[CSE590] MINI PROJECT – 3

[ADITI SINGH SBU ID (110285096) ADISINGH@CS.STONYBROOK.EDU Yelp_academic_data1]

1. Introduction

1.1. Data Set:

The data set was downloaded from the yelp academic data set challenge portal. It comprised of 5 JSON files with Review, Business, Tips, User, Check-in information about 1.5 GB in size in total. Each file has a unique ID which can be used to join different tables together.

2. Problem (or approach) statement

2.1. Problem:

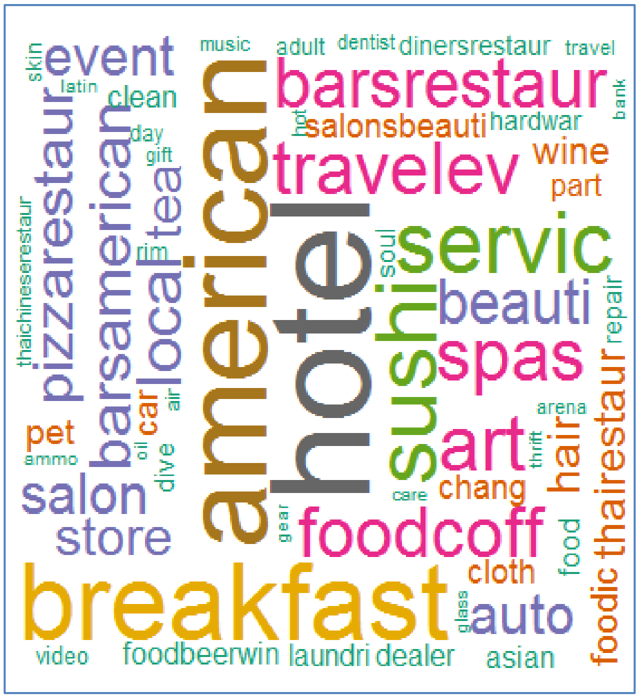
- To find the popular subjects/categories with maximum comments and reviews
- To find the top and bottom 10 successful businesses and their locations
- To find the geographical area where Users are most active and their sentiments' trend over time.

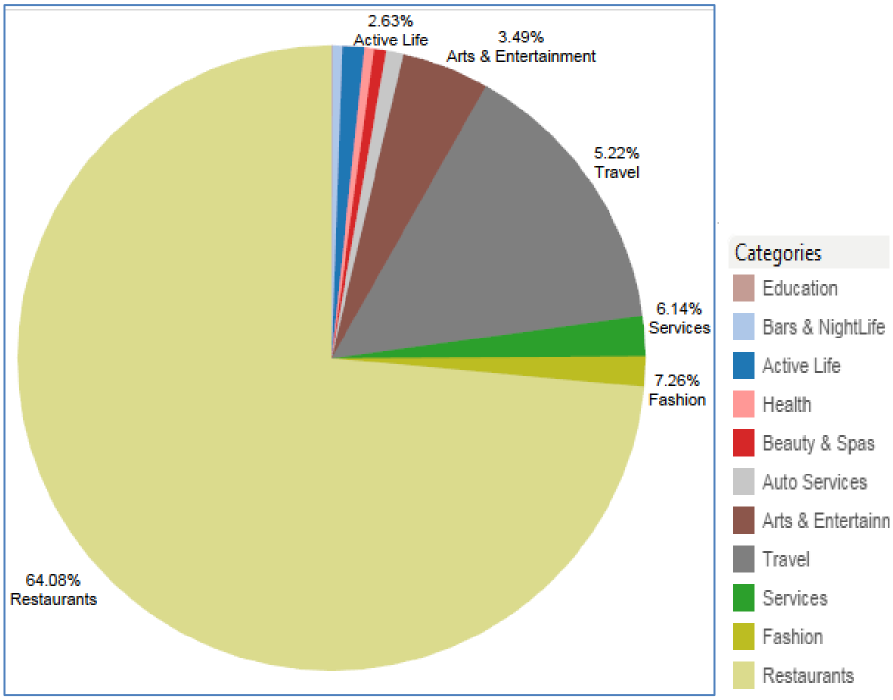
2.3. Approach:

- Data was converted to CSV format using Python, to be ingested in R for the data cleansing phase
- The data was grouped into broad categories (eg. Entertainment, Travel, Food , Health etc) from the given Categories in Business and Review data by hand.
- To find the most popular topics of discussions and the top rated businesses, plots were made for comparison
- To find how the topics vary over time and understand the sentiments of the reviewers, **wordclouds** were created in R using the NLP and Text Mining package.

