Coursera Data Science Capstone

Technophobe1
November 22, 2015

Executive Summary

The purpose of this document is to develop and answer the Coursera Peer Assessment for the Coursera DataScience Capstone Course. The problem and question we wish to ask and answer is

- "*"In what region of the Yelp Dataset Challenge dataset does the highest density of a particular form of restaurant occur (Example Navada) and which particular form of restaurant in the context of a particular event such as Mothers Day gets the best review?"*
- Secondly, ***"What characterists would allow us to predict that restuarants of type X in region Y have a higher probability of a good review?"***

The intent is to combine both location, and temporal data and natural language processing to determine if a correlation exists between location, type of restaurant and event. For example, can we determine if the hypothesis ***"People in region X, have the highest density of Chinese restuarants, yet give the best reviews to Italian restaurants on Mother's Day"*** is true, false or non-answerable based on the data-set we are using?

Data Set Description

The dataset used in this report was provided by Yelp as part of the Yelp Data Set challenge program.

The yelp dataset includes:

Dataset Attributes	Dataset Regions
1.6M reviews and 500K tips by 366K users for 61K businesses	Europe: U.K.: Edinburgh, Germany: Karlsruhe
481K business attributes, e.g., hours, parking availability, ambience.	Canada: Montreal and Waterloo
Social network of 366K users for a total of 2.9M social edges	U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison
Aggregated check-ins over time for each of the 61K businesses	

Our goal was is to do an initial mapping of the geo data, nothing to complex. We simply wanted to explore the data. We started with a macro view; what cities do we have in the data set? From that we were able to then start to look at the geographic data...

Conclusions

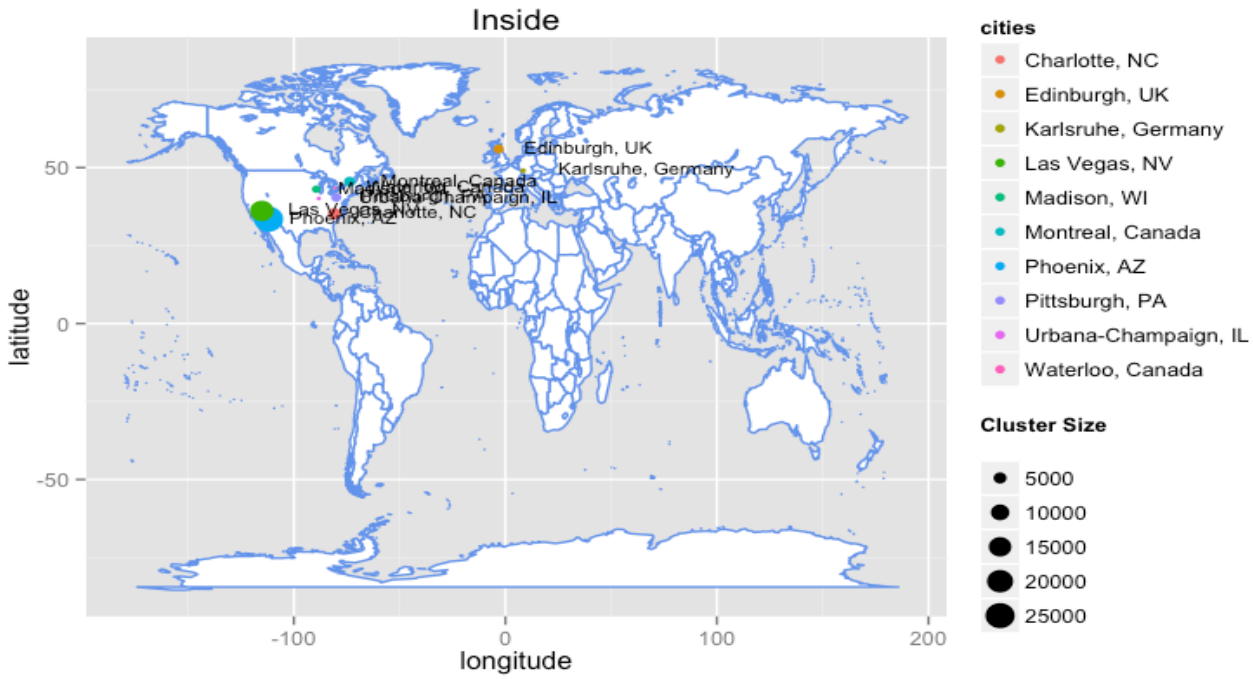
Our analysis conclusion is that it is possible to predict the probability of a good future review for a particular form of restaurant by region. This result was confirmed by the use of **linear regression analysis** of the Yelp: `star rating`, in conjunction with the `Review Length`, incidence of `positive` and `negative` language used in the `reviews` as well as the `overall sentiment` and `total vote count`. Based on the linear model we see a strong Coefficient *`3.608`* in conjunction with a high significance levels (*** are good) for the values selected against the near Variance test.

