2016

ASHISH

Bharath

Mohamed Salihdeen

Surendar

VIgneshwaran

Data Vizualization – Group 10

Contents

[DATASET: 1](#_Toc468735148)

[INSIGHT 1 – ‘WHERE’ 2](#_Toc468735149)

[INSIGHT 2 – ‘WHAT’ 5](#_Toc468735151)

[INSIGHT 3 – ‘HOW’ 7](#_Toc468735152)

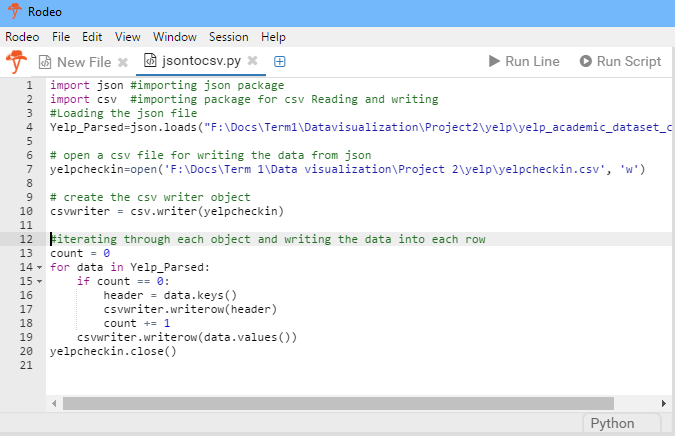
[INSIGHT 4 – ‘WHO’ 11](#_Toc468735153)

00[SUMMARY: 12](#_Toc468735154)

# **DATASET:**

***Link:*** [***https://www.yelp.com/dataset\_challenge***](https://www.yelp.com/dataset_challenge)

* We acquired Yelp database from the Yelp website, which was in “JSON” format.
* We converted these JSON files into CSV files using the below python code

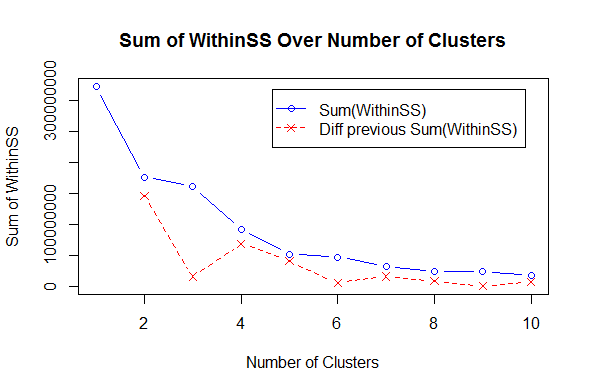


* Since we are only interested in restaurants in USA, we removed business from the dataset that aren’t related to food service business. We did this by creating a restaurants category list and comparing it with business category from the dataset. If the business category matched with our list, the business is kept else it is removed from the dataset.
* We worked on the following:
  + Business: This contained all the information of the restaurant such as:
    - Location
    - Whether it is good for lunch, dinner, breakfast, drinks etc.
    - Ambience
    - Price Range
    - Alcohol information
    - Hours
    - Stars on Yelp
  + Check-in: This contained the check-in data for all the restaurants based on the day and time of the day.
  + User: This contained the reviewer data such as number of followers, how many people voted that reviewer to be useful etc.

