1 Business performance through User Review

The biggest challenge to the yelp datasets are they are big (especially “review” data). Therefore for initial investigation, the first 50 thousand lines from review file are extracted. Upon investigation, “text” column within reviews file is the main contributor to the size and therefore ”text” column was temporary dropped during ETL. Once the questions/problems and approach are determined, only relevant datasets (review & business) and fields are used and extracted. Feature engineering is performed to the new datasets in order to have more useful information.

2 Introduction

Following are the question we like to answer from the datasets:

1) Are we able to help business to assess their performance throughout the year based on user review? (Assuming user review is reliable source on for performance)
2) Is user feedback being address by business?
3) As star rating is subjective, are we able to detect sudden surge of negative feedback which business should take note before it become undesirable “norm” in the business?

3 Methods and Data

A business entity that is widely review by the user should be a potentially candidate for the study. Therefore selection for a business entity will select within the most review categories (in this case is the restaurant category) based on review per month. Assume that users are serious about their unhappiness if they rate the business as “1”. Therefore stars rating of “1” will be studied.

Following are the step performed.

- a. Both datasets are merged together with the following fields: Date, business_id, categories, stars. business_id is used as index key.
- b. New Fields are created for analysis:
  - i. Month (month which the review fall on),
  - ii. Ave (average number of review per month),
  - iii. Rate (classified according to stars with High=5, mid=4-2 and low=1).
- c. Following are be plotted and studied:
  - i. Tread of user review throughout the year.

```
temp.df=data.frame(group_by(reviewSummary.df,Week) %>% ummarise(Total=n()))
ggplot(temp.df,aes(x=Week,y=Total))+geom_point() + geom_smooth(method="loess",color="red")
```

