Relationship Between Number of Reviews and Amount of Travel From Yelp Dataset

Jacob Townson

November 15, 2015

Introduction

I thought it would be interesting to try to find out if there was a correlation between the number of reviews a person has made on Yelp with the amount they have traveled. So my question was: do more frequent reviewers travel more than nonfrequent reviewers? I found the question to be interesting because it raised my curiousity of whether or not these people travel purely for the sake of making reviews on certain organizations. So my hypothesis was that yes, frequent reviewers on Yelp would probably travel more, and their reviews would be more spread out based on city, longitude, and latitude.

Methods and Data

To begin, the data had to be read into R. To download the data, simply go to this link: link. This data needs to be put in the working directory, then read in using the code chunk in **Figure 1 Part A**. Next, a massive amount of data cleaning had to be done. I knew I needed to separate the data by user, the number of cities traveled, and also distance based on latitude and longitude. In order to do this, I ran the (rather large) code chunk seen in **Figure 1 Part B** . The first chunk reads in the data required from the Yelp dataset

Results

Discussion

Appendix

Figure 1

Part A

Reading in the Data

```
if (exists('business_data') == FALSE){
  business_data <- stream_in(file("./yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_busi
}

if (exists('review_data') == FALSE){
  review_data <- stream_in(file("./yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_review
}



if (exists('user_data') == FALSE){
  user_data <- stream_in(file("./yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_user.json
}
```

Part B

Cleaning the Data

```
# Let's Clean it up

users <- user_data[user_data$review_count != 0,] #remove users with 0 reviews
votes <- users$votes$funny + users$votes$useful + users$votes$cool
users <- select(users, review_count, user_id, fans, average_stars)
users <- mutate(users, votes = votes)

businesses <- data.frame(business_data$business_id, business_data$city, business_data$latitude, business
colnames(businesses) <- c("business_id", "city", "latitude", "longitude")

reviews <- data.frame(review_data$user_id, review_data$review_id, review_data$business_id)
colnames(reviews) <- c("user_id", "review_id", "business_id")

full_data <- left_join(reviews, businesses, by = "business_id")
full_data <- left_join(full_data, users, by = "user_id")
full_data <- full_data[complete.cases(full_data),] #get rid of NA's

rm(business_data, businesses, review_data, reviews, user_data) #cleaning up environment
travel <- full_data[,-c(2,3,7,8,9,10), drop = FALSE]
```