|  |
| --- |
| Revealing mysteries behind Yelp ratings |
| MSBA 6310 (Summer Term) |
|  |
| Through the use of Python programming, our group analyzed the ratings data on Yelp.com and found the seasonal rating trends for different business categories and also the current rating trend for selected businesses |
|  |

**Executive Summary**

We analyzed the ratings data on Yelp.com and found seasonal rating trends for different business categories and also the current rating trend for selected businesses.

We found that summer has the highest number of ratings, followed by autumn and spring for most businesses. Surprisingly, we noticed a difference in this trend for “restaurants” and “art and entertainment” businesses categories where winters had a higher number of ratings than in spring.

We found the top 10 controversial stores and this could reveal that either customers are highly satisfied with their service or are extremely disappointed. Another interpretation is that these 5 star ratings are given artificially, to boost flow of customers to the store. Many of these controversial restaurants are in a specific geographic location. Yes, it is Vegas!

We identified businesses based on trends in user rating. Positive trend signifies that the business has constantly improved its performance while negative trend indicates that the business decreased its performance. Read on to dive deeper with us into our analysis of the Yelp ratings data.

**Introduction**

Yelp, an American multinational company, is a popular crowd-sourced local business review and social networking website. It helps customers find businesses based on customers location. In addition to this, Yelp takes ratings (on a 1-5 scale) and text reviews, from its users. Yelp ratings are often considered a good metric to identify the reputation of a business. **[5]**

We found a particular interest in this topic as Yelp is a popular website/app that is used a lot in the consumer market today. The reviews given by users help the business improve its standards and also other users immensely benefit from the ratings and reviews. Our other motivation for selecting this topic was to study and understand the popularity of a business based on its reviews. There are a multitude of questions and analysis that can be derived from this dataset. Extensive research has been done on the review text data available on the Yelp dataset. However, we chose to study the ratings of a business based on the number of stars awarded to the business by the users. From the plethora of questions that can be constructed from the ratings and reviews, we restricted our analysis to studying trends in the businesses.

Our data was collected from the Yelp Dataset Challenge and contains 1.6 million reviews on 61,000 businesses given by approx. 366,000 users. The rest of this paper focuses on:

* [Related Work:](#_Related_Work) A brief summary on other people’s work
* [Process](#_Process): How we arrived at our results
* [Results:](#_Results) Findings on seasonal ratings, controversial stores and trends in user ratings
* [Conclusion:](#_Conclusion) Conclusions drawn based on our analyses

Our current analysis focuses on the following questions:

1. Identify the top 10 popular business categories and study the seasonal trends in user ratings for these categories. Click here to view [results](#_Seasonal_trends_across)
2. Identify the business in the dataset with very high and very low ratings. We can call these businesses as “controversial businesses” as they are either extremely popular or extremely infamous within their user base. Click here to view [results](#_Businesses_which_are)
3. Trend line analysis
4. To study the trend lines for a business, to identify if it would need to take corrective actions to improve the ratings.
5. Geographic trend: Identifying the trends of Restaurants in a popular destination like Las Vegas and comparing it against the overall trend in the US.

We found several interesting insights from our analyses on the dataset. Click here to view [results](#_Identifying_Trend_in)

For seasonal trends, we see in general that summer has the highest number of ratings, followed by autumn and spring. Winter has the least number of ratings for majority of the businesses. This can easily be extrapolated to the number of users visiting businesses in a particular season. Surprisingly, we noticed a difference in this trend for “restaurants” and “art and entertainment” businesses categories where winters had a higher number of ratings than in spring. This could be because some part of winter is a holiday season in the US and we can possibly say that customers tend to visit more restaurants during the holiday season.

The second part of our analysis focused on finding businesses with controversial nature. We were able to find many businesses with this nature. For instance, we found an automobile dealer (Centennial Toyota) in Las Vegas which had the maximum variation in user ratings. This shows that either customers are highly satisfied with their service or they are extremely disappointed. Another interpretation could be that the 5 star ratings are given artificially to the store, so that more customers visit the store. The top 10 businesses by the variation in the ratings are summarized in the results section.

The third part of our study involved analyzing trends of the ratings. This was the most interesting part of the project, in which we studied whether the rating trend of the business is going up or down based on its rating history. We extracted the businesses with the best positive trend and the worst negative trend. We carried out this piece of work for the businesses in Las Vegas as that’s where we observed lots of variability in the user ratings. A more elaborate discussion on the findings is done in the results section.

