## Title

Magma: Detecting Local Changes in Restaurant Review Clusters That May Affect System-Wide Performance

## Author

Allan Spale

## Executive Summary

This paper serves as the culmination of a multi-stage project that studied how changes in time-based dynamic clusters can predict whether there is a more significant change occurring that would affect all of the clusters in a community. The author used restaurant reviews from the public Yelp review dataset focused on Toronto, Ontario in Canada and with supporting data from Phoenix, Arizona in the United States. The author created a waterfall approach that went from the individual review level, whose results flowed into dynamic clusters of reviews within a time frame, and finally aggregated all cluster states that ended in a global state for a time frame. With this global state present for all time frame steps across the dataset, the objective was to determine if the subsequent global state in the next time frame with respect to the current time frame would meet or exceed the threshold for a good review state across an aggregated threshold of clusters. If the state was lower than the aggregated cluster threshold, then a sample action plan may be taken as described in this paper.

## Table of Contents

## List of Tables and Figures (Summarized Titles)

## 1. Preface

We are inundated with data. The scarcity of content in the past made this deluge more manageable. Unfortunately today, it is unavoidable because much of this data deluge is shaping the perception of our world. Social media and mobile devices keep us constantly connected and very dependent the platforms and apps that shape our perception of the world thus affecting our conversations and decision making. Unfortunately, it seems that the people with the best tools to manage this space are marketers and businesses who specifically target our preferences and intent to nudge us toward behavior that is desirable for them and not necessarily us. All the while, we willingly surrender our data as these organizations build bigger profiles of our digital lives. We have reached a tipping point where individuals not only need to have more power with limiting what data is sent into the ether but also to understand what trends are shaping the data that we see so that we understand why we are seeing what we see and can make more informed choices about these perceptions [7].

The author was moved by these events and trends to try to develop a methodology to organize data in such a way that these time-based trends could be detected at an earlier time before wide-scale negative events occur. Since trends may not be fully understood, it is important to provide a means to detect these undefined trends over time across a wide set of attributes across dynamic segments of data. This is why the project is named Magma. Dictionary.com defines magma as "molten material beneath or within the earth's crust, from which igneous rock is formed." [6]. Metaphorically, this describes how a variety of trends at different speeds occur below the surface of our understanding that solidifies into rocks of knowledge or oozes out from the crevices of conflicts within ourselves and among others. While building a tool around such a methodology is a lofty aspirational goal, it needed to be made tangible for a specific dataset. The remainder of the paper will describe this tracking methodology and apply it to a restaurant review dataset.

The author would like to extend heartfelt gratitude to his professor, Dr. Jim Kulich, his wife, and daughter for their support and encouragement over the last few years.

## 2. Introduction

Restaurants are playing a more prominent role in retail spaces as online shopping has reduced the number of visitors and accompanying sales to storefronts. If the difference between shopping online and in-store is barely distinguishable, most people would favor the convenience of shopping at a 24-hour virtual storefront with home delivery than spending the effort to visit a physical store. For many people, especially younger age groups, it is all about the capturing and sharing the experience. With restaurants, the battle of take-out versus dine-in may tip toward dine-in when coupled with events that need to take place beyond one's home. With this shifting consumer behavior, some of the vacant space in retail areas are being replaced with restaurants. As a result, for business communities, recurrent restaurant activity may likely drive more non-restaurant sales. More importantly, reviews affect the revenue and visits of a restaurant; thus it is important to minimize the negative effects [1, 2, 3].

In a large metropolitan city, like Toronto, Ontario in Canada, the city may pride itself on restaurants with the most unique cuisine or visited most often. However, given the trends of restaurants anchoring a commercial retail center and more classically being anchors in neighborhoods within a city. This is why it matters that the big restaurants and small restaurants are doing well because they are another way to measure the lifeblood of a city. Because of the importance of restaurants to a city, it is important to help monitor this critical community component to ensure that restaurants are performing as well as they are able given the clientele they serve [5]. Through monitoring restaurant review data, one could know when more troubling trends are occurring as opposed to just having a "bad week".

The purpose of this project is to determine if the data can be reorganized into a set of dynamic groups for the predictive model to identify patterns more easily that contribute to a negative state. More specifically, the author would like to prove if the aggregated metrics of dynamic groups and the aggregated count of how many good groups exist for a specific time frame can be used as input to a model to more easily predict if the next time frame will be in a negative or positive state.

To validate this idea, a multi-part study was conducted to business review data from the Yelp public dataset [10] for restaurants in Toronto, Ontario and Phoenix, Arizona. Restaurant reviews focused on Toronto, while the Phoenix data was used to build a word pair dictionary for sentiment using Phoenix reviews. A hierarchical approach was taken to organizing the data. First, individual reviews were considered around a set of performance metrics. Next, the individual reviews were aggregated by discrete, overlapping time frames and dynamically grouped by similar characteristics within each time frame. Each group was marked as to whether they fell below or met or exceeded a performance threshold. Then, all of the group performance indicators were aggregated to determine if a threshold of individual groups was exceeded to put the entire system at risk for a problem. While one would expect that the current metrics for each of the groups would provide a clear picture of predicting if there is a problem now, it may be more challenging to see if a model could be created that would predict one time frame in advance of the current time frame. Using only one time may seem rather simplistic; however, it will serve as a basis in the future to determine how much accuracy can be maintained as the number of weeks increases.

The remainder of the paper will include these sections. First there will be section that explain the business understanding and data understanding of the project.  Next, a section about how the data was prepared will follow. The model and evaluation sections will occur afterward and provide insights into the composition of the model and how the model was selected. Finally, based on the decisions, a section will provide guidance about how to deploy the model and take actions based on the model recommendations.

## 3. Business Understanding

The author completed a majority of the work for this stage of the project by through the completion of earlier stages, and each of the stages created a foundational level from which the next stage was built. The prior stages and the current stage use a public Yelp dataset for business reviews which the author further restricted to restaurant data in Toronto with supporting reference information from Phoenix restaurant reviews. The first stage of the project focused on the fundamentals of an individual review— what data is needed to determine if an individual review is "good" or positive. First, the definition of what a "good review" was determined to be when a reviewer assigned 4 or 5 stars. Then, the author scoured the dataset for metadata on reviewers and restaurants and even tried constructing groups of metrics that were intended to emphasize different patterns that would improve the predictive power of the model. However, it turned out that the best model was a very simple one that focused review text sentiment and length, punctuation rates in the review text, and historical count of peer-marked useful reviews.