Regression Results

	Dependent variable:		
	stars.y		highRating
	OLS		*normal*
	(1)	(2)	(3)
reviewLength	-0.00004***	-0.00003***	-0.00002***
	(0.00000)	(0.00000)	(0.00000)
totPos	0.020***	0.020***	0.012***
2	(0.0003)	(0.0003)	(0.0002)
totNeg	-0.033***	-0.034***	-0.019***
	(0.0004)	(0.0004)	(0.0003)
totVotes	0.0001***	0.0001***	0.0001***
	(0.00000)	(0.00000)	(0.00000)
BizReviewCount	-0.0001***		-0.0001***
	(0.00000)		(0.00000)
Constant	3.609***	3.602***	0.767***
	(0.001)	(0.001)	(0.001)
Observations	341,726	341,726	341,726
R^2	0.090	0.087	
Adjusted R^2	0.090	0.087	
Log Likelihood			-135,907.800
Akaike Inf. Crit.			271,827.600
Residual Std. Error	0.507 (df = 341720)	0.508 (df = 341721)	
F Statistic	6,753.605*** (df = 5; 341720)	8,111.536*** (df = 4; 341721)	
Note:			

Note: Sample size by region was determined to have an impact on accuracy based on the cluster analysis.

Exploratory Analysis



What we find is that the size of the relative geospatial datasets sku's the initial result. There is not an even distribution of data by location. The map displayed in the figure above depicts and addresses this problem, the table below shows the cluster sizes. We obtain the data by pulling the geocodes for the cities referenced and use these codes to create the cluster buckets. Note that if we sum the buckets we get: 61184 , which aligns with the sampe row size of `yelpBusinessData`

City	Cluster
Waterloo, Canada	351
Urbana-Champaign, IL	627
Karlsruhe, Germany	948
Madison, WI	2309
Pittsburgh, PA	3041
Edinburgh, UK	3115
Montreal, Canada	3921
Charlotte, NC	5151
Las Vegas, NV	16490
Phoenix, AZ	25231

Cluster Analysis

We created a subset for Las Vegas and created a set of cluster maps that examined the price denisity of restaurants. Looking at the data we saw a cluster of expensive restaurants downtown. Appendix A

vars	n	mean	sd	median	trimmed	mad
1	3262	1.61	0.6881	2	1.518	1.483

Price Density (continued below)

min	max	range	skew	kurtosis	se
1	4	3	1.097	1.444	0.01205

Note that in Las Vegas the median star rating is: **3.5** and the median Price range is **2**. Hence, we see that when we focus interest on higher priced restaurants we see a sku to the the center of town. Specifically, the Las Vegas party strip.

Data Validation

After removal of the superfluous values, and data columns we move to validate the data in preperation for analysis… We first checked for and prepare to filter out near-zero varience predictors, and validate for between-predictor correlations

Near Zero Varience Predictors

To filter for near-zero variance predictors, we use the caret package function nearZeroVar() which will return the column numbers of any predictors that fulfill the conditions outlined.

Why check? There are potential advantages to removing predictors prior to modeling. First, fewer predictors means decreased computational time and complexity. Second, if two predictors are highly correlated, this implies that they are measuring the same underlying information. Removing one should not compromise the performance of the model and might lead to a more parsimonious and interpretable model. Third, some models can be crippled by predictors with degenerate distributions. In these cases, there can be a significant improvement in model performance and/or stability without the problematic variables.

Correlation Predictors

Similarly, to filter on between-predictor correlations, the cor function was used to calculate the correlations between predictor variables:

To visually examine the correlation structure of the data, we used the corrplot package The function corrplot has many options including one that will reorder the variables in a way that reveals clusters of highly correlated predictors.

Methods and Data

The principle method of analysis was to perform a linear regression. To do this used a combination of **dplyr** and **tidyr** to partition and clean the data. The ingest of data was performed using **jsonlite rlist** and **stargazer** packages

A key focus of the exploratory analysis was to look to determine how and where the Yelp data interacts. One aspect of the exploratory analysis has been the stabilization of the reviews by volume over time. As yelp has grown, the accuracy of it reviews appears to improve in part becuase the impact of review outliers is less.

