Yelp Capstone

Peter Blane

November 22, 2015

Intro

This project is from the [Yelp Dataset Challenge] (http://www.yelp.com/dataset_challenge). The Yelp Dataset Challenge is a contest that gives a montetary reward to students based on which Yelp judges "submissions on their technical depth and rigor, the relevance of the results to Yelp, our users, or the field, and finally their novelty, uniqueness, and yes, their Yelpy-ness."

What I have chosen to do is ask the following question about the data: of the businesses that are restaurants, what makes a restaurant with a quiet atmosphere good?

Methods

In order to figure out how to answer the question about quiet restuarants, some exploratory analysis needs to be done to see what types of data is available.

Reading in the JSON files

After downloading the files, the first thing discovered about the data is that the files are in JSON format. This is different than the typical data we see in examples (e.g. CSV), so the read method will be:

- Business

```
library("jsonlite", lib.loc="/Library/Frameworks/R.framework/Versions/3.1/Resources/library")
```

```
## Warning: package 'jsonlite' was built under R version 3.1.3
```

```
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:utils':
##
##     View
```

```
yelp_business <- "yelp_academic_dataset_business.json"
business <- fromJSON(sprintf("[%s]", paste(readLines(yelp_business), collapse=",")))
```

- Checkin

```
library("jsonlite", lib.loc="/Library/Frameworks/R.framework/Versions/3.1/Resources/library")
yelp_checkin <- "yelp_academic_dataset_checkin.json"
checkin <- fromJSON(sprintf("[%s]", paste(readLines(yelp_checkin), collapse=",")))
```

- Review

```
library("jsonlite", lib.loc="/Library/Frameworks/R.framework/Versions/3.1/Resources/library")
yelp_review <- "yelp_academic_dataset_review.json"
review <- fromJSON(sprintf("[%s]", paste(readLines(yelp_review), collapse=",")))
```

- Tip

```
library("jsonlite", lib.loc="/Library/Frameworks/R.framework/Versions/3.1/Resources/library")
yelp_tip <- "yelp_academic_dataset_tip.json"
tip <- fromJSON(sprintf("[%s]", paste(readLines(yelp_tip), collapse=",")))
```

- User

```
library("jsonlite", lib.loc="/Library/Frameworks/R.framework/Versions/3.1/Resources/library")
yelp_user <- "yelp_academic_dataset_user.json"
user <- fromJSON(sprintf("[%s]", paste(readLines(yelp_user), collapse=",")))
```

Exploring the Data

After looking through the files, it became apparent that the "Business" dataset had the information needed to help answer the question.

```
unique(business$attributes$`Noise Level`)
```

[1] NA "average" "loud" "quiet" "very_loud"

So, besides the NA's (which we'll take care of later), there are 4 categories of the businesses reviewed: average, loud, quiet, and very_loud. The question, however, was regarding quiet restuarants, so the data needs to be narrowed to just businesses that are restuarants. This can be accomplished by pulling out the descriptions that categorize each business under the "categories" column:

```
b <- subset(business, grepl("Restaurants", business$categories), select=c(business_id:type))
```

What would be interesting to know is which "Noice Level" got the most reviews and stars.

```
bquiet <- subset(b, grepl("quiet", b$attributes$`Noise Level`), select=c(business_id:type))
bloud <- subset(b, grepl("loud", b$attributes$`Noise Level`), select=c(business_id:type))
baverage <- subset(b, grepl("average", b$attributes$`Noise Level`), select=c(business_id:type))
```

- Stars

```
table(baverage$stars)
```

```
##
##    1  1.5    2  2.5    3  3.5    4  4.5    5
##   16   90  265  809 1844 3285 3342 1100  106
```

```
table(bquiet$stars)
```