The next stage of the project focused upon groups or clusters of reviews bound by a five week time frame with each subsequent time frame stepping forward one week in time. This would allow the time frames to have heavy overlap and possibly moderate the fluctuations that would have occurred without such an overlap. The cluster requires a set of attributes whose points (reviews) will be grouped together in multidimensional space (consider each attribute as a dimension). The set of attributes builds on the success of the individual review model but expands the available attributes to include the following:

- Average bigram sentiment value
- Punctuation rate per number of characters in the review text
    - Exclamation point, period, apostrophe, question mark
- Transformed count of lifetime useful reviews
- Log10 values of the number of lifetime reviews for business and reviewer
- Cluster item count

Each resulting cluster contains the mean value of each of these attributes. Just as there was a metric to determine whether an individual review was “good” (4 or 5 stars), the metric that defined a “good review cluster” needed to incorporate this metric as well. The author decided upon two thresholds that defined two quality levels of cluster. One threshold was that 85% of reviews in a cluster must be “good reviews” in order to classify a cluster as a “good review cluster”. These clusters of reviews were considered to be originating from high-performing restaurants. The other threshold level was that 70% of reviews for a “good review cluster”. Unlike the prior threshold, this threshold was like a “passing grade” for a class. The cluster reviews were minimally acceptable, and failure for a cluster to meet this minimum threshold would imply a cluster was struggling and immediate action would be needed to help the distressed restaurants.

The use of clusters without any preconceived ideas of a target prediction allow the implicit interactions among the attributes to occur more naturally. Additionally, if one retains the cluster number for each individual review, where each review contains ID codes that link to reviewer and restaurant metadata, it is possible to see how the clusters are composed in order to better understand how to treat restaurants in underperforming clusters. Nevertheless, it is important to emphasize that the purpose of the clustering is to focus on the qualitative review aspects of the cluster based on the set of attributes selected. The composition of the cluster is intended to help with targeting decisions for how to help restaurants in underperforming clusters within a given time frame.

The final stage of this project occurred when trying to link together all of the clusters for a given time frame into a single observation (row of data). While the horizontal linking was relatively straightforward because each original cluster observation included time frame number, the vertical linking was hardly straightforward. To facilitate a reasonable method of linking together the clusters, the attributes from the current cluster row and prior cluster row had their cosine similarity values calculated. By rank ordering the pairs current and prior clusters, one could align them more correctly since the earlier decision was made to create a cluster based on the qualitative metrics rather than the composition of restaurants in a cluster. Technical information for how this task was accomplished will be explained in a later section.

To determine if the set of clusters in the current time frame were in a “good overall state”, the author decided to simply count the number of clusters marked as “good review clusters”. If the number of “good review clusters” exceeded the threshold, then the system was in a “good overall state”. Since there were two thresholds for the individual clusters, it was necessary to have two “good overall state” thresholds based on the underlying “good cluster review” thresholds. The 85% “good review cluster” used three “good review clusters” as its threshold, while the 70% “good review cluster” had six “good review clusters” as its threshold. Based on the cumulative distribution of the good reviews for each of the thresholds in Figure. For this aspect of the project, rather than predicting the current state based on the set of attributes being used; instead, the prediction will be for the subsequent time frame.

In summary, the “Magma” methodology is effective because it has a hierarchical method for marking positive events in each layer. The first fundamental layer is the individual review with the most critical set of attributes that allow predictions to occur effectively. The second layer uses the cluster which is the core of the methodology. The cluster aggregates individual reviews around the set of attributes from the first layer within a given time frame and are marked as “good” if the percentage of individual “good reviews” exceeds a threshold. Care should be taken in choosing the length of this time frame as it provides a means of extending out the prediction time with the understanding that predicting farther out will decrease the accuracy of the model. The third layer is the system layer whereby “good” clusters are counted to determine if they are below a threshold which would trigger action encompassing metadata about the reviewers and restaurants captured by the cluster number assigned to individual reviews (see Figure 3-1).

## 4. Data Understanding

As was the case for the last two stages of this project, the author derived the datasets from an earlier version of the Yelp public dataset (only the current version is available) [10]. The overall data covered approximately August 2008 through July 2017. Since the complete dataset across multiple cities consisted of 4.7 million business reviews, 156,639 businesses, and 1.1 million users, it was necessary to filter the data further to only encompass restaurants in the city of Toronto. Toronto data was still significant with 4,475 businesses, and 59,016 users, and 217,620 reviews restricted to the top fifty cuisines. An interesting feature about this dataset (possibly specific to Toronto) is that the number of reviews increased over time. The structure of the overall dataset appears below. A separate dataset for Phoenix, Arizona which was used to create a sentiment bigram dictionary.

Based on the diagram and its relations among tables as shown above, the following list of tables were used for this project:

- **Business:** name, location, historical star rating and review count, and flag for currently in operation.
- **Review:** star rating, review text, date, useful flag
- **User:** reviewer; historical review count, starting year, fan count, historical useful reviews count

The modeling components structured across the earlier projects are summarized in ascending order. The fundamental idea of a “good review” derives from changing the five-star rating to a binary flag where a “good review” is four stars or greater (at most five). With this in place, and from prior analysis, the best model developed during this stage related to review and sentiment attributes. These were described in brief in chapter 3. Next, using the attributes selected for individual reviews, a hierarchical clustering using the squared Euclidean distance for collecting similar reviews and Ward’s method for building smaller clusters into larger clusters. A total of fifteen clusters will be formed for each time frame consisting five weeks with one week forward increments. This was because in the later part of the dataset, some time frames contained 6,000 reviews and using fifteen clusters would, on average, allow most clusters to average around 400 reviews. In the final step, the overall collection of clusters for a given time frame wee joined together into a single observation. At this level, a "good review cluster system state" would be based on the accumulated number of "good review clusters" for all of the clusters in the time frame. If the accumulated number exceeded the threshold number of "good review cluster" for each lower level "good review" threshold (i.e. 70% and 85% of "good reviews") as described in chapter 3. The specific threshold was based on attempting to get a near 50-50 balance of "good review cluster system states". The charts in Figure 4-2 illustrate this further that shows the frequency for each threshold and the temporal distribution across the dataset.

Nevertheless, the creation of the dynamic clusters at each time frame made it difficult to align together the exact clusters for each time frame. Thus it was necessary to construct a method that would align the clusters together as accurately as possible over each of the time frames. As stated earlier, a conscious decision was made to link clusters together based on a qualitative similarity rather than a content-based similarity, that is, a similarity based on similar restaurants, reviewers with their cuisines and locations. Ultimately, for each time frame, each cluster calculated its cosine similarity with another cluster from the time frame resulting in $15^2$ (or 225) calculations. With the elimination of some duplicates to come close to $½n^2$. The most similar pair was used as the first link, followed by the second most similar pair, etc. The next section describes the algorithm in additional technical detail. It seemed that after aligning the clusters across the time frames, changes were generally stable. In Figure 4-3, there is a heatmap by cluster highlighting the actual cosine similarity between the adjacent clusters. Excel was used for the conditional formatting coloring that generated the heatmap. The 3-color scale is 0.7 (and below) for orange, 0.85 for light yellow, and 1 for green. While natural deviations in the data would create significant differences, it seems that with the dominance of dark green that many of the clusters have adjacent similarity and remain stable over time. It is this final dataset of fifteen individual cluster attribute metrics that are used to predict whether the next time frame is in a “good overall state”.

