Analyzing success in the Restaurant Industry

Piyush Pritam Sahoo, Malvania Smeet Saunil, Rhea Chandra and Li Xiang
Singapore Management University, ppsahoo.2012@business.smu.edu.sg, ssmalvania.2012@business.smu.edu.sg, rheac.2012@economics.smu.edu.sg, xiang.li.2012@sis.smu.edu.sg

Abstract - The popularity of crowd-sourced review sites and the rapid advancement in communication technologies have resulted in an astronomical increase in the collection of customer data. Every day, more and more people are sharing their experiences, likes, dislikes and needs. This explosion of data sharing provides a valuable opportunity for businesses to understand their customers and streamline their offerings to match the desires of their markets using data analytics. We utilize the data available on Yelp and design our project with an aim to enable businesses investigate the factors affecting user ratings, explore customer sentiments and attitudes over time, as well as predict ratings. The scope of our analysis is limited to restaurants in the Phoenix Metropolitan Area that have been active in the last two years.

Index Terms – Multiple Linear Regression, Sentiment Analysis, Text Mining, Time Series Analysis, Gower Clustering.

INTRODUCTION

Consumerism fuels a need to consistently see differences in products and services and find the best alternatives before making a decision. With this rising need for expert opinion and recommendations, crowd-sourced review sites have brought forth one of the most disruptive business forces of modern age. Since Yelp was launched in 2005, it has been helping customers stay away from bad decisions while steering towards good experiences via a 5-star rating scale and written text reviews. With its vast database of reviews, ratings and general information, Yelp not only makes decision making for its millions of users much easier but also makes its reviewed businesses more profitable by increasing store visits and site traffic.

The Yelp Dataset Challenge provides data on ratings for several businesses across 4 countries and 10 cities to give students an opportunity to explore and apply analytics techniques to design a model that improves the pace and efficiency of Yelp's recommendation systems. Using the dataset provided for existing businesses, we aim to identify the main attributes of a business that make it a high performer (highly rated) on Yelp. Since restaurants form a large chunk of the businesses reviewed on Yelp, we decided to build a model specifically to advice new restaurateurs on how to become their customers' favorite food destination.

With Yelp's increasing popularity in the United States, businesses are starting to care more and more about their ratings as "an extra half star rating causes restaurants to sell out 19 percentage points more frequently". This profound effect of Yelp ratings on the success of a business makes our analysis even more crucial and relevant for new restaurant owners. Why do some businesses rank higher than others? Do customers give ratings purely based on food quality, does ambience triumph over service or do geographic locations of businesses affect the rating pattern of customers? Or is the old adage "location, location, location" indeed an important factor for the success of a business on Yelp? Through our project we hope to analyze such questions and thereby be able to advice restaurant owners on what factors to look out for.

MOTIVATION AND OBJECTIVES

Our personal interest in the topic has motivated us to choose this as our area of research. When planning trips abroad, we explore sites like HostelWorld and TripAdvisor that make planning trips a lot faster and easier; not only is this helpful to customers planning trips but also to the businesses that have been given honest ratings. Since the team consisted students from a management university, our motivation when choosing this project was more business focused. Our perspective on recommendations was more catered towards how a business can improve its standing on Yelp, and thereby improve its turnover through more visits by customers.
We believe that our topic of analysis is crucial for the following reasons:

- It can encourage low quality restaurants to improve in response to insights about customer demand by changing some of the key features offline and online.
- The rapid proliferation of users trusting online review sites and incorporating them in their everyday lives makes this an important avenue for future research.

Prospective restaurant openers (or restaurant chain extenders) can intelligently decide the location based on the proximity factor to other restaurants around them.

ANALYSIS

The approach adopted in this paper is primarily incremental. Authors have tried to add additional variables into the analysis in order increase the amount of variation explained by the

model. Apart from the restaurant attribute level data already provided in the Yelp Academic Dataset, additional variables derived from the following sets of analysis were also added:

- Clustering information
- Overall sentiment from reviews to a restaurant
- Category information

We have further sought to aggregate all interesting findings made along the way to suggest recommendations for business owners. We have also extended into trend analysis that allows business owners to gain a dynamic picture of the restaurant industry and its evolution.

I. Clustering

Various methods of clustering were employed to understand the inherent groupings in the dataset:

- K-Means Clustering
- K-Medoids Clustering
- Mixed Clustering Method

We started with K-Means clustering, but realized quickly that outliers would skew the output. Furthermore, the presence of a significantly large number of binary variables rendered the output unsuitable for analysis.

The next method we used is called K-Medoids Clustering. This method was similar to K-Means, but differed in that it took actual points as center points. While this method rendered the results less swayed by outliers, it still did not address the binary nature of the data.

The third and final method employed for this analysis is what we termed a Mixed method. It was Partitioning around Methods (PAM) with Gower's Dissimilarity Matrix. As our dataset is a combination of different types of variables, we chose this as a more robust method which did not require the variables to be converted into numeric form.