3. Results and discussion

3.1. Categories: The categories are shown below in a Pie chart format and Word Cloud format (11 broad categories)

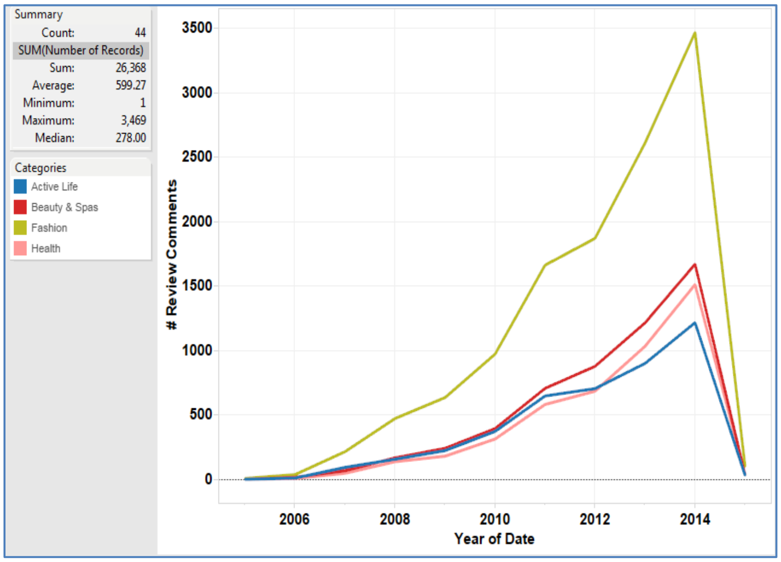




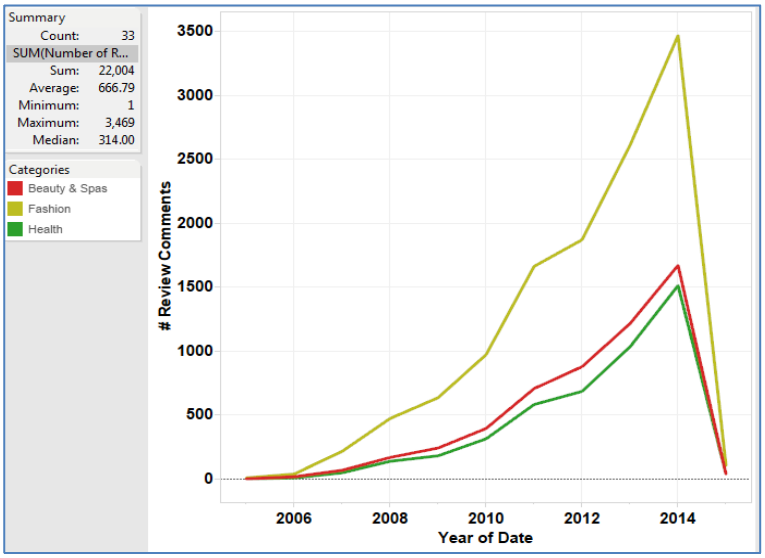
- The WordCloud defines the most discussed topics (in this case : the categories over all). The Categories were converted to a normal text file and then cleaned (with white space , punctuations, numbers, non- english words and common english words like(“the”, “a”, “to” etc) removed). This was then converted to a word-cloud in R using their frequency as size.
- Result: As shown in the above two figures, **Food and Restaurants** are the most reviewed categories with 64% of the total frequency. In the word cloud also we could see **“breakfast”, “american -cuisine”,”Bars”, “Pizza”** and **“sushi”** standing out followed by **“travel”** and **“hotel”.**

3.2. Categories trending over time: #Review comments for a category Vs. Year

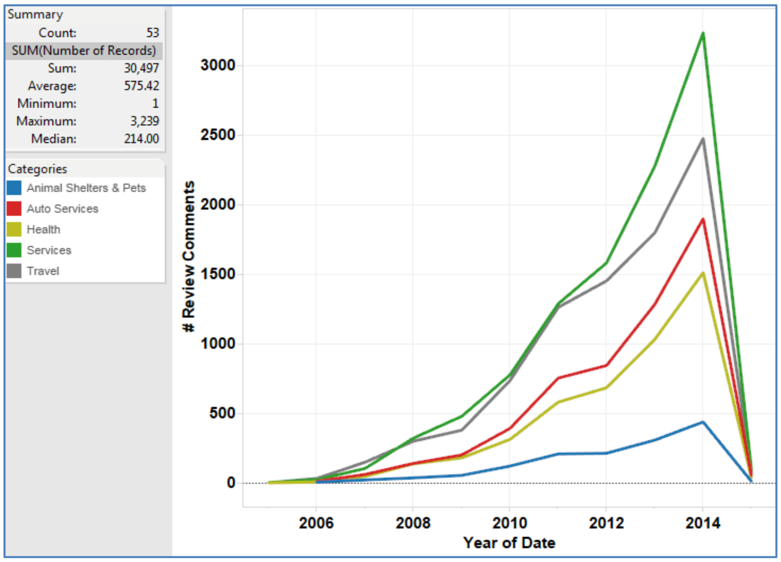
The Following 3 plots show the trend the various categories follows over the years. The different colored lines show the different broad categories and the Y axis shows the frequency of their occurrences in the reviews.



Summary	
Count:	44
SUM(Number of Records)	
Sum:	26,368
Average:	599.27
Minimum:	1
Maximum:	3,469
Median:	278.00



Summary	
Count:	33
SUM(Number of R...	
Sum:	22,004
Average:	666.79
Minimum:	1
Maximum:	3,469
Median:	314.00



Summary	
Count:	53
SUM(Number of Records)	
Sum:	30,497
Average:	575.42
Minimum:	1
Maximum:	3,239
Median:	214.00