**Related Work**

Discrete data can be analyzed in multiple ways. Our analysis is mainly focused around three questions. Based on the reviews from Yelp we derived some interesting analyses. After reading through several works which have been done using Yelp data, we figured that there are a variety of ways in which a certain question can be perceived and interpreted. All the study/research mentioned have been done on Yelp dataset. For instance, here are some examples in which similar questions can be analyzed based on the same dataset:

1) Identifying the top 10 business categories and then plotting the seasonal trends (winter, spring, summer and autumn) across these top 10 categories. We considered review counts as a metric for our evaluation. Alternately, this analysis can also be done for particular holiday seasons to see if people tend to change their habits. Study [4] focuses on the Christmas holiday season (Dec-Jan) and thanksgiving week to find the food and beverage preferences of the users.

2) Finding controversial stores which have unreasonably high and low ratings. At a broad level, we have done this analysis using standard deviation (to capture the variance between ratings). Although our approach to this question is novel, we have a recommendation based on the research that is already done. For instance, study [1] tries to identify quality phrases from the reviews.

We can map these quality phrases with the ratings provided to see if the stores are controversial (a quality phrase is supposed to have a high rating, but a low rating (given that the review has a quality phrase) implies that there is something unusual about the store.). Also, study [4] shows the cluster of words that are given by users for every rating. So, we have different word clouds for every rating. These words can be used to spot any correlations between ratings and reviews. Here is an overview of how this advanced phrase technique works:

Advanced Phrase Mining: The primary idea is to automate the identification of quality words within a review using text analytics. This is done using advanced Natural Language Processing (NLP) algorithm features such as dependency parser to improve accuracy. Also this work tries to minimize the difference between phrase quality and phrase segmentation by using statistical measures as one of the parameters to enhance the quality.

3) Understand the performance of businesses (whether they have an increasing trend or decreasing trend). We used the moving average technique to identify the trend of a business category. We looked at a study on Yelp data [2], which uses collective factorization technique to understand the correlations between several metrics (for ex: length of review, rating, location etc.). Here is a brief description on how collective factorization technique works:

Collective Factorization Model: Recommendation systems were initially based on matrix factorization method (dot product of factors) which is a more generalized method. However, using latent factors in collective modeling makes it easier to find multiple connections for the same relation. This research focuses on developing a model that can understand the connections between all metrics (by interpreting the relations between these factors). This model assigns a latent low dimensional vector for every variable in data and then tries to predict the relations between variables.

Also, another study [3] uses LDA (Latent Dirichlet Allocation) technique to find the demands of customers from the reviews which have high dimensionality. Traditionally, Latent Semantic Indexing (LSI) is a widely used technique for dimensionality reduction. LSI uses singular value decomposition (SVD) to reduce the data to a latent space representation, allowing for more reliable estimation. But LSI faced several issues due to the formulation of the probabilistic model.

Recently, a new model called Latent Dirichlet Allocation (LDA) has been implemented. This model treats the probability distribution of each document over topics, as a K-parameter hidden random variable rather than a large set of individual parameters (K is the number of hidden topics.

**Process**

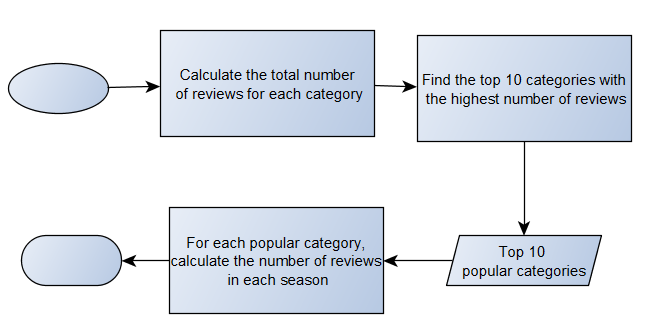
We coded our Python program using the pandas, numpy, matplotlib, and scipy libraries to analyze and visualize Yelp data. We used a combination of two datasets - Reviews and Business, both of which were obtained from the Yelp website. The Reviews dataset constitutes the main part of our data and contains 1.6 Million records in JSON format (each record has one review) with the ratings and text reviews given by users over a span of 2004 to 2015. The Business dataset contains the 783 categories defined by Yelp in which a user can categorize each business. The Reviews dataset contains one JSON object per review. We first extracted the data we needed from JSON and converted it into CSV format, with the purpose of easily converting it to a pandas DataFrame. However, instead of using comma or other regular separators to separate values, we used a Chinese character, “霽”, because the user review data had unexpected combinations of special characters including comma.

**Finding seasonal trend of top 10 popular categories:**

To study the seasonal trend of ratings, we calculated the total number of reviews for each category and selected the top 10 popular business categories with the highest number of ratings. We next encoded the 12 months to represent the seasons as below.