## 5. Data Preparation

A large majority of the data preparation occurred during the earlier stages of the project. While Three significant tasks will be described: important attribute transformations, the creation of a sentiment dictionary, and construction and alignment of review clusters.

As a review, the attributes present in each cluster group of columns are: sentiment, pct_exclam, log10_usr_review_count, pct_quest, review_size, pct_periods, pct_apost, useful_review_count, log10_biz_review_count, item_count, and good_review (more specific definitions appear in chapter 3). Most of these attributes are used as is. Attributes beginning with “pct” for punctuation rates **per character** in the review text are already in an appropriate form that keeps the value between 0 and 1; although these are probably too small if the value range magnitude was more critical. The review count attributes had their value ranges minimized using the log10 transformation. However, cluster item count and review size (number of characters in the review) were kept as is. While these two last attributes were transformed for improving cluster cosine similarity calculations (using the 11th root of their value: $x^{1/11}$ ), they were retained with their original values [11]. As this was not ideal, a random forest technique was used for modeling which would minimize the impact of these attributes having a much more significant value range than any of the other cluster-based attributes.

One of the major items that required construction in the previous project was a sentiment bigram dictionary. Rather than use the same dataset as the restaurant reviews in Toronto, the author decided it would be better to use review text from a separate dataset from

Phoenix. Keep in mind that there are 231,448 reviews for Toronto and 259,950 reviews for Phoenix. The final dictionary had about 1.7 million bigrams and 4.9 million entries in the Toronto bigram table. If the review was "good", all bigrams were marked as 1; otherwise 0. As the bigram values accumulated in the dictionary, a sentiment rate resulted. After the construction of this dictionary based on Phoenix reviews, it was used for bigram sentiment calculation on the text of the Toronto reviews where the mean value of sentiment was calculated based on the bigrams in the review text divided by the total number of bigrams **found in the Phoenix bigram sentiment dictionary.** Excluding missing bigrams and removing "stopwords" were helpful with improving the accuracy of the sentiment calculations.

Aligning the clusters proved to be theoretically simple but somewhat challenging to implement. The underlying concept is straightforward. The cluster numbers that are generated by the clustering algorithm seem to not have any direct relation from one time frame to the next time frame. For this reason, it is necessary to find prior and current clusters between time frames which are statistically similar to each other using cosine similarity for the aggregated metrics of the cluster. Given $n$ clusters ($n=15$ for 15 clusters in this dataset), each cluster similarity pair will be exhaustively calculated ($n * n$ *calculations)* even though the ordering of the elements in the pair does not matter (a more careful set of calculations would result instead in $(n*n)/2$ cluster similarity calculations). In calculating the cosine similarity for the cluster metrics, there were two additional transformations that occurred. The author transformed the attribute for the number of items in the cluster and the number of characters in the review to be $x^{1/11}$ so as to keep the range of potential values no higher than 5 or 10. These transformations helped make things more manageable and improved the quality of clustering.

Once these transformations completed, the attribute data needed to be organized into a matrix of similarity calculations. Rows consisted of the prior time iteration clusters, while columns contained the current time iteration clusters. For each time frame and for each unique cluster pair, the top similarity value is selected. Afterwards, the row of the prior cluster number, and the column of the current cluster number are “zeroed out” to ensure a unique prior cluster number and current cluster number pair do not partially recur within a single time frame.

One other aspect to consider is that the physical assignment to a cluster section within an observation was not straightforward. At first, the initial cluster assignment for the very first time frame is “as-is” (e.g. cluster 1 belongs to cluster 1, cluster 2 belongs to cluster 2, etc.). At the next time iteration, the first round of cosine similarity calculations creates the first link. However, at the third time frame and beyond, it is necessary to link back to the original cluster. This is done by keeping a map (dictionary data structure) that consists of key-value pairs. In a slightly counterintuitive manner, the values are the original clusters, while the keys are the current cluster numbers that link back to the original cluster numbers. Because of this somewhat inverted structure, it is necessary to make a copy of this data structure from the prior time frame before changes are made to the structure in the current time frame. See Figure 5-1 for an example of how this works. Furthermore, as each new link is made, that data is inserted into the appropriate group of columns for each cluster in the observation. Additionally, for reference purposes, three new attributes are appended to each cluster: current cluster number, prior cluster number, cluster similarity value. This set of attributes is repeated fifteen times for the number of clusters used in the dataset.

## 6. Modeling

The data split was done with 80% for training and 20% for validation. Unlike in the prior project, the author did not split based on the time frame. This is because the good cluster count across the dataset when ordered by time frame number produced uneven distributions for the selected good cluster count for each of the 70% and 85% good review threshold (see Figure 4-2). Instead, a random sample was taken across the entire dataset for creating the training and validation datasets. The training and validation models utilized cross-validation with five-fold validation repeated five times.

Because of the nature of the nonlinearity of the data rearranged using the clusters, a suitable technique to use seemed to be random forests. This technique was used for both variable selection and modeling. The author decided to use multiple passes random forest with one pass for variable selection and the other pass for modeling. The author expected that even with a random forest modeling method, the variable importance lists should have been generally similar; however, this turned out not to be the case. Furthermore, sometimes the distribution of importance values in a variable list would drop at different rates. It seemed that when variable importance values had quick drop-offs in their distribution, this seemed to provide a trained model with good accuracy and Kappa value. Perhaps the variability in the lists should have been expected given the nature of the algorithm, yet this made the author question if there was some very odd nonlinear pattern occurring with the many clusters in the datasets with identical attributes despite the data in each cluster being disjoint from the others. Based on some second-hand experience and with this article, perhaps it would have been better to conduct an iterative process for variable selection that utilized taking the top *n* variables or values above some threshold at each iteration or looking for overlap across multiple variable selection iterations.

The author tried to do a manual check of this process by creating a correlation matrix for selected attributes related to target variables (see Figure 6-1) for one of the threshold levels. Then, sorting the last column which was the final target attribute “good system”, there seemed to be some consistency with attributes with absolute values of correlation that were high. As a follow-up test, a model with the top ten (most positive) and bottom ten (most negative) correlation values to see what kind of training model would occur. Surprisingly, the model was generally comparable in accuracy to the other models generated. This somewhat reassured the author that the random forest process for selecting variables was fairly random and, ideally, should be done in an iterative manner to ensure that the strongest attributes come through and are not quirks of randomness. It should be noted that the default random forest parameters were used since no testing was done with optimizing model parameters [4].

