



RECOMMENDATIONS MATTER TO US!



Li Xiang, Piyush Pritam Sahoo, Rhea Chandra, Malvania Smeet Saunil

TEAM ACCURO

Contents

Executive Summary.....	2
1. Project Overview.....	3
1.1 Introduction and Background.....	3
1.2 Review of Similar Work.....	3
1.3 Motivation for the Project.....	4
1.4 Project Scope and Methodology.....	5
1.4.1 Key Questions.....	5
1.4.2 Primary requirements (for “restaurants” and one city only):.....	5
1.4.3 Secondary Requirements.....	6
1.4.4 Future research.....	6
1.5 Limitations and Assumptions.....	7
1.6 Risks and Mitigation.....	7
2. Roles & Responsibilities.....	8
3. Project Execution.....	8
3.1 Work Scope.....	8
3.2 Deliverables.....	8
3.3 Tools Used.....	9
3.4 Project Timeline.....	10
3.5 Findings.....	11
3.5.1 Step 1: Exploratory Data Analysis and Data Manipulation.....	11
3.5.2 Step 2: Understanding salient groupings through Clustering.....	13
3.5.2 Step 3: Evaluating the importance of salient features through Regression.....	19
3.5.3 Step 4: Spatial Lag Analysis.....	23
4. References.....	27
5. Appendix.....	28

Executive Summary

The purpose of this study is to understand the salient features of the restaurant market through an analysis of the Yelp Academic Dataset. We wish to uncover inherent groupings within the restaurant industry and in so doing help business owners make better decisions. We also seek to understand some of the reasons why businesses are rated higher than others, and we wanted to see if there is a significant effect of location on the variables. If so, we would try to see the mechanism of this effect and hence recommend the importance of location of a new restaurant. We will be employing procedures like Clustering, Stepwise and Multiple Regression, and Spatial Lag regression. We will subsequently develop a data visualization tool through Tableau in order to visualize all the insights for ease of decision making.

The midterm report has developed key insights in understanding the inherent groupings of restaurants in Arizona by seeing the high performers, average performers, low performers and the "hit & miss" restaurants. Upon further inspection we found that average and low performers generally tend to higher number of reviews and that high performers tend to have quieter restaurants.

The report has also developed on understanding the mechanism of this effect where we found that good performing restaurants generally had little deviation in their rating and were generally more talked about on Yelp than low performing restaurants. We further found that the level of noise plays a big role in contributing to the rating on yelp. Subsequently, we also found that tagging your business with more high performing categories will leave you with a higher chance of doing well (maybe by virtue of getting discovered more).

The paper also elaborates on the logic behind the spatial lag analysis that will be conducted after the submission of this document. Currently we have confirmed that there is spatial autocorrelation and hence spatial dependencies do exist.

1. Project Overview

1.1 Introduction and Background

We live today, in what could be best described as the age of consumerism, where, what the consumer increasingly looks for, is information to distinguish between products. 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' favourite 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 analyse such questions and thereby be able to advice restaurant owners on what factors to look out for.

1.2 Review of Similar Work

1) Visualizing Yelp Ratings: Interactive Analysis and Comparison of Businesses:

The aim of the study is to aid businesses to compare performances (Yelp ratings) with other similar businesses based on location, category, and other relevant attributes.

The visualization focuses on three main parts:

- a) Distribution of ratings: A bar chart showing the frequency of each star rating (1 through 5) for a single business.

- b) Number of useful votes vs. star rating A scatter plot showing every review for a given business, with the x-position representing the “useful” votes received and y-position representing the for the business.
- c) Ratings over time: This chart was the same as Chart 2, but with the date of the review on the x-axis

The final product is designed as an interactive display, allowing users to select a business of interest and indicate the radius in miles to filter the businesses for comparison. We will use this as a base and help expand on some of its shortcomings in terms of usability and UI. We will further supplement this with analysis of our own using other statistical methods to help derive meaning from the dataset.

2) Your Neighbours Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction

This study focuses on the influence of geographical location on user ratings of a business assuming that a user’s rating is determined by both the intrinsic characteristics of the business as well as the extrinsic characteristics of its geographical neighbours.

The authors use two kinds of latent factors to model a business: one for its intrinsic characteristics and the other for its extrinsic characteristics (which encodes the neighborhood influence of this business to its geographical neighbours).

The study shows that by incorporating geographical neighborhood influences, much lower prediction error is achieved than the state-of-the-art models including Biased MF, SVD++, and Social MF. The prediction error is further reduced by incorporating influences from business category and review content.

We can look to extend our analysis by looking at geographical neighbourhood as an additional factor (that is not mentioned in the dataset) to reduce the variance observed in the data and improve the predictive power of the model.

3) Spatial and Social Frictions in the City: Evidence from Yelp

This paper highlights the effect of spatial and social frictions on consumer choices within New York City. Evidence from the paper suggests that factors such as travel time, difference in demographic features etc. tend to influence consumer choice when deciding what restaurant to go to.

“Everything is related to everything else, but near things are more related than distant things” (Tobler 1970).

1.3 Motivation for the Project

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:

- 1) It can encourage low quality restaurants to improve in response to insights about customer demand.
- 2) The rapid proliferation of users trusting online review sites and incorporating them in their everyday lives makes this an important avenue for future research.
- 3) Prospective restaurant openers (or restaurant chain extenders) can intelligently decide the location based on the proximity factor to other restaurants around them.

1.4 Project Scope and Methodology

“How to dominate the restaurant scene in a city?”

1.4.1 Key Questions

We will seek to answer 3 main questions for the purpose of our analysis:

- 1) What are the salient features of these inherent groupings within restaurants?
- 2) How do these features that contribute to good/bad overall ratings for restaurants?
- 3) How important is location within all of this?

1.4.2 Primary requirements (for “restaurants” and one city only):

Step 1: Descriptive Analysis - Analysing Restaurants specifically for what differentiates High performers, low performers and Hit or Miss restaurants. For each of the 3 segments mentioned, the following analysis will be done:

- Clustering to analyse business profiles that characterize the market. Explore various algorithms and evaluate each of the algorithms to decide which works best for the dataset.

Step 2: Key factors identification for prescriptive analysis (feature selection) for new restaurants by region, in order to succeed. Regression will be used to identify the most important factors and the model will be validated so that we can analyse how good the model is. This will constitute the explanatory regression exercise.