* December, January and February as Winter
* March, April and May as Spring
* June, July and August as Summer
* September, October and November as Autumn

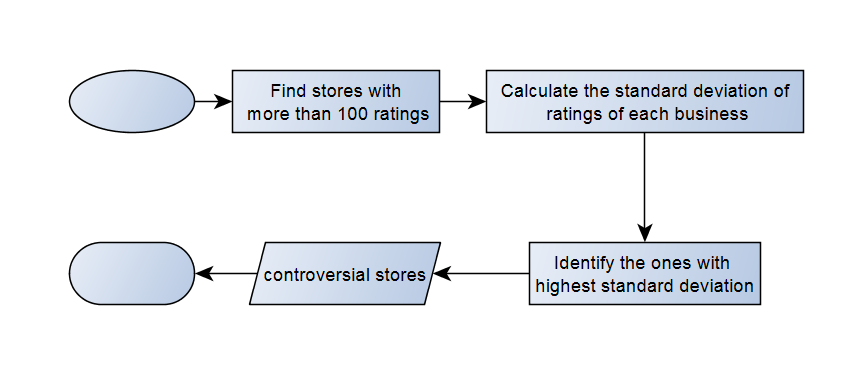
For each popular category, we calculated the number of reviews in each season to study the seasonal trend in the number of ratings. Each of the top 10 business categories contained at least 5000 reviews (a significant number) over the years. So it was reasonable for us to assume that the number of reviews of a business represented the number of customers that the business had. We were able to identify certain seasonal patterns which are highlighted in the results. The process flow of the code we used to arrive at this analysis has been shown in Fig. 1.



*Fig. 1: Process Flow showing our analysis of seasonal trends*

**Businesses which are controversial in nature:**

For our study, we defined controversial businesses as the ones with a large variation in the ratings. We calculated the standard deviation of ratings for each business in our dataset and identified the ones with highest standard deviation. For the purpose of removing the bias in the stores which have not been rated too many times (typically stores with lower number of aggregate ratings can have large variation), we segregated the data based on the number of ratings that the stores got. For our analysis, we considered only the stores which had more than 100 ratings. The flowchart in Fig.2 describes details of our approach.