Gower's method is able to handle mixed data types while clustering, and the following formula is used to calculate the similarity matrix:

$$S_{ij} = \sum_{k=1}^{*} s_{ijk}\ \delta_{ijk} \Big/ \sum_{k=1}^{*} \delta_{ijk}$$

After the similarity matrix is created, we used an elbow plot to determine the optimal number of clusters in the dataset. Subsequently, the results are computed where cluster membership is assigned to each data point.

II. Regression Analysis

We will begin with some data preparation for regression analysis, followed by execution of the regression model(s), findings from the results, and assessing the assumptions.

Existence of a large number of independent variables, with ordinal, categorical and measure variables necessitates the use of a multiple regression model on predicting mean ratings between 2013 and 2015. The large number of independent variables also necessitates reduction in variables, and we will hence employ the subset selection as one of the steps in picking the best variables.

- All-subset regression to find the best combination of variables
- Standard Least Squares Regression

After model development, the Standard Least Squared model will follow iterations to remove variables with insignificant values until all variables remaining are significant.

Should any of the assumptions for the multiple regression be violated, we will do some data transformation and manipulation, and redo the analysis.

- Linearity
- Multivariate normality
- No or little multicollinearity
- Statistical Independence (No auto-correlation)
- Homoscedasticity

Data Preparation for Regression

Creating a plot of actual vs predicted residuals, we found the existence of outliers in the dataset. Further analysis suggested that the deviation in these points differed primarily on their lower ratings and low variance in the dataset. Furthermore, the review count ranged between 5 to 8 reviews, suggesting user activity was limited for these restaurants. Since roughly 400 businesses skewed the distribution greatly, we concluded that they may not have been part of the same population, and must be studied independently.

To allow for predictive analysis later on, we conducted this analysis on 60% of the dataset, leaving 20% for test and 20% for validation.

The dependent variable for the regression was recent mean rating for the business (2013 and 2014), calculated from the review dataset. The independent variables were as given in the business dataset.

Methodology for Regression Analysis

Iteration 1:

Within the first iteration of the revised dataset, the Adjusted R-square was found to be 0.52.

We thought of improving this by adding additional variables to improve the accuracy of the prediction. Therefore we used clustering results from the analysis before to add cluster membership into the equation. We realized that the analysis for regression required a transformation of Review Count as a variable in the dataset. We used JMP Pro to create the transformation. We also added two additional variables which captured the number of top 5 high performing and low performing categories.

Iteration 2:

The second iteration yielded an Adjusted R-square of 0.55, which is an improvement after the first Iteration. We felt that there could still be more room to increase the explained variation in the model by including additional information

about the restaurants. Hence, we used results from sentiment analysis to enter additional variables into the equation.
Among competing methods, we chose to use the lexical affinity method in extracting sentiment from reviews. We decided on this because of time constraints and knowing the polarity for a review would suffice in providing information about overall positive or negative sentiment for a business. Of course, a better method to improve on this method would be to weigh recommendations based on users' influence level.
The level of sentiment analysis here was very basic, taking the difference in positive words and negative words and creating a sentiment score to reflect the information from the review in explaining mean ratings. The list of positive and negative words chosen were from papers that have researched on opinion analysis and review studies (Hu, 2004).
We further entered information on the average word count for reviews in a particular restaurant to serve as an additional proxy of level of conversation for restaurants. With that, we proceeded to the final iteration of the analysis.

Iteration 3:

The third iteration yielded an Adjusted R-square of 0.66. This significant improvement was driven primarily by the sentiment variable. The results of this final increase are seen in Table I below. The overall summary of fit for the model can be found in Table II. The Assumptions for this final model were tested for and the outputs can be seen from the Appendix.

Predictive Analysis:

After completing the analysis, a prediction equation was generated. The predicted values were calculated. Predicted values were hence used to test and validate on the rest of the dataset. The output can be seen in Table III.

III. Spatial Lag Analysis

Methodology

In order to make our regression models more robust, we forayed into exploring the spatial lag model for our project. Two data points are said to be spatially autocorrected when their dependent variables seem to move together due to a larger stimulus like neighborhood effect acting upon them. 'Tobler's first law of geography encapsulates this situation: ''everything is related to everything else, but near things are more related than distant things.'' In context of our project, we suspected that the average rating of a neighborhood affects the star rating of any restaurant within that area.
The methodology adopted to conduct this analysis can be summarized in five steps. *Firstly*, a neighborhood criterion is set which is critical to build the weights matrix; we have three criteria to choose from namely, distance, contiguity and kernel. *Secondly*, we create the weights matrix which summarizes the relationship between n spatial units after being row standardized and setting the diagonal to zero. *Thirdly*, spatial autocorrelation is gauged by testing for the Moran's Index. *Fourthly*, the appropriate model is chosen based on the Lagrange Multiplier test and *finally*, the spatial regression model is built.

