Yelp Dataset: Capstone Report for Data Science Specialization

Betty Yeo

Sunday, 22 November 2015

1. Introduction

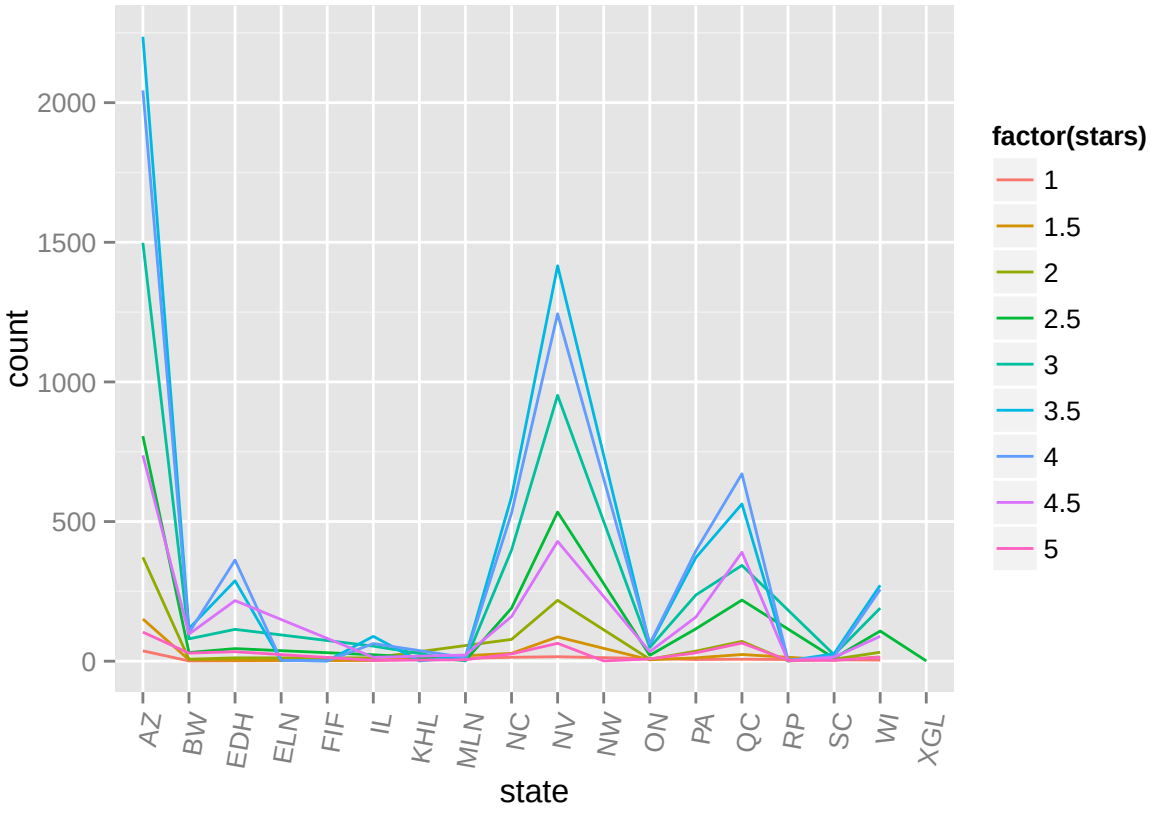
Considering where to setup or expand business is the question to answer for the capstone project. This could interest entrepreneur in their business set up or expansion. Typically, this could be done through data collection either through surveys or personally walk the ground to collect data but this would be too awesome a challenge given the constraint of time and resources. Fortunately, in these days, we found social media strewn wtih reviews and comments which could be used for our discretions. Yelp extensive data would be useful to dig into to give business user a sensing on how viable the venue and the features its offering features/facilities in their business venture.