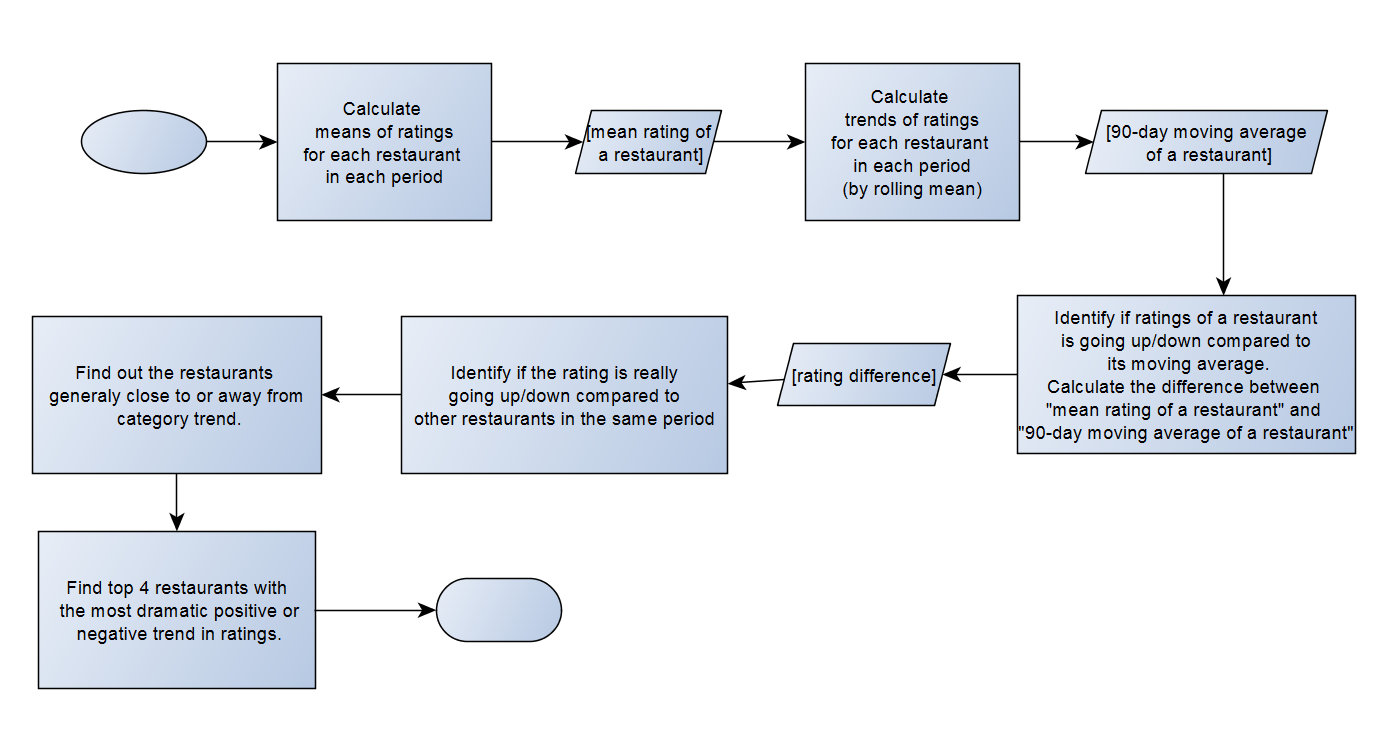


*Fig.2: Process Flow for our analysis to identify controversial stores*

**Identifying Trend in User Ratings:**

We approached this question with two sub questions: 1) Based on a business’s rating history, is the rating trend of the business going up or going down? Can we find out the businesses with the best positive trend and the worst negative trend? 2) What is the trend in ratings compared to other businesses in the same category and during the same period?

We used a *90 day moving average (90 DMA)* of the ratings to identify the rating trend based on a business’s rating history. If a business’s actual average rating is greater than its moving average in the same period, we identify this as a positive trend, and vice versa. We then calculate the difference between actual average and 90 day moving average, and standardize the differences for all businesses in the same category and the same period to compare a trend with other trends. The flowchart in Fig.3 describes details of our approach.

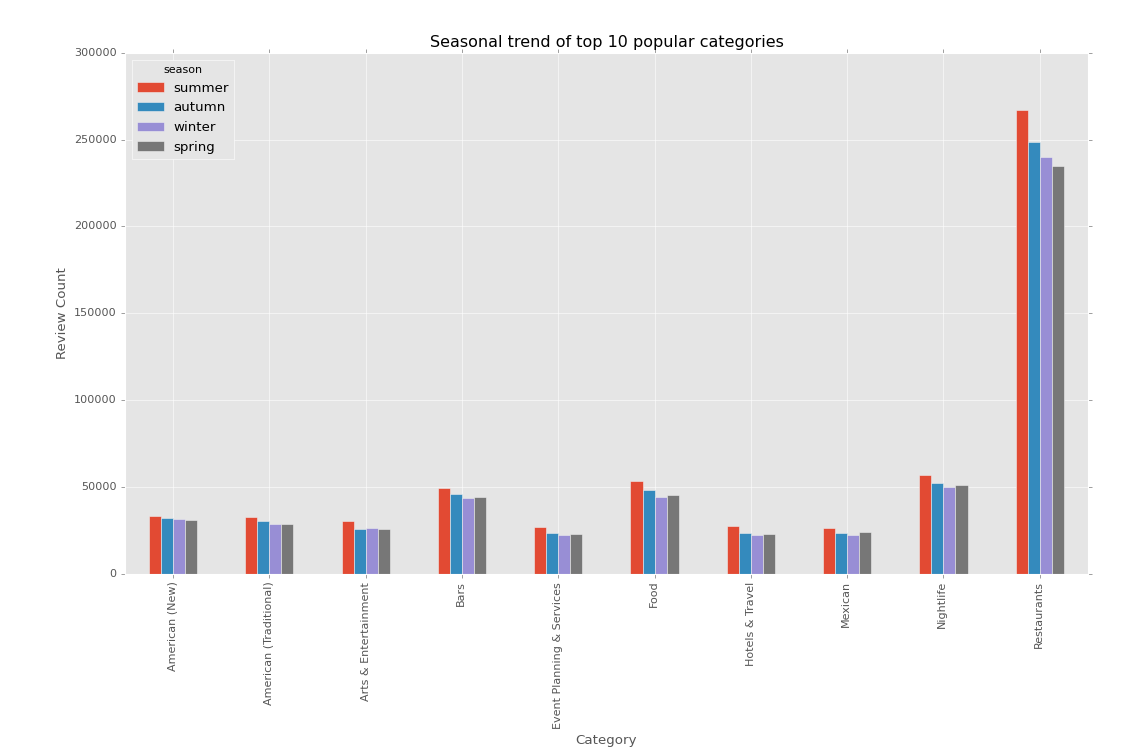


*Fig.3: Process Flow for our analysis on user rating trends*

**Results**

**Seasonal trends across top 10 categories**

From Fig.4, we can interpret that restaurants are the most visited categories when compared to other businesses. The number of people rating or visiting other businesses is way less as compared to the number of people rating and visiting restaurants. Another observation is that for any of the top 10 categories the number of people visiting in summer is greater than other seasons. The number of restaurants in the Yelp dataset is 21,892 when compared to total businesses of 61,184. As more than 33% of the businesses in the Yelp dataset are restaurants we see that it is the most dominant category in Fig.4.