IV. Time Series Analysis

Overview:
Over the years, Yelp has served as an efficient influencer for restaurants all over the world. This escalates the need for businesses to avoid negative reviews and attract positive ones. To do so, it is important for businesses to know when people are reviewing the most, the period when people are reviewing the least, the sentiments of users about certain topics at different periods, customer preferences over time and so on. Knowing this will help businesses avoid service failures during periods of high reviews, to associate themselves with words or topics that correlate with higher ratings and avoid the words with lower ratings.
Our team delved into time series to:

- Examine the structure of the Yelp reviews and identify relevant patterns in the same.
- Explore possible patterns in sentiments and ratings for key topics over time
- Conduct basic forecasting.
- Simplify the above by building an Application using R Shiny to let the customer understand patterns in reviews for any topic (input by the user) and visualise the corresponding forecasts.

Analysis Scope:
The analysis involves two sections:

1. Decomposition: In this section, we decompose the time series into three major components that influence the observed values:

$$Y_t = S_t + T_t + E_t$$

Y_t is the time series observed value at period t,
S_t is the seasonal component at period t,
T_t is the trend cycle component at period t,
E_t is the remainder component at period t.

2. Forecasting lets the businesses to estimate the values of the time series for the next two years based on the previous year's values. By sing the Naïve method of forecasting, we forecast the time series separately for the de-seasonalized and the seasonalized components and then add the two to get the forecasted values.[1]

Our analysis involved exploration and forecasting of the time series by type of review and by four attributes related to reviews, ratings, and sentiments.

[1] Please refer to Appendix for description of Decomposition, Forecasting, and the assumptions used in the analysis

Types of Reviews

- All
- Cool Reviews
- Funny Reviews
- Useful Reviews

The Cool, Funny, and Useful Reviews were calculated using a 25% rule. As users vote for reviews based on the three criteria, we assume that if the number of votes for a type (let say, cool) is at least 25% of the total votes received (cool votes + funny votes + useful votes), it is considered as a Cool Review. This is true for Funny and Useful Reviews.

Attributes:

- Total Review Count: Sum of the reviews by month and year
- Proportion of Total Reviews: Proportion of the reviews containing the input word (topic) by month and year
- Average Stars: Mean of rating by month and year
- Average Sentiments: Mean of Sentiments by month and year

Yelp Time Series App

Link 1: https://thisppsguy.shinyapps.io/Timeseries
OR
Link2: https://goo.gl/bzGyiC

The Yelp Time Series App is an interactive visualisation and analysis tool developed using R Shiny package to conduct the above Time Series Analysis. The Application contains two tabs, "Decomposition" and "Forecast".

The App lets users input key words or topics through text input. They may also select the Types of Reviews as well as the measures as shown in the picture.

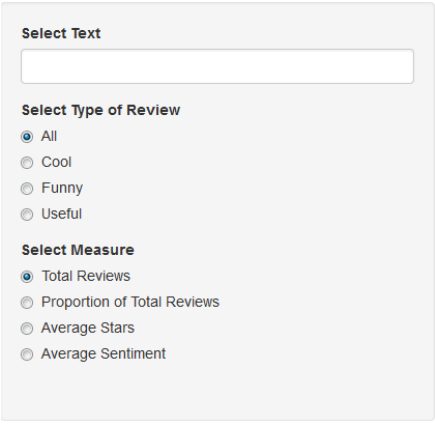


Figure I - Controls on Yelp Time Series App

Depending on their selections, the users can visualize the decomposition of the time series through time series plot and table outputs on the Decomposition Tab.

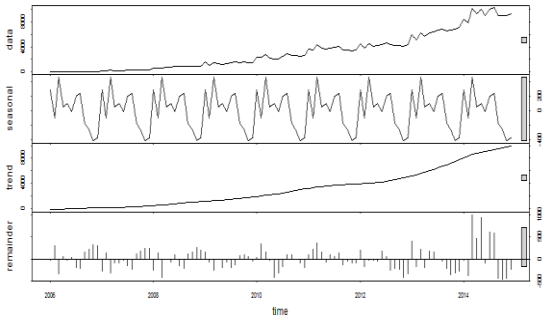


Figure II - Decomposition of Time Series

Components

	seasonal	trend	remainder
Jan 1	-115.586180	113.64593	89.9402484
Feb 1	-382.029535	104.00926	336.0202746
Mar 1	171.170515	94.37259	-233.5431044
Apr 1	-127.940483	88.76069	69.1797953
May 1	-25.505820	83.14879	-31.6429663
Jun 1	-55.140023	80.68324	0.4567844
Jul 1	178.953106	78.21769	-221.1707961
Aug 1	280.828860	77.14915	-283.9780067

Figure III - Table Outputs for Decomposition

The bars on the right side of the decomposition graphs show the relative contribution of the component towards the observed value. The bigger the bar of the component, the lesser is the contribution of the component.

