YELP TSD

David Djernaes
Tatiana Luchian
Shruti Nuchhi

TABLE OF CONTENTS

1. ABOUT THE PROJECT
2. PROJECT DESCRIPTION & OBJECTIVES
3. PROJECT PLAN
4. WHAT IS IN THE YELP DATASET?
   (Hint: not only reviews and not all of them are for restaurants!)
   - ❖ Data composition and quality
   - ❖ Data composition by category of business
   - ❖ Data composition by location
   - ❖ Data composition by count of restaurants state and city
   - ❖ Data composition by count of restaurant reviews
5. INTERPRETING THE DATA BY GEOLOCATION
6. INTERPRETING THE DATA BY CATEGORY
7. INSIGHTS FROM NATURAL LANGUAGE PROCESSING (NLP)
8. INTERPRETING USER/REVIEW BEHAVIOR
   - ❖ Searching for trends and insights from group behavior
9. CONCLUSION

ABOUT THE PROJECT

In today's world, the success of every business is decided on customers' reactions. Data Analytics can help businesses identify potential opportunities to streamline operations or maximize their profits. It helps identify potential problems, eliminating the process of waiting for them to occur and then take action on the same. Hence, we decided to choose a topic that would help businesses decide on what factors should be considered and how customer reaction can boost or kill their business.

We plan to use the Yelp Dataset and provide insights into the data that would help us in deciding the success of a restaurant or any other business that is on Yelp. The Yelp dataset contains various data pointers such as star ratings, text review, user-check in, tips by users and details about the users who review such as the number of comments, location, active since etc. Using these features provided by Yelp, we plan to analyze various factors such as user reviews, star ratings, geo locations etc. via Tableau, R, Pandas and Seaborn in order to gain useful insight in how business are rated.

PROJECT DESCRIPTION & OBJECTIVES

01 DESCRIPTION

This project exploring via visual tools like Tableau and analytics tools like Python Pandas and Seaborn from which various data points from Yelp in order to understand better how various criteria (location, type of business) affect business and its reviews.

02 OBJECTIVES

Gain insight into business review and its effects on business by analyzing various factors such as user reviews, star ratings, geo locations etc. via Tableau, R, Pandas and Seaborn.

03 DATA

Yelp is a which is a business directory service and publishes crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations. The company also trains small businesses in how to respond to reviews, hosts social events for reviewers, and provides data about businesses, including health inspection scores. YELP dataset challenge publicly available data was used in this project.The dataset contains information on certain businesses, their location, reviews and users that rate businesses.

PROJECT PLAN

- Clean and prep data (required Python script to identify all invalid entores (e.g., 0 representing null)
- Transfer Yelp datasets to BigQuery & Connect Tableau Desktop to BigQuery
- Translate Project Objective into specific analytics objectives
- Repeat the previous step with at least 5 other Project Objectives until you just randomly pick any, for sake of it (just few hours before the deadline)
- Translate analytics goals into SQL queries and visualization charts in Tableau
- Spend the same amount of time that it took the team to complete step 1 - 4, to find a design for the final presentation (but the efforts were worth it!)
- Consolidate the project into a final presentation while sipping Quarantinies!

Quarantini

01

WHAT IS IN YELP DATASET?

Getting to know what we have been served

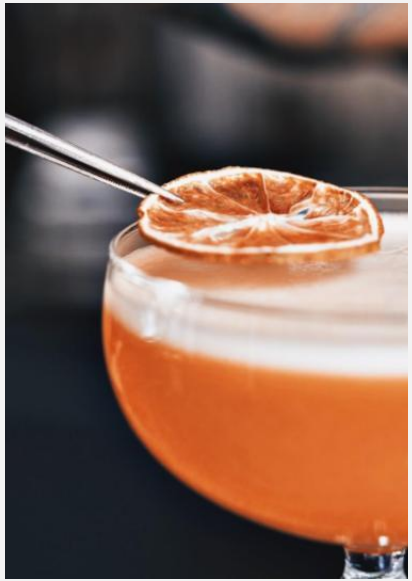
DATA: COMPOSITION & CHALLENGES

Size

YELP dataset is 8GB, it consists of 6 relations: user, businesses review, checkin, tip, photo.

Quality

Inconsistent tagging and categorization of data (e.g., restaurant category has 1615 different names), also some categories use boolean 0 instead of NULL to indicate that there is not data for a particular data point.



Limited Data

Despite large volume of data, it actually gives data points on relatively small number of geo locations and categories of businesses.

Tools Used

Python (Pandas) was used in order to clean us the data and perform analytics on the reviews
BigQuery was used as a shared analytical group tool
Tableau and Seaborn were used for gaining visual insight that sometimes were not very clear from just querying the data (e.g., limited data location is very evident of map graph).
Seaborn for gaining visualisations of the analytics done using pandas

Data Composition

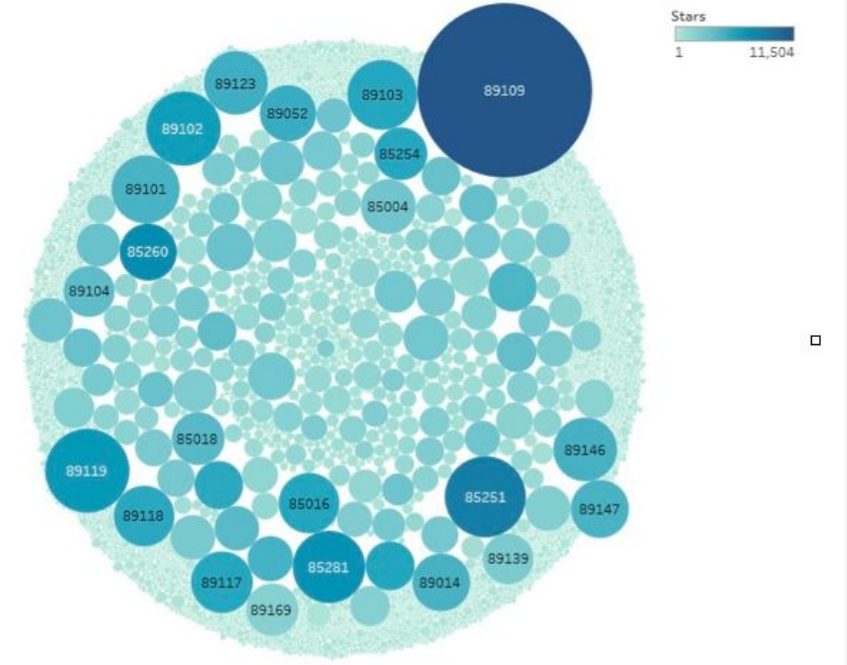
By Category

DATA COMPOSITION BY GEO LOCATION

01 Visual presentation of zip codes in YELP reviews from the dataset zip codes of all reviews.

Pros: Mesmerizing chart, a piece of art, harmonious color scheme, and look at those bubbles… tons of them!