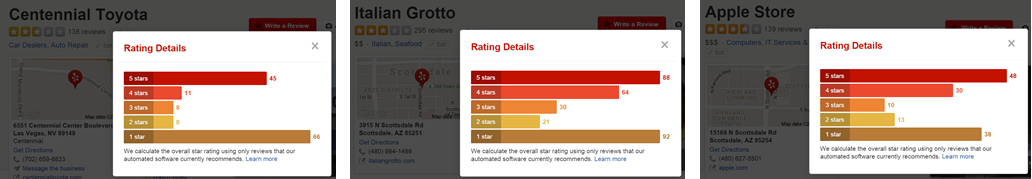
*Fig.4: Seasonal trend across top 10 categories showing summer as the dominant season*

**Businesses which are controversial in nature:**



*Fig.5: List of top 10 controversial stores across US*

6 out of the top 10 controversial stores are in Las Vegas. This may be because the tourists Las Vegas are one time visitors. This is not the case with other locations. However, surprisingly the other 4 stores are located in Arizona.

We compared 3 such stores (top 3 by the standard deviation - Centennial Toyota, Italian Grotto and Apple Store) with the actual ratings that are in the Yelp website. For these 3 businesses the variation in ratings (too many high and low ratings) is clearly visible, which substantiates our findings.

*Fig.6: Actual store ratings from Yelp for these controversial stores. Look at the high number of 5 star as well as 1 star ratings. Controversial!*

**Identifying Trend in User Ratings:**

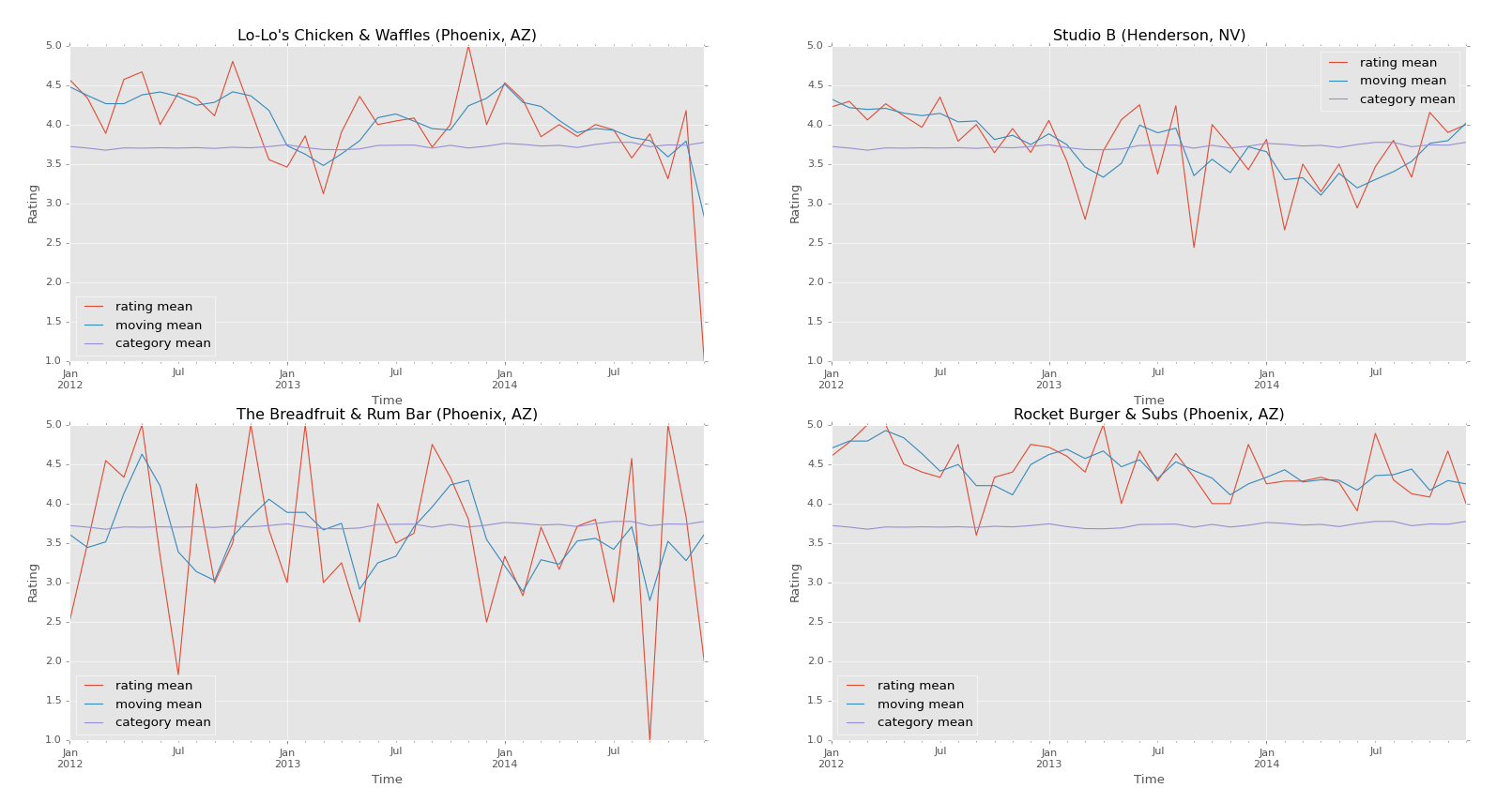
Trends in restaurant ratings were studied for the years 2012 to 2014 and we classified these into 3 types of restaurants:

1. **Positive Trend:** These were the restaurants that have grown positively over the three years. These restaurants are in the safe zone and are possibly benefitting from the improvement in either their service quality or product offering over the last few years.



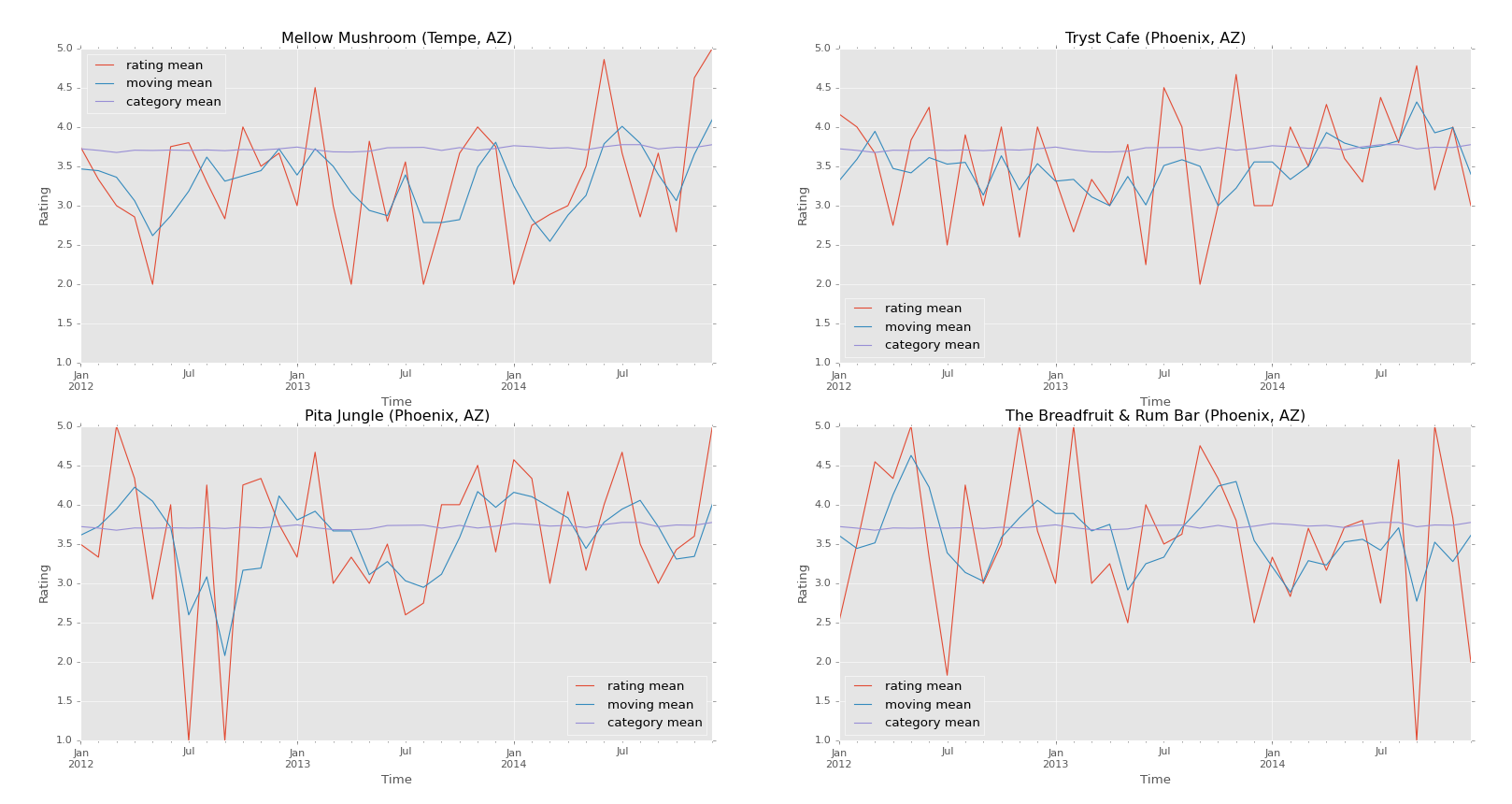
*Fig.7: Restaurants across US with strongest positive trend in ratings*