# **INSIGHT 1a – ‘WHERE’**

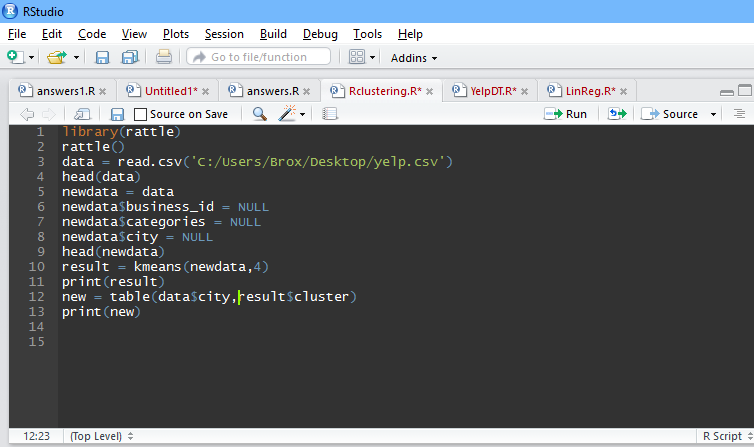
Location based Clustering:

* We decided to use KMeans Clustering to find patterns in location that has the highest rating among the users to decide **where** to open a restaurant.
* We first need to find the optimal number of clusters for which we used the rattle() package in R Studio to iterate with increasing number of cluster values to find the best one which turned to be 4 from the below graph.



Code to Cluster based on location:

* Below is the code to cluster the dataset based on cities and their average ratings.



Clustering results:

* Based on the below results, we can infer that the city that has most restaurants under cluster 3 will have the highest rating. Las Vegas has 86% of restaurants under cluster 3, the one which has the highest rating. Even for the second-best rating, Las Vegas topped the list with 45% of the restaurants in cluster 2.

K-means clustering with 4 clusters of sizes 2180, 1216, 38, 22856

Cluster means:

latitude longitude review\_count stars

1 54.29863 -0.4405512 14.28991 3.758716

2 35.47422 -110.2142404 387.60197 3.833059

3 35.83192 -114.8420091 2156.23684 3.921053

4 36.99211 -100.3371124 32.44146 3.438528

Interpretation:

As highlighted below, Las Vegas has the most restaurants under higher ratings cluster.

