# Data Science Capstone Project Final Report

*Kenneth Lim*

*Sunday, November 22, 2015*

## Yelp Dataset Challenge

### 1. Introduction

This Coursera Data Science Course Capstone Project requires course participants to use the datasets provided by Yelp Dataset Challenge to come up with their own data science ideas and implement them.

Yelp is a business review online platform where users share reviews with other users or businesses. The datasets provided consists of information on the businesses being reviewed ('business', 'check-in'), the reviews ('review'), the reviewers ('user') and any helpful tips provided ('tip').

The initial targeted question this project attempts to answer is - **If a specific business is looking to expand i.e. setting up a new outlet, which location/neighbourhood/city would likely be most favourable?**

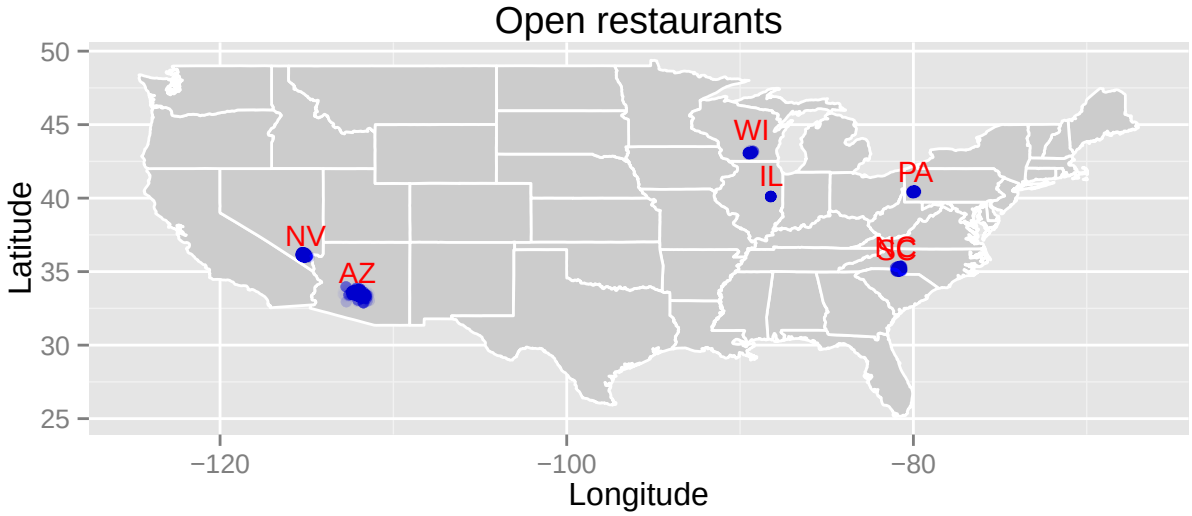
### 2. Methods

**Get & Prepare the Data** The datasets are downloaded via the link provided in the Coursera Capstone project instructions. Once downloaded and unzipped, the 5 pseudo-JSON files are read and flattened as R data tables for easy manipulation. These objects are also saved as RDS for easy reload in subsequent R sessions.

**Exploratory Analysis** From preliminary data exploration and past experience with relational database (RDBMS), it becomes immediately apparent that the datasets are inter-related with business_id and user_id used as primary/foreign keys.

An examination of the business dataset reveals that each business is tagged by multiple categories. Since the target question calls for a specific business, a business category has to be selected. A quick tabulation as well as word cloud shows 783 possible candidates. "Restaurants" jumps out as clear winner with 21892 records out of 61184.

Using the business dataset, initial plotting on world map quickly reveals that most open restaurant businesses in the Yelp dataset are US-based (13753 out of 17558) with the remaining 3805 residing in European continent. For the purpose of this project, we assume the focus is in the US. Upon narrowing onto US map, it becomes apparent where restaurant businesses in Yelp's dataset are clustering around.