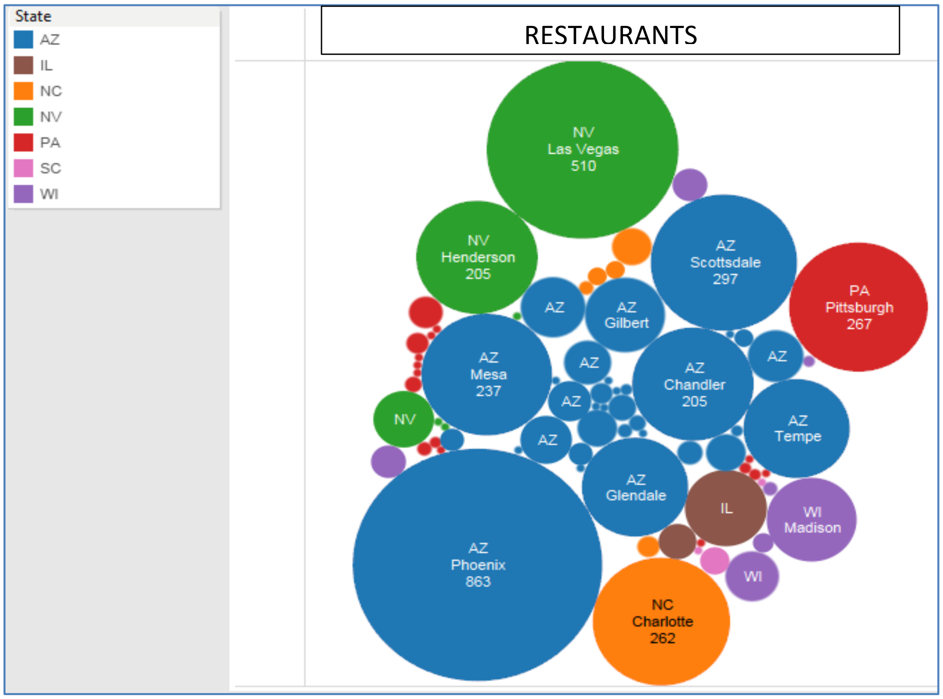
ABOUT CATEGORIES:

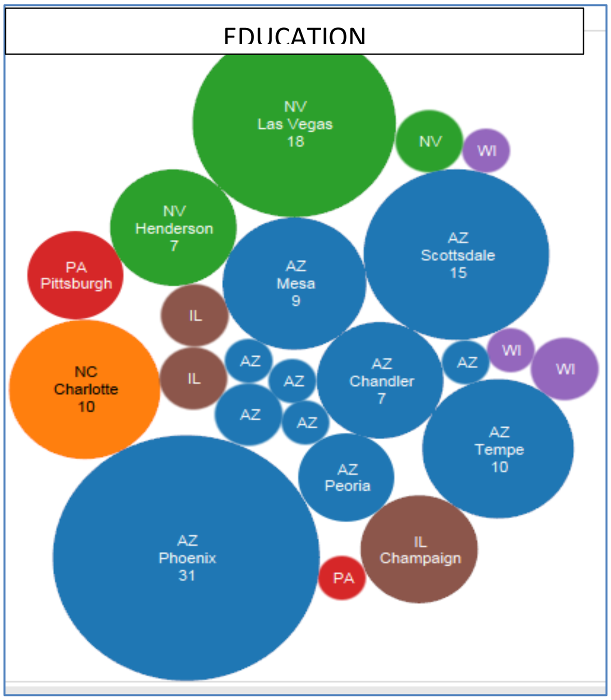
- The broad categories plotted above show a definite rise with small dips in between, peaking at 2014.
- This trend is visible over the whole set, showing that the number of yelp users increasing over time, thus a proportional increase in the reviews per category.
- The sudden dip in 2015 is there because the data is only till the end of January (size is very small compared to the huge volumes in the previous year).
- **Deduction: The general trend is a upward rise, signifying the increase in users using the yelp portal to write reviews about the services they received.**

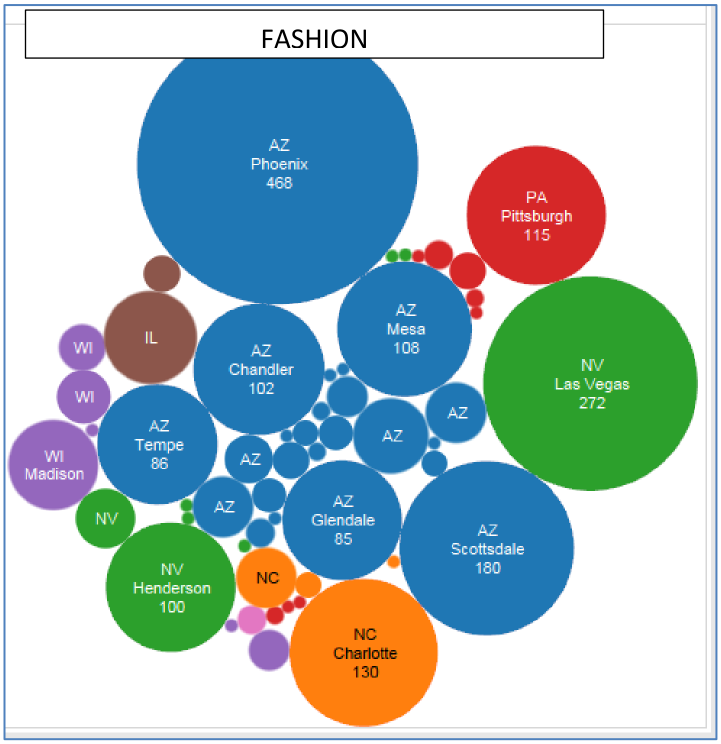
3.3. The States leading in their Businesses: [4 Major Categories discussed]