On the Forecast tab, the users can see the Naïve Forecast as well as the table outputs:

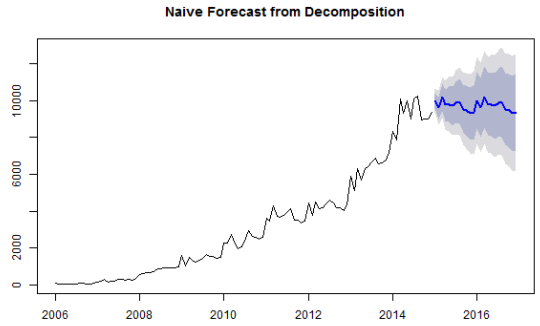


Figure IV - Forecast of Time Series

	Point Forecast
Jan 2015	10007.504
Feb 2015	9618.813
Mar 2015	10187.122
Apr 2015	9768.346
May 2015	9813.569
Jun 2015	9711.964
Jul 2015	9917.137
Aug 2015	9957.039
Sep 2015	9542.275

Figure V - Table Output for Forecast of Time Series

FINDINGS

The regression output is summarized as shown below in Table 1 and Table 2.

Table I

Term	*Estimate*	*Prob>|t|*	*Std Beta*	*VIF*
Intercept	4.150	<.001	0.000	.
Recent_Rate_Variance	-0.348	<.001	-0.401	1.714
attributes.Ambience.casual	-0.080	<.001	-0.064	1.316
attributes.Ambience.trendy	-0.154	<.001	-0.048	1.205
attributes.Delivery	0.064	<.001	0.041	1.046
attributes.Good.For.lunch	-0.054	<.001	-0.046	1.190
attributes.Noise_loud	-0.107	<.001	-0.044	1.053
attributes.Outdoor.Seating	-0.063	<.001	-0.053	1.116
attributes.Price.Range	-0.063	<.001	-0.062	1.426
Total.Opening.hours.with.mfm.	-0.007	<.001	-0.059	1.096
No.ofHighperformingcategories	0.048	<.001	0.056	1.114
Review.Count.log.transformation	0.159	<.001	0.114	1.526
wordcount	-0.006	<.001	-0.282	1.476
score	0.176	<.001	0.439	1.929

Some interesting observations from Table I:

- Highly rated restaurants tend to not have high variance in their ratings, and are generally stable. There could be other factors leading to this consistency that can be discovered in the future.
- The coefficient for the sentiment "score" variable is also significant, and comparing standardized betas could be the single most important factor in contributing to better ratings in comparison to all other factors.
- Among other factors that explain the variation are word count and review count. Highly rated restaurants tend to have a lower average word count and a higher review count.
- Higher number of high performing categories can possibly render you to have higher ratings, but the causality of that cannot be established based on the regression alone.
- Loud restaurants and restaurants with a casual ambience tend to not have a higher rating.
- Opening for a longer time period also may not be good for a restaurant's rating, owing possibly to longer shifts that compromise on service quality.

Table II

R-Square	*0.659*
Adjusted R-Square	0.657
Root Mean Square Error	0.342
Mean of Response	3.620
Observations	2818

Results from Table II show that the amount of variance explained by the model in the dataset is close to 65%.

Table III

Training		Validation		Test	
RSquare	0.65	RSquare	0.66	RSquare	0.65
RSquare Adj	0.65	RSquare Adj	0.66	RSquare Adj	0.65
Root Mean Square Error	0.35	Root Mean Square Error	0.33	Root Mean Square Error	0.35
Mean of Response	3.62	Mean of Response	3.65	Mean of Response	3.64
Observations (or Sum Wgts)	2818	Observations (or Sum Wgts)	951`	Observations (or Sum Wgts)	926

Table III shows the comparison of models among all three segments confirming the uniformity of variance explained by the prediction formula.

Table IV

```
$observed
[1] 0.02872399

$expected
[1] -0.0001958864

$sd
[1] 0.00673187

$p.value
[1] 1.73935e-05
```

The above figure shows the Moran's I generated using a distance criteria of 1/d to construct the weights matrix, to test for spatial autocorrelation. The output suggests that the p-value is < 0.05, suggesting that the results are significant, and that there is no spatial autocorrelation given the observed output is close to 0.