Our supposition is that a key driver of Yelp review usage is the **star rating**. Hence, we want to look at what drives this rating and look at how we might predict future **star ratings**. Based on the data so far reviewed, our sense is that the review text of past reviews may be a strong predictor of future reviews. Our asusmption here is that **People like to be right together, or wrong together; people never want to be right alone, or wrong alone**

Thus, our hypothesis is that positive star ratings are driven by positive textual reviews. Hence, we can look at whether positive and negative review language has a material impact on the star rating given in a review. To do this we perform a linear regression of stars on the number of positive words in the review, the number of negative words in the review. We also look at the length of the review in the context of the language used. i.e. The amount of words in the context of positive or negative language implies stronger feelings.

Thus, our approach is to initially perform a regression of # stars in a Yelp review on # positive words, # negative words, and # words in review, returns these results.

Results / Conclusions

The initial model used for the capstone I originally felt would have a low predictive outcome in reference to future reviews. I was pleasantly surprised by the outcome. Based on the linear model we see a strong Coefficient *3.608* in conjunction with a high significance levels for the values selected against the near Variance test. Expectation is to expand upon this work and build out a ***Random Forest model*** *in conjunction with* ***Stochastic Gradient Boosting*** *for submission* ***to the Yelp challenge.***

Thus, in conclusion the NLM (Natural language model) selected whilst simple gave good results. One of the most interesting aspects of the analysis is the observation that Yelp reviews are stabilizing over time and become thus becoming more accurate.

It should be noted however that some aspects of the dataset create problems as Yelp does not track changes in situation. For example, when dogs are allowed the date might be updated but no record of when is recorded. Thus, there is a historic gap in the data set.

References

- *Mathematical Statistics with Resampling and R*
  - By: Laura Chihara; Tim Hesterberg Publisher: John Wiley & Sons Pub. Date: September 6, 2011 Print ISBN: 978-1-11-02985-5
- Yelp developer documentation

Appendix A

As part of the exploratory analysis the following reports were created. These are provided purely for reference.

Restaurant Density by Type?

We have observed the star ratings and price range interact.

What types of restaurant exist, and which restaurants get the best reviews? The Yelp developer documentation is invaluable in this case as it contains a breakdown of the category types of restaurant listed by yelp.

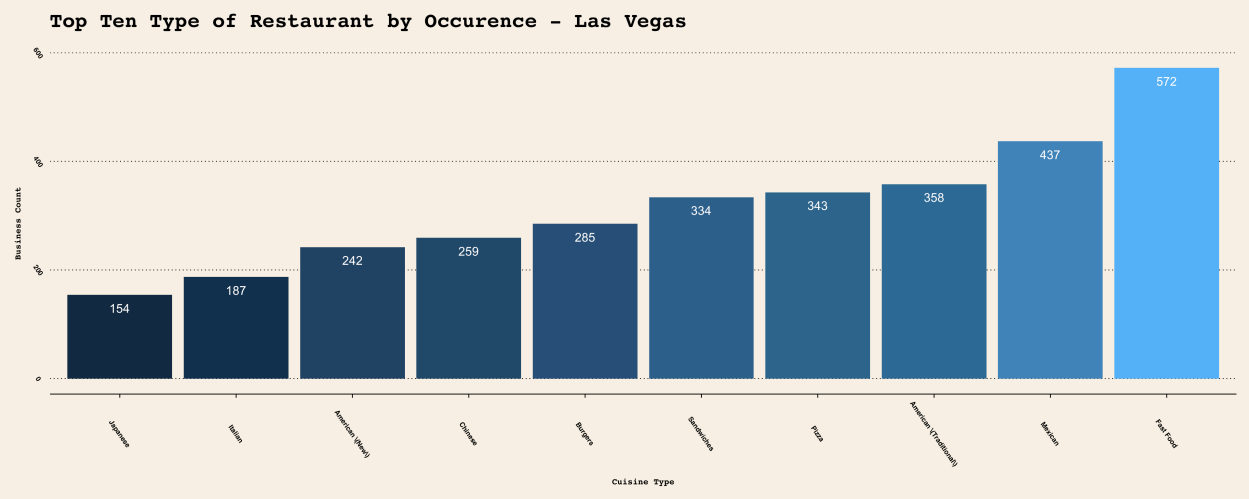
We use the yelp categories data to extract the restaurant data for las vegas for exploration...