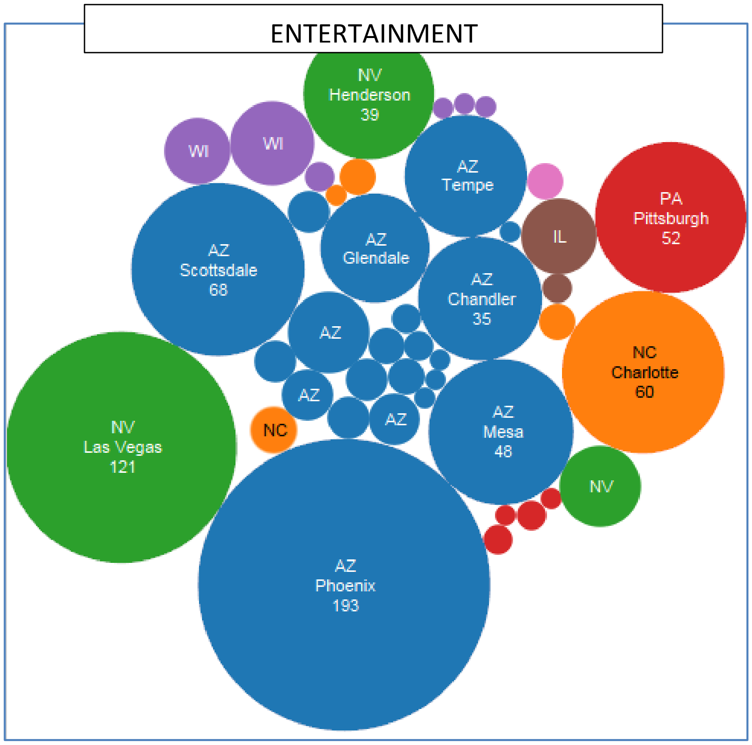
Assuming that the number of user reviews signifies the popularity of a particular industry, Bubble charts below show the leading States and Cities for a particulat category.

As shown in the figure below Arizona's Phoenix City leads in all major catogories , followed by Nevada's Las Vegas.