Another consideration was the "look ahead" component. As a first pass, the model was built to predict if the current set of clusters was in a bad state. Models for this type of prediction for both the 70% and 85% thresholds did well. The more important step was to be able to see how accurate a model could be at predicting a system state. For one threshold, models were still tolerable beyond one week, and this could be because history is already included in the cluster since each time frame contains five weeks of review data. While the author decided to make models that predicted the system state just one week in advance, experimenting with the creating multiple models consisting of clusters of different time lengths and models that predicted out different lengths could provide the essence of "Magma" whereby multiple layers of prediction could guide the timing and strength of actions depending upon the level of sophistication required.

Figure 6-2 for the results that produced the selected variables for each model. Figure 6-3 lists the training and validation based on the selected variables from the prior step.

## 7. Evaluation

The models did fairly well. By plotting the ROC, as seen in Figure 6-3, it determined the best threshold to optimize the model for the 70% and 85% good reviews which calculated 0.422 and 0.495, respectively [8, 12]. The 70% model produced an accuracy of 77% with a Kappa of 0.544, while the 85% model produced an accuracy of 85% with a Kappa of 0.698. The 85% ROC plot shows good performance in the earlier specificity ranges, while the 70% is not quite as good as the 85% model but still has decent performance in these same areas. After reviewing these accuracy measures within the project charter model goals, the 65% accuracy of predicting a good system for the 85% good review model was met. However, the predicting a good system for the 70% good review model was slightly below the 80% accuracy recommendation goal. The stakeholders can reassess if this is an acceptable gap; meanwhile, some additional attempts to increase model performance can be investigated to boost the accuracy amount higher.

Since no pilot plan occurred, a good next step for real-world evaluation would be to deploy the 85% threshold model, since it is the stronger of the two models, which would focus on monitoring high-performance restaurants and providing detection when something disruptive happens in this group. The 70% threshold group can be deployed as well but used in an evaluation context to see how it performs with live data.

However, one of the key aspects of this project is that it is not just the model that will be deployed. In addition to the models, the underlying database of restaurant and reviewer metadata will be provided as well as the data for individual reviews with cluster assignment information. As part of the action plan, when a bad system state is detected, one would query the datasets and use the tools as specified by the author to come up with initial segments who could benefit from some assistance from the chamber of commerce. (if time permits, provide an example).

## 8. Deployment

The deployment process will be somewhat more involved because of additional parts that require maintenance besides the models. Updated data is necessary to power the system at weekly aggregated levels. If Yelp is the preferred dataset to use, there are two options for getting new data. First, Yelp provides a number of data services for professional use. The Yelp Knowledge product covers local analytics and sentiment. This would provide the data necessary to maintain what was already available in the trail dataset. Additionally, Yelp has a number of partner businesses to help develop solutions. Minimally, purchasing access to Yelp Knowledge is a necessary first step. Additionally, the code as provided, if run in batch on a weekly basis, should be sufficient to run in a proof-of-concept stage; however, for a dependable solution that restaurants throughout Toronto rely on, the data platform needs to be more solid. If other solutions architects from one of  the listed partners needs to extend the capabilities of what was already provided, we could engage with them to share what we have and drive a better solution forward.

Second, the underlying dataset needs to be accessible to a marketing analytics team to evaluate the best way to handle clusters of restaurants that require varying levels of intervention to improve their ratings. At this stage, there is no dashboard or tool available and the data would just be as-is. However, to make the data more accessible, pre-packaged, basic queries and reports can be run (sent via email) that would provide insight into what is happening. Training would be provided to ensure that the analysts could operate in a self-sufficient manner. Existing database and programming tools would be used with the assumption that analysts who use the system will have an intermediate level knowledge of how to use these tools to run queries and reports and be able to do basic debgging if things are not looking correct.

Finally, the weekly data batch jobs that run will have data quality monitoring built-in as one set of outputs produced by this process. The reports will use traditional monitoring techniques around static metrics for data about aggregated reviews, restaurants, and reviewers. This aggregated data will be used to run segment comparisons over time using z-scores and t-tests to ensure that things are remaining stable. Additionally, the system itself can use a version of Magma without the modeling component to monitor the restaurant models and quickly assess if there are problems and, with the clusters, quickly find solutions to the problems.

## 9. Conclusion

The Magma methodology is an effective method for organizing data based on a multi-tiered approach to detect signature behaviors across the system and use the underlying data to construct strategies for action. While this is an unsupervised learning method, the intended target attribute that defines the primary event should be included in the cluster. First, find the attribute set that best allows for the prediction of the fundamental target attribute. Then, decide on a time range to collect the individual observations or records using the fundamental dataset. Next, perform hierarchical clustering using Euclidean distance and Ward's method for building larger clusters from smaller clusters and obtain the number of clusters needed to monitor and describe the set of observations. Afterwards, align the clusters using a short-term approach whereby the prior cluster and the current cluster have the highest possible cosine similarity. Finally, decide on the threshold of good clusters, build a model around the fundamental attribute set used to describe an individual record. Since the cluster has a collection of records that spans a specified time frame, the system is capable of predicting when the bad system will occur of at least the minimum aggregation unit out. Moving farther out in time will significantly decrease the accuracy of the model but must be weighted depending on what needs to be accomplished.

For monitoring restaurants using the public Yelp dataset for the city of Toronto, the fundamental target attribute is the definition of a good review, simplified as a star rating of four or five stars. The best attribute set was focused on sentiment details of a review some of which was selected punctuation rates, review size, lifetime review counts of reviewers and businesses, etc. This attribute set was used to aggregate reviews over a five week time frame with one week increments to generate fifteen clusters. Two models were made based on the good review thresholds, with a prediction of one week ahead. The 70% good review threshold was the minimum viable restaurant cluster where if six clusters met or exceeded the threshold, then the system would be in a good state one week from now. If the 85% good review threshold model were used instead, the system will be in a good state one week from now if the number of good review clusters in this time iteration is three or more. Should the system be in a bad state, then analysts can use the dataset to identify the problematic clusters to obtain not only the cluster metrics but also the contents of the reviews aggregated in these clusters and similarly the metadata for the restaurants and reviewers to find trends for what other driving factors may be problematic within the cluster. With additional attributes and a good framework to query the data, this system will be powerful in helping detect nonlinear trends for struggling restaurants and empower analysts to develop strategies to help them get back on track.