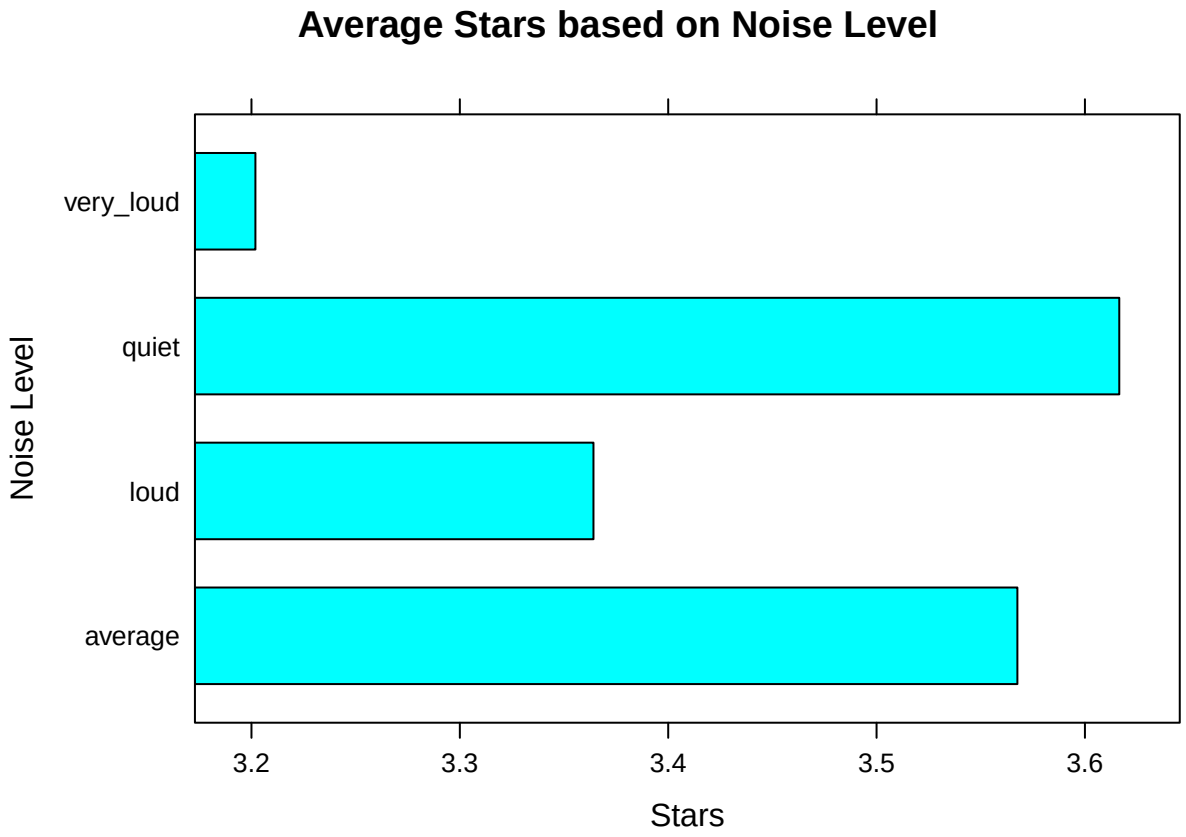
```
##
##    1  1.5    2  2.5    3  3.5    4  4.5    5
##   10   38  114  298  607  974 1122  547  108
```

```
table(bloud$stars)
```

```
##
##   1 1.5   2 2.5   3 3.5   4 4.5   5
##   7  38 115 271 487 563 355  90  12
```

Stars as a chart relative to Noise Level

```
# creating average stars relative to noise level and
bd <- data.frame(business$attributes$`Noise Level`, business$review_count, business$stars)
bd <- bd[complete.cases(bd),]
names(bd)[1] <- "Noise_Level"
names(bd)[2] <- "ct_Review"
names(bd)[3] <- "Stars"
library(lattice)
MeanBDS <- aggregate(Stars ~ Noise_Level, data = bd, mean)
barchart(Noise_Level ~ Stars, data = MeanBDS, main = "Average Stars based on Noise Level", xlab = "Stars
```