Figure VI

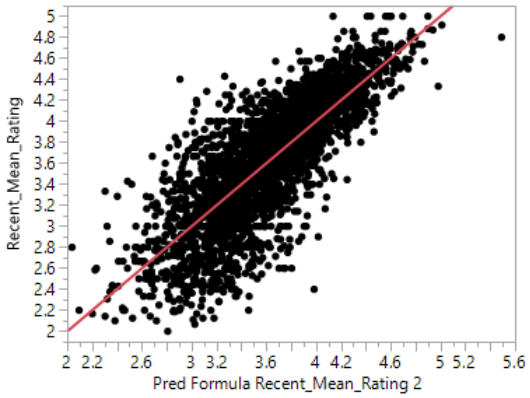


Figure VII

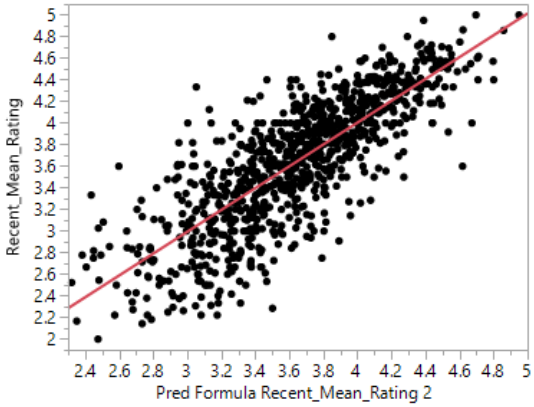


Figure VIII

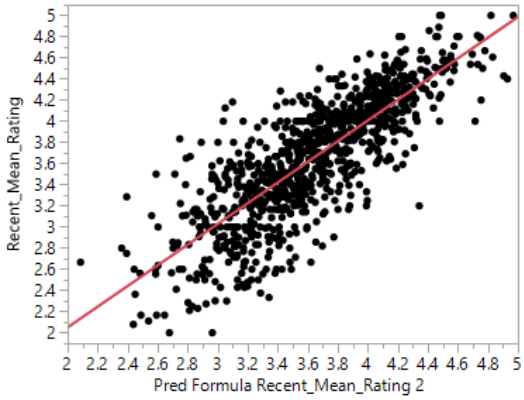
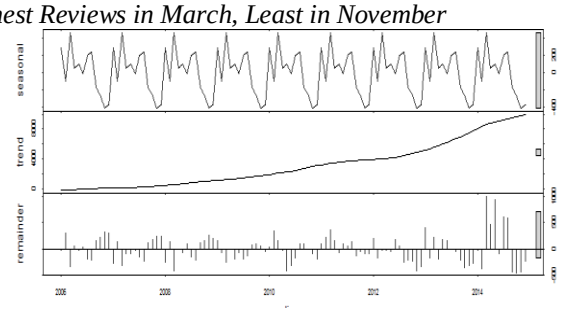


Figure 1 to Figure 3 show the distribution of actual mean rating with the predicted mean rating. As the graphs show, the actual mean rating and the predicted mean rating vary together for all three sets suggesting that the model is robust.

Figure IX

Highest Reviews in March, Least in November



From our exploration of the App, we found that although the total number of reviews was most influenced by the trend component, some seasonality did exist. There was a spike in number of reviews in March and a dip in November as can be seen below.

Figure X

Components	seasonal
Jan 2006	288.031100
Feb 2006	-100.659574
Mar 2006	467.649464
Apr 2006	48.873191
May 2006	94.096640
Jun 2006	-7.508134
Jul 2006	197.664448
Aug 2006	237.566643
Sep 2006	-177.197727
Oct 2006	-264.376772
Nov 2006	-414.666992
Dec 2006	-369.472597

On an average, there were nearly 800 more reviews in March compared to November.

Figure XI

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2015	10007.504	9580.519	10434.49	9354.487	10660.52
Feb 2015	9618.813	9014.965	10222.66	8695.307	10542.32
Mar 2015	10187.122	9447.563	10926.68	9056.063	11318.18
Apr 2015	9768.346	8914.376	10622.32	8462.312	11074.38
May 2015	9813.569	8858.802	10768.34	8353.379	11273.76
Jun 2015	9711.964	8666.069	10757.86	8112.406	11311.52
Jul 2015	9917.137	8787.441	11046.83	8189.416	11644.86
Aug 2015	9957.039	8749.344	11164.73	8110.028	11804.05
Sep 2015	9542.275	8261.320	10823.23	7583.224	11501.33

The seasonality and trend component can be used to forecast the number of reviews for the next year to give the businesses an idea of number of reviews that they may expect on Yelp as shown below.

Figure XII

Components	seasonal	trend
Jan 2006	0.09975027	2.978355
Feb 2006	0.21598048	3.131763
Mar 2006	-0.08697640	3.285171
Apr 2006	-0.17022098	3.433845
May 2006	0.08265969	3.582520
Jun 2006	-0.01365311	3.730643
Jul 2006	0.04835841	3.878767
Aug 2006	0.03819753	4.032082
Sep 2006	0.03687140	4.185397
Oct 2006	-0.15657875	4.318869
Nov 2006	-0.07671774	4.452341
Dec 2006	-0.01767066	4.534236

Components	seasonal	trend
Jan 2006	0.004190324	3.093632
Feb 2006	0.208599537	3.325641
Mar 2006	-0.089804471	3.557649
Apr 2006	-0.345950449	3.778183
May 2006	0.038365731	3.998718
Jun 2006	-0.169569062	4.213114
Jul 2006	0.089952447	4.427510
Aug 2006	0.201259013	4.639198
Sep 2006	0.244619175	4.850887
Oct 2006	-0.115086839	4.995881
Nov 2006	0.013997142	5.140875
Dec 2006	-0.080572470	5.198487

Figure: Average Sentiments for All Reviews and Cool Reviews respectively

Positive Sentiments on Cool Reviews

There were positive average sentiments on Reviews voted as Cool compared to All Reviews.