While the Magma methodology was used to predict a future system state across all clusters, it can be used as a general purpose monitoring system (even for monitoring the performance of the restaurant prediction models). The monitoring would focus on tracking individual clusters, which along with cluster alignment, be able to distinguish when a cluster is starting to fail. If this were combined with a classifier algorithm that was trained to recognize signatures that indicate potential problems with the system, one could enhance the dataset to make recommendations based on the cluster's signatures as evidenced by the clustering technique I described earlier. With the limited success for this particular domain, it seems likely Magma could be applied to other domains requiring monitoring and action.

## 10. List of Figures and Tables

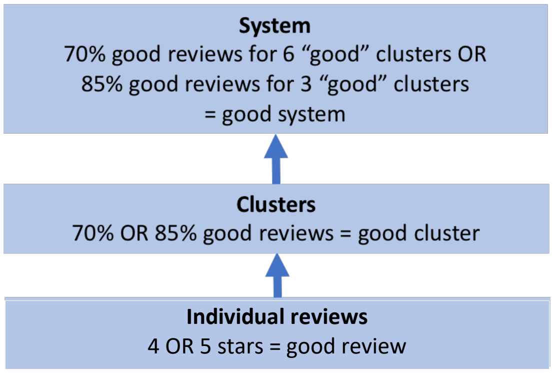
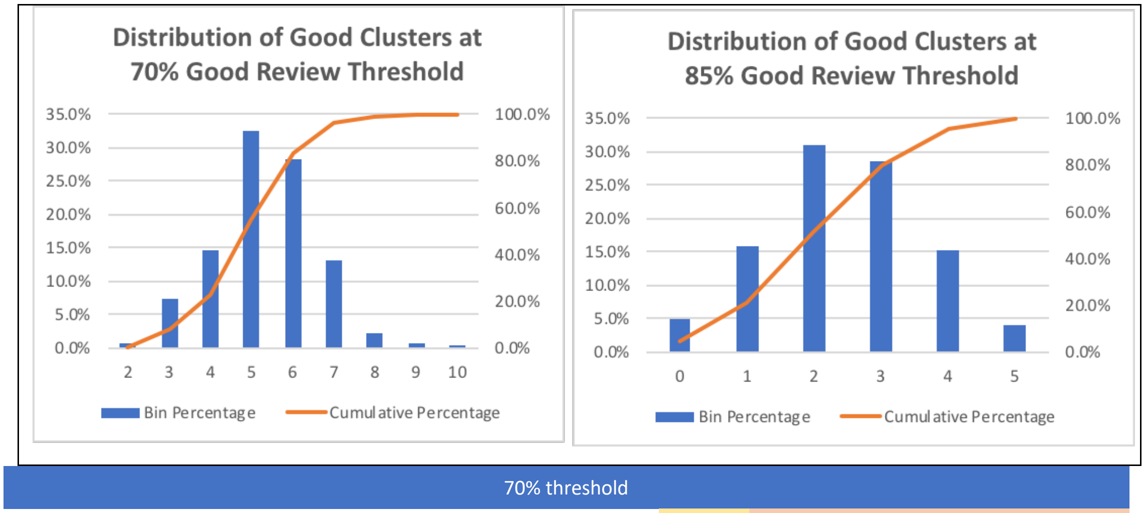

System
70% good reviews for 6 “good” clusters OR
85% good reviews for 3 “good” clusters
= good system
Clusters
70% OR 85% good reviews = good cluster
Individual reviews
4 OR 5 stars = good review

Figure 3-1. Magma target attributes in each layer



| 70% threshold | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **good_cluster_count** | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **bin percentage** | 0.6% | 7.5% | 14.7% | 32.4% | 28.4% | 13.0% | 2.3% | 0.9% | 0.2% |
| **cumulative percentage** | 0.6% | 8.1% | 22.8% | 55.2% | 83.6% | 96.6% | 98.9% | 99.8% | 100.0% |

| 85% threshold | | | | | | | |
|---|---|---|---|---|---|---|---|
| **good_cluster_count** | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| **bin percentage** | 4.9% | 15.8% | 30.9% | 28.6% | 15.1% | 4.1% | 0.6% |
| **cumulative percentage** | 4.9% | 20.7% | 51.6% | 80.2% | 95.3% | 99.4% | 100.0% |

Figure 3-2. Determine the thresholds for good clusters at the 70% and 85% threshold. It was preferable to get close to a 50% occurrence of a good cluster event; therefore, using 6 and 3 good clusters were used as a threshold.