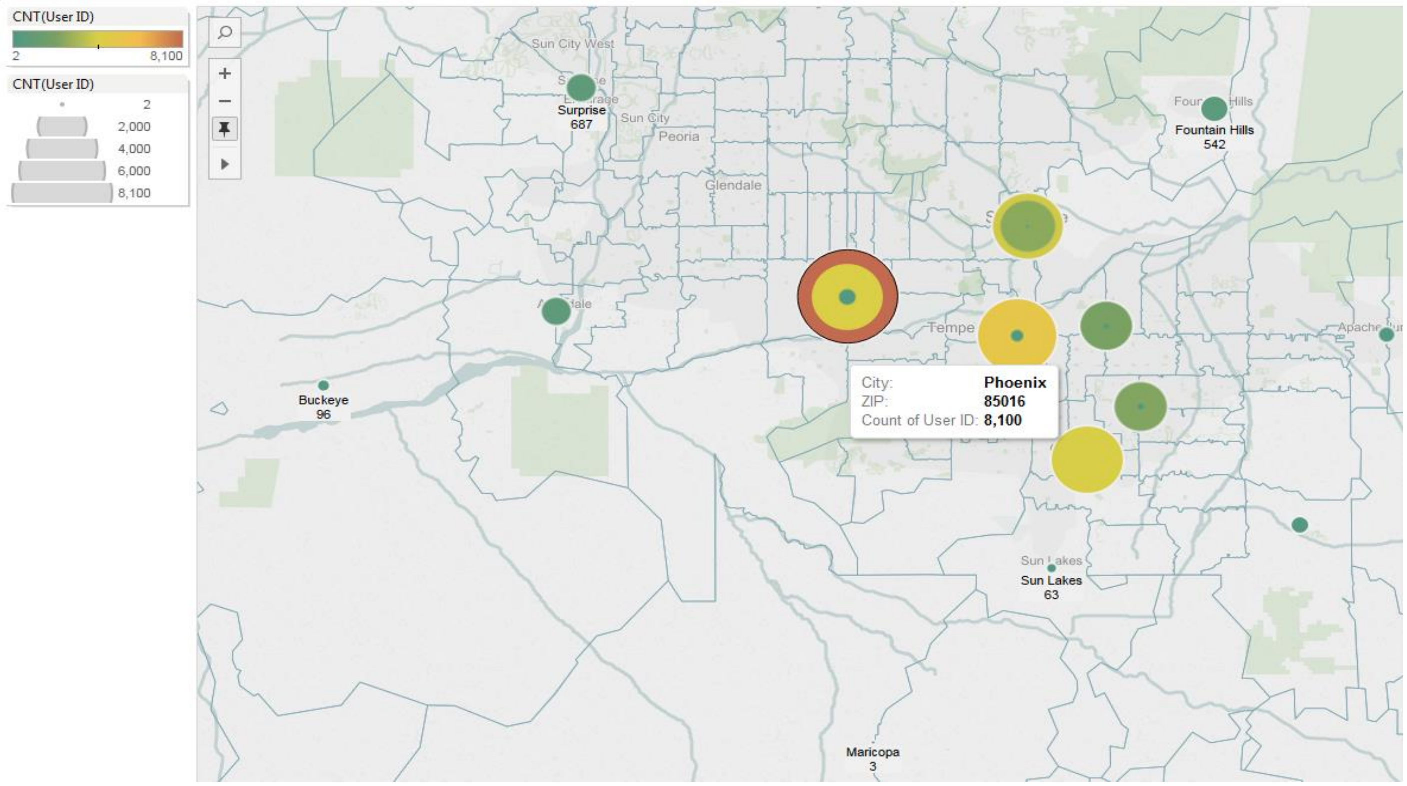








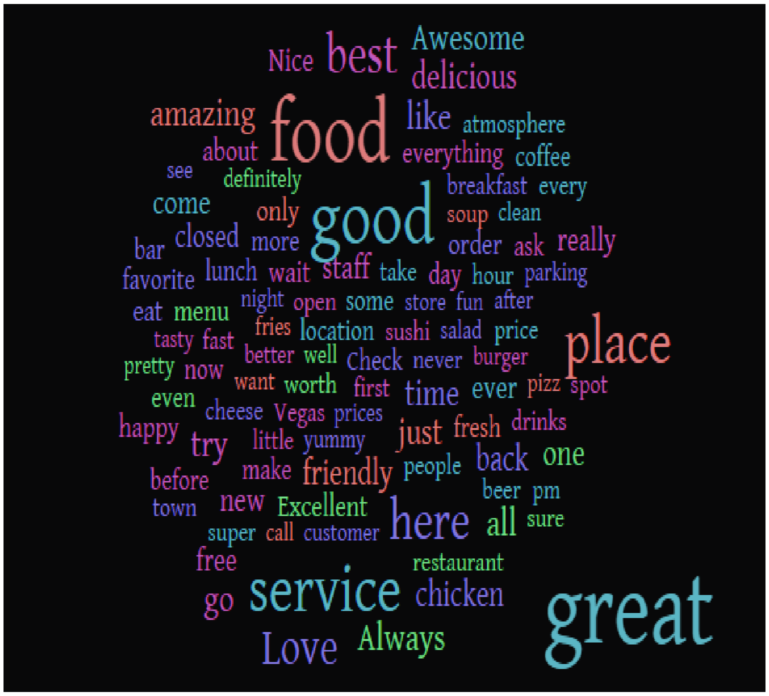
3.4. The States leading in their User Count: [All Combined]

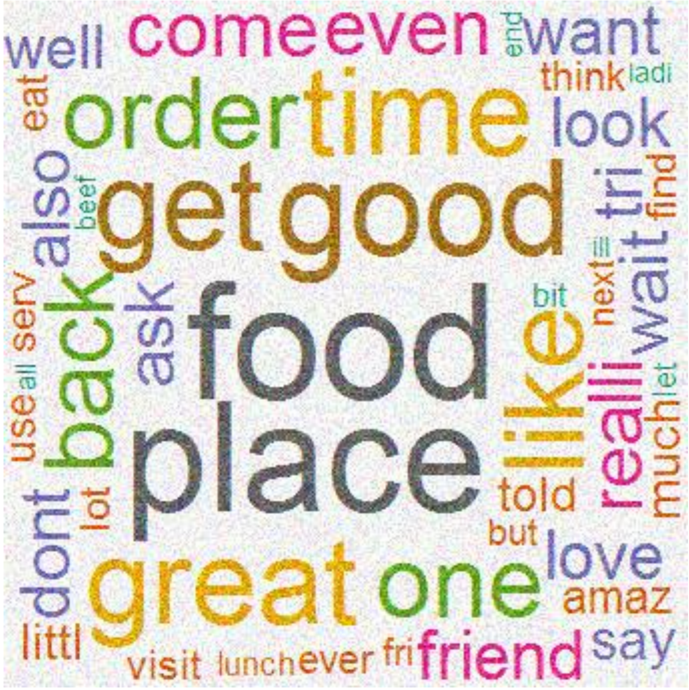


From the above deduction that Arizona's Business industry is the most reviewed on Yelp, particularly in Phoenix, I wanted to see if it is because most of the users originate from Phoenix.

As shown in the figure above, Arizona's Phoenix City leads in the count of the number of distinct users as well , followed by Tempe and Scottsdale. The Zip was exctracted from the full address and then the count was calculated.