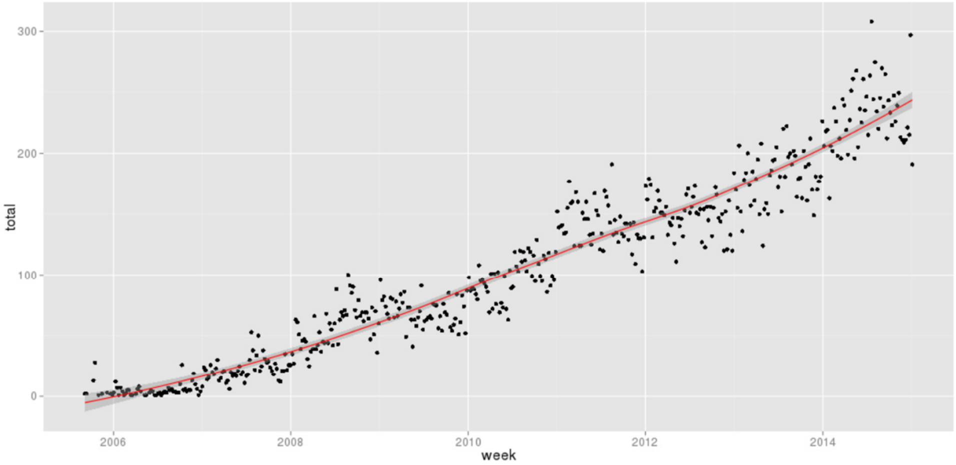


Figure 1: Review per week

ii. Tread of user review group by business categories throughout the year

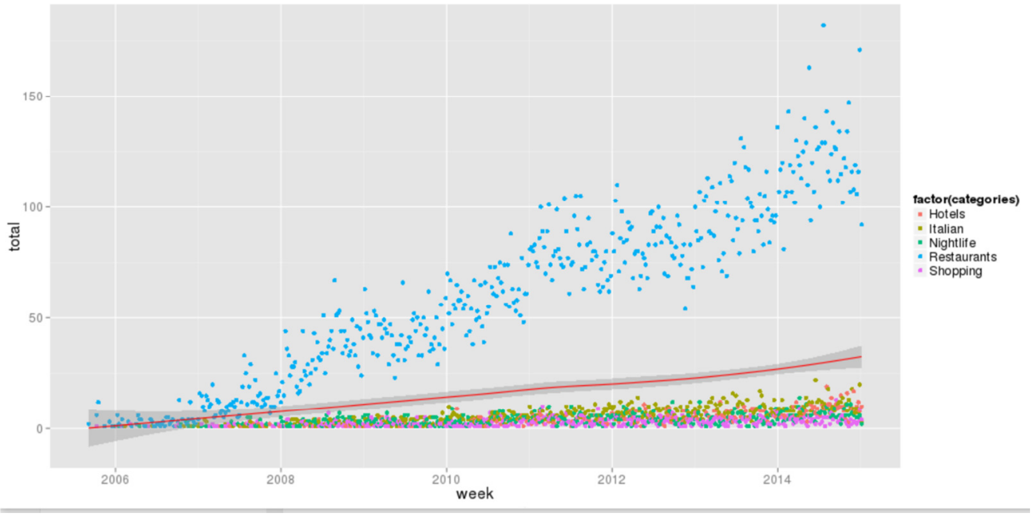
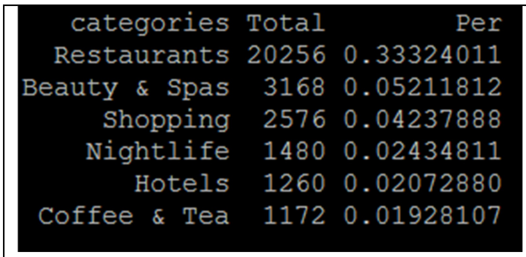


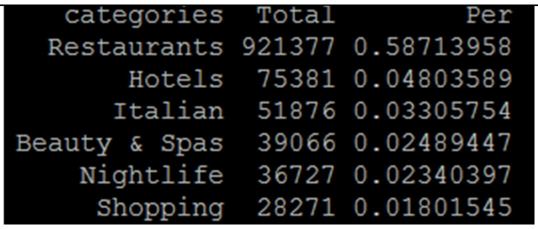
Figure 2: User Review group by business category

d. Top 5 business category composition and top 5 most reviewed business category are tabulated. From Figure 4, restaurant category has the major user review. Therefore a business entity is selected from within restaurant category.



```
    categories Total        Per
   Restaurants 20256 0.33324011
Beauty & Spas   3168 0.05211812
      Shopping  2576 0.04237888
     Nightlife  1480 0.02434811
        Hotels  1260 0.02072880
  Coffee & Tea  1172 0.01928107
```

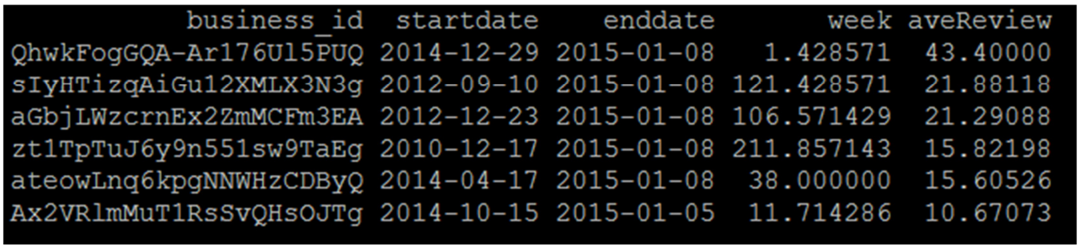
Figure 3: Top 5 business category composition



```
    categories  Total        Per
   Restaurants 921377 0.58713958
        Hotels  75381 0.04803589
       Italian  51876 0.03305754
Beauty & Spas   39066 0.02489447
     Nightlife  36727 0.02340397
      Shopping  28271 0.01801545
```

Figure 4: Top 5 Most reviewed business category

e. From Figure 5 the first business entity had 43 review/week (aveReview). Further investigation shows that this business had short review period (around 1.5 week which most likely a new startup business) and therefore distorted the figure. Therefore this entity is rejected. For the next top three entity, “zt1TpTuJ6y9n551sw9TaEg" (Restaurant A) is selected since it has the longest business review period (211weeks) with around 15 review/week.



```
             business_id  startdate    enddate       week aveReview
QhwkFogGQA-Ar176Ul5PUQ 2014-12-29 2015-01-08   1.428571  43.40000
sIyHTizqAiGu12XMLX3N3g 2012-09-10 2015-01-08 121.428571  21.88118
aGbjLWzcrnEx2ZmMCFm3EA 2012-12-23 2015-01-08 106.571429  21.29088
zt1TpTuJ6y9n551sw9TaEg 2010-12-17 2015-01-08 211.857143  15.82198
ateowLnq6kpgNNWHzCDByQ 2014-04-17 2015-01-08  38.000000  15.60526
Ax2VRlmMuT1RsSvQHsOJTg 2014-10-15 2015-01-05  11.714286  10.67073
```