Cons: Missing a unicorn, and user need to memorize all zip codes and their location in order to understand it (so unicorn would certainly help!)

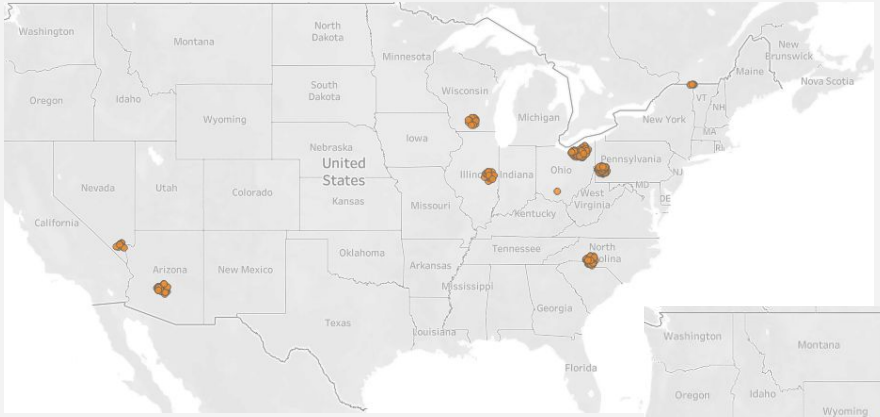


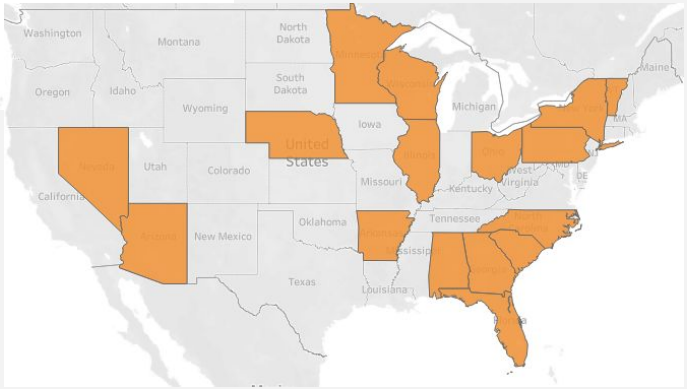
DATA COMPOSITION
BY GEO LOCATION

02 Visual presentation of locations of businesses (by city upper chart and by state) in YELP dataset.

Pros: Very insightful graph that clearly outlines that 8GB of YELP review data just is related to a few cities.

Cons: In dire need of any unicorn (or anything to make it exciting!), not as breathtaking as charts from the previous slide



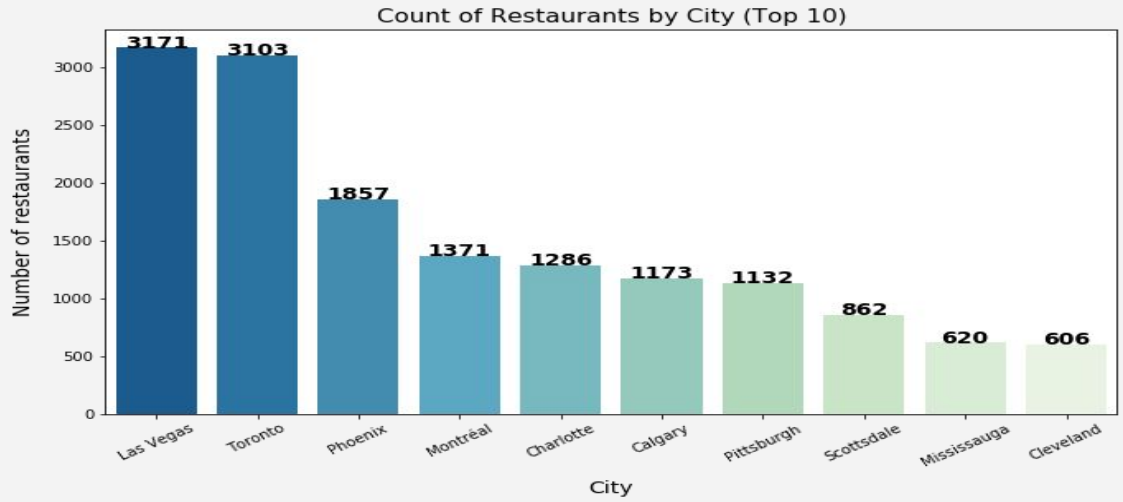


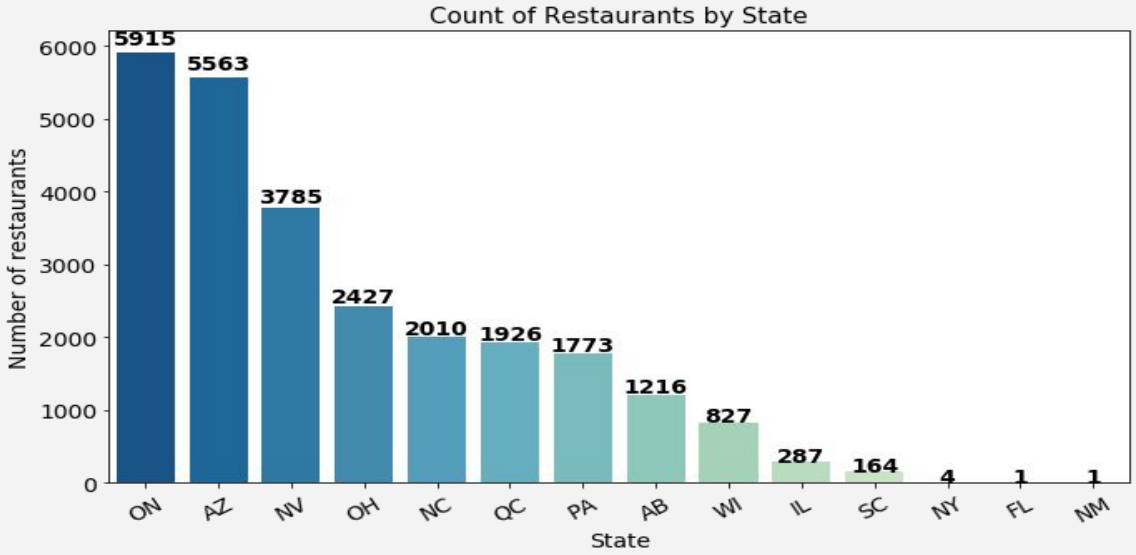
DATA COMPOSITION FOR RESTAURANTS

03 Visual presentation of locations of Restaurants (by city upper chart and by state) in YELP dataset.

Pros: Very insightful graph that clearly outlines that 8GB of YELP review data just is related to a few major cities which are generally expected.

