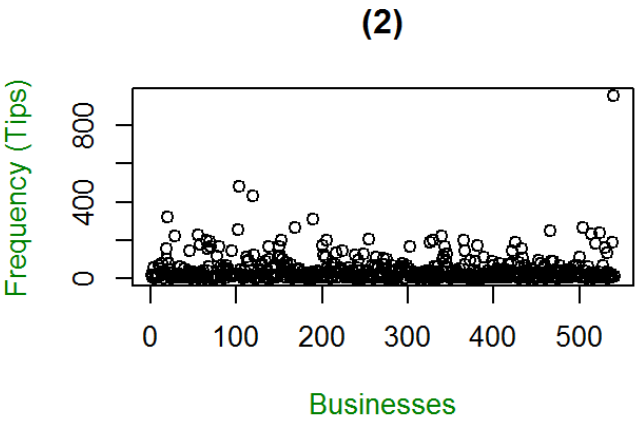
```
##                    business        tips        reviews
##  --pOlFxITWnhzc7SHSIP0A:  1   2      : 25   14     :  8
##  -6j-KVPPX2xKjCruNO2HnQ:  1   6      : 24   16     :  8
##  -741QDj3PPLD4Jii2yMU-w:  1   4      : 21   17     :  8
##  -9pVS__IliMA2aNEYzrQrg:  1   7      : 20   27     :  8
##  -appMO8r0lE0clUwl5b_HA:  1   5      : 18   20     :  7
##  -GR4Dvxhx6ddaDiUICOSFQ:  1   15     : 17   37     :  7
##  (Other)               :547   (Other):428   (Other):507
```

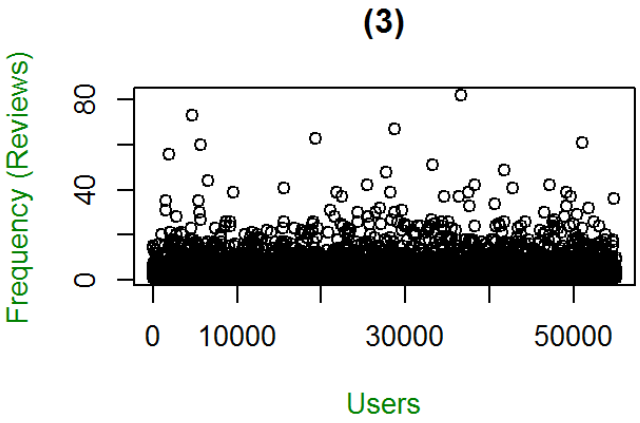


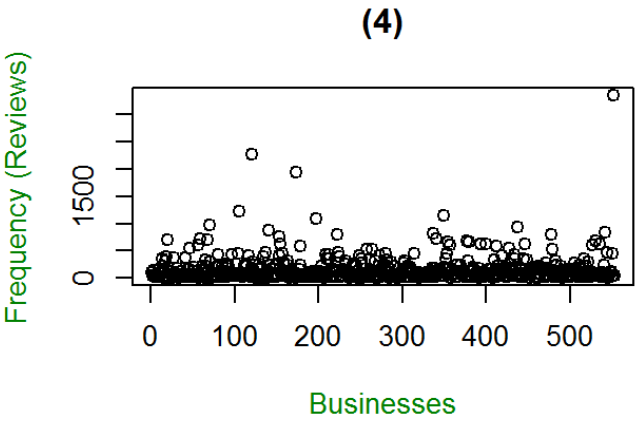
Here is the result plot of *bus_map*, it gives an overview of the relations between businesses, tips, and reviews. The data shows some clustering around tips & reviews indicating users providing both.

Graph (1) below is looking at the **12,554** observations of users providing tips on trendy restaurants. In graph (2) we are looking at **55,067** observations of users providing reviews on trendy restaurants. Graph (3) is looking at users providing reviews on trendy restaurants. Finally, in graph (4) we look at each trendy business and the frequency that the reviews come in. We again see a concentration of the distribution less dense and more scattered. In conclusion, the data supports the majority of users post tips in the single digits and reviews in the lower double digits. While trendy businesses recieve tips in the double digit range and reviews in the triple digits.