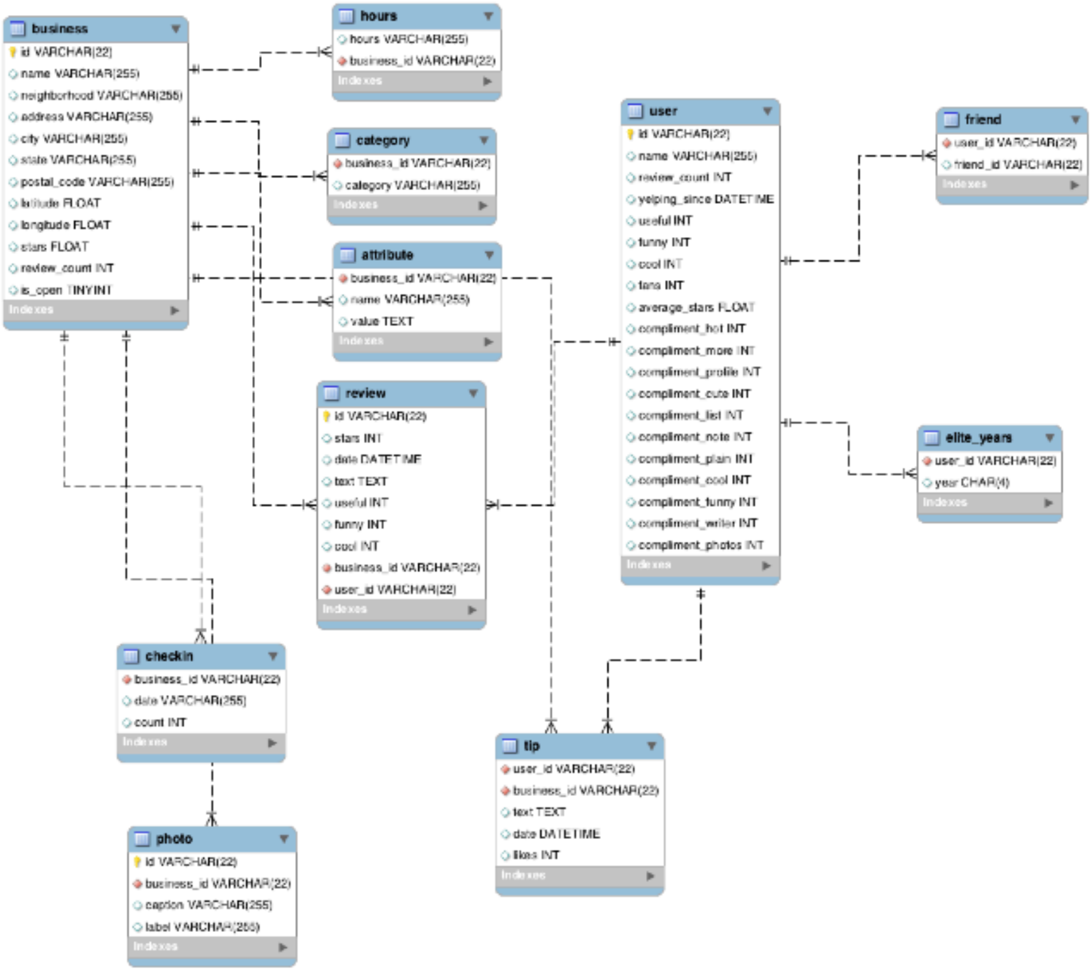


Figure 4-1. Data structures in the Yelp dataset

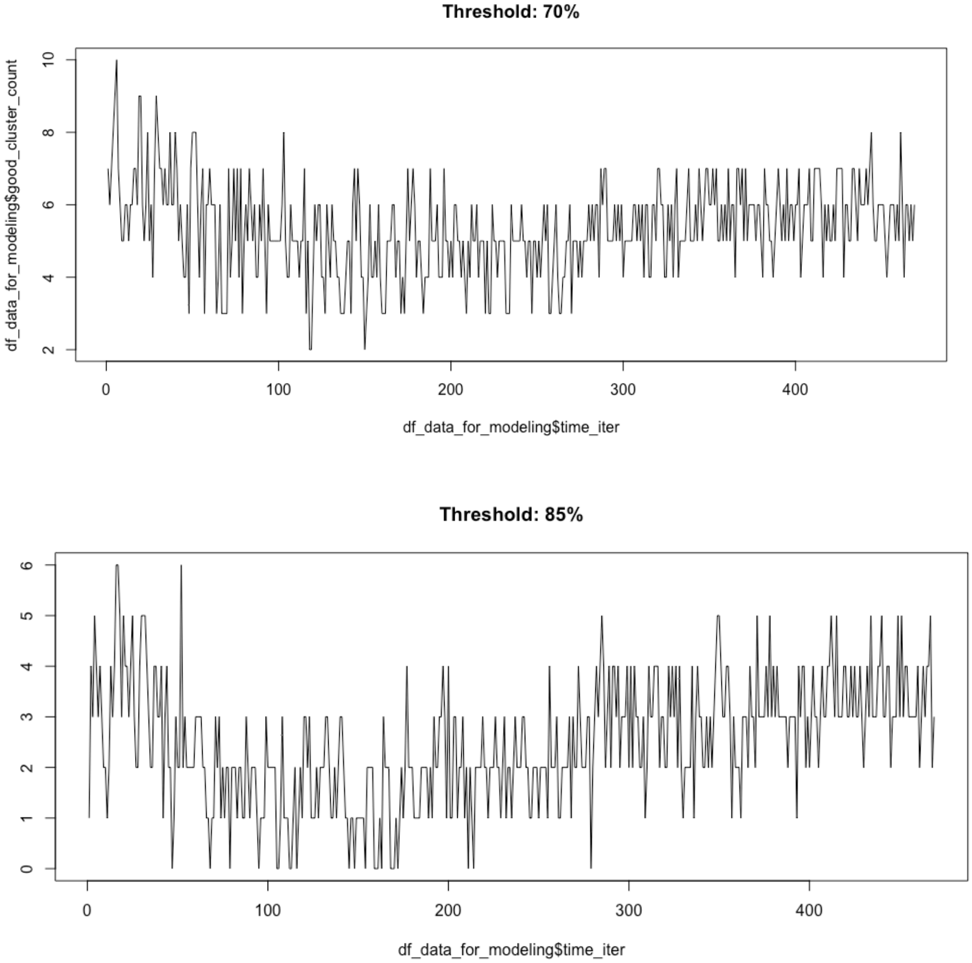
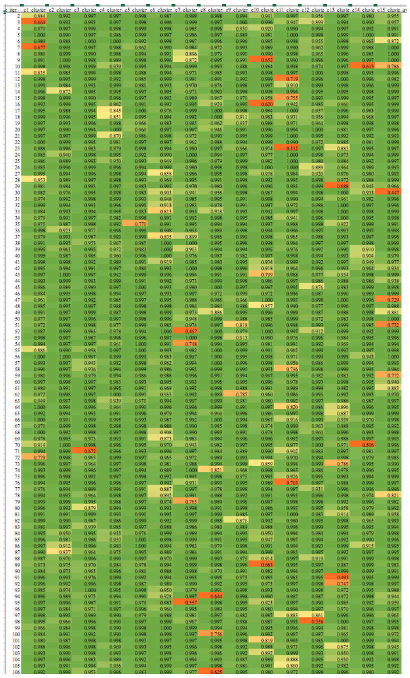
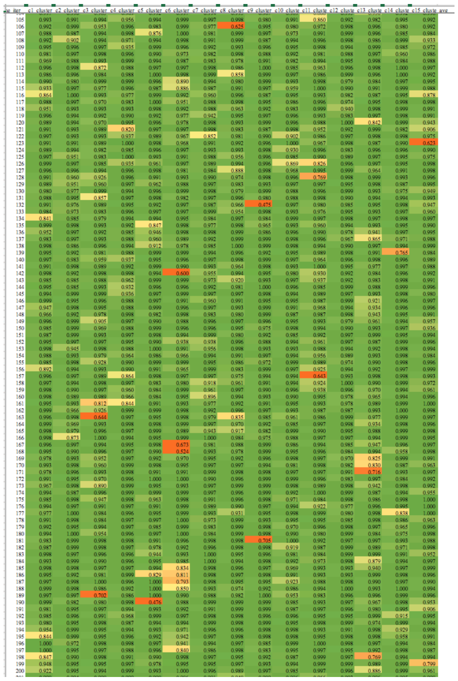
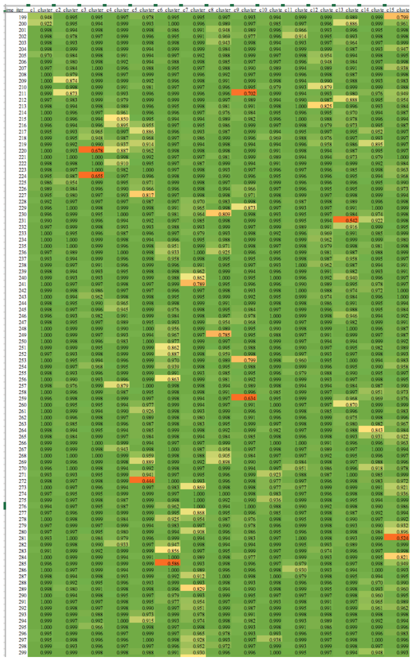


Figure 4-2. Accumulation of “good review clusters” across the dataset by “good review” threshold.