Cons: Does not contain data for many cities ideally



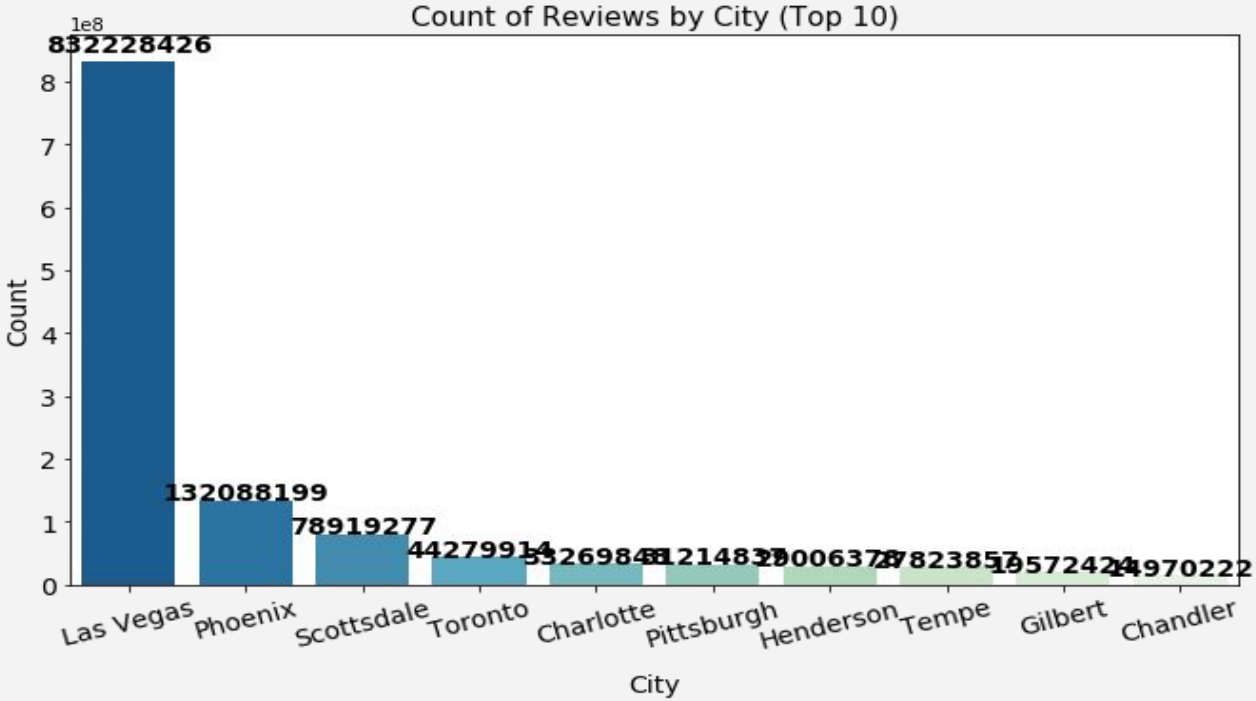


DATA COMPOSITION FOR RESTAURANTS

04 Visual presentation of number of reviews by city in the YELP dataset

Pros: Clearly indicates the city with maximum number of reviews and can quickly see the overall weighting in velocity of total rating per city.

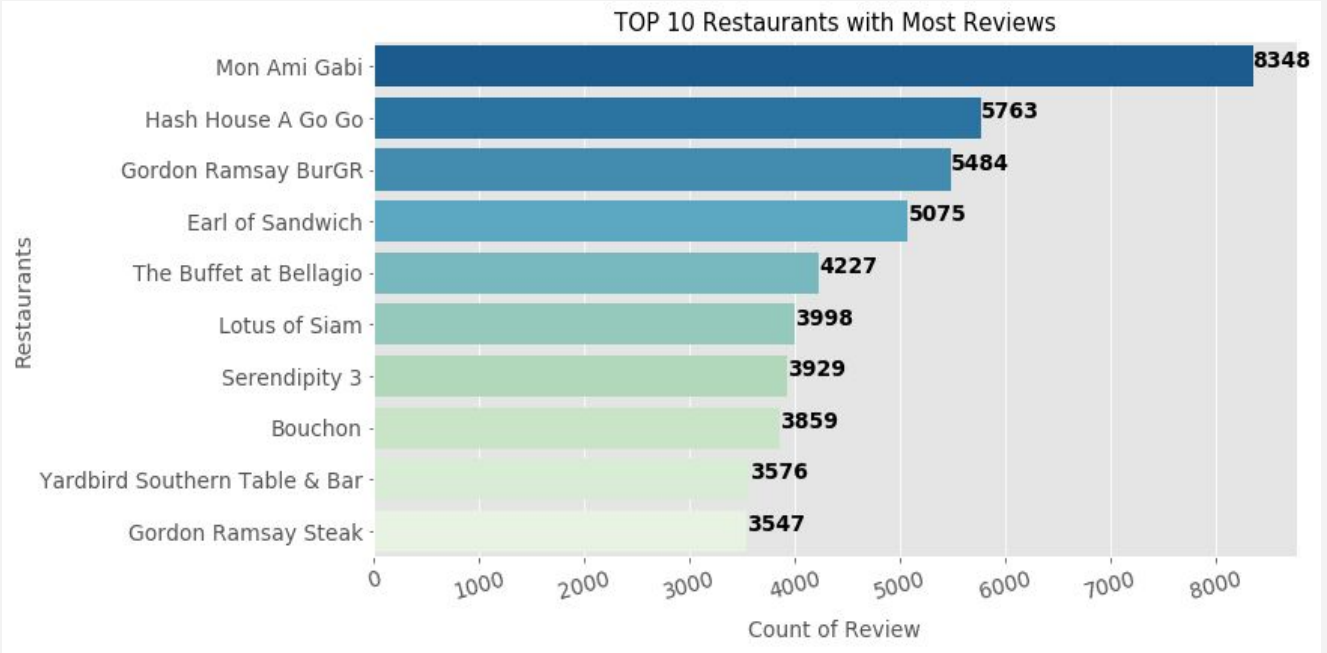
Cons:The seems to mainly focus on Vegas, and does not have good data for other major cities such as New York, Boston, Chicago or other major hubs.