Figure XIII

Customers are more price sensitive during September, November, and January

Overall, the average sentiments for words relating to expense have been on the decline indicating that the customers may be getting more price sensitive.

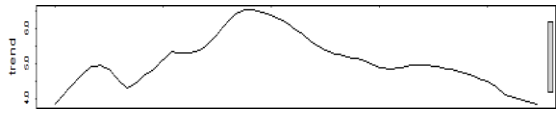


Figure: Trend for average sentiments on the topic "Pricey"

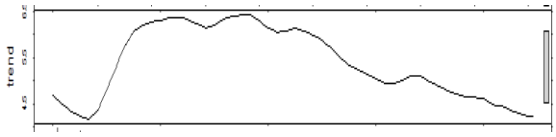


Figure: Trend for average sentiments on the topic "Expensive"

Moreover, we observed that there is a dip in average sentiments for high costs during the festive season: November, December, and January. The strength of seasonality was, however, not very significant.

Components	seasonal
Jan 2006	-0.51970947
Feb 2006	0.62843434
Mar 2006	-0.04824007
Apr 2006	0.21214411
May 2006	0.46151855
Jun 2006	0.01977108
Jul 2006	0.81763809
Aug 2006	-0.29393441
Sep 2006	-0.46866021
Oct 2006	0.25090790
Nov 2006	-0.74513330
Dec 2006	-0.31473617

Figure: Seasonality for average sentiments on the topic "Expensive"

The Yelp Time Series App lets users play with the data and see results as above in seconds and without working knowledge of any statistical software.

DISCUSSION

Spatial Lag

As can be seen from Table IV, the Moran's I test using the distance criteria 1/d (inverse of distance in km) to construct the weights matrix yielded no autocorrelation, proving that there may not be a spatial relationship in ratings among regions in Arizona. In order to further explore our results, we changed the criteria for the distance matrix several times in order to check for spatial dependencies. Following are some of the criteria used:

- inverse of distance squared
- inverse of distance raised to the power of 6
- contiguity matrix

Despite changing the weights criteria, Moran's index only increased marginally. When distance was punished more, the index rose from 0.02 to 0.08. Essentially, this meant that the star ratings of Yelp restaurants were spatially independent of their neighbor's ratings. To completely rule out any chances of spatial interaction we tested for correlation in other measures like count of ratings, ratio of high/low ratings, and variance of ratings. We tested the Moran's I for two other cities with the largest data points in the dataset but found similar results, strongly suggesting that a spatial relationship doesn't exist in the restaurant industry.

Regression Analysis and Predictive Analytics

Recent ratings tend to suggest that high performing restaurants tend to also not vary in consistency. Whether consistency in service contributes to this is topic for future research that may expand into actual service offerings.

The safest bet for a business owner to do well on Yelp tends to be with a fine dining restaurant. At the same time, business owners must know that reviews are the most important determinant of ratings on Yelp. Involving influencer analysis may shed more light on how important some reviews are compared to others.

Time Series

Some months denote stark differences in user activity, suggesting that business owners can benefit from great reviews if they enhance their performance during months preceding higher user activity.

Time of the year greatly determines the kind of conversation people have. For example, festive seasons where you typically see hiked prices tend to garner lower sentiment among reviewers. This might spell an opportunity for a business owner to position their restaurant with promotions and "gifts" for reviewers. It also benefits a restaurant to have cool reviews and attracting such reviews can lead to higher ratings. This can be done by inviting active yelp users with customized invitations who can later write a review for the restaurant after the visit.

CONCLUSION

Our project reveals that, although food quality is a must, many other factors such as noise levels and overall sentiment in the Yelp reviews play a vital role in affecting the popularity of a restaurant. It also reveals that the location of a restaurant is not a major contributor towards the popularity of a restaurant. Moreover, the project enables businesses to understand customer preferences, attitudes, as well as realize their pattern over time. Understanding these will definitely help businesses not just to enhance their popularity and maintain their competitive edge, but also to cut costs and make wise investments.

At this stage, the project offers various rooms for improvement. As the scope is limited to Phoenix Metropolitan Area, similar analysis could be replicated for other cities as well. A more sophisticated method of quantifying customer sentiments may be employed to relate reviews to ratings in a more efficient manner. Furthermore, automation of the above steps can help businesses understand new data without having to replicate the entire analysis manually.