Step 3: Spatial Lag regression model. This section will focus on Geospatial Analysis to examine the effect of location of a business on its rating. The goal of this will be to modify the regression model in Step 2 by adding the geospatial components as additional variables to the model. This section will explore the three spatial regression models and use the model that best fits the dataset:

- Checking for Spatial Autocorrelation: Spatial dependencies existence will be checked using Moran’s I (or any other spatial autocorrelation index) to see if they are significant.
- Spatial lag model (for regression) will be used if the dependent variable, the business rating, is spatially auto correlated i.e. the ratings of businesses in one location are

correlated to the ratings in nearby locations. Spatial proximity will be defined using an $n \times n$ matrix and various weight matrices (to test validity) will be used in the estimation of spatial regression.

- Spatial cross-regressive model will be used in place of Spatial lag model if the independent variables (or business attributes) in the regression are also spatially auto correlated.
- Spatial error model will be used if the residuals of the OLS regression are spatially auto-correlated.
- Results from the exercise will be interpreted to recommend salient features of regions to describe to businesses typical characteristics of similarly rated, close-by restaurants.

Step 4: Build a visualization tool for client for continual updates on business strategy. Focus will be to build a robust tool that helps the client actively visualize all insights developed during the project.

1.4.3 Secondary Requirements

- 1) The team will try to answer questions regarding any emerging trends for the restaurant industry, and perhaps even compare cities to see some differences.
- 2) The team will try to build some predictive tools based on the Spatial Lag output to test to robustness of our model.

1.4.4 Future research

- Evaluating the importance of review ratings for restaurants – Are they effective to improve ratings? Do restaurants that utilize recommended changes succeed?
- Can the ratings and reviews of local experts be assimilated in feature extraction to help improve the predictability of ratings success? We realize that people are social entities and can be heavily influenced by reviews from local experts in their criticism on Yelp. Future research in this area can enrich our analysis for a business as well.

1.5 Limitations and Assumptions

In doing our analysis, we have overall concluded below some of the major limitations we can foresee from this project:

Limitations	Assumptions
Limited data points on businesses and cities	Project methodology will be scalable for looking at regional trends
Limited action-ability of insights since companies may not care about Yelp ratings	Project findings will help set priorities for improvement for business owners
Businesses attribute may not be completely accurate	Assuming that data has been updated as accurately as possible
Defining business categories	Assuming business tags under categories are comprehensive for the competitive set

Future projects can further seek to mitigate some of these by adopting larger datasets and actually partnering with a real business to assess the impact of the recommendations in terms of a profitability analysis to recommend the best solutions.

1.6 Risks and Mitigation





Risk Assessment Metric:

	Likelihood			
		Low	Medium	High
Impact	Low	C	C	B
	Medium	C	B	A
	High	B	A	A

Risks	Level	Mitigation
Insufficient statistical knowledge	B	Consult with supervisor and online course materials
Lack of actionable business insights	A	Continuous literature search on meaningfulness of insights for businesses according to each city

Dashboard UI design may not be intuitive or extensive	A	User testing and consistent updates with the supervisor
---	---	---

2. Roles & Responsibilities

			
Li Xiang <i>Data Engineer</i>	Piyush Pritam <i>Data Visualizer</i>	Rhea Chandra <i>Data Scientist</i>	Smeet Malvania <i>Project Manager</i>
Key Responsibilities: <ul style="list-style-type: none"> ○ Exploratory Data Analysis ○ Data Cleansing ○ Data Manipulation and Integration ○ Algorithm Execution 	Key Responsibilities: <ul style="list-style-type: none"> ○ Exploratory Data Analysis ○ Literature and research expert ○ Visualization of insights 	Key Responsibilities: <ul style="list-style-type: none"> ○ Exploratory Data Analysis ○ Algorithm Design and Evaluation ○ Methodology Design & Evaluation 	Key Responsibilities: <ul style="list-style-type: none"> ○ Project Flow Direction ○ Liaison between team and Supervisor ○ Documentation ○ Wiki page update

3. Project Execution

3.1 Work Scope

Through this project we are hoping to build to an interactive dashboard as a solution to the ratings and recommendations system Dataset Challenge by Yelp. Some research methods and machine learning techniques we would like to look into are:

- Cultural & Cultural Trends
- Location Mining
- Change-points analysis
- Hierarchical and Non-Hierarchical Clustering
- Explanatory & Predictive Regression analysis
- Spatial Lag Regression Analysis

3.2 Deliverables

- ~~Project Proposal~~
- ~~Mid-term presentation~~
- Mid-term report
- Final presentation
- Final report
- Project poster

- Visualizations of findings and insights hosted on Tableau
- [Wiki page](#)

3.3 Tools Used

Tool	Purpose
Excel	Data Cleaning/Manipulation
JMP	Regression Analysis
Java	Data Manipulation
R	Clustering, Spatial Lag Analysis
Tableau	Data Visualization

3.4 Project Timeline

Task	I/C	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13
Research														
Project scope exploration	All													
Tools Familiarity	Rhea, Piyush, Smeert													
Exploratory Data Analysis	Li Xiang, Piyush													
Proposal Development	Rhea, Smeert													
Project Proposal														
Proposal Document	All													
Wiki Update	Smeert													
Step 1: Descriptive Analysis														
Data Preparation	Li Xiang													
Advanced Data exploration	Li Xiang, Piyush													
Clustering Business Profiles	Rhea, Smeert													
Step 2: Feature Extraction														
Model Development	Rhea, Piyush, Li Xiang													
Model Evaluation	All													
Factor Interpretation	All													
Spatial autocorrelation assessment	All													
Midterm Report														
Interim Report Preparation	All													
Interim Presentation Preparation	All													
Wiki Update	Smeert													
Step 3: Spatial Lag Regression														
Spatial Lag model development	Rhea, Li Xiang													
Alternative model testing and eval	Piyush, Smeert													
Predictive Analysis	All													
Trend Analysis	All													
Step 4: Data Visualization														
Tool development	Piyush, Li Xiang													
User Testing	Rhea, Smeert													
UI Improvement	Li Xiang, Smeert													
Step 5: Final Presentation														
Final Report	All													
Final Presentation	All													
Wiki Page	Smeert													
Poster	Smeert													

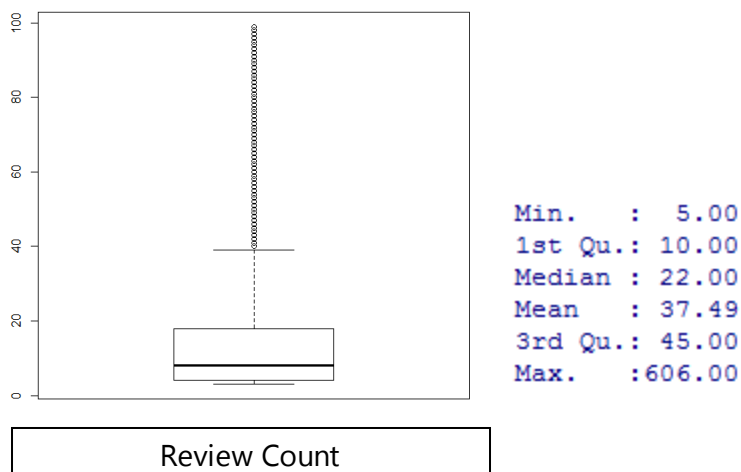
3.5 Findings

As mentioned in the proposal, the analysis started with taking a subset of restaurants in the state of Arizona using the yelp_academic_dataset_business.csv file (business data). This was done through extracting the category text "Restaurant" and selecting all businesses tagged as such. Our first step was to perform some Exploratory Data Analysis, and do some data cleaning and manipulation to suit our analysis.