Figure 5: Top 5 restaurant business reviewed

2) With restaurant A selected as target business to study and based on assumption that stars rating should be a good gauge on how business is performing, feature engineering was performed to classify the rating as high(stars = 5), mid(stars = 4-2) and low(stars = 1) for each week. Rate trending was plotted for Restaurant A.

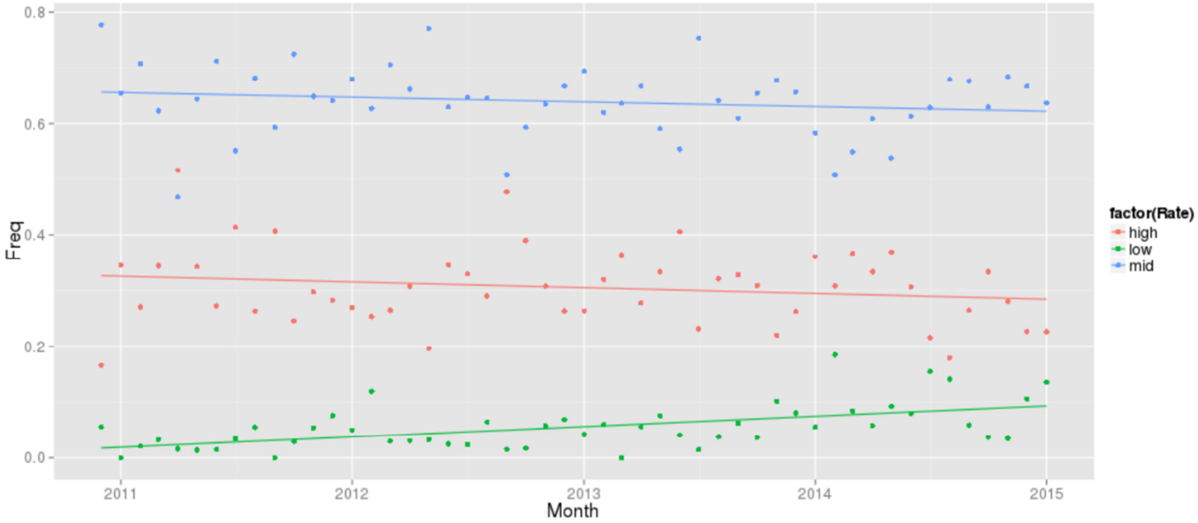


Figure 6: Restaurant A’s rate trending

3) Distribution of the rating was investigated and “low” rating appears to have gamma distribution (green). Whereas high (red) and mid (blue) rating, they appear to have normal distribution which skewed right or left respectively. Since we are interested in “low” rating, further fit plot (Figure 8) was performed. The fit plot indicates that gamma distribution (shape=1.88685, rate=34.07838) fit well on low rating distribution.