Juniper Green 3 0 0 0

Kahnawake 0 0 0 1

Karlsbad 3 0 0 0

Karlsruhe 434 0 0 0

Kirkland 0 0 0 12

Kitchener 0 0 0 122

L'’\_le-Bizard 0 0 0 1

La Prairie 0 0 0 8

Lachine 0 0 0 14

Lake Wylie 0 0 0 4

Las Vegas 0 574 33 4250

Las Vegas 0 0 0 2

Lasalle 0 0 0 29

Lasswade 2 0 0 0

Laval 0 0 0 125

Laveen 0 0 0 33

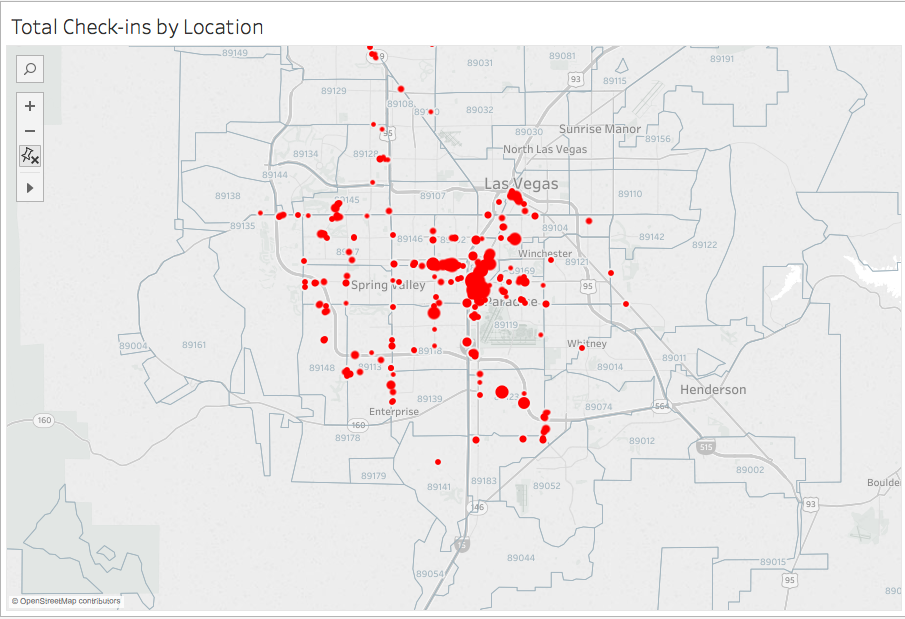
Lawrenceville 0 0 0 2

Leith 2 0 0 0

Litchfield Park 0 0 0 34

# **INSIGHT 1b**

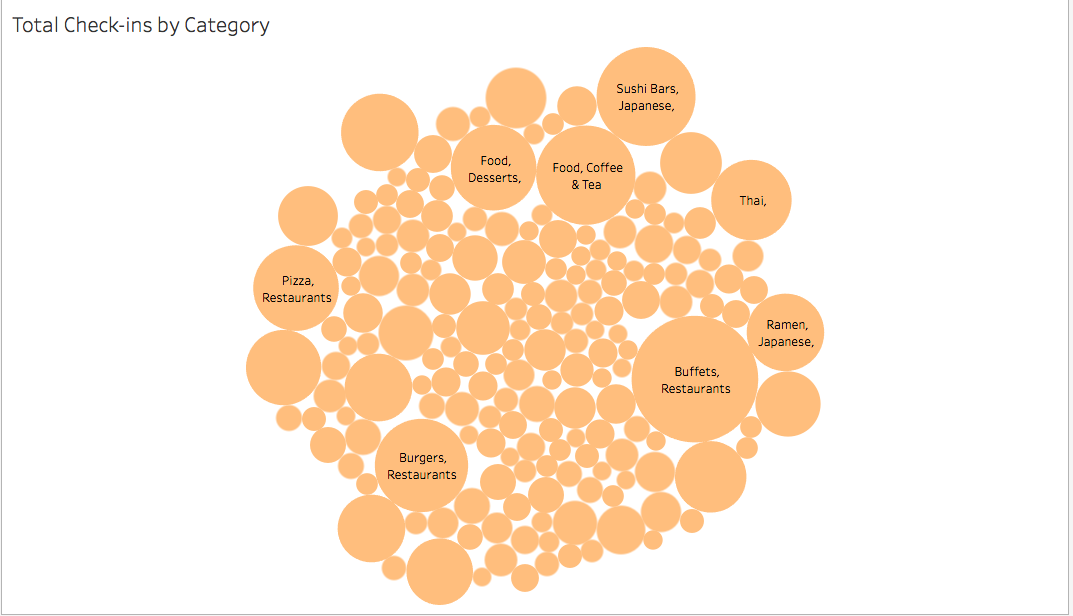
* As explained above, we decided to focus on Las Vegas. This made sense to us as well, because Las Vegas has a lot of influx of tourists and it would have great potential to open a restaurant.
* We decided to focus on check-ins in the city of Las Vegas.
* We wanted to focus on areas with good restaurants and ones that were popular.
* Therefore, we filtered with the Average stars to be at least 3. And minimum number of check-ins to be 1000.
* We created a calculation that calculated the total check-ins of the restaurants after adding the hour wise check-in information.
* The result is the map below.



* Quite visibly there are certain areas that have more check-ins and better rated restaurants.
* In order to find the zip code the map narrowed into, we created a dimension to filter the location with zip code 89109(Zipcode\_In\_Focus). This happens to be the Las Vegas Strip.

# **INSIGHT 2 – ‘WHAT’**

* We then wanted to focus on **what** categories of restaurants do better than the others.
* Again, we used the same filters as mentioned above (Average stars = 3 and Minimum check-ins = 1000)
* The following is the resulting bubble chart.