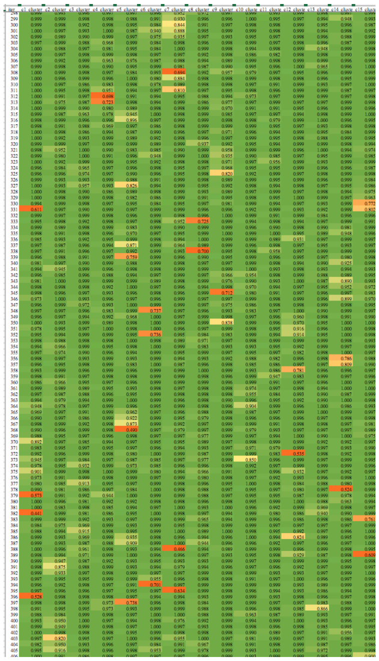


Table 4-1. Selection of heatmap illustrating the absolute value of cluster similarity; approximately 400 time frames; cluster 1 is on the left and cluster 15 is on the right

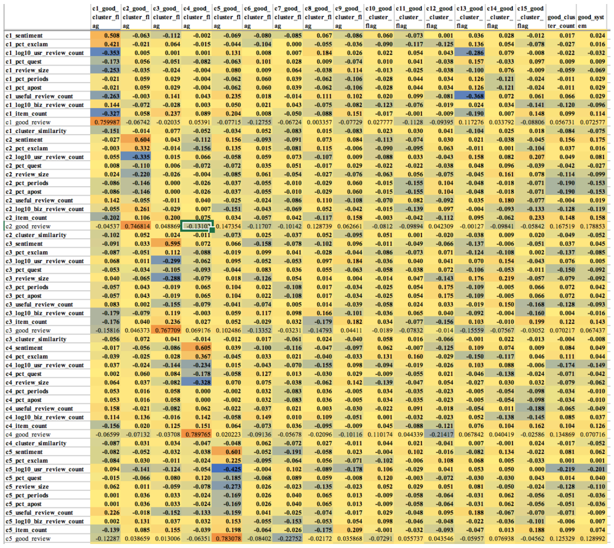
| | | | | |
|---|---|---|---|---|
| 0.8391 | 0.5448 | 0.3453 | 0.5206 | 0.179 |
| 0.1262 | 0.0097 | 0.3235 | 0.5648 | 0.5471 |
| 0.875 | 0.7605 | 0.2688 | 0.4565 | 0.6538 |
| 0.5955 | 0.316 | 0.4261 | 0.9349 | 0.4217 |
| 0.4258 | 0.9331 | 0.9458 | 0.4557 | 0.5609 |

| | | | | |
|---|---|---|---|---|
| 0.8391 | 0.5448 | 0.0000 | 0.5206 | 0.1790 |
| 0.1262 | 0.0097 | 0.0000 | 0.5648 | 0.5471 |
| 0.8750 | 0.7605 | 0.0000 | 0.4565 | 0.6538 |
| 0.5955 | 0.3160 | 0.0000 | 0.9349 | 0.4217 |
| 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

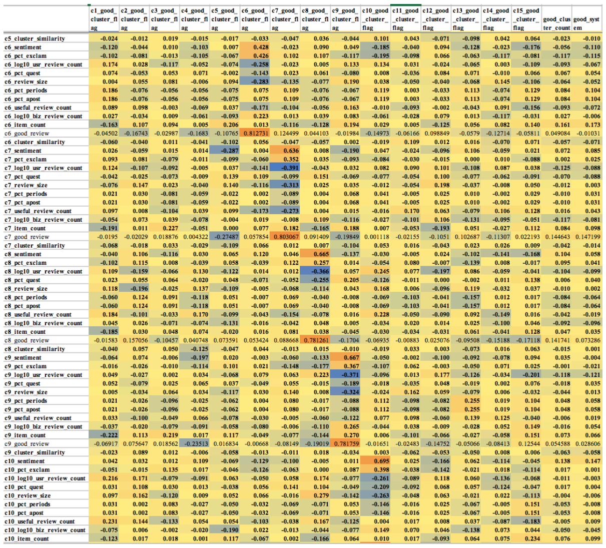
| time_iter | cluster_entry_num | prior_cluster_num (source) | current_cluster (current) | similarity |
|---|---|---|---|---|
| 2 | 15 | 1 | 14 | 0.88424774 |
| 2 | 9 | 2 | 3 | 0.99156839 |
| 2 | 4 | 3 | 8 | 0.99747687 |
| 2 | 6 | 4 | 2 | 0.99689249 |
| 2 | 3 | 5 | 6 | 0.99797178 |
| ... | ... | ... | ... | ... |
| 3 | 15 | 14 | 15 | 0.66378923 |
| 3 | 10 | 3 | 5 | 0.99231887 |
| 3 | 8 | 8 | 4 | 0.99507363 |
| 3 | 5 | 2 | 8 | 0.99726415 |
| 3 | 3 | 6 | 11 | 0.99760239 |
| ... | ... | ... | ... | ... |

| **time_iter = 1** | **time_iter = 2** | **time_iter = 3** |
|---|---|---|
| Current: Source<br>1: 1<br>2: 2<br>3: 3<br>4: 4<br>5: 5<br>… | Current: Source<br>14: 1<br>3: 2<br>8: 3<br>2: 4<br>6: 5<br>… | Current: Source<br>15: 1<br>5: 2<br>4: 3<br>8: 4<br>11: 5<br>… |

