The Impact of Celebrity Reviewers on Businesses

Bobby Tan

19 November 2015

Introduction

In this report, we study the impact of what we call celebrity reviewers on businesses, the results of which business owners will undoubtedly be interested in. The analysis is based on the Yelp Dataset which may be downloaded in the zip format here.

While in general, the definition of a celebrity reviewer may vary from individual to individual, we will use a very specific definition of the term. We adopt the following definition of a celebrity reviewer:

> (Definition) A celebrity reviewer is a reviewer who is amongst the top 10 percent (i.e. is in the top decile) in terms of the total number of votes they have received from their reviews.

We believe that this is a good reflection of the popularity of a reviewer, moreso than other metrics you can obtain from the Yelp dataset such as the number of comments they have received or the number of reviews they have written. The reason being it is a good measure of how many people their reviews have impacted the Yelp usership.

Based on this definition of a celebrity reviewer, we are seeking to address 2 questions: How much effect does a celebrity reviewer and their reviews have on a business overall? Are they more influential than the average Yelp reviewer?

Methods and Data

The following are the libaries required for the analysis:

```
library(jsonlite)
library(dplyr)
library(ggplot2)
```

We first read extract the data from the relevant files:

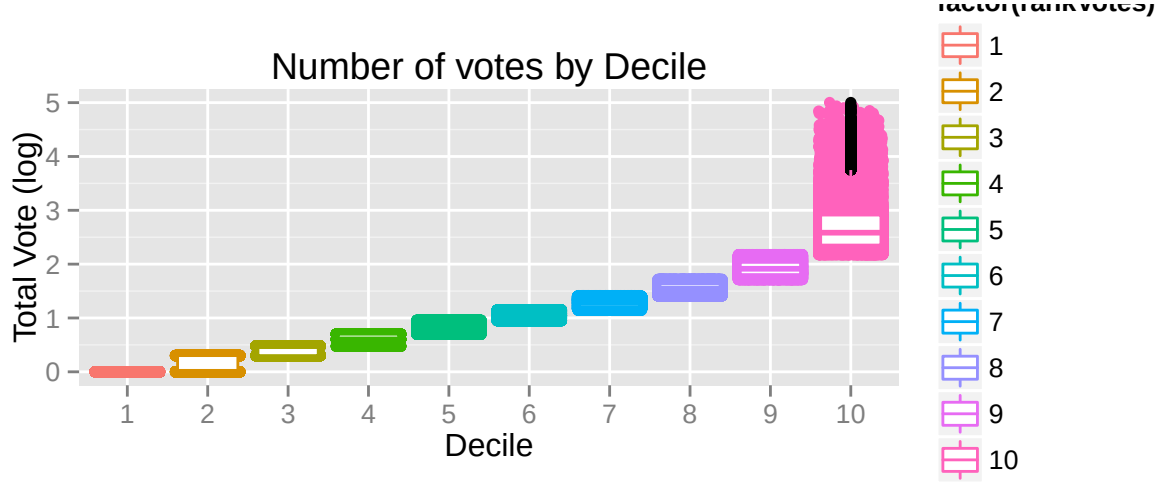
```
# Data files' locations
fileBusiness <- "./yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_business.json"
fileCheckIn <- "./yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_checkin.json"
fileReview <- "./yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_review.json"
fileTip <- "./yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_Tip.json"
fileUser<- "./yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_user.json"

# Function to prepare and read the JSON data files
ReadData <- function(fileLocation, numberOfLines = -1L) {
  con <- file(fileLocation, open = "r")
  readJSON <- readLines(con, n = numberOfLines)
  readJSON <- paste("[", paste(readJSON, collapse = ","), "]")  # To make JSON array
  close(con)
  fromJSON(readJSON)
  }
```

```
# Reads in data
dataBusiness <- ReadData(fileBusiness)
dataCheckIn <- ReadData(fileCheckIn)
dataUser <- ReadData(fileUser)
```

The votes for each users are tabulated and sorted by deciles:

```
totalVotes <- rowSums(dataUser$votes) ## Computes the total number of votes
rankVotes <- ntile(totalVotes, n = 10) ## Ranks data according to deciles
topDecile <- rankVotes == 10
```

```
qplot(x=factor(rankVotes), y= log10(totalVotes+1), geom = "jitter", color = factor(rankVotes))+
  geom_boxplot()+
  xlab("Decile")+
  ylab("Total Vote (log)")+
  ggtitle("Number of votes by Decile")##Plots the data according to decile
```



Note the log scale in the vertical axis. From the above, it is clear that the decile ranking of the users and their total votes have an exponential relationship. In this case the top decile commands hundres of times more votescumulatively than the lowest decile. The top decile also contains the most outliers as you can see from the vertical spead in the 10th decile.

The scores, business IDS, User IDs and the review dates are also read in from the Yelp reviews dataset.

```
# Find of business IDs, user_ids, dates and star rating of all reviews.
reviewsUserID <- rep("",1569264 )
reviewsBusinessID <- rep("",1569264 )
reviewsDate <- rep("",1569264 )
reviewsStars <- rep(0,1569264 )

con <- file(fileReview, open = "r")
line <- 0

while(line < 1569264) {

  temp <- data.frame(fromJSON(readLines(con, n = 1)))
  line <- line+1
```

```
  reviewsUserID[line] <- as.character(temp$user_id)
  reviewsBusinessID[line] <- as.character(temp$business_id[[1]])
  reviewsDate[line] <- as.character(temp$date[[1]])
  reviewsStars[line] <- temp$stars[[1]]
}
close(con)
```

In addition, we also tabulate the total number of checkins for every business with check in data provided.

```
# Tabulates total checkins.
totalCheckin <- {}
for(i in 1:length(dataCheckIn$business_id)){
  totalCheckin <- c(totalCheckin,
                    sum(na.omit(unlist(dataCheckIn$checkin_info[i,]))))
}
```

We then pick a random sample of reviews, where the sample size is the same as the number of reviews made by celebrity reviewers. This will represent the average "non-celebrity" review, which will be the basis for our comparison.