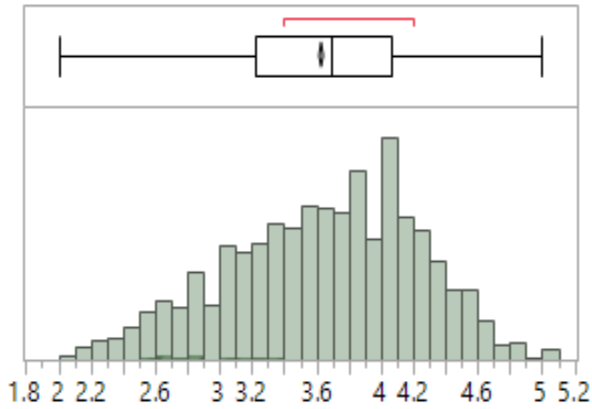
3.5.1 Step 1: Exploratory Data Analysis and Data Manipulation

We realized that the dataset actually contained records beyond the past 10 years. Since we did not want our model to be skewed by factors that were only important in the past, we chose to narrow down the dataset by only taking companies with greater than 5 reviews in the past 2 years (from 2013 to 2015), and changed the dataset to reflect that. Given that the mean rating was a rounded average for the ratings for all years, we had to compute the recent mean rating by combining the dataset containing reviews and filtering it by recent ratings, and subsequently mapping it back to the businesses dataset to develop a more recent and precise variable in mean ratings.

We suspected that it is likely for us to see a variance in the ratings and including that within our analysis would in fact allow us to see if highly rated restaurants get ratings high consistently. For that purpose, we again used the user review dataset and calculated the variance in rating for each business between 2013 and 2015 according to how users rated it.



Review Count as a variable was also manipulated to reflect number of reviews for a particular restaurant between 2013 and 2015, and as mentioned above, only restaurants with greater than 5 reviews were included in the dataset.



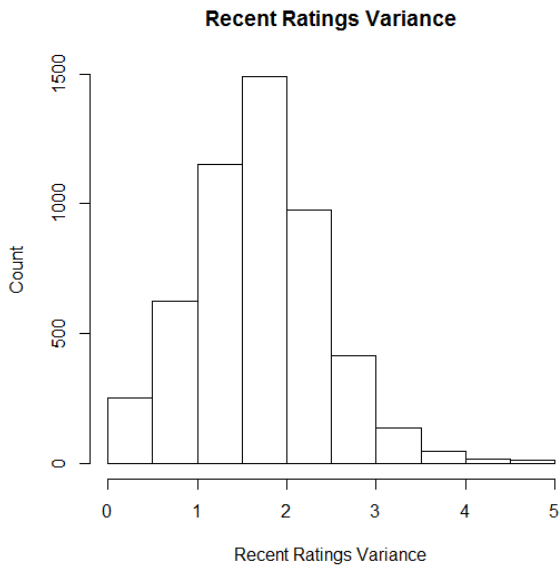
The Recent Mean Ratings shows a slightly right skewed distribution with 75% of the restaurants having a rating of at least 3. 50% of the restaurants had a rating in between 3 and 4.

```

Min.      :1.000
1st Qu.   :3.091
Median    :3.609
Mean      :3.508
3rd Qu.   :4.022
Max.      :5.000

```

Recent_Mean_Rating



The recent rating variance of the restaurants is normally distributed as can be seen in the diagram.

```

Min.      :0.000
1st Qu.   :1.209
Median    :1.689
Mean      :1.697
3rd Qu.   :2.158
Max.      :4.800

```

Recen_Ratings_Variance

iven that there was a substantial number of missing values (>50%) for some of the variables, we decided that we needed to remove these variables. The variables that were removed from our analysis were as follows:

attributes.Happy Hour	attributes.By Appointment Only
attributes.Order at Counter	attributes.Dietary Restrictions.kosher
attributes.Hair Types Specialized In.kids	attributes.Dogs Allowed
attributes.BYOB	attributes.Drive-Thru
attributes.Payment Types.mastercard	attributes.Dietary Restrictions.vegetarian
attributes.Corkage	neighborhoods
attributes.Payment Types.amex	attributes.Open 24 Hours
attributes.Music.live	attributes.Music.jukebox
attributes.Dietary Restrictions.dairy-free	attributes.DietaryRestrictions.vegan

attributes.Music.background_music	attributes.Smoking
attributes.Music.karaoke	hours.Thursday.open
attributes.Good For Dancing	hours.Friday.open
attributes.Good For Kids	hours.Tuesday.open
attributes.Payment Types.cash_only	hours.Friday.close
attributes.Music.video	hours.Thursday.close
attributes.Dietary Restrictions.halal	hours.Saturday.open
attributes.Ages Allowed	hours.Wednesday.close
attributes.Payment Types.discover	hours.Monday.close
attributes.Dietary Restrictions.gluten-free	hours.Tuesday.close
attributes.Payment Types.visa	hours.Saturday.close
attributes.Music.playlist	hours.Sunday.open
attributes.Coat Check	hours.Sunday.close
attributes.Accepts Insurance	hours.Wednesday.open
attributes.Music.dj	hours.Monday.open
attributes.Dietary Restrictions.soy-free	attributes.BYOB/Corkage

Overall, as can be seen from the table above, we removed the variables with Music, payments, hair types, BYOB, and other miscellaneous variables. Opening hour variables were computed into two new variables for Weekday opening hours and Weekend opening hours. As can be seen, many salient attributes that could contribute to how customers view the restaurant have been removed from the analysis due to bad data quality.

Since most of the fields consisted of binary data and still did not have all the fields, we decided that replacing missing values was essential for clustering and regression analysis. Therefore we proceeded with imputing missing values with the average score for each category. Since binary variables were changed to continuous data, we essentially took the average and imputed the values as such.