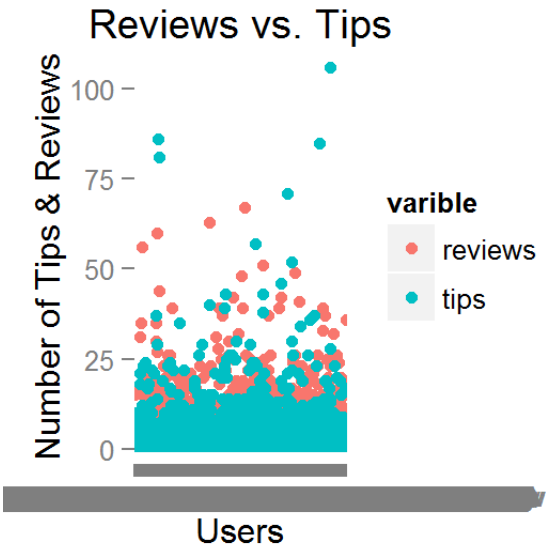


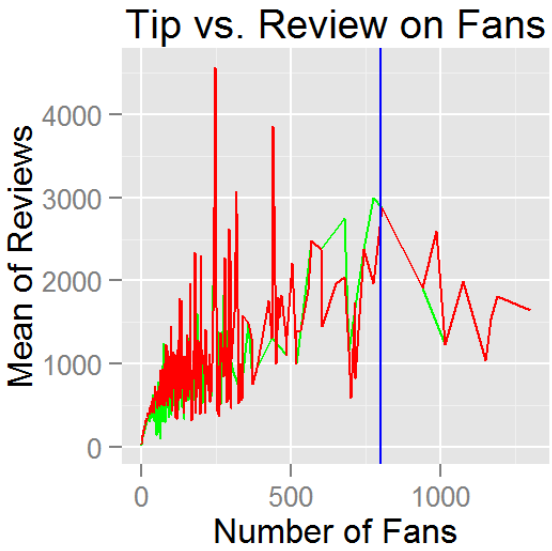






On this **Reviews vs. Tips** chart, the data demonstrates users providing more reviews than tips on the each of the trendy food establishments. The **Tip vs. Review on Fans** chart includes the **55,067** reviews and **12,554** tips, almost (4) times the number of reviews recorded to tips. The graph uses the mean of reviews for our trendy reviewers (green) and tip posters (red), the number of fans increases and drops off sharply around 800 (blue line).





To investigate the user fan's correlation further, we'll build a linear model. First here is a summary of the reviewers and details of the *review_profile* model file. Build the model and look at the top 5 count relations. Now perform the correlation test set with a 95% confidence level.

```
summary(review_profile)
```

```
##    user_id           review_count       fans
##  Length:55067       Min.   :   1   Min.   :   0
##  Class :character   1st Qu.:   7   1st Qu.:   0
##  Mode  :character   Median :  19   Median :   0
##                     Mean   :  74   Mean   :   4
##                     3rd Qu.:  65   3rd Qu.:   2
##                     Max.   :4573   Max.   :1298
```

```
lmodel <-aggregate(fans~review_count, data = review_profile, mean)
head(lmodel, 5)
```

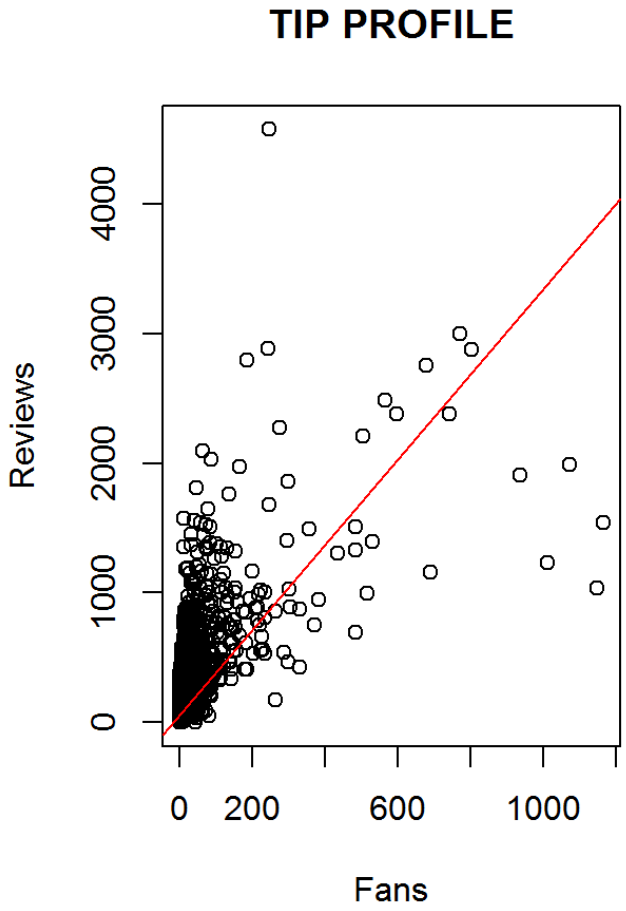
```
##   review_count   fans
## 1            1 0.0433
## 2            2 0.0930
## 3            3 0.1417
## 4            4 0.1374
## 5            5 0.1963
```