Using the above information, we will perform the following analysis:

- A comparison between a random sample against the group of celebrity reviewers in terms of checkins.
- A student's t test to determine if the celebrity group increased the number of reviews in the 3 months following the review when the review is good
- A student's t test to determine if the celebrity group decreases the number of reviews in the 3 months following the review when the review is bad

Results

The first observation is that the top 10% of yelp reviewers are responsible for 586695 out of the total 1569264 which is 37.3866348 percent of all reviews given in the Yelp dataset. So they are clearly prolific, and already more influential on this basis.

We also compare the review scores of both the random sample and the celebrity group:

```
summary(topDecRevStars) ## review scores for celebrity group
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.731   5.000   5.000
```

```
summary(randSampRevStars) ## review scores for random sample
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    4.00    3.74    5.00    5.00
```

The tastes of the average Yelp reviewer and the celebrity group is virtually identical.

We also compare the number of check ins per business reviewed:

```
summary(topDecileCheckIn)  ## checkins per business in celeb group
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    14.0    38.0   150.8   118.0 62650.0
```

```
summary(randSampleCheckIn) ## checkins per business in random sample
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    12.0    35.0   145.7   112.0 62650.0
```

Again, the difference in means is not large. We perform a t test to decide if this difference is statistically different:

```
t.test(topDecileCheckIn, randSampleCheckIn, paired = FALSE)
```

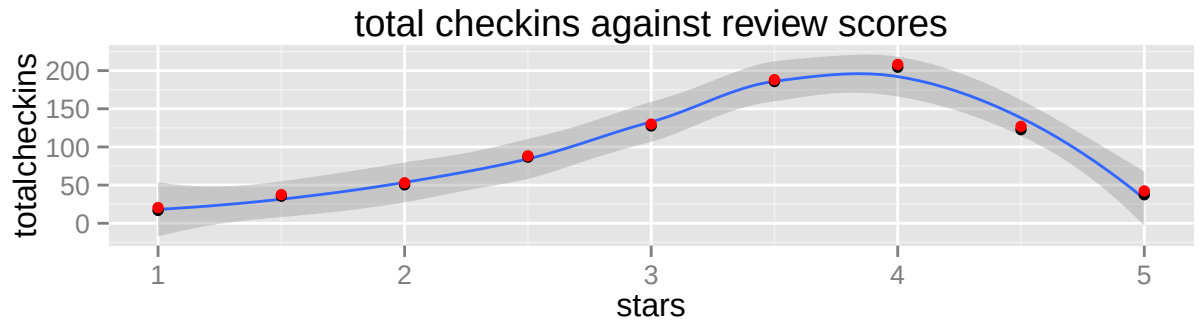
```
##
##  Welch Two Sample t-test
##
## data:  topDecileCheckIn and randSampleCheckIn
## t = 1.1842, df = 83077, p-value = 0.2363
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.341153 13.541892
## sample estimates:
## mean of x mean of y
##  150.8190  145.7187
```

Since the p-value of 0.2363>0.05 is large, the Yelp data does not support that there is a statistical difference at the 95% confidence level.

We also sort the total number of check ins by their averaged review scores and compare both groups.

```
library(ggplot2)
#plots rand sample in black, celebrity group in read
qplot(x = stars, y = totalcheckins, data = randSampSummary, geom =  "smooth") +
  geom_point()+
  geom_point(x = topDecSummary$stars, y=topDecSummary$totalcheckins, color = "red")+
  ggtitle("total checkins against review scores")
```

```
## geom_smooth: method="auto" and size of largest group is <1000, so using loess. Use 'method = x' to cl
```



Again, there does not appear to be a significant difference between both groups are well within each other's error bars and the averaged data practically overlaps each other.

Finally, as a final measure of impact, we also tabulate the number of reviews before and after a review from the celebrity group and sorted the data into good reviews (4 or 5 stars) and bad reviews(1 or 2 stars). A 2 sided t test is then performed to see if the mean difference in zero. The number of reviews is used as a proxy for how well a business is doing because the checkin is not detailed enough for analysis. For good reviews, the results are:

```
##
##  Paired t-test
##
## data:  goodDF$before and goodDF$after
## t = -7.1596, df = 28839, p-value = 8.286e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2341273 -0.1334871
## sample estimates:
## mean of the differences
##              -0.1838072
```

And for bad reviews:

```
##
##  Paired t-test
##
## data:  badDF$before and badDF$after
## t = -3.3846, df = 6884, p-value = 0.0007168
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.23188914 -0.06179278
## sample estimates:
## mean of the differences
##               -0.146841
```

Which suggests that there is a difference between the celebrity group and the random sample, but it amounts only to less than 0.19 increase in the number of reviews in the 3 months after a review by a celebrity Yelp reviewer is posted.

Discussion and

In summary, even though celebrity Yelp reviewers are highly prolific,with roughly 40% of the reviews in the dataset commanded by celebrity reviewers, on an individual review basis, the difference between a celebrity review and the average Yelp review is not significant. For instance, businesses reviewed celebrity reviewers do not experience a statistically different numbe of check ins as compared to the "average" (i.e. randomly sampled) yelp user. There is a statistically significant increase of 0.15-0.18 reviews in the 3 months following a review posted by a celebrity reviewer, but it is still only a minor increase which business owners are not likely to be concerned about. The results suggest that Yelp users' usage patterns do not overweigh the opinions of celebrity reviewers, but are likely to look at the overall opinion of the business across all reviews posted.