Restaurants were tagged under a string variable called "Categories". This variable consisted of tags for a particular business and consisted of fields like "Greek", "Pizzas", "Bars", etc. We found that these categories might be useful in determining the level of success or failure for restaurants. Unfortunately, since we had 192 different categories, we grouped categories according to high performing ones and low performing ones, and created two numerical variables titled "high performing categories" and "low performing categories". This will hopefully lend greater credibility to the level of analysis and provide a better explanation for the performance of restaurants.

3.5.2 Step 2: Understanding salient groupings through Clustering

Data Preparation for Clustering:

- a) For K-means and K-Medoids Clustering, all variables must be in numeric form. Therefore, the following changes were made to the different variable types to convert them to numeric form.

Variable Type	Cleaning Steps	Example of variables
Binary (True/False)	True = 1 False = 0	attributes.Ambience.touristy attributes.Waiter Service
Nominal Categorical	- Dummy variables were created for each level containing values 1/0. - Some variables were removed due to lack of meaningfulness in clustering.	Dummy variable example: attributes.Wi-Fi Removed variables: names, latitude, longitude
Ordinal Categorical	Categorical variables were replaced with based on their order.	attributes.Price.Range
Numerical variables	Left as is	Recent stars

- b) For Mixed Clustering, no data conversions were required as the algorithm recognises all types of data. Missing values are also acceptable. However, due to lack of meaningfulness of some variables in the clustering process, such as name, business id, the variables were assigned a weight of 0 to exclude them from analysis.

Clustering Methods Used:

1) K-Means Clustering:

After converting all variables into numeric form and imputing the missing values with average value, k-means clustering technique was used to cluster the businesses.

However, due to the nature of the data, k-means clustering is not be the most ideal clustering algorithm. The issues with the technique are as follows:

- As binary variables were converted into numeric, the resulting clustering means may not be as representative.
- Due to presence of outliers in the data, the clustering will be skewed.

2) K-Medoids Clustering: Partitioning around medoids (PAM)

After converting all variables into numeric form and imputing the missing values with average value, k-medoids clustering technique was used to cluster the businesses.

K-Medoids clustering is a variation of k-means clustering. In K-Medoids clustering, the cluster centres (or "medoids") are actual points in the dataset. The algorithm begins in a similar way as k-means by assigning random cluster centres. But, in k-medoids the cluster centres are actual data points. A total cost is calculated by using the summing up the following function for all non_medoid-medoid pairs:

$$\text{cost}(x, c) = \sum_{i=1}^d |x_i - c_i|$$

, where x is any non-medoid data point and c is a medoid data point.

In each iteration, medoids within each cluster are swapped with a non-medoid data point in the same cluster. If the overall cost is less (usually defined by Manhattan distance), the swapped non-medoid is declared as new medoid of the cluster.

Although, k-medoids has protects the clustering process from skewing caused by outliers, it still has other disadvantages. To see the results, click [here](#).

The issues with the K-Medoid technique are:

- a) As binary variables were converted into numeric, the resulting clustering means may not be as representative.
- b) The computational complexity is large.

3) Mixed Clustering: Partitioning around medoids (PAM) with Gower Dissimilarity Matrix

As our dataset is a combination of different types of variables. Therefore, a more robust clustering process is needed which does not require the variables to be converted to numeric form.

Gower dissimilarity technique is able to handle mixed data types within a cluster. It identifies different variable types and uses different algorithms to define dissimilarities between data points for each variable type.

The following formula is used to calculate a similarity matrix (S_{ij})

$$S_{ij} = \frac{\sum_{k=1}^d s_{ijk} \delta_{ijk}}{\sum_{k=1}^d \delta_{ijk}}$$

The dissimilarity matrix is calculated by $1 - S_{ij}$. In the above equation, s_{ijk} represent similarity of individuals i and j based on variable k. δ_{ijk} represents the weight variable and denotes the importance of the variable in the matrix calculation. An important variable can be assigned a higher weight than others whereas if two individuals are incomparable the weight is 0.

For dichotomous variables, if the attribute is "present" (or True) for two individuals (restaurants) a dissimilarity of 0 is assigned. If the attribute is "absent" (or False), no dissimilarity is assigned as the restaurants are incomparable. The rules can be shown in the following table.

Individual <i>i</i> <i>j</i>	Values of character <i>k</i>			
	+	+	-	-
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

,where s_{ijk} represent similarity between individuals i and j based on variable k and d_{ijk} is the weight of the variable k for individuals i and j

For categorical variables, if the values for two data points are same, dissimilarity is 0 (or similarity is 1) and vice versa.

For numerical variables, distance is calculated using the following formula:

$$s_{ijk} = 1 - |x_i - x_j|/R_k$$

, where S_{ijk} is the similarity between data points x_i and x_j in the context of variable k , and R_k is the range of values in variable k

The daisy() function in the cluster library in R is used for the above steps. Weights of 3 and 2 were provided to Recent Variance and Recent Mean Ratings respectively. This was done in order to differentiate clusters with a higher importance given to differences in variance and ratings.

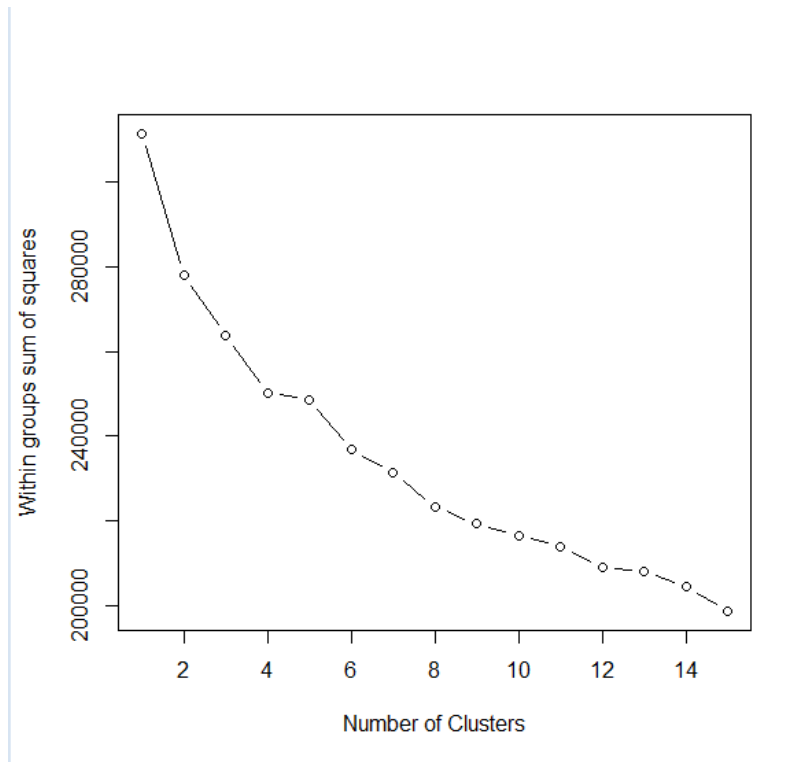
The dissimilarity matrix generated is used to cluster with k -medoids (or PAM) as described earlier. The dissimilarity matrix obtained serves as the new cost function for k -medoids clustering.