REFERENCES

[1] Predicts the ratings of the business based on the review text provided by the user. http://arxiv.org/abs/1401.0864

[2] Is there a correlation between the business' ratings and the neighbours ratings? http://dl.acm.org/citation.cfm?id=2557526

[3] Spatial and Social Frictions in the City: Evidence from Yelp. http://faculty.chicagobooth.edu/jonathan.dingel/research/DavisDingelMonrasMorales.pdf

[4] M. Anderson and J. Magruder. "Learning from the Crowd." The Economic Journal. 5 October, 2011.

[5] Author's Last name, First initial, Middle initial, "Title," *Journal or book (italics)*, Vol, No #., date, pp.

[6] Author's Last name, First initial, Middle initial, "Title," *Journal or book (italics)*, Vol, No #., date, pp.

[7] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA,

[8] Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

ACKNOWLEDGEMENT

We would like to thank our Professor Seema Choksi and Professor Prakash Chandra Sukhwal for their assistance in guiding us with relevant advice for the project. Their support steered us into the right direction and helped us come up with concrete findings that can be useful for any business looking to expand or improve their value proposition.

AUTHOR INFORMATION

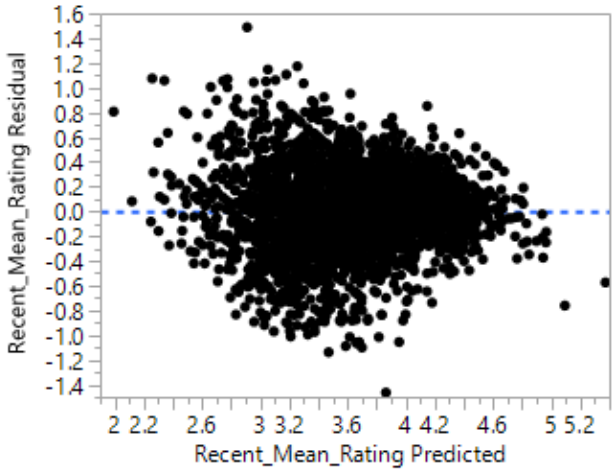
Piyush Pritam Sahoo, Student, Bachelor of Business Management, Singapore Management University.
Malvania Smeet Saunil, Student, Bachelor of Business Management, Singapore Management University.
Rhea Chandra, Student, Bachelor of Science (Economics), Singapore Management University.
Li Xiang, Student, Bachelor of Information Systems, Singapore Management University.

APPENDIX

Testing for Assumptions to the Regression Analysis

As mentioned in the report, there are multiple assumptions to a regression equation:

1) Linearity and Additivity
2) Homoscedasticity
3) No Multicollinearity
4) Statistical Independence of observations
5) Multivariate Normality



The above graph is a plot of actual residuals vs predicted residuals. Given the relatively even distribution of the dots throughout the line, it seems that the assumption of linearity more or less holds constant. Given also that the points bounce around the horizontal band evenly, homoscedasticity should be a fair assumption to make.

The following table shows a test of autocorrelation through the Durbin-Watson test. Since close to no autocorrelation is observed, it can be safe to conclude that there is statistical independence of observations.

Durbin-Watson	Number of Obs.	AutoCorrelation
2.0217401	2818	-0.0114

All metric variables were tested for normality. Given the exponential nature of the Review Count variable, a log transformation was applied to reduce the positive skew of the distribution.

Multicollinearity was checked by assessing the pairwise correlations of all variables and excluding variables which were highly correlated with each other.

Understanding Patterns – Decomposition[2]**:**
Decomposition is a commonly used method in Time Series Analysis that aims to separate the underlying patterns in the data from randomness. Generally, decomposition procedures separate the time series into three major components that influence the observed values over time:

- Trend: Represents the increase, decrease, or stationarity of the time series
- Seasonal: Represents the variation of the time series by seasons (usually, months)
- Randomness: The remaining unexplained component of the time series after removing trend and seasonality

The time series is decomposed as follows:

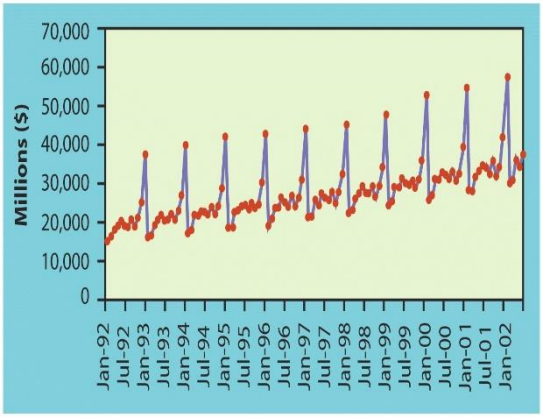
Data = Pattern + Error = f (trend-cycle, Seasonality, error)

$$Y_t = f(S_t, T_t, E_t)$$

Y_t is the time series observed value at period t,
S_t is the seasonal component at period t,
T_t is the trend cycle component at period t,
E_t is the remainder component at period t.