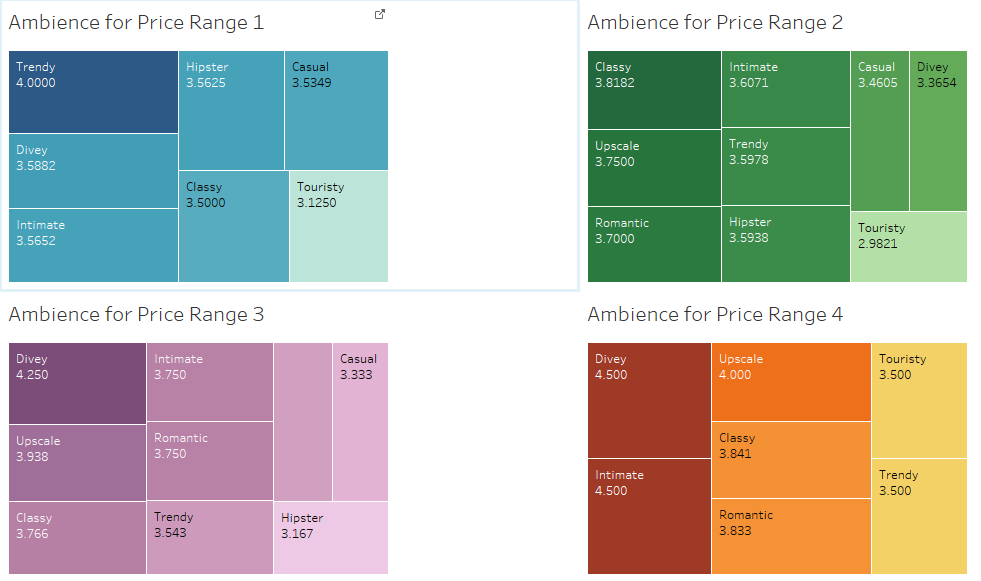
* According to the visualization above, it was clear that Restaurants with category “Buffet” were the most popular and highly rated. Followed by Cafes and Sushi Bars.

# **INSIGHT 3 – ‘HOW’**

* Another important and interesting visualization was to compare different factors that a restaurant in different price ranges need to be successful. This explains **how** restaurant can be successful.
* Therefore, we visualized these factors according to the Average stars they received.
* We filtered those restaurants that did not have this information available.

**Ambience:**

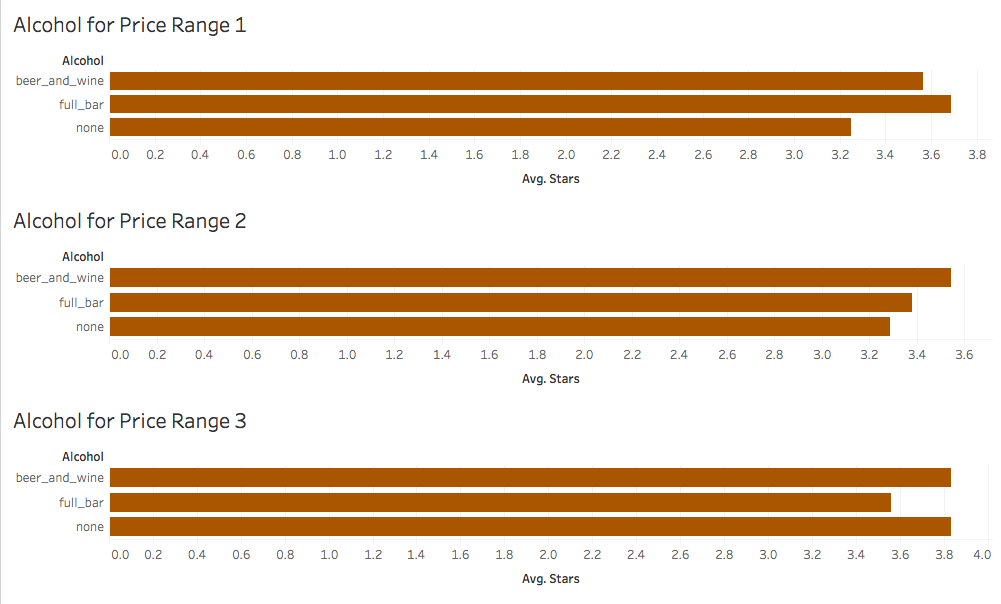
* The first factor was Ambience.
* There were 9 different attributes defining ambience, hence we decided to combine it to one field known as “Ambience”