We call this two-step process "Mixed Clustering". This method has a number of datasets:

- As k -medoids method is used, the clustering is not affected by outliers.
- Clustering can be done without changing the data types.
- Missing data can also be handled by the Gower dissimilarity algorithm.

Elbow Plots:

The following elbow plot was generated using R.

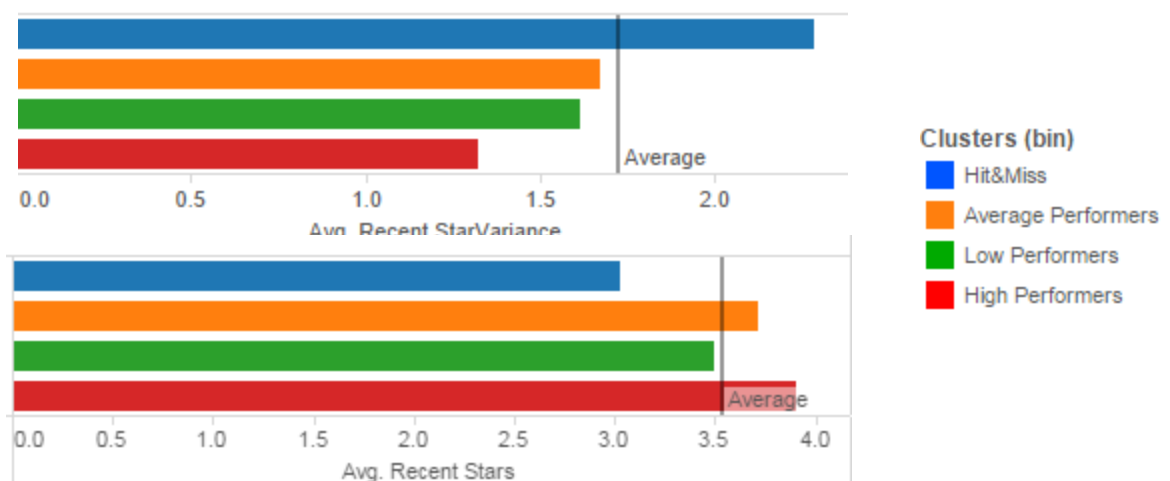


As there is a clear break at 4 number of clusters, we proceeded to carry out clustering with 4 clusters.

Key Findings from Mixed Clustering

Please follow this [link](#) to access the dashboard on Tableau Public:

The clustering using Mixed Clustering Method gave the following results:



The Hit and Miss cluster has a high variance of stars received by its patrons representing a mixed responses about the quality of such restaurants. Average performers are the ones who constantly received ratings in the middle range. Similarly, low performers are those with consistently low stars and High Performers have received consistently high stars.

The Mixed Clustering method generated clusters with very similar populations. This is shown in the following table.

Cluster Count

Clusters	
Hit&Miss	21.407%
Average Performers	24.803%
Low Performers	26.622%
High Performers	27.168%

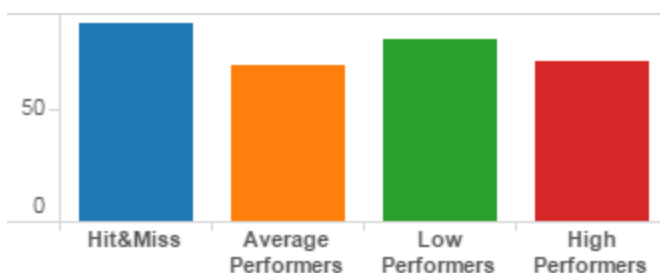
The various profiles of the clusters based on different attributes can be explored using the Tableau Dashboard. Some of the major profiles are as follows:

Average number of Reviews

Clusters	
Hit&Miss	11.68
Average Performers	57.43
Low Performers	53.43
High Performers	39.03

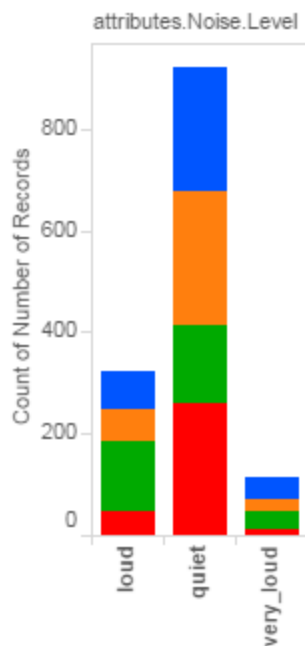
The restaurants in average performers and low performers clusters received highest number of reviews on an average than the other clusters.

Weekly Opening Hours:



On an average, Hit and Miss cluster is open for longer hours than other clusters.

Noise Level



It is interesting to note that most restaurants in the High performer cluster have a quieter environment. This suggests that Noise level may be one of the major factors affecting ratings. This can be supported by the fact that the largest proportion of restaurants with “loud” noise levels are low performers.

3.5.2 Step 3: Evaluating the importance of salient features through Regression

Methodology for Regression

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.

- 1) All-subset regression to find the best combination of variables
- 2) 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.

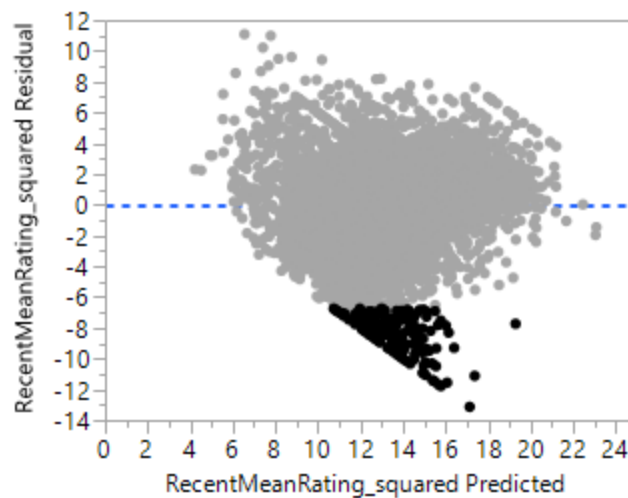
Should any of the assumptions for the multiple regression be violated, we will do some data transformation and manipulation, and redo the analysis.