DATA COMPOSITION FOR RESTAURANTS

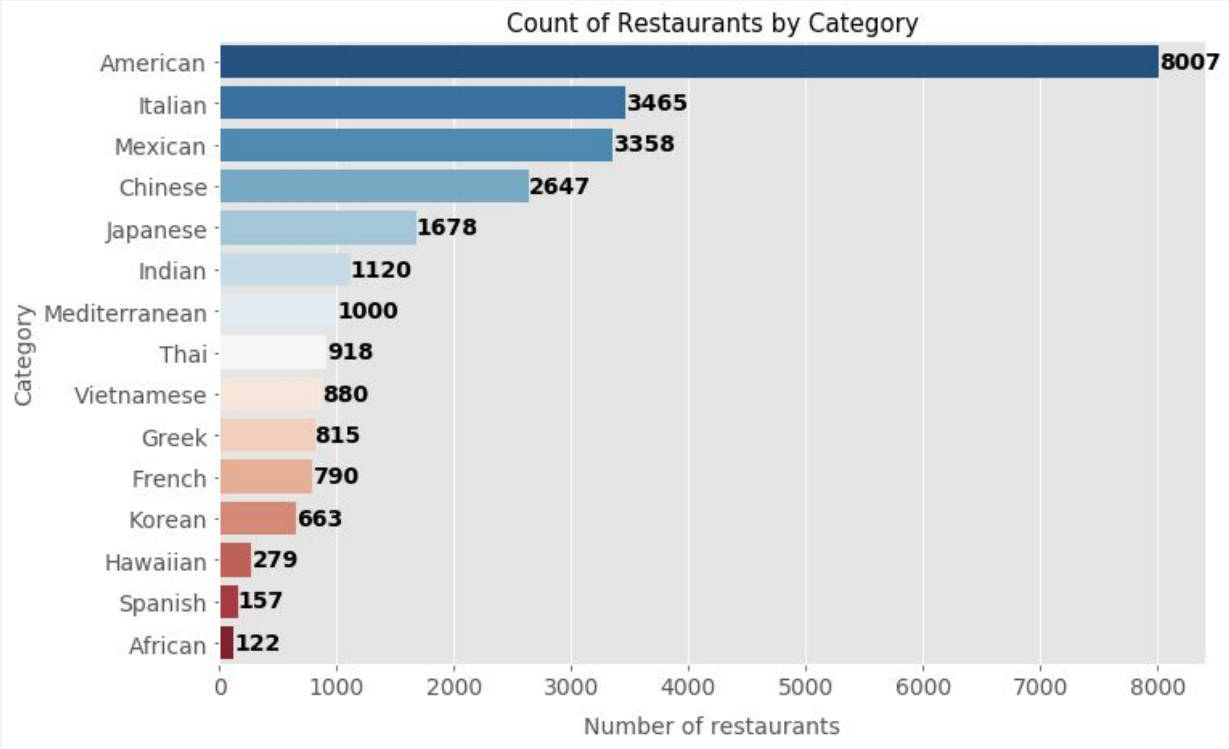
05 Visual presentation of number of top 10 restaurants with maximum reviews in the YELP dataset

Pros: Clearly indicates the top 10 restaurants that with most reviews which is clearly Vegas.
Cons: The outcome of looking at the dataset in this way is that it illustrates only that the city with the highest review velocity will naturally have restaurants with the highest review velocity.



DATA COMPOSITION FOR RESTAURANTS

06 The table to the right shows the top 15 types of restaurants in the dataset. From looking at the data in this way it is a snapshot into the general taste preferences of the consumers. While *American* is listed as the top category, it is worth noting that this is truly a broad term with American cuisine being strongly influenced by regional appetites and the specific flavors unique to that area.

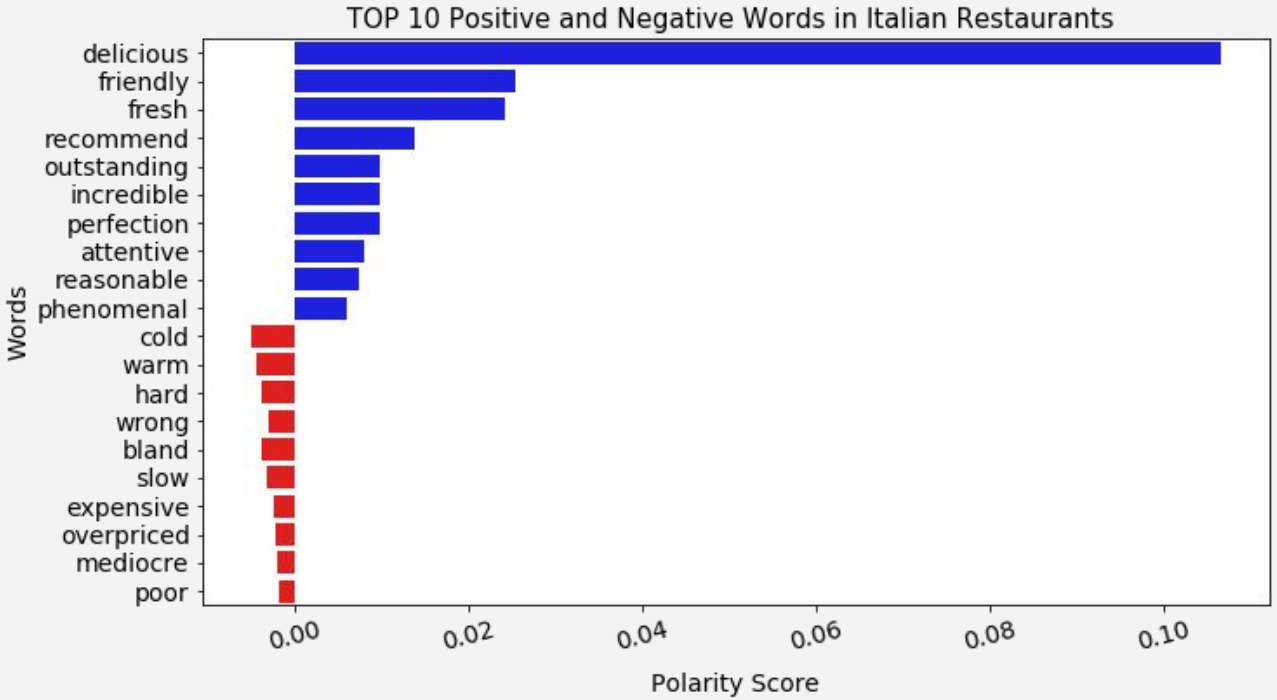


WORD TREND
TOP 5 CUISINES

WORD TREND

ITALIAN CUISINE

07 Once the top food categories were created (American, Japanese, etc.), we were able to assess the language of user reviews. Removal of common stop words and punctuations cleared the way to search for positive or negative words to calculate the overall polarity for a word grouped within the cuisines and trained the model using linear SVC Support Vector Machine (SVM).