3.5. Sentiments Illustrated From the User's Tips and Reviews:



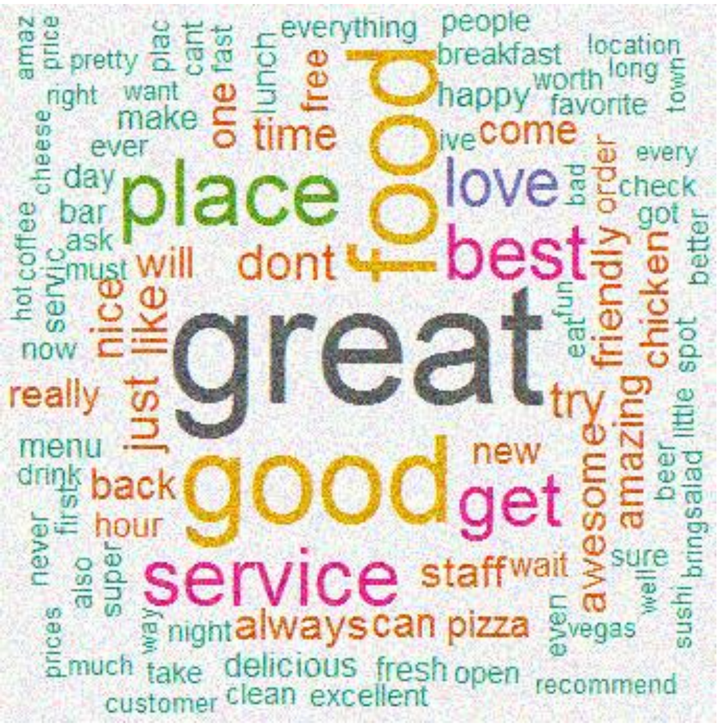


USER TIP'S FOR YEAR- 2015

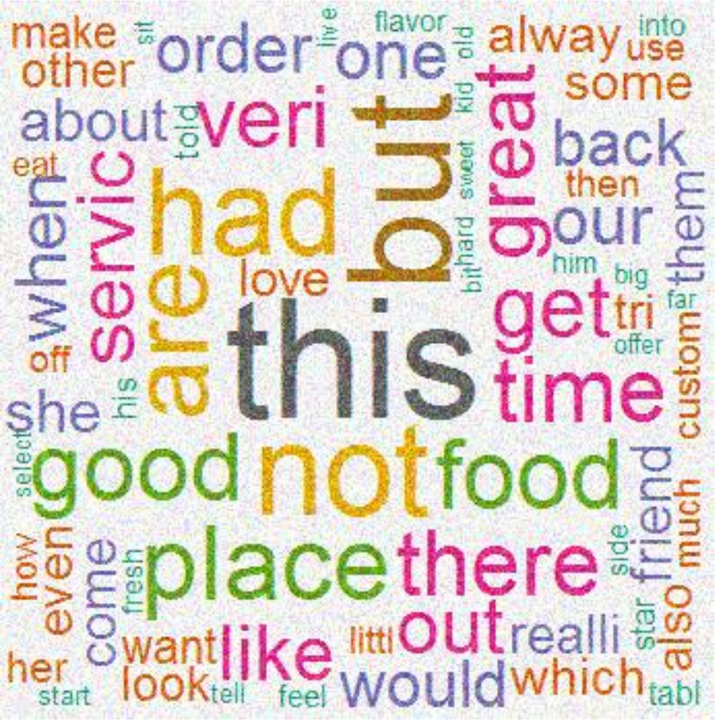
REVIEWS FOR YEAR- 2015

- **The tips show a certain positivity, with Great and Good overpowering the rest of the sentiments of the customers.**
- They talked more about the customer service ,restaurants and food more often than the other topics.
- The most tips were given for Sushi, Salad, Coffee, Chicken ,burger, Cheese, Pizza and drinks.

- Similarly for Reviews , most of the users gave <u>**positive comments**</u> about the Food at restaurants and bars.
- This can be determined by the ,words "food" and "Place" appearing most frequently. So the review is mostly about the **food and the place** .
- Most have described about the time and the services they received as well. Since the words like "Order", "Time" and "Service" have come up pretty frequently in the review cloud.
- The words "great" , "good" , "like" show that most were positive in nature. **So the users generally on an average tended to post more positive comments than negative** (eg. The word "don't " appears smaller)
- The second observation I had was that people tended to go out with **friends**, since the word "friends" appears quite dstinctly in the wordcloud.



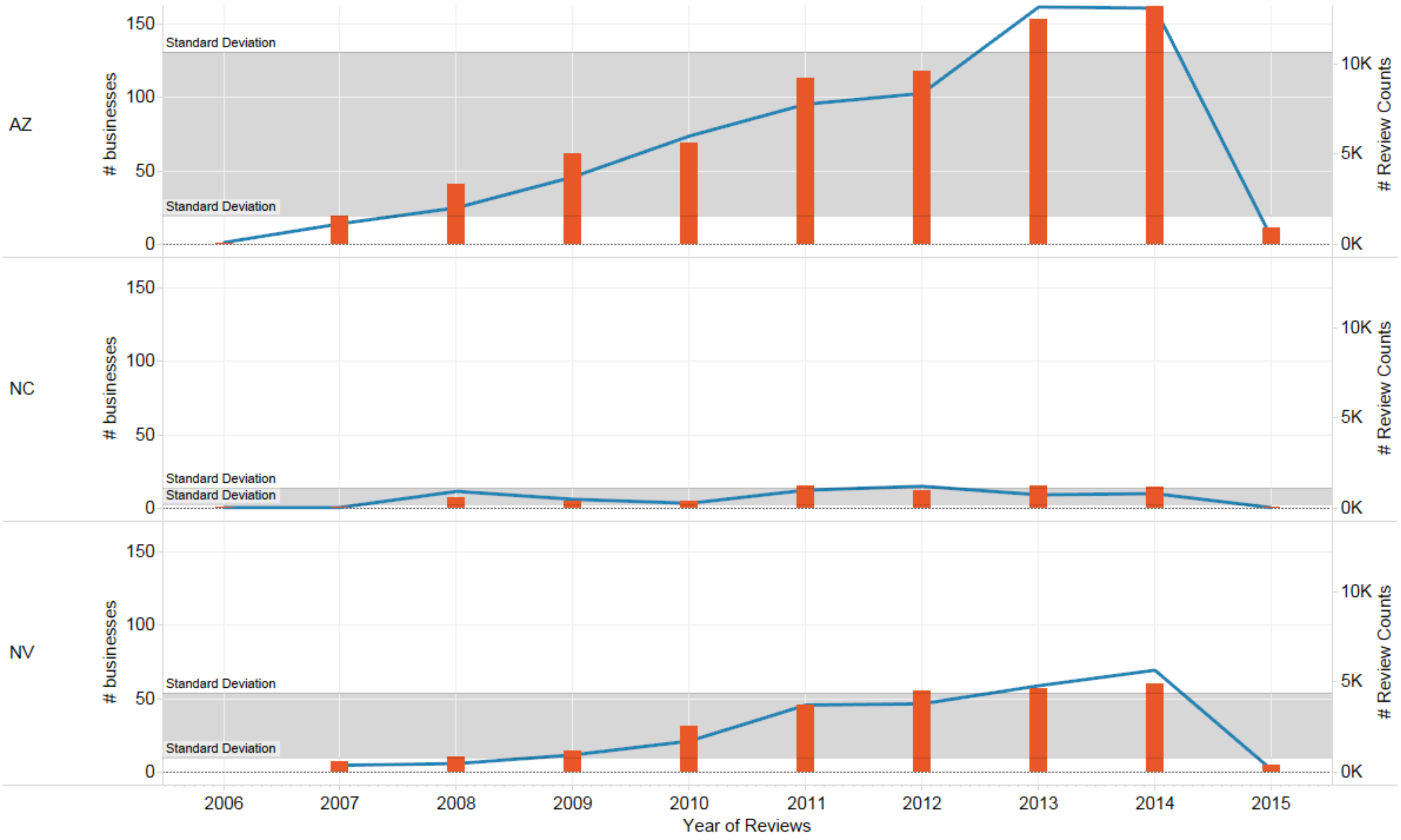
USER TIP'S FOR YEAR- 2014



REVIEWS FOR YEAR- 2015

- A similar trend could be seen in the reviews for 2014 too.
- But in reviews, the word "but" and "not" appear frequently too showing that some negative comments and reviews were posted as well.