- 1) Linearity
- 2) Multivariate normality
- 3) No or little multicollinearity
- 4) Statistical Independence (No auto-correlation)
- 5) Homoscedasticity

Data Preparation for Regression

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](#).

At the same time, we realized that a significant number of outliers dominated the dataset. Upon doing the first round of regression (with backward Min BIC regression followed by standard least squares regression) we found that the distribution Residuals to the predicted mean ratings showed that points did not randomly bound around the horizontal line, hence suggesting that there were outliers in the dataset.



Further analysis of these outliers indicated that they were primarily data points with very low variance and very low mean ratings. In the interests of limiting our analysis to the general population, we further reduced the dataset with these data points.

We will further investigate the best treatment for these outliers in using the final model for our analysis.

Findings

For subset selection, we chose to run a backward stepwise regression using a Minimum BIC criteria. We chose this since we were primarily looking to remove variables from the model, and the approach for elimination than exploration. The stepwise regression model revealed the following variables in the final model:

Current Estimates							
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Recent_Rate_Variance	-0.5366987	1	564.0549	3642.613	0
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Review Count log transformation	0.19208254	1	21.32146	137.692	2.3e-31
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Total Opening hours(with mfm)	-0.0109982	1	14.14523	91.349	1.9e-21
<input type="checkbox"/>	<input checked="" type="checkbox"/>	attributes.Noise_loud	-0.2505792	1	13.69295	88.428	8e-21
<input type="checkbox"/>	<input checked="" type="checkbox"/>	attributes.Noise_very_loud	-0.3243741	1	9.62703	62.170	3.9e-15
<input type="checkbox"/>	<input checked="" type="checkbox"/>	No.ofHighperformingcategories{0-1&2&3&4&5}	-0.0831549	2	18.79238	60.680	9.6e-27
<input type="checkbox"/>	<input checked="" type="checkbox"/>	attributes.Price Range{1-1.55&2&3&4}	0.04529109	1	8.345642	53.895	2.5e-13
<input type="checkbox"/>	<input checked="" type="checkbox"/>	attributes.Caters	0.07835312	1	5.406096	34.912	3.69e-9
<input type="checkbox"/>	<input checked="" type="checkbox"/>	attributes.Ambience.casual	-0.0734943	1	5.036723	32.527	1.25e-8
<input type="checkbox"/>	<input checked="" type="checkbox"/>	attributes.Noise_average	-0.0690695	1	3.314935	21.408	3.81e-6
<input type="checkbox"/>	<input checked="" type="checkbox"/>	No.ofHighperformingcategories{1-2&3&4&5}	-0.0540029	1	2.757528	17.808	2.49e-5
<input type="checkbox"/>	<input checked="" type="checkbox"/>	attributes.Delivery	0.05826177	1	2.006652	12.959	0.00032
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Open_binary	0.0599751	1	1.364891	8.814	0.003

Subsequently, we created the model using the Standard Least Squares method and found the following results:

RSquare	0.546926
RSquare Adj	0.545668
Root Mean Square Error	0.393509
Mean of Response	3.630041
Observations (or Sum Wgts)	4696

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	4.5698954	0.040534	112.74	<.0001*	.
Open_binary	0.0599751	0.020201	2.97	0.0030*	1.0423462
Recent_Rate_Variance	-0.536699	0.008893	-60.35	<.0001*	1.1123052
attributes.Ambience.casual	-0.073494	0.012886	-5.70	<.0001*	1.1041475
attributes.Caters	0.0783531	0.013261	5.91	<.0001*	1.0925828
attributes.Delivery	0.0582618	0.016185	3.60	0.0003*	1.0853382
attributes.Noise_very_loud	-0.324374	0.041139	-7.88	<.0001*	1.0905936
attributes.Noise_loud	-0.250579	0.026647	-9.40	<.0001*	1.2837857
attributes.Noise_average	-0.069069	0.014928	-4.63	<.0001*	1.4565038
attributes.Price Range{1-1.55&2&3&4}	0.0452911	0.006169	7.34	<.0001*	1.1494602
Total Opening hours(with mfm)	-0.010998	0.001151	-9.56	<.0001*	1.0511728
No.ofHighperformingcategories{0-1&2&3&4&5}	-0.083155	0.007569	-10.99	<.0001*	1.2395288
No.ofHighperformingcategories{1-2&3&4&5}	-0.054003	0.012797	-4.22	<.0001*	1.1197436
Review Count log transformation	0.1920825	0.016369	11.73	<.0001*	1.4175674

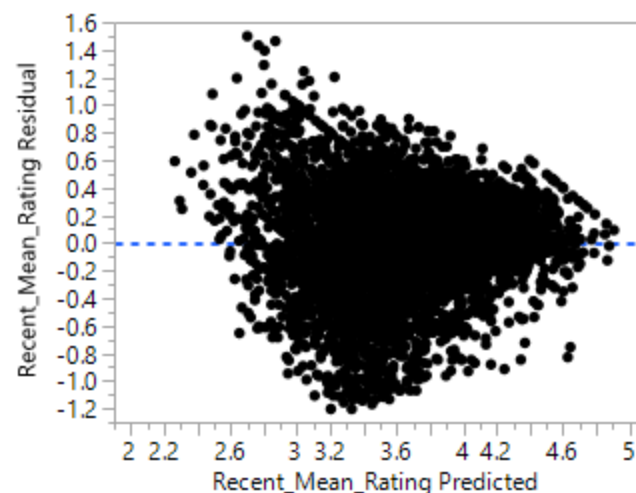
The results confirm the following findings:

- 1) The Adjusted R square has a moderate effect. Hence there is scope for further exploration on what contributes to a rating on Yelp eg. Location. There could be other variables such as quality of food or any of the other variables excluded that may contribute to lack of good fit for the model.

- 2) Noise level plays a part in determining how a restaurant is rated. It helps if the restaurant is tagged to be quiet.
- 3) Higher the review count and lower the variance, the better your restaurant generally tends to perform.
- 4) Catering and Delivery tend to be helpful assets for a restaurant.
- 5) In our EDA we noticed that restaurants were predominantly tagged with casual ambience, and that these results suggest that may not be a great idea.
- 6) Being tagged with high performing categories tends to have higher rating.
- 7) Being open for longer may in fact decrease the rating on yelp. This result may be due to a hidden variable such as longer working hours for staff in diners etc. which may cause service quality to suffer.

Having done the regression analysis, we must analyze the assumptions of multiple regression so that we don't over-estimate the robustness of our results.