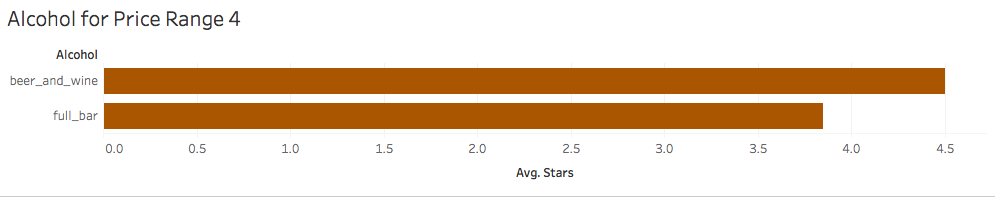


* The Ambience that did best with cheaper restaurants were “Trendy” (Price Range 1) and “Classy” (Price Range 2).
* The Ambience for Price Range 3 and 4 that did best was “Upscale” and “Divey”. This result was contradicting and hence we explored the dataset. We found Price Range 3 and 4 had 2 and 1 restaurant data respectively which was too small to recommend to our client.

**Alcohol:**

* + The second factor we compared was Alcohol

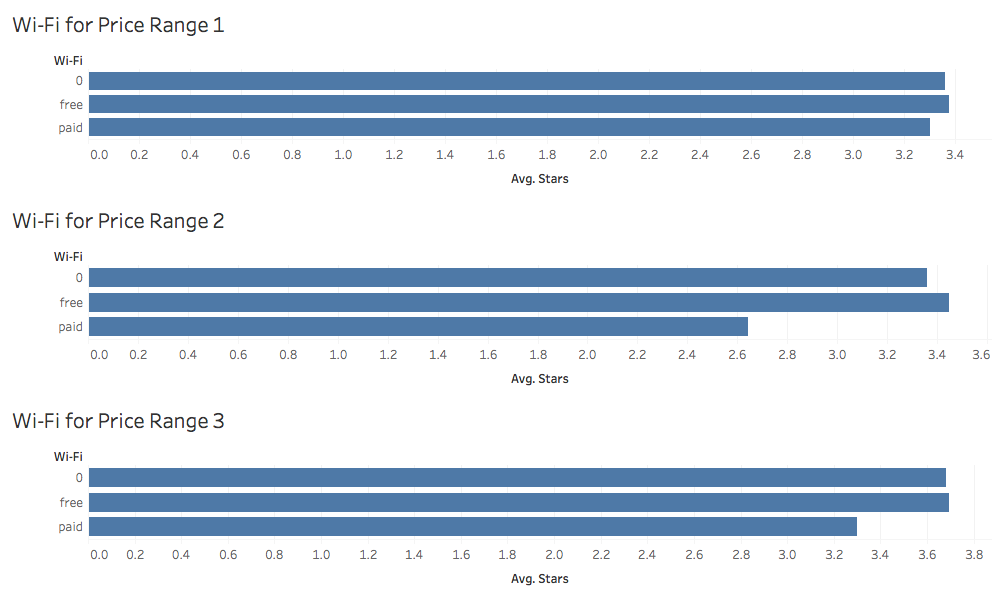


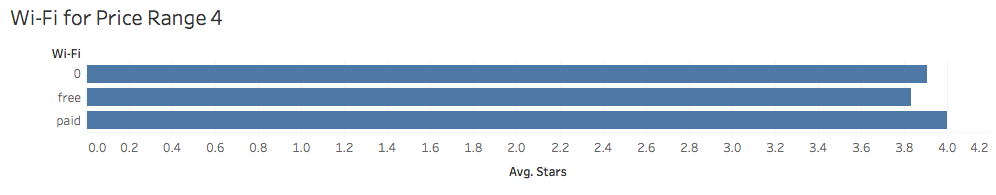


* The findings of this visualization suggested that for cheaper restaurants had better ratings with a Full bar (Price range 1) or Beer and Wine (Price range 2). Price range 3 had equally good ratings for No alcohol and Beer and Wine. While, If it is very expensive restaurant, it was expected to serve alcohol, and just beer and wine restaurants did better.

**Wi-Fi:**

* The third factor we compared was Wi-Fi

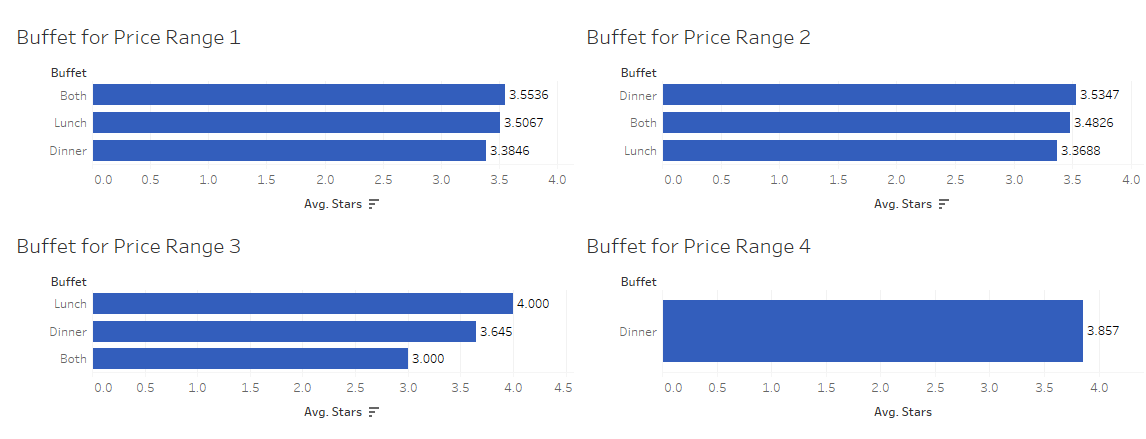




* It was clear that restaurants with no or free Wi-Fi performed better up to price range 3. Which means, it was better to have no Wi-Fi at all as compared to Paid Wi-Fi.
* But, for the really expensive restaurants, Paid Wi-Fi did better than the other two. Which meant if it was a really expensive place, people did not mind paying for Wi-Fi.

**Buffet:**

* The fourth and final factor was Buffet.
* We had two fields Good\_for\_lunch and Good\_for\_dinner. We decided to combine it to one field called as ‘Buffet’
* We wanted to explore when restaurants did better if there was buffet going on in the restaurant. Like better for lunch or dinner.



* Restaurants with Price range one performed best if they served both Lunch and dinner Buffets. Price Range 3 restaurants did best for Lunch buffets. And very expensive restaurants served buffet only for dinner.

Regression based on the above 4 attributes:

* Regression estimates the variable worth and when ran we came up with the results below.

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.12077 0.06731 46.362 < 2e-16 \*\*\*