Methods

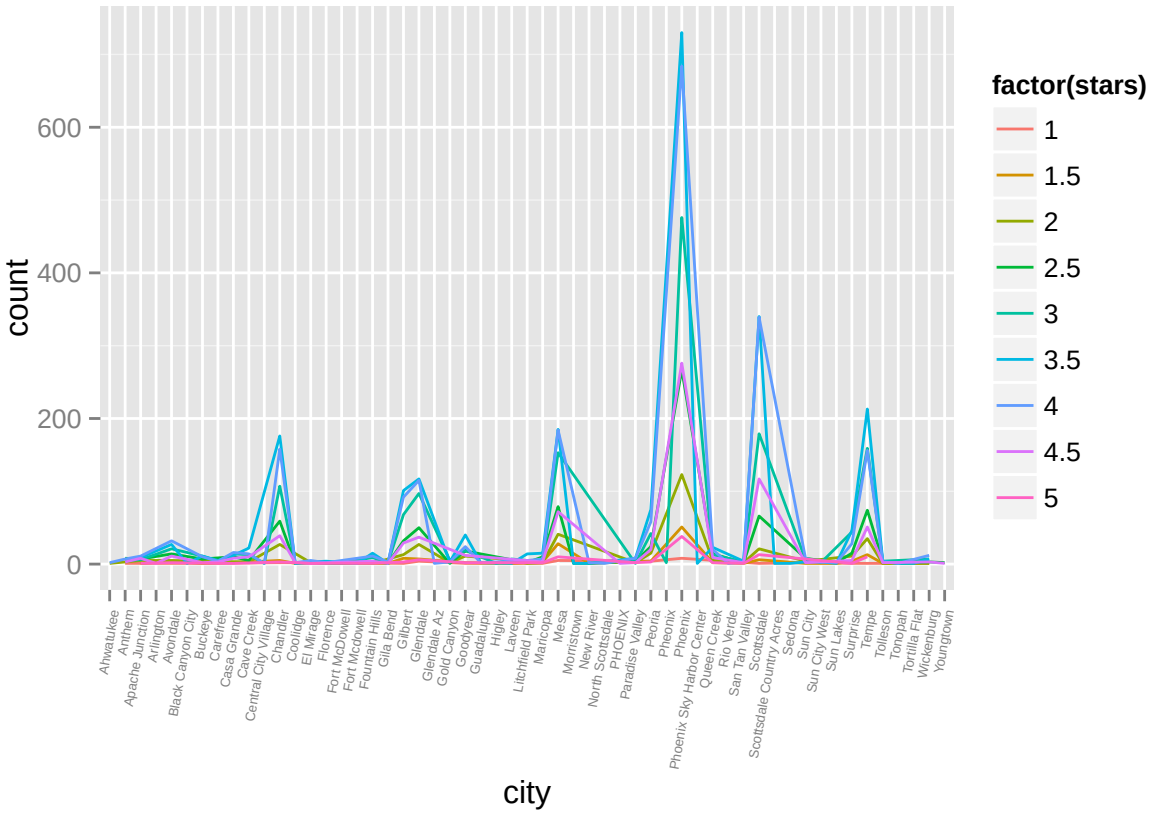
The data is downloaded from Yelp Data Challenge at http://www.yelp.com/dataset_challenge. It is then save into my working directory as RDS format to save the time of downloading the huge data set.

For this exercise, i'm looking at business data which contain business unit, types of business, review rating and the features that the business adopts. There is a total of 61k businesses from America and Europe. For this exercise, i'll be looking at restaurant business as earlier exploration with data set shows that restaurant or food business is one of the highest business featured in Yelps.