This gives the process and answer in the context of Las Vegas as to:

1. In what region of the Yelp Dataset Challenge dataset does the highest density of a particular form of restaurant occur
2. Who gives the most reviews by restaurant type by region?
3. Finally in the context of a particular event such as Mothers Day - ***"Can we predict that restuarants of type X in region Y have a higher probability of a good review?"***"

Highest density of a particular form of restaurant

In the context of Las Vegas we determine that the highest occurance of restaurant is fast food, followed by mexican based on occurance.

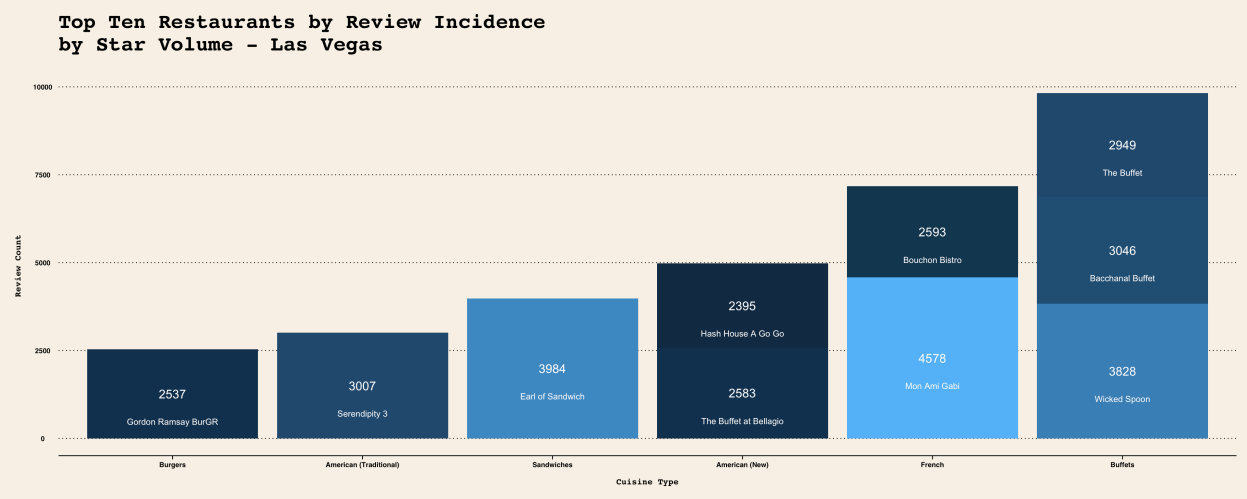


Who gives the most reviews by restaurant type by region?

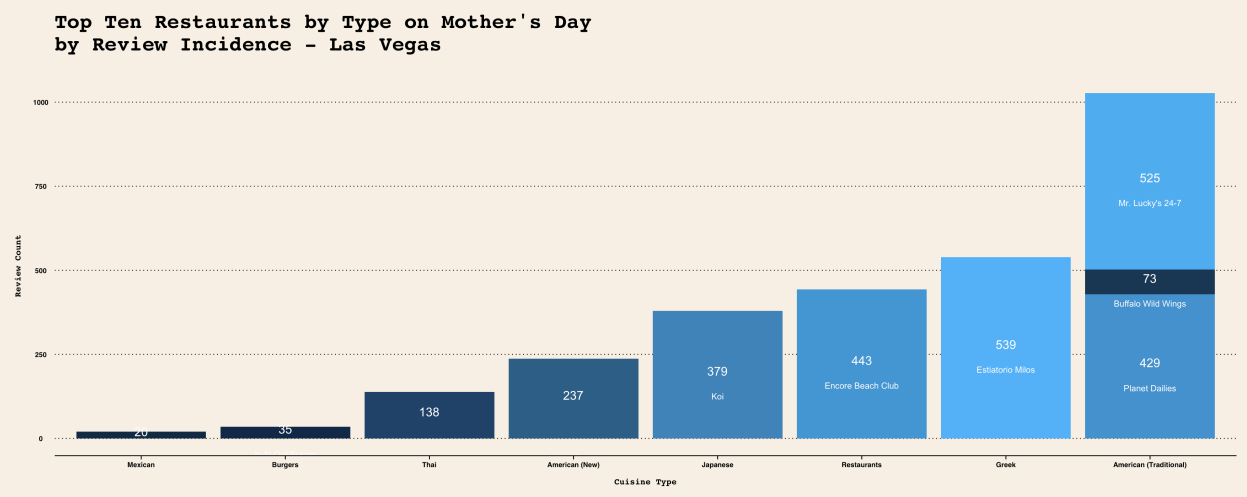
In the context of Las Vegas we determined that **'Buffet'** restaurants recieve the highest review count by Star volume. Note that this also hilights a basic strength and weakness in the Yelp review system...