**Approach**

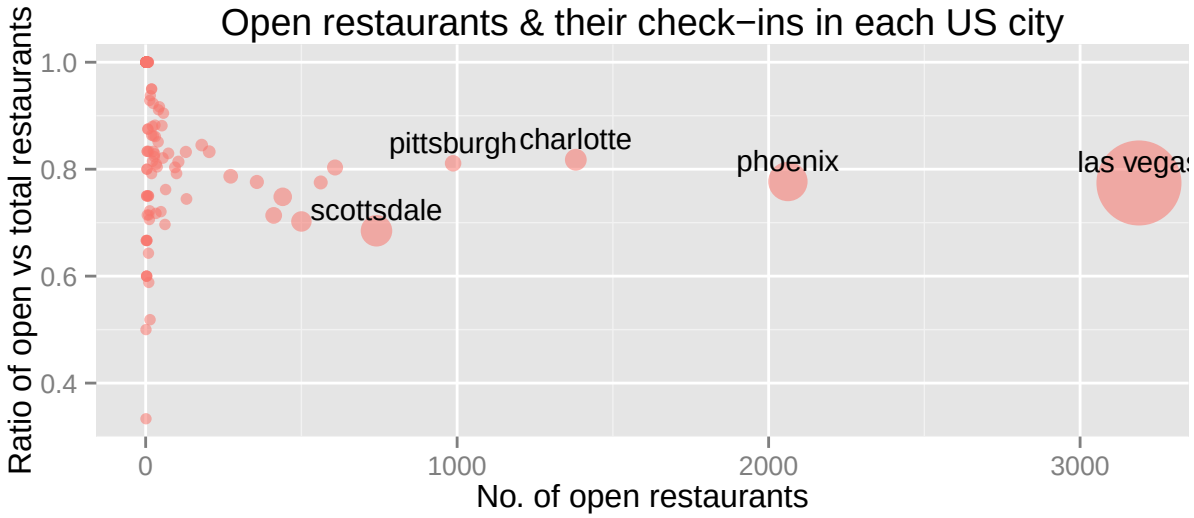
Now to address the targeted question:

1. For a specific business category, we look for cities where there are high ratio of open businesses to closed ones as well as high volume of transactions proxied by no. of check-ins provided by the check-in dataset. This reflects that market demand there are able to support the specific business and market supply have not gone past market saturation point.
2. Once we have selected the city, we shall make a preliminary plot of the restaurants' locations for any obvious pattern or clusters.
3. Using data clustering algorithms (e.g. K-means), we attempt to identify locations of high market demands (proxied by check-ins volume) and conducive consumer market (proxied by high review ratings and high volume of reviews).

**3. Results**

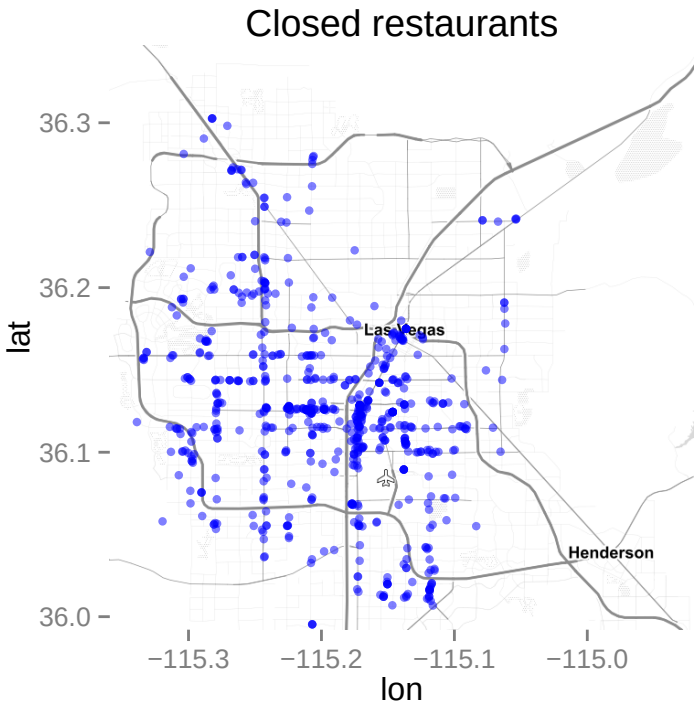
**Candidate City** To look for our target city, we plot the ratio of open restaurants vs all restaurants vs volume of check-ins in each US city. It should be noted that:

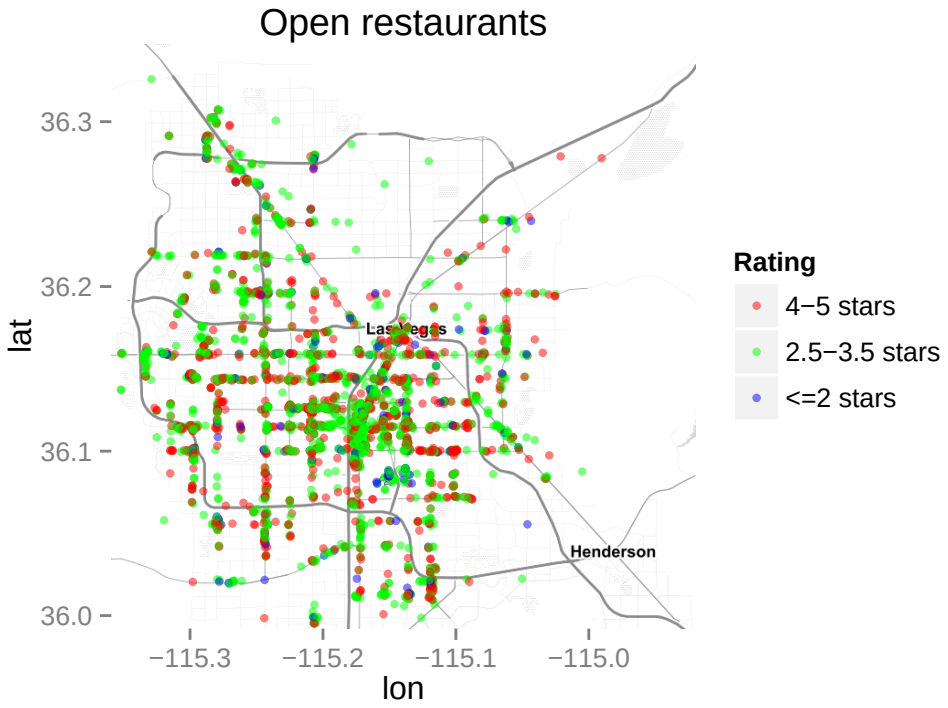
- The value of the textual field "city" is not consistently maintained (e.g. spelt differently) and this is addressed as best as possible e.g convert all to lower case.
- Some 16018 out of total of 61184 businesses do not have check-in data. In turn, 627 open US restaurants out of 13753 do not have check-in data.




From the plot, we have an obvious candidate city - **Las Vegas** - which overshadows every other US city in terms of no. of restaurant businesses (23% or 3190 out of all 13753 open restaurants in US) as well as the no. of check-ins (40% or 1261764 out of total 3138591 check-ins in all open restaurants in US). In addition, the ratio of open restaurants to both open and closed restaurants in Las Vega stands at 77%, which rightly or wrongly as a proxy reflects reasonably good survival rate of restaurant business in Las Vegas.