The Approach

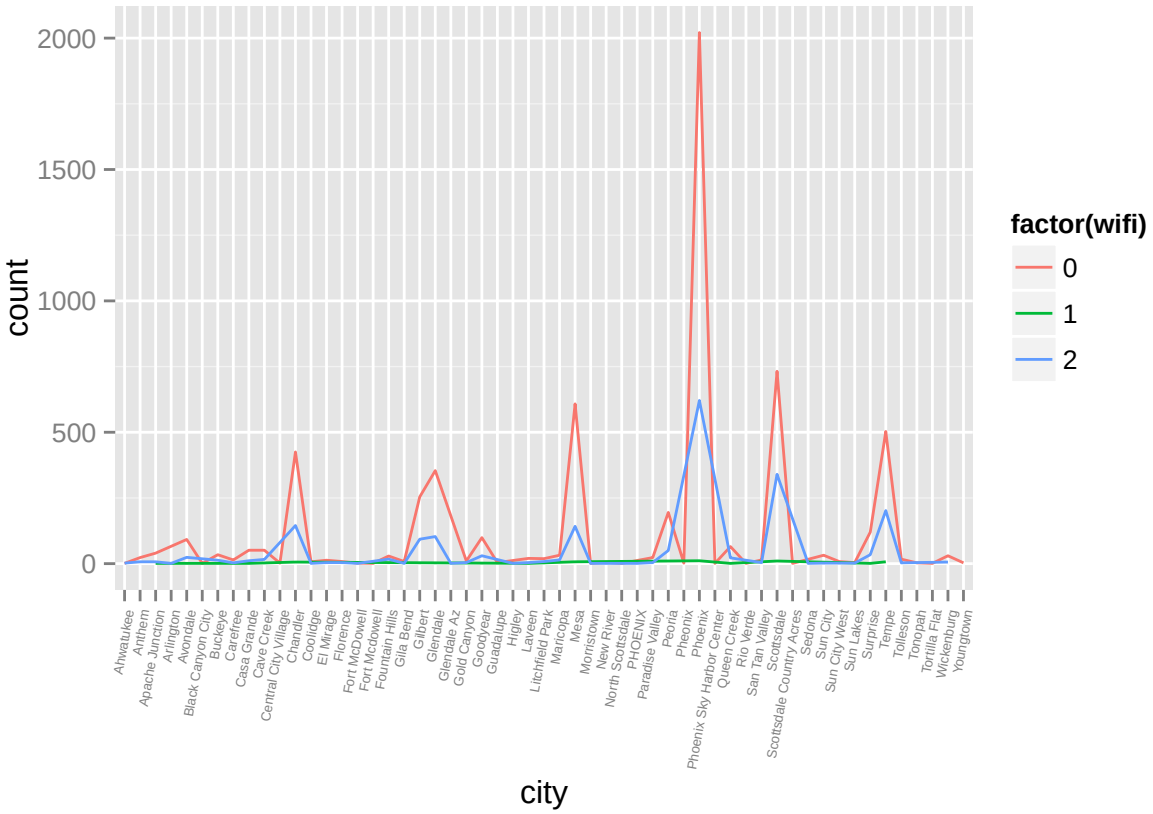


From qplot, we notice that Arizona (AZ) is the highest state. Hence, we will look at the cities which has the highest number of registered restaurants in Arizona given that it's one of populated state and one which is the highest rated State.