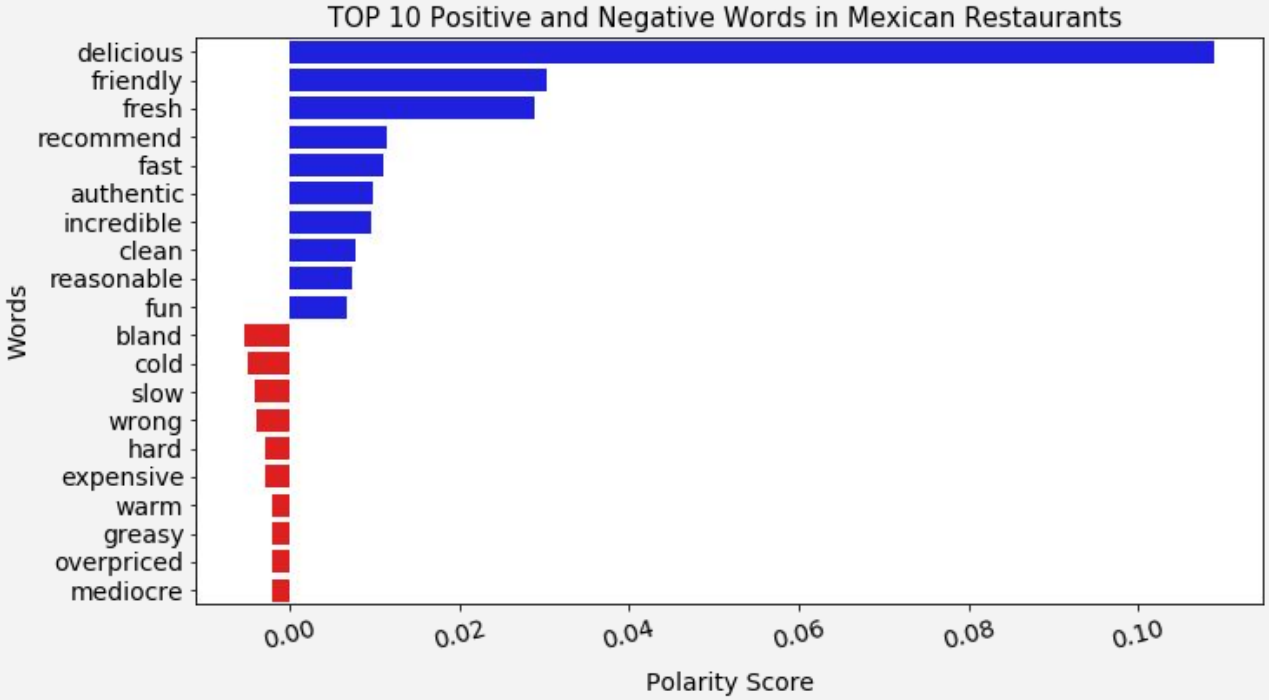


WORD TREND

MEXICAN CUISINE

08 Once the model was created, we filtered the top 10 words with highest polarity and bottom 10 words with lowest polarity for each of the top 5 cuisine types to determine what people in general like or dislike in particular for each cuisine types.

For the purposes of this project we focused on the top 5 after American cuisine.

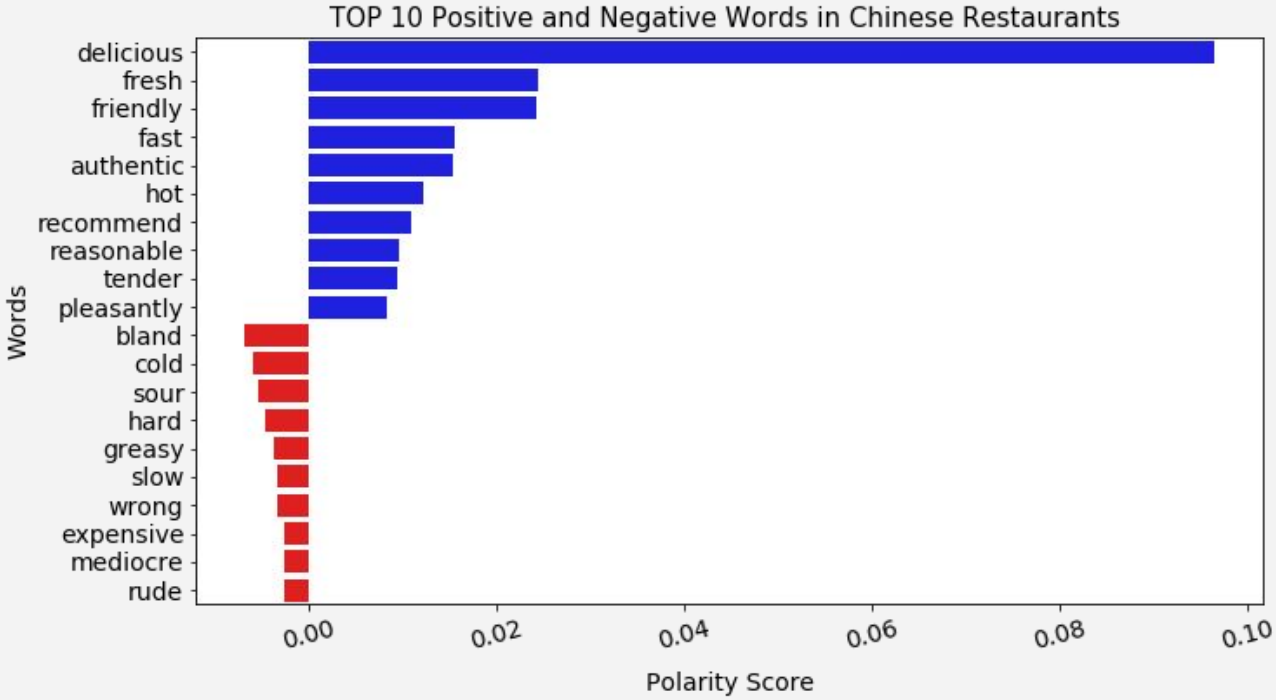


WORD TREND

CHINESE CUISINE

09 Words excluded from the tables are common words like 'good', 'bad', 'love', 'hate', 'wonderful', 'happy', 'impressed', 'disappointed', etc.

Regarding the word distribution in this way, there is a clear correlation between considering food delicious, fresh and authentic and giving the restaurant a high star rating. The most common words to show dislike of a restaurant were bland ,cold, expensive, hard and greasy.