- 1) Linearity – This is done through analysis of the following residuals vs predicted ratings plot.



- 2) Multivariate Normality

As covered in the data preparation part, metric variables were changed to reflect a more normal distribution. At the same time, review count log transformation was not a normal curve, hence the interpretation of the results is limited. Given that the object of our regression is to test for significance of predictors instead of testing model fit, satisfying strict normality is not necessary.

- 3) No or little Multicollinearity

Since the VIF factor as shown in the table above is not more than 10 (which is generally considered to be a threshold for significant multicollinearity), we can safely conclude that there seems to be no multicollinearity in the variables.

- 4) Statistical Independence

To determine statistical independence, we conducted a Durbin-Watson test and the results are as below:

Durbin-Watson	Number of Obs.	AutoCorrelation
2.0052003	4696	-0.0029

Since the results show little autocorrelation, that assumption is satisfied.

5) Homoscedasticity

Since the points on the residuals vs predicted ratings plot is fairly evenly distributed, and that the variables in the model are primarily binary, we can assume that the homoscedasticity assumption is reasonable.

Having covered regression to understand the salient features of restaurants and their contribution to the overall model, we will now proceed to understand the importance of location and how to add that as a factor within this equation.

3.5.3 Step 4: Spatial Lag Analysis

Social and physical phenomenon are often observed to be highly clustered in space. Regional voting patterns, racial segregation, poverty belt, crime rates, soil chemistry, animal habitats etc. are all examples of spatially clustered observations. Often such spatial relationships are ignored which weakens our ability to generate meaningful inferences about the processes we study. Spatial regression models include relationships between variables and neighboring values by including the values of error terms, x or y values in surrounding areas as explanatory variables. This allows us to examine the impact that one observation has on the other proximate observations.

If we ignore spatial similarities we violate certain regression assumptions in our model therefore our estimated regression coefficients are biased/inconsistent and our R^2 is exaggerated. Therefore, if spatial effects are present and are not accounted for, and then the model in question is inaccurate.

Approach

We believe that neighbourhood and location have a role to play in the overall star ratings of a restaurant in the Yelp dataset. This is why our group has forayed into exploring the spatial lag model for our project. '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 the context of our project, we suspect that the average star rating of a neighbourhood affects the star rating of any restaurant within that area.

Step 1: Set Neighbourhood criteria

Deciding the neighbourhood criteria is critical for building the weights matrix. We have chosen distance as our criterion, which takes distance between two data points as a relative

measure of proximity between neighbours. So the Weights Matrix is populated by values in terms of miles or kilometres or any other unit of distance. On the other hand, the contiguity criteria divides the data points into blocks and creates binary values for the weights matrix with 1 referring to 'neighbours' sharing a common boundary (adjacency factor) and 0 referring to distant businesses or 'not neighbours'. The third criteria is a more complex version of the first two which must only be set if the first two do not work.

In the distance criteria, we would want the weight attached to far-off businesses to be less than the weights attached to neighbouring businesses. This means that the values for the weights matrix must be inversely proportional to the distance. Keeping this mind, we chose '1/d' as a formula to populate our matrix. In order to increase the effect of distance we may edit the formula to '1/d²' or '1/d⁵' or any similar values in order to get a higher significance of the spatial coefficient parameter.

Once such a criterion has been decided based on the needs of the dataset we move on to the next step.

Step 2: Create Weights Matrix

The Weights Matrix summarizes the relationship between n spatial units. Each spatial weight W_{ij} represents the "spatial influence" of unit j on unit i. In our case the row and columns of our square matrix will have each unit on the two axis as being a restaurant with the diagonal being zero. Once the matrix has been created, it needs to be row standardized. Row standardization is used to create proportional weights in cases where businesses have an unequal number of neighbours. It involves dividing each cell unit in a row by the sum of all neighbour weights (all values in that row) for that business.

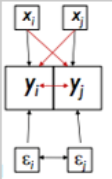
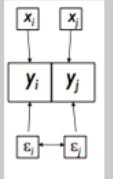
Step 3: Check Spatial Autocorrelation

Next step involves checking the need for a spatial model. When do we decide that a Linear regression is not enough to predict our ratings and that our dependent variables may be spatially lagged? We use the Moran's Index or Geary's C to make this decision. The index of spatial autocorrelation we use is Moran's I which involves the computation of cross-products of mean-adjusted values that are geographic neighbours (i.e., covariation), that ranges from roughly (-1, -0.5) to nearly 0 for negative, and nearly 0 to approximately 1 for positive, spatial autocorrelation, with an expected value of $-1/(n - 1)$ for zero spatial autocorrelation, where n denotes the number of units.

We used R (library ape) to compute the index (= 0.9409), which turns out to be significant for our model. Thus we can conclude that there is some spatial interaction going on in the data.

Step 4: Choose the appropriate Model

Now that we know for sure that we have strong spatial autocorrelation we must choose an appropriate model to explain it. The table below summarizes the main differences between the Spatial Lag and Spatial Error Model. The Lagrange Multiplier Test is used to

Spatial Lag Model	Spatial Error Model
<ul style="list-style-type: none"> • Spatial Dependence- similarity of nearby observations that is functionally linked to their proximity through an active process. • Dependent variable has dependence. Used when we know the structure of spatial dependence. • <u>Eg</u> – Mean star ratings being affected by the ratings of restaurants nearby, through social externality effects <div style="text-align: center;">  </div> <div style="text-align: center; margin-top: 10px;"> $y = \rho W y + x \beta + \varepsilon$ </div>	<ul style="list-style-type: none"> • Spatial Heterogeneity- similarity of nearby observations since they are similarly affected by stimuli acting on a larger region • Error term has dependence. Used when structure of dependence is unknown. No interaction assumed. • <u>Eg</u> – Mean star ratings being affected as a result of business owners in <u>neighbourhoods</u> learning from one another about store décor etc. <div style="text-align: center;">  </div> <div style="text-align: center; margin-top: 10px;"> $y = x \beta + \varepsilon$ $\varepsilon = \lambda W \varepsilon + \xi$ </div>

mathematically compute the significance of using each model. So far, we suspect that the Spatial Lag Model will be more relevant for our project.