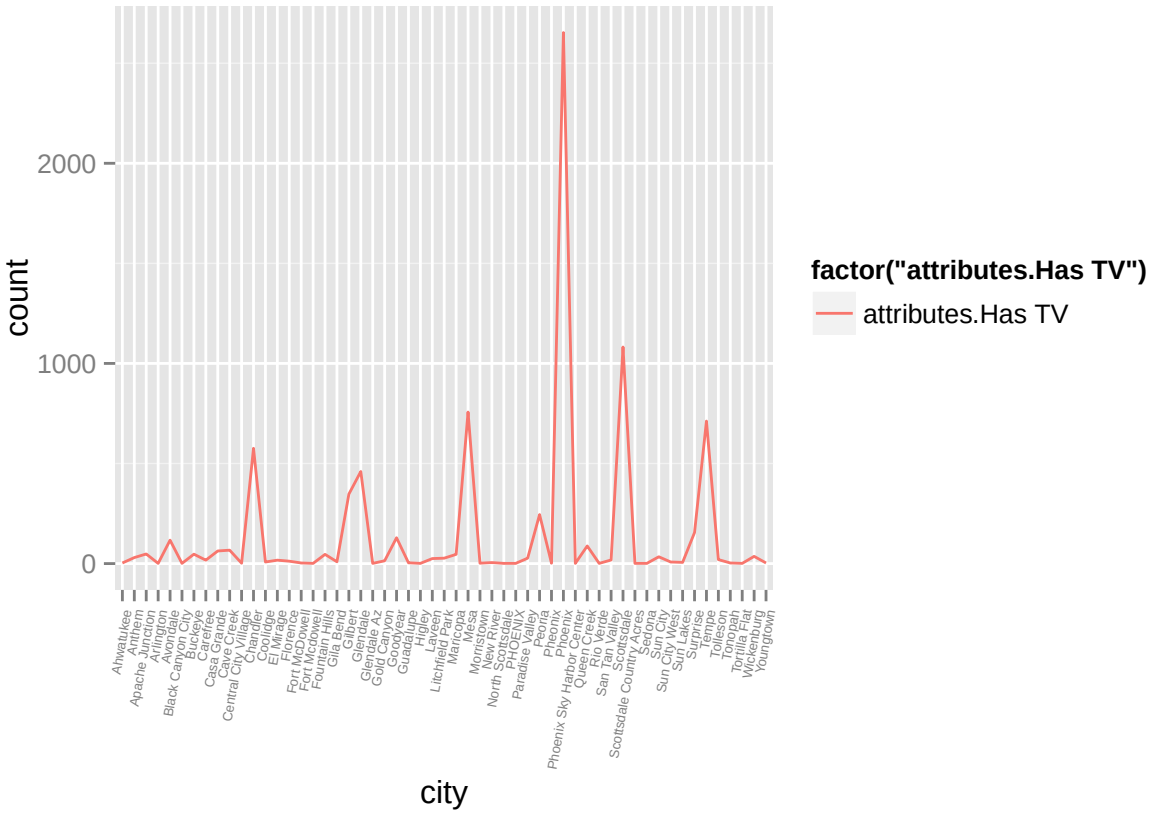


From the city graph, it appears that **Phoenix**, **Scottsdale** and **Tempe** peaked above the others in terms of the number of restaurants found which is more than 200 restaurants. That does not mean that these 3 cities have the highest rated restaurants in the States. But from the data, it does suggest that Phoenix is thei highest in AZ. Not only did it have the largest restaurants, its rating peaked above the others as well. Well, one could really scrutinise even to the types of cuisine or the features that makes Phoenix a popular choice.