A review is defined by the context of the review attributes of the observer... thus we get very different answers based on what we value... For example, do we order by review volume, star rating and price? Or do we order by price, star rating and review volume. In each case we get a different view of the restaurants that buble to the top.

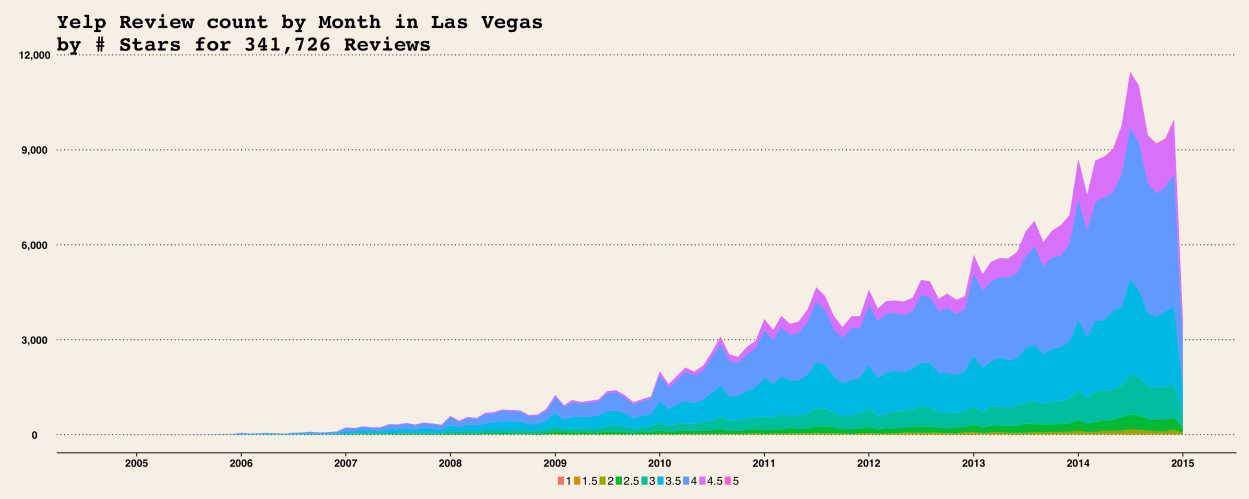
Thus, we can observe our social context will define our search criteria and impact on our results.



Which restaurant type gets the best review in the context Mother's Day?



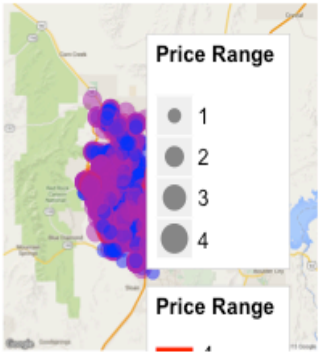
What is the volume of review in the context of star ratings?



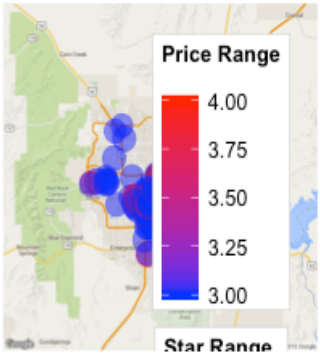
Where are the expensive restaurants situated?...

Here, we compared price range to review count... Note what is interesting here is that the median review count is: **33**, whilst the mean is: **116.49**, the max is **4578**

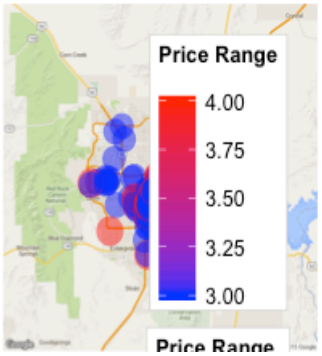
Plot 1: Price Density - Las Vegas



Plot 3: Star Range [3 to 4] - Las Vegas



Plot 2: Price Range [3 to 4] - Las Vegas



Plot 4: Price by Review - Las Vegas