**Preliminary Plot** Now we plot all restaurants in Las Vegas city against Google maps.



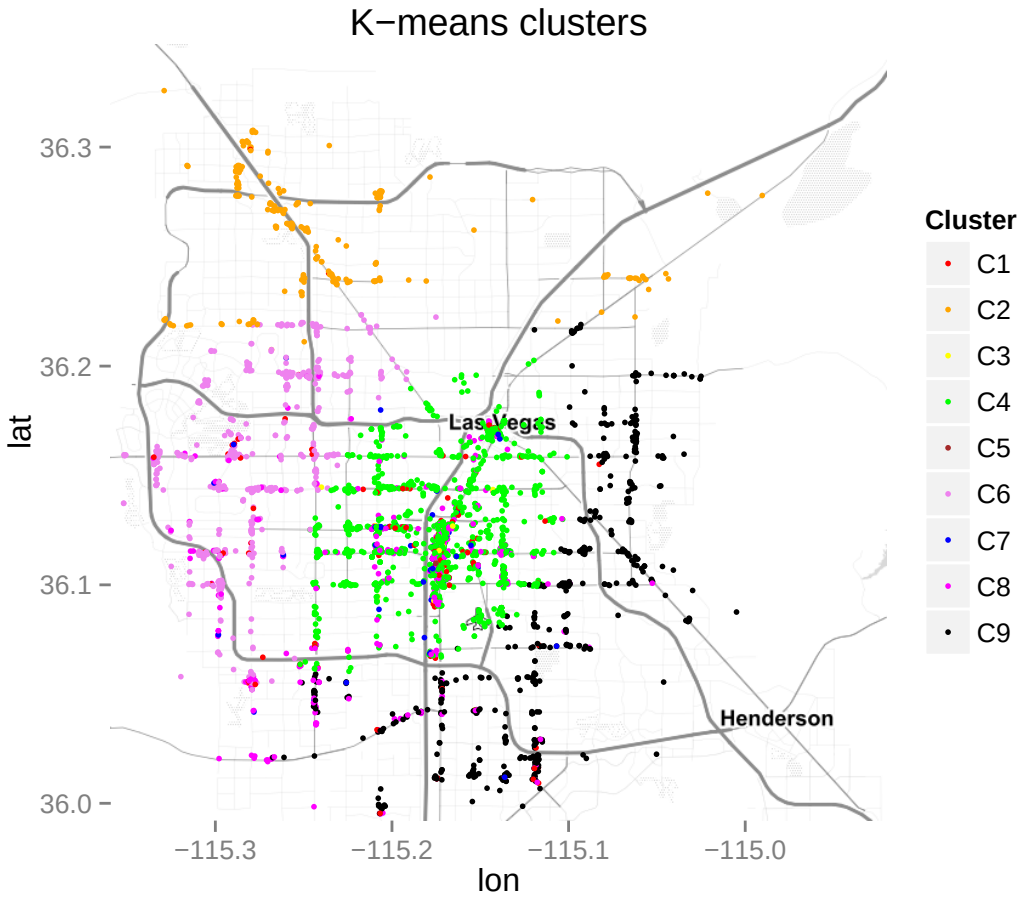



Comparison between the clusters of closed restaurants vs clusters of open restaurants shows both their locations to be almost identical. Notably, the largest clusters are along the stretch of route through the city centre which a search on Wikipedia reveals it to be known as the Las Vegas Strip, internationally known for its concentration of resort hotels and casinos as well as art district. This points to the likelihood that restaurants' closures there are due to rental hikes and tenancy renewal rather than poor business or human traffic/footfalls.

Also can be seen from the plots are that restaurants always line-up against major roads and concentrated in the city centre and thin out towards the outskirts. Interestingly, the cone-shaped area north of city center bounded by the 2 highways is quite bare. Based on Google satellite map, that area is populated by North Las Vegas Airport, golf course, a few parks, and schools. This hints at the possibility that the area is low-income residential area, thus not favourable for setup of restaurants despite being well-connected by roads.

Based on the plot on Google map, there is no immediate discernible difference in location-based clustering of restaurants that are highly rated, average, or poorly rated; they do not deviates from the general pattern of clustering of restaurants in Las Vegas.

**Data Cluster Analysis** Now we use K-means clustering to try to further identify locations of high market demands and conducive consumer market based on 3 variables - check-ins volume, star ratings and volume of reviews (9 clusters) - together with longitude and latitude.




- Cluster C5, C3 has the highest mean check-ins (11079, 5630 respectively), star ratings (4.0, 3.8 respectively) and volume of reviews (3859, 2096 respectively) but only their cluster sizes are very small (4, 17). While worth taking into consideration, they may not be true reflection of of the market condition and consumer sentiments at their locations.
- Cluster C7 has the next highest mean check-ins (2729), star ratings (3.9) and volume of reviews (842). It is unsurprising that it is situated around the renowned Las Vegas Strip and would be considered prime candidate location for restaurant business.
- Clusters towards the outskirts of the city have the lowest mean check-ins, star ratings and volume of reviews, thus should be avoided.

**Assumptions** It should be noted that certain assumptions were made:

- We are not looking for first mover advantage in places where the market have yet to be established. Otherwise, the work carried out would have taken a totally different direction.
- We assume there is no astroturfing but that every review submitted is genuine.
- We initially assumed locations with high rate of closure indicate market saturation, thus to be avoided but as the work continues, it need not be so as seen above.
- We are not factoring in the effect of entering a location where there is highly successful competitors which can have its disadvantages (as well as advantages).