Types of Decomposition Models:

- Additive Model: This model is used when the magnitude of the seasonal fluctuation does not vary with the level of the series. In the example shown below for the sale of general merchandise in the US, the magnitude of variation remain the same over the years.



U.S. retail Sales of general merchandise stores

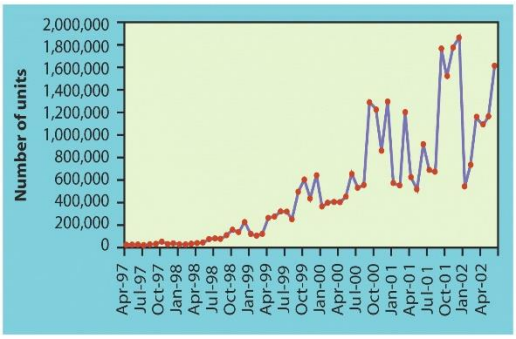
Hence, it is an additive model.

For additive models, the time series is a sum of its components.

$$Y_t = S_t + T_t + E_t$$

- Multiplicative Model: This model is used when the magnitude of the seasonal fluctuation varies with the level of the series. In the example shown below for the Number if DVDs sold in the US, the magnitude of variation varies over the years. Hence, it is an multiplicative model.

For multiplicative models, the time series is a product of its components



Number of DVD players sold in US

$$Y_t = S_t \times T_t \times E_t$$

Forecasts from Decomposition – Naïve Method[3]*:*
Decomposition not only helps to understand the composition of the time series but also helps to project the time series into the future and forecast. For our forecast, we used the combination naïve method and seasonal naïve method.

- Forecast of the De-Seasonalised Component: The Naïve method first conducts forecasting for the de-Seasonalised component. De-Seasonalised Component refers the remaining time series after removing the seasonal component. The forecast can be done either by random walk drift method or Holt method.

$$Y_t = S_t + T_t + E_t$$

Additive

$$Y_t = S_t \times T_t \times E_t$$

Multiplicative

[2] https://onlinecourses.science.psu.edu/stat510/node/69

[3] https://www.otexts.org/fpp/6

- Forecast of the Seasonal Component: The Seasonal Naïve method does forecasting for the Seasonal Component based on the past values of the seasonal data. It assumes that the seasonality is constant or is changing very slowly.

$$Y_t = S_t + T_t + E_t$$

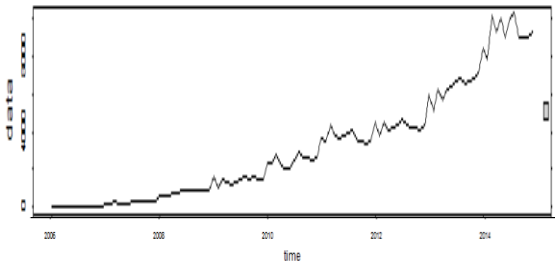
Additive

$$Y_t = S_t \times T_t \times E_t$$

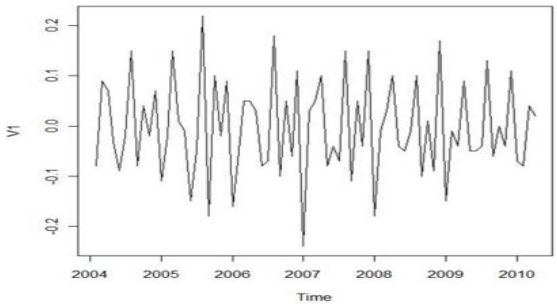
Multiplicative

Assumptions:
We have made two major assumptions in pour analysis of the time series.

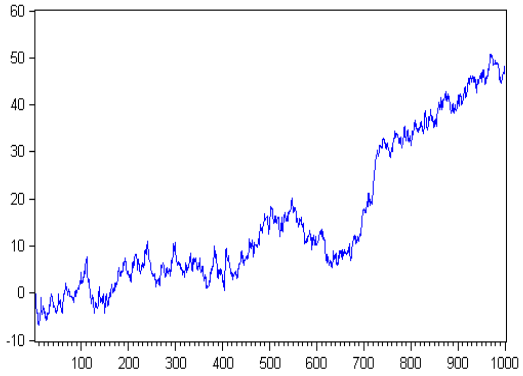
- The analysis assumes that all the time series are Additive. This is because, in our initial exploration of the time series, most of the time series were additive



- The analysis assumes that the time series is Non-Stationary. Non-Stationary time series refers to the time series which contains perceptible trends and seasonality over time. A time series is Stationary when there is no such perceptible trends or seasonality.



Stationary



Non-Stationary