3.5. State Wise Business and review counts:

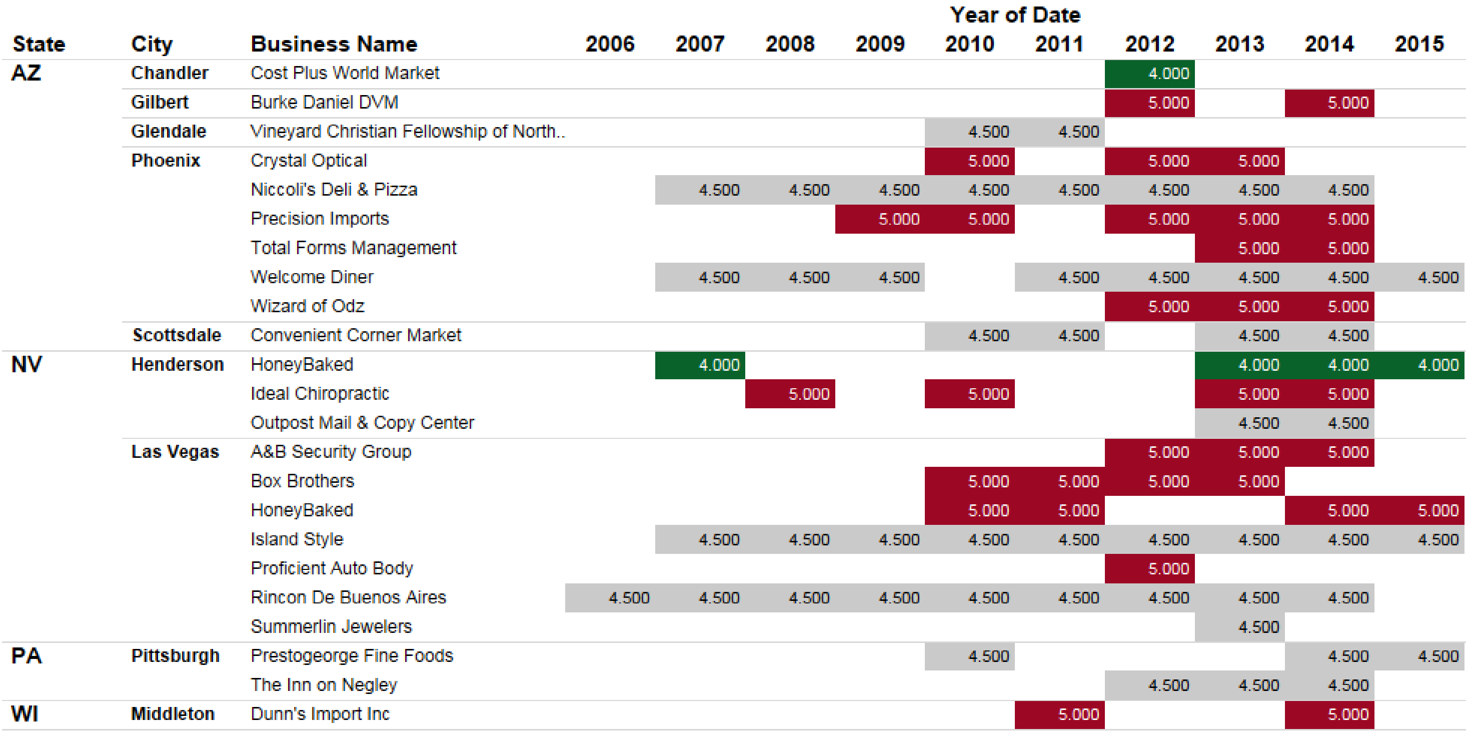


The figure above show the plot of # of businesses over time and how they compare to the sum of the review counts over time. The Blue line is the Review count and the Orange bar describes the # businesses the reviews were posted about.

Both the attributes mimic each other except in a few places . eg. Arizona in 2011 etc.