Step 5: Build Spatial Regression Model

The final and conclusive step would be to build the Spatial Regression model, which incorporates a spatial dependence. This is done by adding a 'spatially lagged' dependent variable on the right hand side of the regression equation. The model now looks like this: $y = \rho W y + x \beta + \varepsilon$

Or $(1 - \rho W)y = x \beta + \varepsilon$, where

y = restaurant rating

ρ = spatial correlation parameter

W = Spatial weights

x = other attributes

β = coefficient of correlation

ε = error term

Spatial Autocorrelation

The output for the Moran's I are as shown alongside:

As you can see, the number is close to one showing a positive autocorrelation, and the p-value is close to 0, which suggests that the test is significant. This will help us carry on with our analysis for the spatial dependencies and then develop the model subsequently.

```
$observed
[1] 0.9409711

$expected
[1] -0.0001958864

$sd
[1] 0.01902652

$p.value
[1] 0
```

Next Steps

Our team will continue developing and testing the spatial lag model and assessing the impact of location. We will also look to generate a data visualization tool for a prospective restaurant owner so that they can dynamically generate all necessary information.

4. 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>
- 3) M. Anderson and J. Magruder. "Learning from the Crowd." The Economic Journal. 5 October, 2011.
- 4) The Yelp Dataset Challenge: http://www.yelp.com.sg/dataset_challenge
- 5) Link to K-Medoids wikipedia: <https://en.wikipedia.org/wiki/K-medoids>
- 6) Link to Gower's Method:
http://www.jstor.org/stable/2528823?seq=1#page_scan_tab_contents
- 7) Link to the Tableau Public Dashboard to analyze results of Clustering:
<https://public.tableau.com/profile/piyush.pritam.sahoo#!/vizhome/ClusteringenTableau/Story1>

5. Appendix

5.1 Results for Clustering:

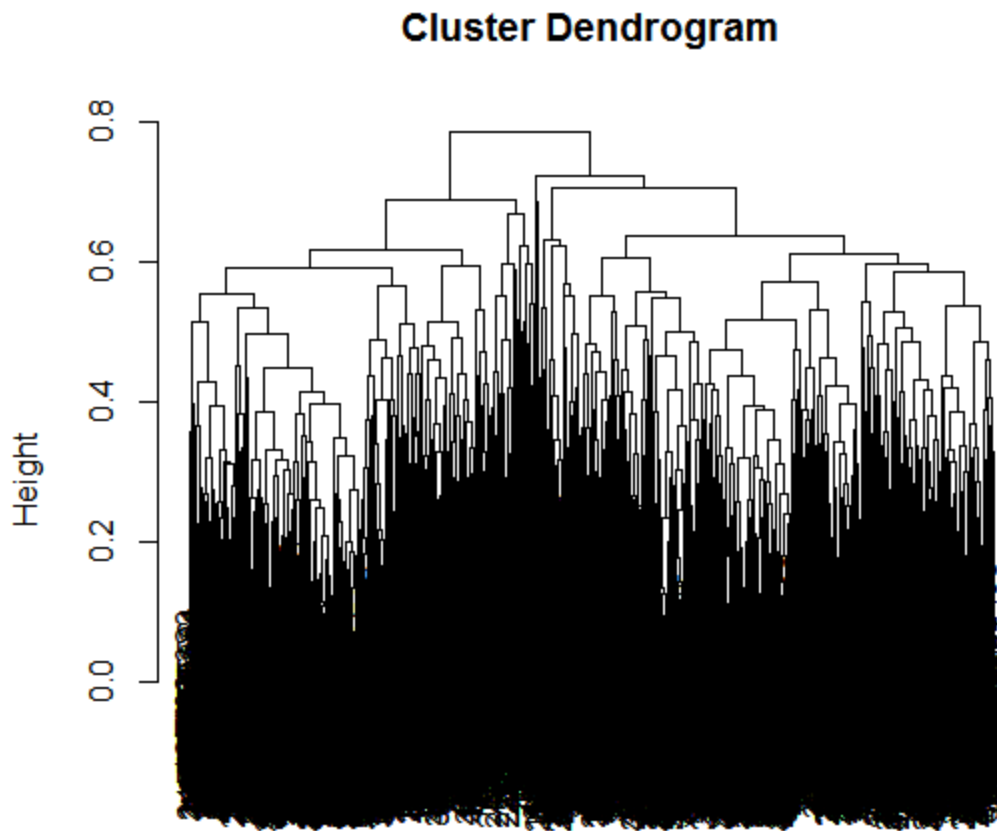
5.1.1 K-Means Clustering:

Clusters	Cluster Number	Count of ID	Recent_StarVariance	Recent_Stars	Recent_ReviewCount
High Performers	1	1795	-2.582805804	1.086712268	-0.034829995
Hit&Miss	2	1716	2.805558697	-1.234171824	-0.416702905
Average Performers	3	83	-0.516569358	0.967951815	0.453850644
Low performers	4	1512	-0.089502017	0.057440693	0.489360067

5.1.2 K-Medoids:

Clusters	Cluster Number	Count of ID	Recent_Star Variance	Recent_Stars	Recent_ReviewCount
Hit&Miss	1	1496	3.291491486	-1.206391055	-0.391265605
Average Performers	2	1005	0.236734962	0.173784684	0.507824983
Low Performers	3	1493	-0.843024453	-0.542618252	-0.136957916
High Performers	4	1112	-3.510210784	2.194457249	0.251301624

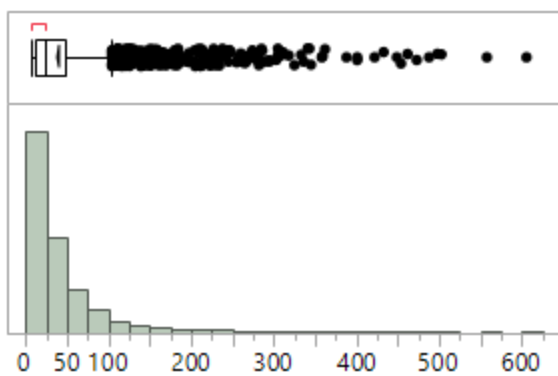
5.1.3 Cluster Dendrogram:



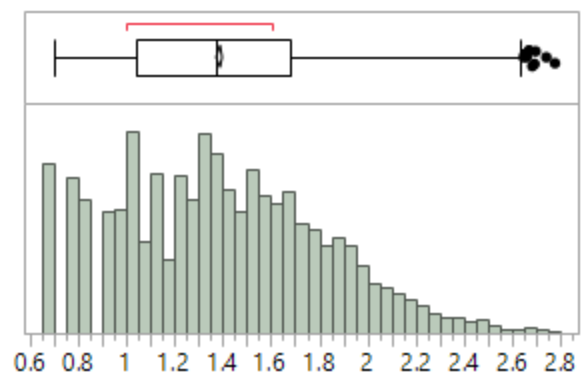
5.2 Results for Regression

5.2.1 Data Cleaning transformations:

Transforming Review Count by taking a log of the values:



Review Count



Review Count log transformation