2. **Negative Trend:** These were the restaurants that have shown a decline in their ratings over the last 3 years. These restaurants could benefit from paying attention to the falling ratings by either improving their service quality or by improving their product offerings.



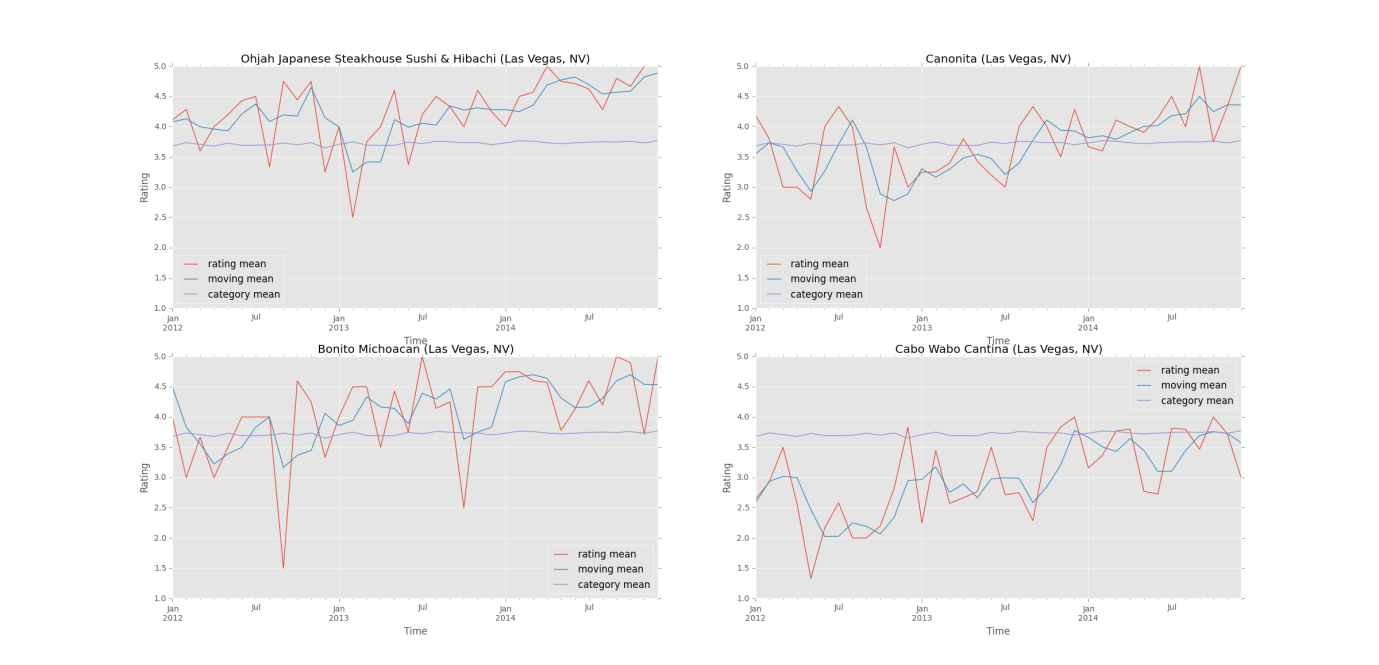
*Fig.8: Restaurants across US with strongest negative trend in ratings*

3) **Highly variable trend:** The restaurants that have ratings that fluctuate highly over the last two years. The trend is difficult to identify as for some period the trend in ratings has been positive and for other periods it has been negative. These restaurants also need to find out the reason for such high variability and take corrective actions for improvement.