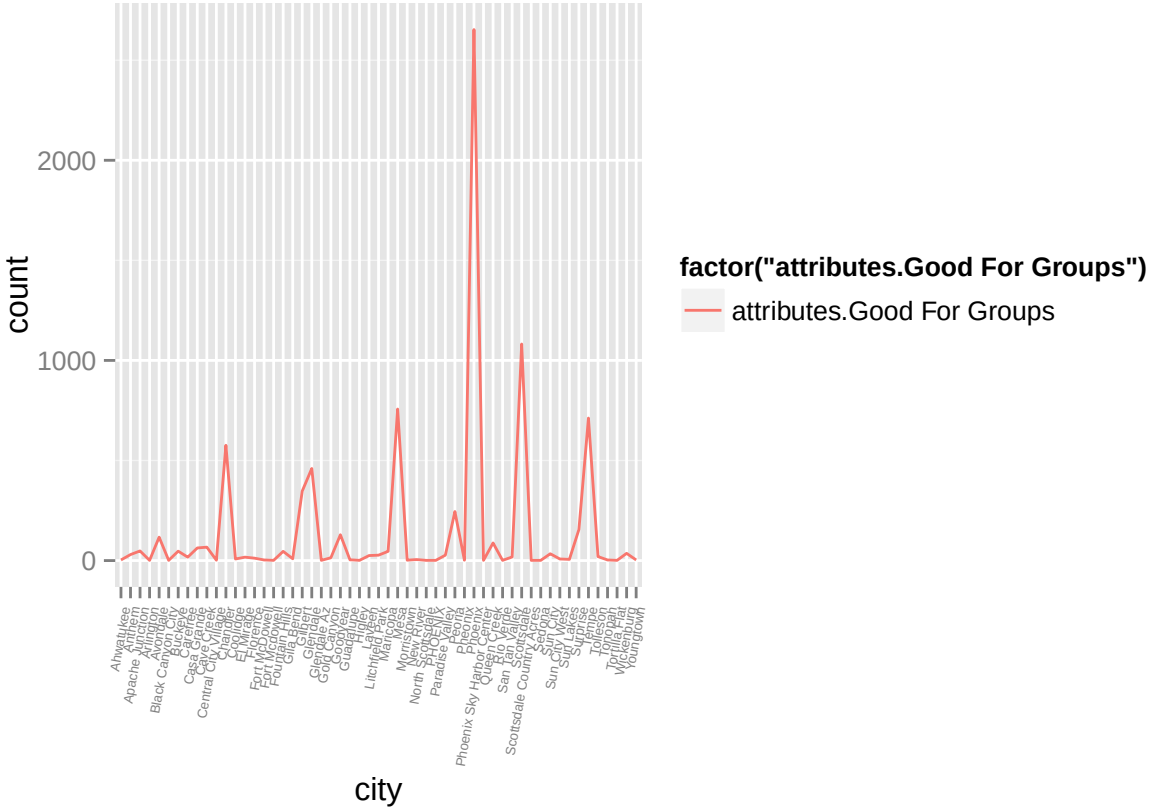
We now examine the four features performance in AZ.



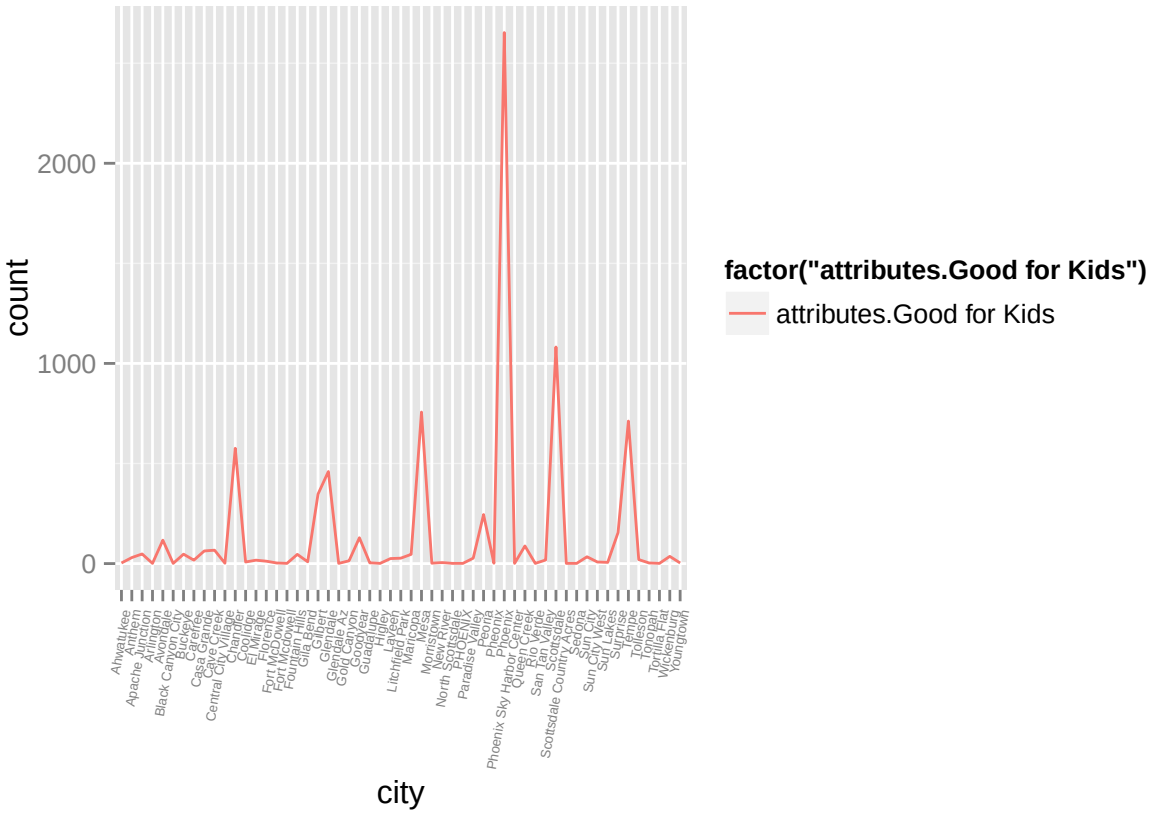
There was some transformation done at the data as the data value included “no”, “NA”, “free” and “paid”. As a typical consumer, one would typically give a higher rating if the service is provided free. As such a rating of 2 is given. No wifi is given score of 0 and paid wifi is given a score of 1. There will be dispute if score of 1 should be given for paid wifi, but the reasoning is that at least the feature is available and one could then decide to pay if there is an urgency needed. Looking at top 3 cities, Phoenix, Scottsdale and Tempe, it is quite consistent that the 3 of them are the highest in terms of free wifi service given.