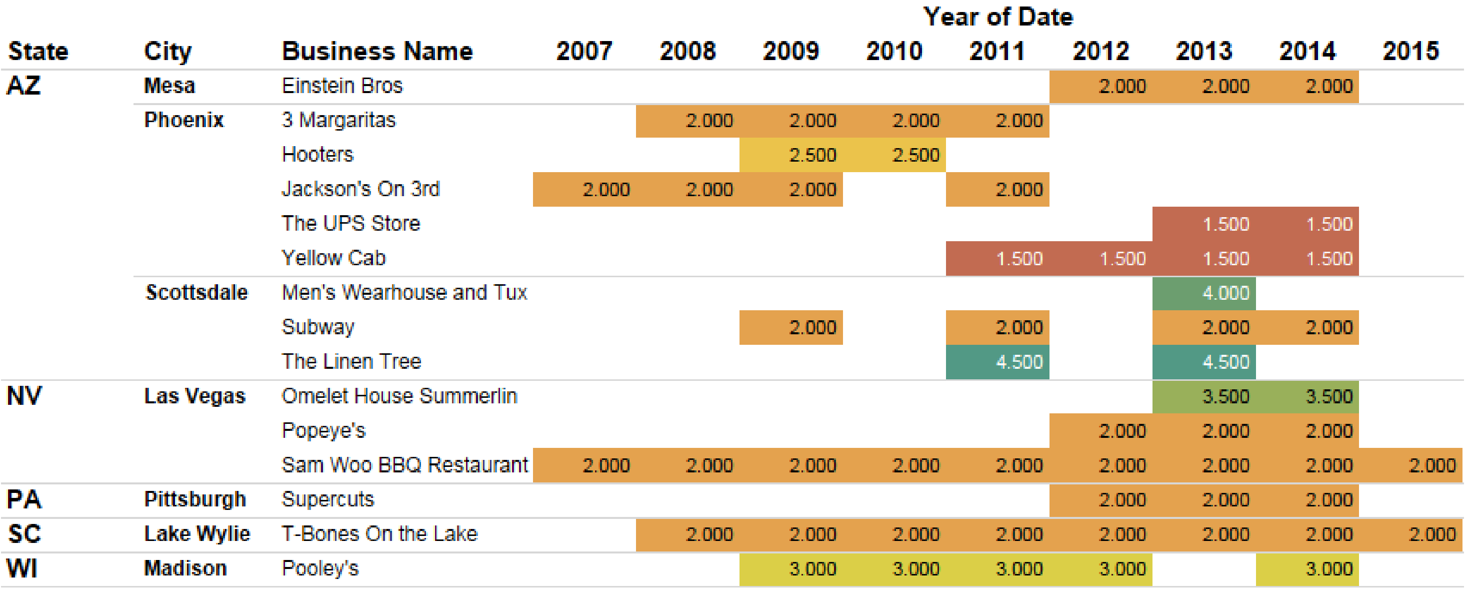
Here too we can see that Arizona is leading in both aspects followed by Nevada.

3.6. State and City Wise Business's User Rating (Top 10):



			Year of Date									
State	City	Business Name	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
AZ	Chandler	Cost Plus World Market							4.000			
	Gilbert	Burke Daniel DVM							5.000		5.000	
	Glendale	Vineyard Christian Fellowship of North..					4.500	4.500				
	Phoenix	Crystal Optical					5.000		5.000	5.000		
		Niccoli's Deli & Pizza		4.500	4.500	4.500	4.500	4.500	4.500	4.500	4.500	
		Precision Imports				5.000	5.000		5.000	5.000	5.000	
		Total Forms Management								5.000	5.000	
		Welcome Diner		4.500	4.500	4.500		4.500	4.500	4.500	4.500	4.500
		Wizard of Odz							5.000	5.000	5.000	
	Scottsdale	Convenient Corner Market					4.500	4.500		4.500	4.500	
NV	Henderson	HoneyBaked		4.000						4.000	4.000	4.000
		Ideal Chiropractic			5.000		5.000			5.000	5.000	
		Outpost Mail & Copy Center								4.500	4.500	
	Las Vegas	A&B Security Group							5.000	5.000	5.000	
		Box Brothers					5.000	5.000	5.000	5.000		
		HoneyBaked					5.000	5.000			5.000	5.000
		Island Style		4.500	4.500	4.500	4.500	4.500	4.500	4.500	4.500	4.500
		Proficient Auto Body							5.000			
		Rincon De Buenos Aires	4.500	4.500	4.500	4.500	4.500	4.500	4.500	4.500	4.500	
		Summerlin Jewelers								4.500		
PA	Pittsburgh	Prestogeorge Fine Foods					4.500				4.500	4.500
		The Inn on Negley							4.500	4.500	4.500	
WI	Middleton	Dunn's Import Inc						5.000			5.000	

3.7. State and City Wise Business's User Rating (Bottom 10):



			Year of Date								
State	City	Business Name	2007	2008	2009	2010	2011	2012	2013	2014	2015
AZ	Mesa	Einstein Bros						2.000	2.000	2.000	
	Phoenix	3 Margaritas		2.000	2.000	2.000	2.000				
		Hooters			2.500	2.500					
		Jackson's On 3rd	2.000	2.000	2.000		2.000				
		The UPS Store							1.500	1.500	
		Yellow Cab					1.500	1.500	1.500	1.500	
	Scottsdale	Men's Wearhouse and Tux							4.000		
		Subway			2.000		2.000		2.000	2.000	
		The Linen Tree					4.500		4.500		
NV	Las Vegas	Omelet House Summerlin							3.500	3.500	
		Popeye's						2.000	2.000	2.000	
		Sam Woo BBQ Restaurant	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
PA	Pittsburgh	Supercuts						2.000	2.000	2.000	
SC	Lake Wylie	T-Bones On the Lake		2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
WI	Madison	Pooley's			3.000	3.000	3.000	3.000		3.000	

4. Conclusion:

- The above analysis shows a lot of things. The word clouds show the general sentimentality of the users being in general **more positive** in their reviews. They prefer to post positive tips and reviews than negative ones. 2015 shows that clearly , while some **ambivalent emotions** existed in the reviews about 2014 reviews, which weren't totally positive but not totally negative too
- Secondly, the State of Arizona (city: Phoenix) were the **most reviewed places**. In terms of the number of businesses and number of users, they clearly stood out. Next to stand out was Nevada.

- Not only was the above true, Arizona's places were the **highly rated places** too. With most of its cities coming up in the **top 10 highly rated places**. Here too **Phoenix** was a clear winner.
- The number of users on the Yelp portal has also increased dramatically over the years, showing that people do like to write tips and reviews about the services for the benefit of other. The general trend shows a comfort with the online portals. Majority of the reviews were from the USA than from some other country.
- Lastly, people tended to post more about **food and restaurants** followed by travel and beauty than about **financial institutions and education.** Here too the topics of food reviewed were **American and sushi** being the most favored cuisine.

5. References:

1. http://www.yelp.com/dataset_challenge/

2. https://georeferenced.wordpress.com/2013/01/15/rwordcloud/