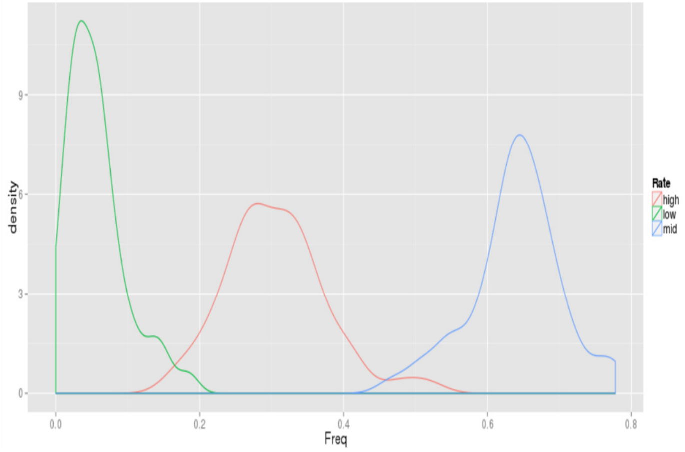


Figure 7: Frequency Distribution based on rate

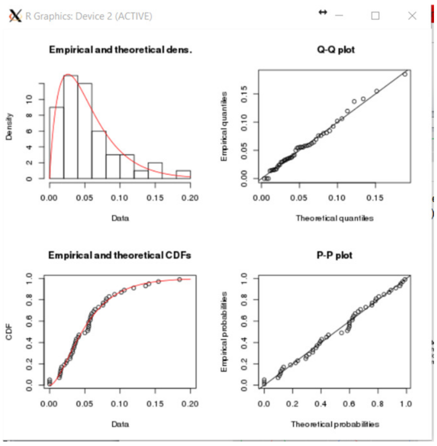


Figure 8: Fit plot

4) Therefore gamma distribution was used as model to provide confidence level to identify surge in low rating. To be 90% confidence that there is a serious surge in negative feedback, we need to have around 15% of the total review for that month with rating of 1.

5) In the month of Feb & July 2014, it hit 15%. Upon investigation into review for these months, Aug 2014 is near to 15% (14%) and therefore text analysis (Figure 9) is performed on Feb, July and Aug 2014. Food and Buffet seems to be the main focus of complaint.

6) A baseline (Figure 10) was established by performing text analysis for the whole review period. Buffet & Food are still the main focus.

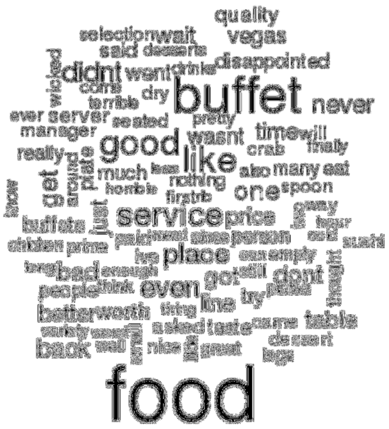


Figure 9: Word cloud for surge in stars rating "1"

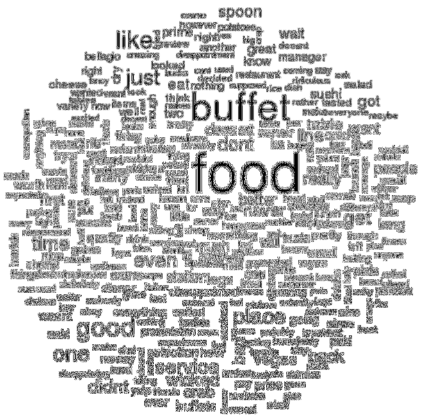


Figure 10: Word cloud for whole period

4 Results

From Figure 1, Yelp user review base had started to increase linearly from 2007 onward. From Figure 2 restaurant category's review grows linearly. In line with yelp user grow. Even though total business classified under restaurant stand at 33%(Figure 3) but for some reason, user are more willing to provide feedback to restaurant and occupy near to 60% of total review (Figure 4).

Are we able to help business to assess their performance throughout the year based on user review? (Assuming user review is reliable source on for performance)

From Figure 6, restaurant A does not seem to be doing well. In fact, the rating is declining throughout the year with more user (base on %) rate restaurant A as “1”. Both mid and high ratings too are declining which is not a good sign.

Is user feedback being address by business?

From Figure 9 & Figure 10, Food and Buffet have being the main complaint and it amplified in certain months. Reading into the review, more and more user feel that the price does not justify for the quality of food provided and in fact it got worsen. The restaurant A obviously is not addressing it. From the review, the fact that there are other restaurant nearby which was highlighted to be able to provide better value to their money, the business owner should seriously look into it.

As star rating is subjective, are we able to detect sudden surge of negative feedback which business should take note before it become undesirable “norm” in the business?

The surge in the complaint can be identified with gamma distribution model. There might be special event (eg. surge of crowd or changes to the food manual) that amplified the negative aspect of the food and this has started to become a “norm” to the restaurant A.

5 Discussion

This approach can potentially be used for further exploration into other business entity as a form of measurement on how their business is performing from customer prospective. It also help business owner to focus on area of improvement and also to study the impact of any promotion or activities that was conduct during that period and how customer response to them. But regardless, if the fundamental problem is not address (in the case of restaurant A), the problem will be amplified and this will have more negative impact to the business.