Alcoholbeer\_and\_wine 0.10062 0.08011 1.256 0.209247

Alcoholfull\_bar 0.02570 0.07656 0.336 0.737130

Alcoholnone -0.03504 0.07280 -0.481 0.630346

Wifi0 -0.04516 0.04471 -1.010 0.312542

Wififree 0.04977 0.04913 1.013 0.311136

Wifipaid -0.53566 0.11221 -4.774 1.91e-06 \*\*\*

lunch 0.10735 0.02738 3.921 9.05e-05 \*\*\*

dinner 0.12124 0.03058 3.964 7.55e-05 \*\*\*

upscale 0.37265 0.09691 3.845 0.000123 \*\*\*

touristy -0.34463 0.10214 -3.374 0.000751 \*\*\*

intimate 0.32098 0.09748 3.293 0.001005 \*\*

hipster 0.40004 0.12727 3.143 0.001689 \*\*

casual 0.26388 0.03171 8.321 < 2e-16 \*\*\*

trendy 0.35201 0.07222 4.874 1.15e-06 \*\*\*

classy 0.55911 0.06764 8.265 < 2e-16 \*\*\*

divey 0.30830 0.06363 4.845 1.34e-06 \*\*\*

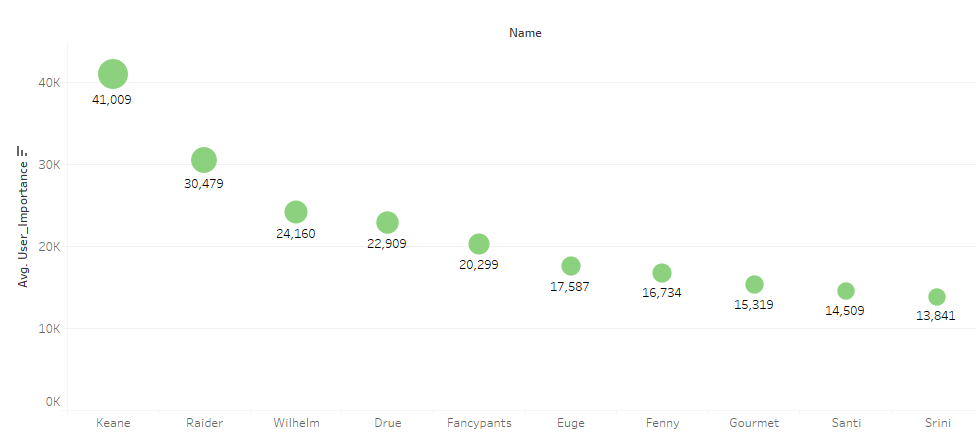
romantic 0.38484 0.10202 3.772 0.000165 \*\*\*

* Classy ambience restaurants did better by a factor of 0.55 (Price Range 1) and upscale ambience restaurants did better by a factor of 0.37 (Price Range 3)
* Beer and Wine did better by a factor of 0.10 compared to full bar and no alcohol restaurants. (Price Range 2 to 4)
* Restaurant which provides free Wi-Fi did better by a factor of 0.049 than paid and no Wi-Fi. (Price Range 1 to 3)
* Both Lunch and Dinner did better by a factor of 0.10 and 0.12, hence having any one would increase the average rating of the restaurant.

The above shows that regression results align with our insights drawn from the visualization.

# **INSIGHT 4 – ‘WHO’**

* We then wanted to determine **who** would be best reviewers to get restaurants reviewed by such as to create the most buzz.
* We ran Regression using R on the various factors that cause the user to be important.
* Using this regression result we created a calculated field for user importance attaching a weight to the factors:
  + Number of Fans
  + Number of times reviewer was voted useful
  + Number of times reviewer was voted funny
  + Number of times reviewer was voted cool
* We then created a set that included only the 10 reviewers based on their User importance.



* Therefore, we got a list of the top 10 reviewers that should be contacted and invited to review a restaurant.

# **SUMMARY:**

* + From our analysis of the Yelp dataset, we have gleaned certain valuable insights which could explain how a restaurant could be profited if it has certain characteristics explained in the above report.
  + We have explained the following key indicators:

1. ‘Where’ – Where the restaurant can be opened(Location)
2. ‘What’ - What category of restaurant bring in more revenue(Category)
3. ‘How’ – How the ratings could be improved (Characteristics and features)
4. ‘Who’ – Among the reviewers, who are more influential. (Social traction)