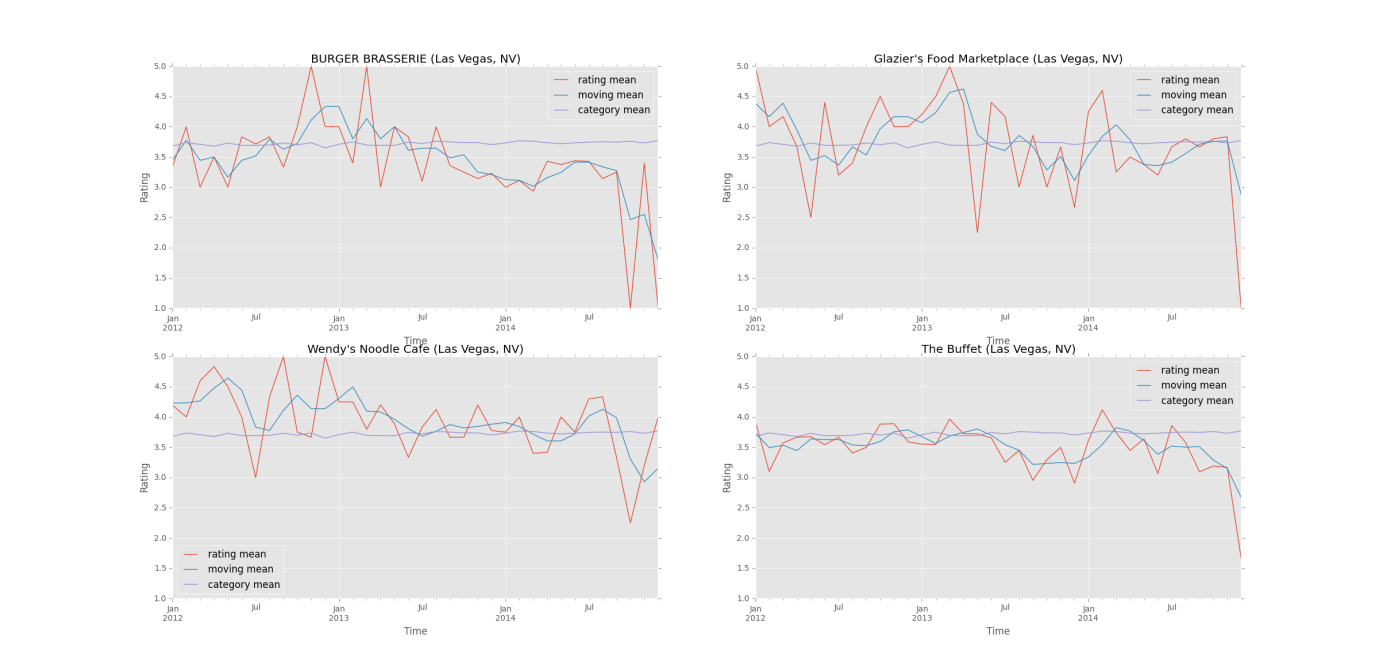


*Fig.9: Restaurants across US with highly variable trend in ratings*

We carried out the same trend analysis for the restaurants located in Las Vegas and we classified them into the above mentioned groups. It was interesting to see this large variation in trends for Restaurants only in Las Vegas. This could be attributed to the huge tourist population that flows into Vegas and how these restaurants match up to their expectations.



*Fig.10: Restaurants in Las Vegas with Positive Trend in ratings*



*Fig.10: Restaurants in Las Vegas with Negative Trend in ratings*



*Fig.11: Restaurants in Las Vegas with highly variable Trend in ratings*

**Conclusion**

Based on our analysis of seasonal trends, we see that summer is the most dominant season for many businesses as more people are review in summer. Least number of ratings (for the top businesses) were given in winter. However, for businesses in arts and entertainment category, winter had more ratings than spring. All of the top 10 controversial stores are in Las Vegas (majority) and Arizona. As Las Vegas is a famous destination the variance in ratings given by users can be understood. Trends in user ratings can be used by restaurants to take corrective measures to improve their service quality.

Many more interesting insights can be derived from Yelp data which can range from mining review text to predict business ratings or finding out any trend setter in a particular new category, which we would like to unravel at a later stage.

**References**

1) Liu, Jialu, et al. "Mining Quality Phrases from Massive Text Corpora."*Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015.

2) Gupta, Nitish, and Sameer Singh. "Collective Factorization for Relational Data: An Evaluation on the Yelp Datasets."

3) Huang, James, Stephanie Rogers, and Eunkwang Joo. "Improving restaurants by extracting subtopics from yelp reviews." *iConference 2014 (Social Media Expo)* (2014).

4) http://kevin11h.github.io/YelpDatasetChallengeDataScienceAndMachineLearningUCSD/

5) M. Potamias, “The warm-start bias of Yelp ratings,” arXiv e-print 1202.5713, Feb. 2012