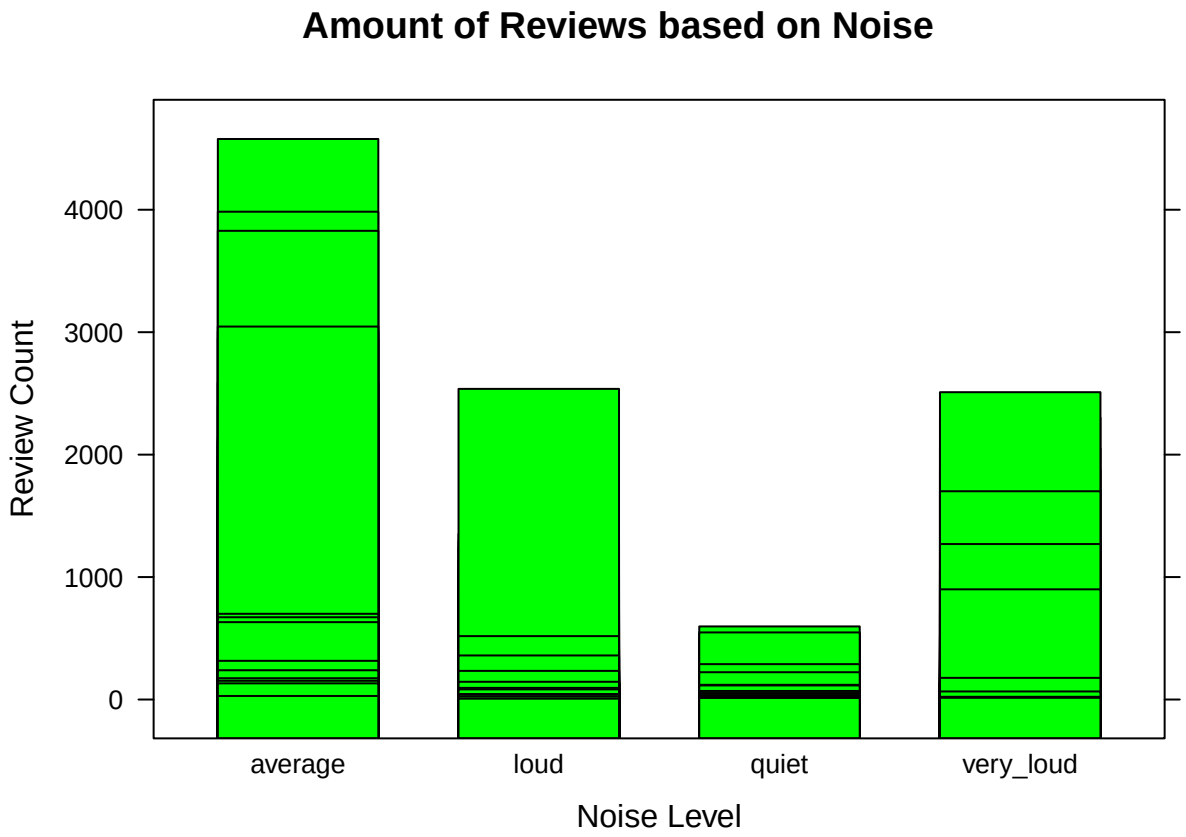
From the above chart, we can see the restuarants with a quiet attribute associated with the review, on average, gets the highest star rating.

Now let's look at the review counts:

Reviews

```
# Creating data frame from Business data to extract Noise Level without NA's
bd <- data.frame(business$attributes$`Noise Level`, business$review_count, business$stars)
bd <- bd[complete.cases(bd),]
names(bd)[1] <- "Noise_Level"
names(bd)[2] <- "ct_Review"
names(bd)[3] <- "Stars"

# Looking at the number of reviews relative to noise level
library(lattice)
barchart(ct_Review ~ Noise_Level, data = bd, main = "Amount of Reviews based on Noise", xlab = "Noise L
```



What is interesting here, based on the above chart, the quiet restuarants receive the fewest reviews compared the the other noise levels.

Statistical Model

In the past, when looking for the "best" of something online, the number of stars plus the number of reviews is typically what I use to evaluate a good product or service.

Hypothesis

Since I normally use the amount of stars and total reviews to evaluate decisions, I believe the best way to see what makes the best quiet restuarant will be by looking at the relationship between Stars and Reviews.

```
Call:
lm(formula = Stars ~ ct_Review, data = rmQuiet)

Residuals:
     Min       1Q   Median       3Q      Max
-2.54664 -0.54389 -0.04389  0.44511  1.46710

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.5246510  0.0141299  249.45   <2e-16 ***
ct_Review   0.0027486  0.0002881    9.54   <2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7065 on 3816 degrees of freedom
Multiple R-squared:  0.02329,   Adjusted R-squared:  0.02304
F-statistic: 91.01 on 1 and 3816 DF,  p-value: < 2.2e-16
```

Results

From the above results, we can see that the Star rating and the amount of Reviews is highly correlated. Thus the overall factor in determining what makes a quiet restuarant good is the amount of ratings and stars.

Discussion

While the overall answer seems a bit obvious, it did not occur to me until I was developing the model. I was looking a large number of other attributes of quiet restuarants, but saw that the main consistancy was the aforementioned correlation. Here are some other interesting findings about these quiet restuarants:

*They tend to be casual (FALSE = 592, TRUE = 614)

*Good for groups (FALSE = 271, TRUE = 995)

*Cheap (1 = 712, 2 = 439, 3 = 73, 4 = 4)

*Kid Friendly (FALSE = 138, TRUE = 1113)

*Little or no alcohol (beer_and_wine = 49, full_bar = 131, none = 1052)

What is a bit silly, but makes sense, is that the Loud/Very Loud restuarants were pretty much the same (even kid friendly) except for Alcohol and Music (dj's, video, live music, etc.)