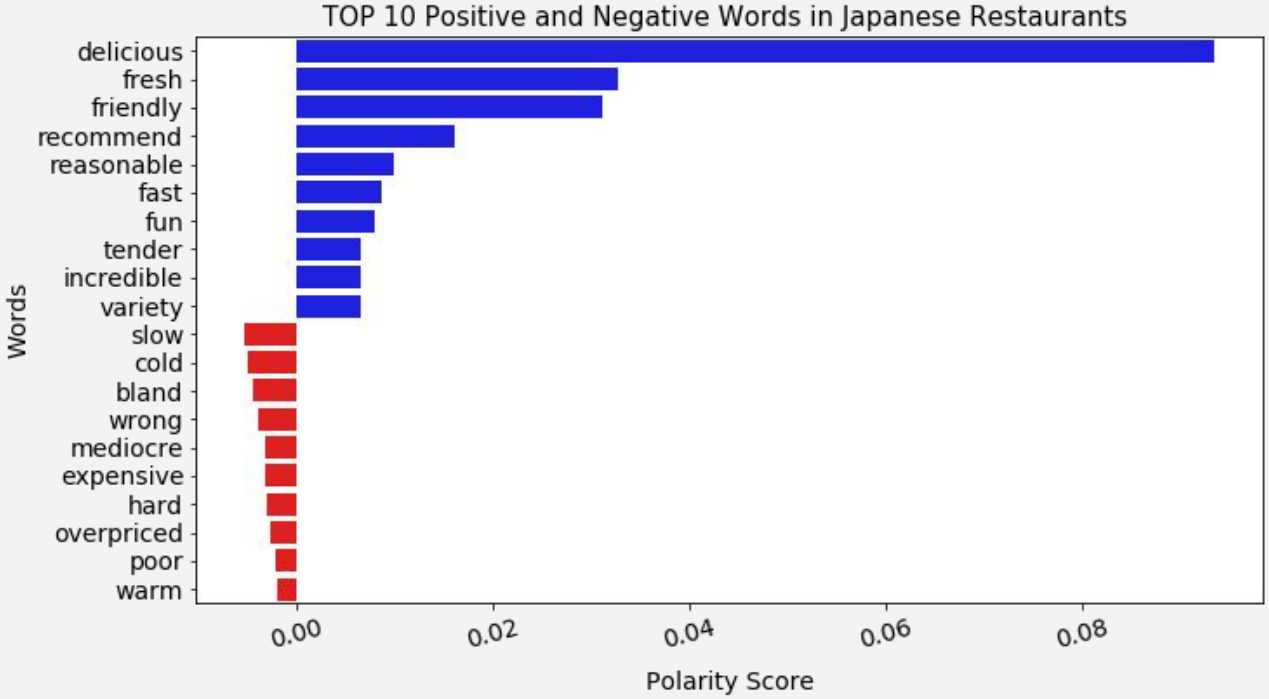


WORD TREND

JAPANESE CUISINE

10 Also, it is interesting to note that people consider a friendly staff and environment to be one of the top indicators of their experience and will give positive reviews for this non-food related parameter of their dining experience.

For most cuisine types, the word friendly ranks higher than the word reasonable, which shows that friendly service is more likely to be the reason for the high score rather than reasonable price.

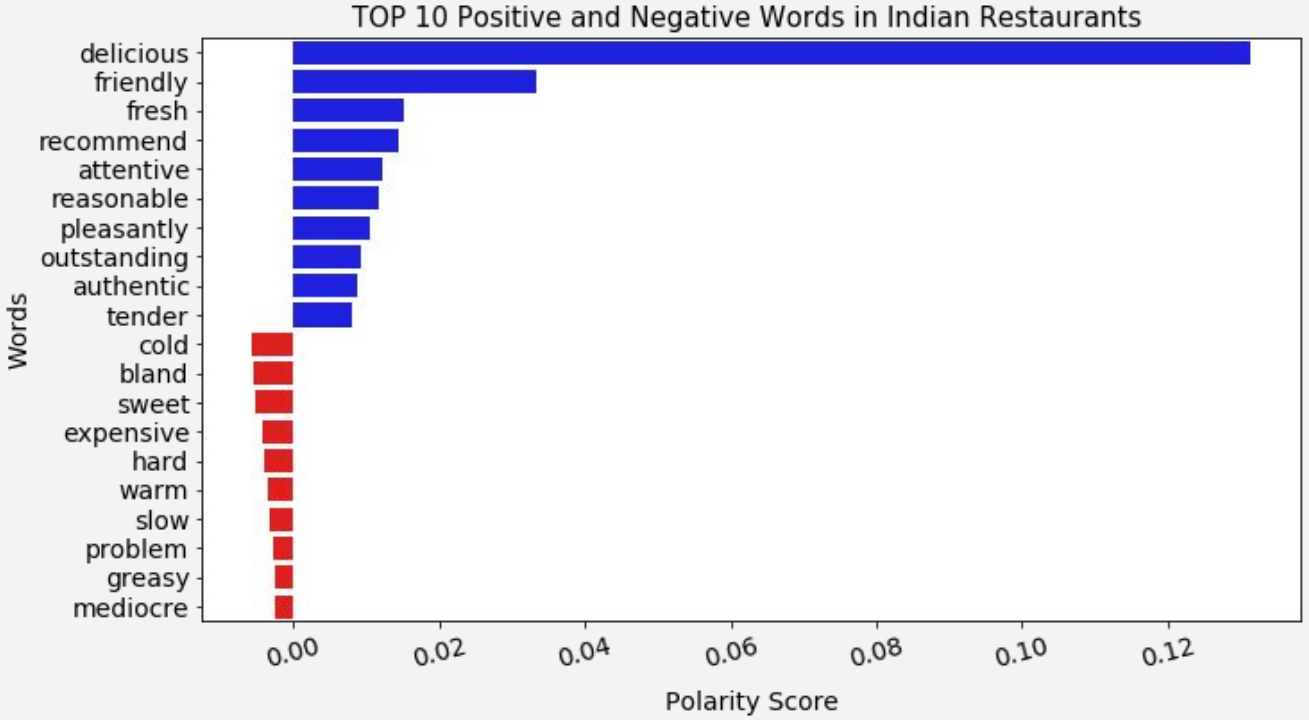


WORD TREND

INDIAN CUISINE

11 As a final thought regarding the language used by the reviewers, it is striking that the top ten positive words are nearly identical in all the reviews, with the top three being *delicious, fresh,* and *friendly*.

As far as the most off-putting qualities in a restaurant are bland flavors, cold food or a cold environment and customers not finding a good value for the money spent.




“Having a successful restaurant
is about creating community.
Without community, it’s just eating.”
—Helen Rosner, food writer, *The New Yorker*

REVIEWER DETAILS

01

The community of Yelp reviewers is very egalitarian in that it is comprised by anyone with an inclination to post an opinion or experience.