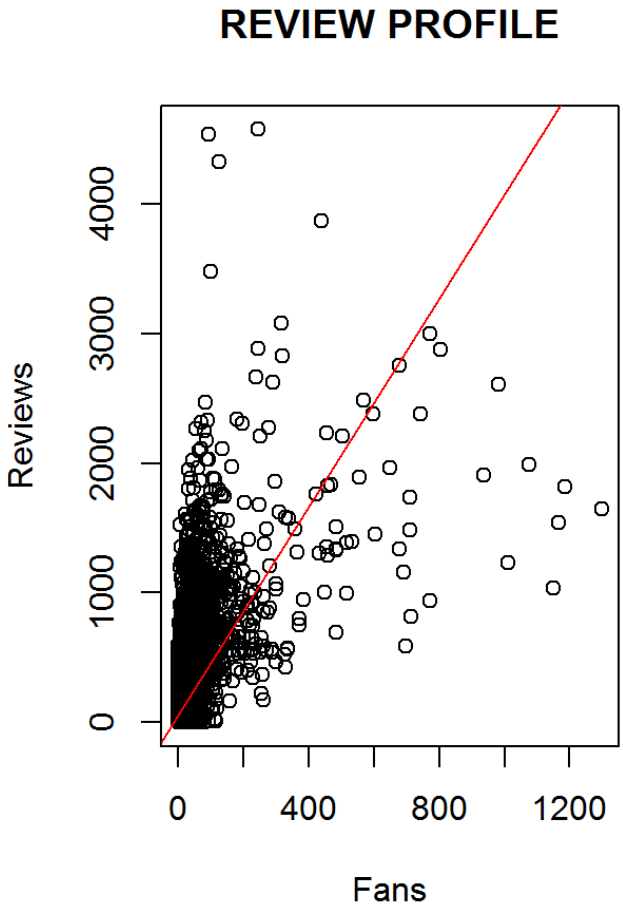
```
with(review_profile, cor.test(review_count, fans, alternative="greater", conf.level=.95))
```

```
##
##  Pearson's product-moment correlation
##
## data:  review_count and fans
## t = 176, df = 55065, p-value < 2.2e-16
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.595 1.000
## sample estimates:
## cor
## 0.6
```

A positive significance of .5996685.

Now for the tip posters using the *tip_profile* in the background. Build the model again for tip posters and post it in the plot. Plot the positive correlated result for tip posters and for reviewers now with *review count* vs. *fans*.





Discussion

To add more focus to the Users of these reviews and tips, we find the top 100 reviewers on trendy food places is 31.87 reviews and the top 100 people providing tips averages at 27.74. If we look at the overall review & tip count on users that provided trendy reviews and tips, we see a significant average in the overall review counts and tips provided. The average is almost identical with 73.647 average reviews and 75.463 average tip posts. This provides some insight into the group providing tips with their overall average reviews matching the group primarily doing reviews lending to the conclusion that tip posters are primarily high reviewers.

For the total sample set of users in the Yelp profile data, we find the average review count is only 32.215 Our trendy user population is more focused on providing feedback. But as a result does that increase their popularity?

Again if we look closely at the yelp profile data, there is a variety of positive enforcers for posting reviews. For example, lets look a the *fans* attribute for our Trendy posting users. The average number of *fans* from the total sample set of users in the Yelp profile data is 1.575 a much lower figure than for our trendy reviewers 4.272 or our tip posters 5.97. This begs the question, could their be a correlation between *fans* and our group of trendy reviewers and tip posters? Absolutely as the data shows.

The significance to Yelp and many of their business partners is that a focused data analysis in the area of marketing to users with a high number of fans could be used to target a variety of business with the ambience attributes. A further study to include the yelp data and data on business traffic may help to identify the income potential in targeting ads, reviews, other social media to increase a businesses success.