Same observation as above applied to the availability of TV found in restaurant.



Same observation for the 3 cities in terms of attribute.Good for groups.



3. Results

Regression modeling can be used to quantify the relationship between the features/facilities and the ratings that the restaurant gets.

Call: lm(formula = stars ~ GoodForGroups + GoodForKids + HasTV + WiFi, data = bizFinal)

Residuals: Min 1Q Median 3Q Max -2.55992 -0.53728 -0.03728 0.44896 1.66406

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.375054 0.040179 84.000 < 2e-16 ***GoodForGroupsTRUE 0.201339 0.031409 6.410 1.58e-10*** GoodForKidsTRUE -0.002712 0.027367 -0.099 0.921073
HasTVTRUE -0.022641 0.018699 -1.211 0.226007
WiFino -0.013761 0.019128 -0.719 0.471924
WiFipaid -0.355340 0.095395 -3.725 0.000197 *** — Signif. codes: 0 '*** 0.001 **' *0.01* '' 0.05 '.' 0.1 '' 1

Residual standard error: 0.6521 on 5257 degrees of freedom Multiple R-squared: 0.01056, Adjusted R-squared: 0.009617 F-statistic: 11.22 on 5 and 5257 DF, p-value: 8.88e-11

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.38	0.04	84.00	0.00
GoodForGroupsTRUE	0.20	0.03	6.41	0.00
GoodForKidsTRUE	-0.00	0.03	-0.10	0.92
HasTVTRUE	-0.02	0.02	-1.21	0.23
WiFino	-0.01	0.02	-0.72	0.47
WiFipaid	-0.36	0.10	-3.72	0.00

The model is suggestive of the corelations. In particuar, the P-value for coefficients shows the 2 features kids friendly environment and the TV is significant compared to the rest.

4. Discussion

Hence, we can infer that having a restaurant set up with kids friendliness in mind could be pivotal for a restaurant business. From the data exploration, we also see that the city fared best with all the four features present. Vice versa for the other cities that do not have the high rating. On data exploration, if you were to examine the categories offered by the 3 top cities, they are almost similar.

On the whole, expanding the restaurant in any of the 3 cities with the provision of any of the features/facilites woud not impact the rating as the rating is already high. But one thing for sure, because of the postiive and high rating, it could also benefit as the reviewers would pick the popular spot and so it is good to be located in the popular belt.

Assumptions for this analysis

- We examine common attributes of similar businesses which achieved good rating to learn and adopt their model.
- We are not concerning ourselves with saturation of market, otherwise, other source of data would have to be used like the population census.