02

The reviewers considered for the following slides are all very active in the community having contributed at least 500 reviews with some as high as 12,000 each.

03

Thus, their continued dedication in contributing to the platform was considered as a unit of measure for the the graphics.

WHAT CAN BE MEASURED

The following tables illustrate the partitions for how specific groups of users tended toward a limited level of apparent satisfaction as indicated by the average number of stars they gave.

TRENDS IN MEASUREMENT

The insights that come from reviewing data trends can help illustrate some larger truths at play within the group. Essentially, those users that are more engaged tend toward higher averages.


REVIEWER TRENDS

REVIEWER BREAKDOWN

Grouping by Review Segmentation

Reviewers

Unique users were partitioned by how many reviews they contributed.

Partitions

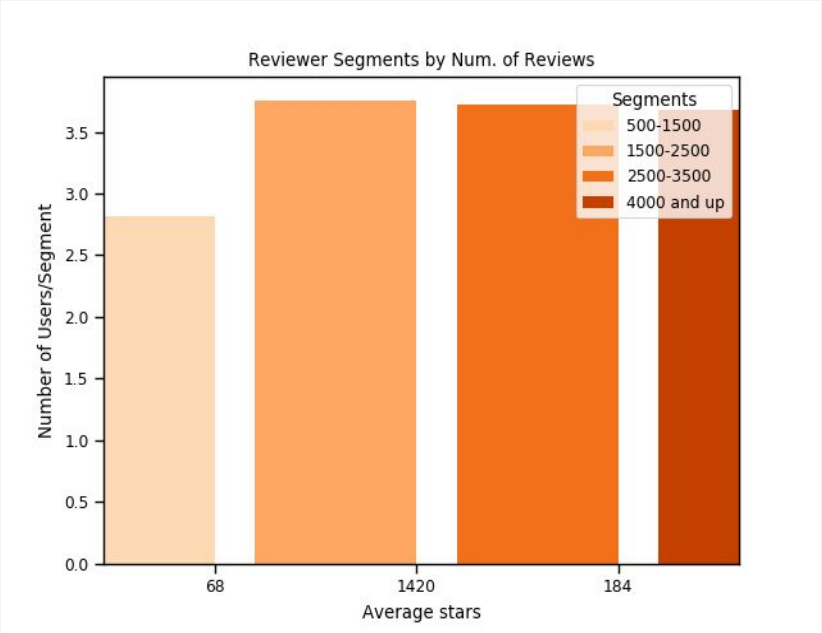
The low end of the group were those that provided 500 to 1500 reviews, with approx. 68 reviewers.

The largest group

Comprised of approx. 1420 users who had written between 1500 and 2500 reviews.

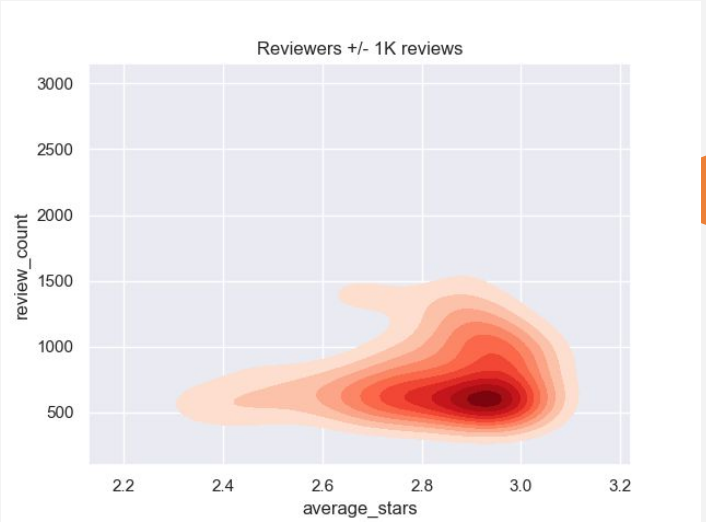
The final groups

This was comprised of approx. 230 users who had written between 2500 reviews on up to 12,000.



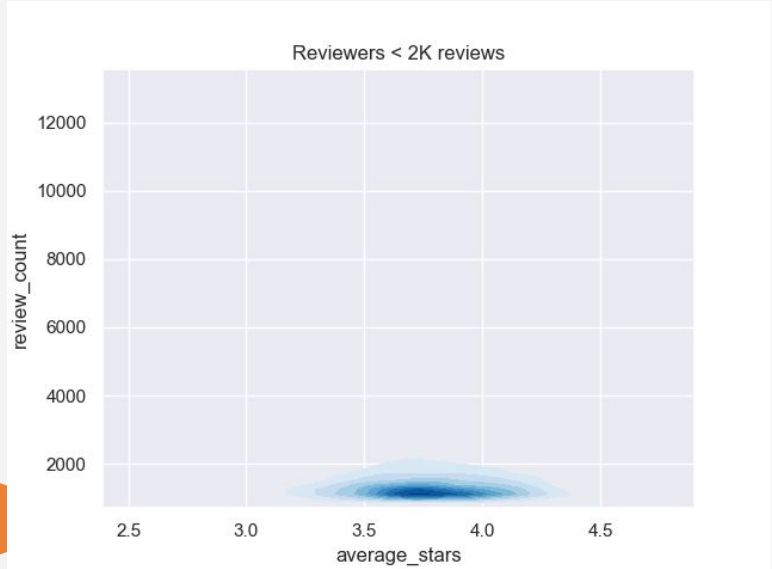
The graph illustrates a leveling off of average stars for the most prolific reviewers to be approx. 3.75 stars and comprise 6331 unique users.

SPOTTING TRENDS



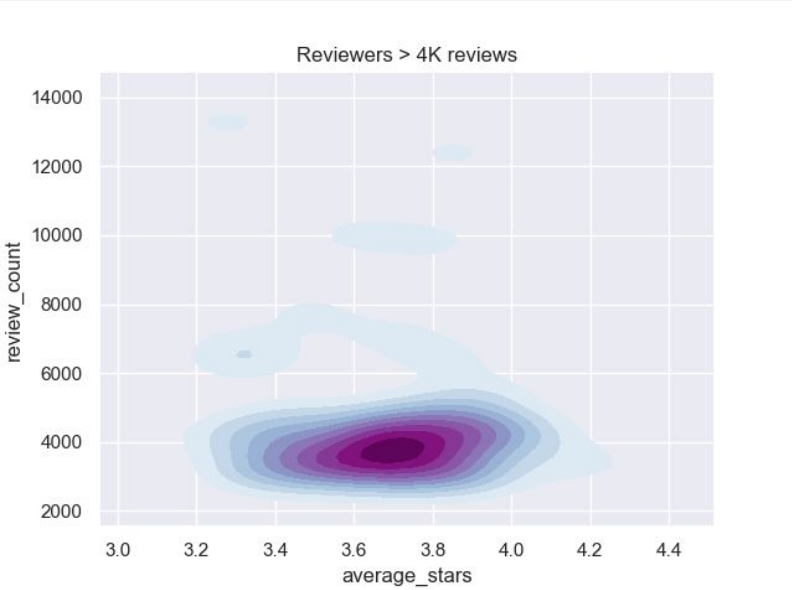
The table to the left shows the density of average stars given by users who left between 500 and 1500 reviews. Note, the highest density is approx. 2.9 stars and varies from 2.3 to 3.1 stars.

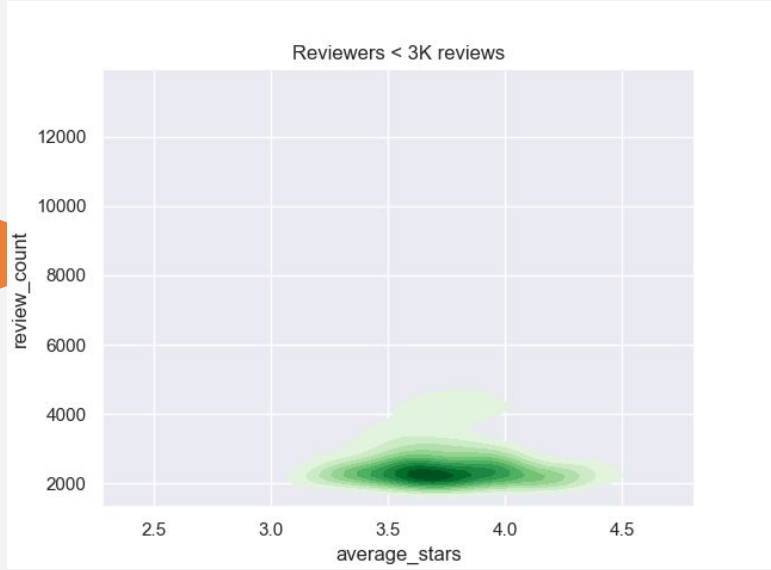
The table to the right illustrates an upward shift and even a coalescing between average stars given by users who left between 1500 and 2500 reviews. Note, the highest density is approx. 3.75 stars.



BECOMING DEVOTED TO SUCCESS?

The upward trend appears to level off at approx. 3.75 stars, although there is a significant proportion in the > 4-star range



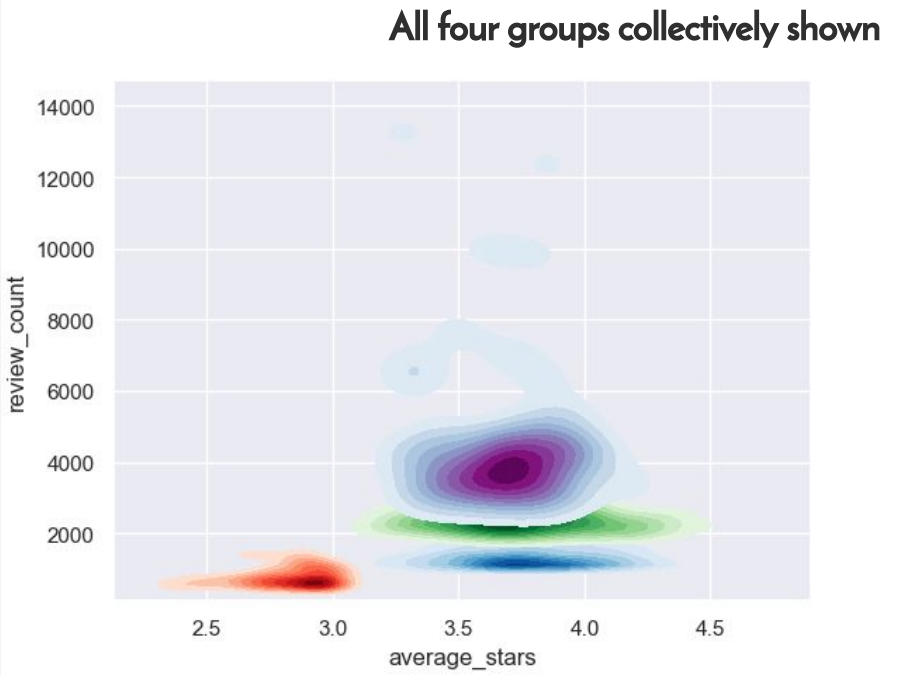


While in the > 4K reviews group a bit of a stronger coalescence is seen by a tighter spread of average stars with the highest density being between 3.4 and 4.0 stars. .

BECOMING LESS CRITICAL?

Group Tendencies

As noted in the previous slides, and as shown to the right, there is a distinct increase in the average number of stars given by the highest reviewing group when compared to lowest level. To be sure, any user providing 500+ reviews to the site could be considered a 'power user,' but this measurement seems to indicate that the more users invest their time in the platform by writing reviews, then the more likely they are to either a) only review restaurants and businesses they like and are personally invested in the success of that business or b) have become less critical of the businesses they review, in general.



Key learnings

Conclusions and Takeaways

The prevailing takeaway from working on this project was learning that the analyst must have a deep understanding of what information the data holds before one begins considering any conclusions drawn about that data. It is somewhat counterintuitive, since this type of project analysis is often treated as if it were a mathematical proof where you start with a conclusion and either prove or disprove the conclusion. However, when dealing with raw data, this can oftentimes lead the analyst down many rabbit trails without any useful insights. Thus, a thorough reading of the data and the connections that one can make from the disparate sources must be understood from the beginning of the project.

Key learnings

Conclusions and Takeaways

The takeaways from the data itself were interesting in that there is such a strong narrowing of opinions in what makes a restaurant good. From the language processing we discovered that the top 3 words people used to describe a positive experience were fresh, delicious and friendly. These words transcended the type of cuisine and speak mainly to the experience of the customer.

While turning the lens onto the reviewers themselves, we discovered that for the most part, a vast majority of reviewers did not leave bad reviews, they mostly only took time out of their day to post a positive review or a somewhat neutral review. And to that point, the reviewers who were most engaged in the platform only tended to leave an average of 3.75 stars, which is approximately 1 star better than those with only a few reviews to their name.

THANKS

Thank you Dr. Tufte, all the guest presenters and to the entire Data Analytics cohort.
The input and contributions provided by the whole group provided helpful insights and key learnings throughout the quarter.