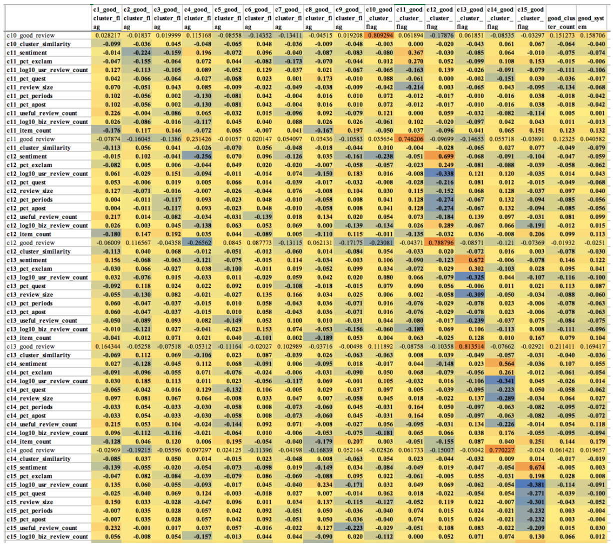
Table 5-1. (Top) A simulated similarity matrix where a row and column represent the prior and current clusters, respectively. After the maximum value is located in the matrix, the coordinates are recorded for tracking purposes and all values in the adjacent cells in the row and column are assigned zero. The process is repeated selecting the next highest value in the matrix, assigning zero to the row and column, and continuing until all cells are zero. (Middle) Based on the activity over each time frame with its similarity matrix, rows are created for the cluster tracking table. (Bottom) Based on the contents of the cluster tracking table, the cluster key-value map for tracking the new clusters to their original source clusters is done. The value is the source cluster and new keys are assigned for each new cluster in the time frame as given by the cluster tracking table.



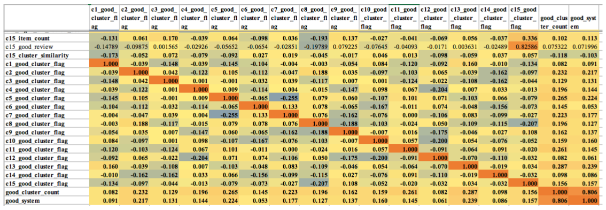
| | c1_good_cluster_flag | c2_good_cluster_flag | c3_good_cluster_flag | c4_good_cluster_flag | c5_good_cluster_flag | c6_good_cluster_flag | c7_good_cluster_flag | c8_good_cluster_flag | c9_good_cluster_flag | c10_good_cluster_flag | c11_good_cluster_flag | c12_good_cluster_flag | c13_good_cluster_flag | c14_good_cluster_flag | c15_good_cluster_flag | good_cluster_count | good_system |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c1_sentiment | 0.508 | -0.063 | -0.112 | -0.002 | -0.069 | -0.080 | -0.085 | 0.067 | -0.086 | 0.060 | -0.073 | 0.001 | 0.036 | 0.028 | -0.012 | 0.017 | 0.024 |
| c1_pct_exclam | 0.421 | -0.021 | 0.064 | -0.015 | -0.044 | -0.104 | 0.000 | -0.055 | -0.036 | -0.090 | -0.117 | -0.125 | 0.136 | 0.054 | -0.078 | -0.027 | 0.016 |
| c1_log10_usr_review_count | -0.353 | 0.005 | 0.001 | 0.001 | 0.131 | 0.008 | 0.007 | 0.184 | 0.026 | 0.022 | 0.054 | 0.043 | -0.286 | 0.079 | -0.008 | -0.022 | -0.032 |
| c1_pct_quest | -0.173 | 0.056 | -0.051 | -0.082 | -0.063 | 0.101 | 0.028 | 0.009 | -0.074 | 0.010 | 0.041 | -0.038 | 0.157 | -0.033 | 0.097 | 0.009 | 0.009 |
| c1_review_size | -0.253 | -0.035 | -0.024 | -0.004 | 0.080 | 0.009 | 0.064 | -0.038 | 0.114 | -0.013 | -0.025 | -0.038 | -0.100 | 0.076 | -0.009 | -0.059 | -0.069 |
| c1_pct_periods | -0.021 | 0.059 | 0.029 | -0.004 | -0.062 | 0.060 | 0.039 | -0.062 | -0.106 | -0.028 | 0.044 | 0.034 | 0.126 | -0.121 | -0.024 | -0.011 | 0.029 |
| c1_pct_apost | -0.021 | 0.059 | 0.029 | -0.004 | -0.062 | 0.060 | 0.039 | -0.062 | -0.106 | -0.028 | 0.044 | 0.034 | 0.126 | -0.121 | -0.024 | -0.011 | 0.029 |
| c1_useful_review_count | -0.263 | -0.003 | 0.141 | 0.043 | 0.235 | 0.018 | -0.014 | 0.111 | 0.102 | 0.020 | 0.099 | -0.081 | -0.368 | 0.072 | 0.061 | 0.066 | 0.029 |
| c1_log10_biz_review_count | 0.144 | -0.072 | -0.028 | 0.003 | 0.050 | 0.021 | 0.043 | -0.075 | -0.082 | -0.123 | -0.076 | -0.019 | 0.024 | 0.034 | -0.141 | -0.120 | -0.096 |
| c1_item_count | -0.327 | 0.058 | 0.237 | 0.089 | 0.204 | 0.008 | -0.050 | -0.088 | 0.151 | -0.017 | -0.001 | -0.009 | -0.190 | 0.007 | 0.148 | 0.099 | 0.114 |
| c1_good_review | 0.759987 | -0.06742 | -0.02035 | 0.05391 | -0.07715 | -0.12755 | -0.06724 | 0.003357 | -0.07729 | 0.027777 | -0.1128 | -0.09395 | 0.117276 | 0.033792 | -0.08806 | 0.056731 | 0.072577 |
| c1_cluster_similarity | -0.151 | -0.014 | 0.077 | -0.052 | -0.034 | 0.052 | -0.083 | 0.015 | -0.083 | 0.023 | 0.030 | 0.041 | -0.104 | 0.025 | 0.018 | -0.084 | -0.075 |
| c2_sentiment | -0.027 | 0.604 | 0.043 | -0.112 | 0.156 | -0.093 | -0.091 | 0.073 | 0.084 | -0.113 | -0.074 | 0.030 | 0.021 | -0.038 | -0.045 | 0.156 | 0.175 |
| c2_pct_exclam | -0.003 | 0.332 | -0.014 | -0.156 | 0.135 | 0.015 | -0.081 | 0.115 | -0.006 | -0.090 | -0.095 | 0.063 | -0.011 | 0.001 | -0.104 | 0.037 | 0.016 |
| c2_log10_usr_review_count | 0.055 | -0.335 | 0.015 | 0.066 | -0.058 | 0.059 | 0.073 | -0.107 | 0.009 | -0.088 | 0.033 | -0.043 | 0.158 | 0.082 | 0.207 | 0.049 | 0.081 |
| c2_pct_quest | 0.008 | -0.110 | 0.006 | -0.072 | -0.072 | 0.035 | 0.051 | -0.017 | 0.029 | -0.022 | -0.022 | -0.038 | 0.048 | 0.096 | -0.039 | -0.042 | -0.027 |
| c2_review_size | 0.024 | -0.220 | -0.026 | -0.004 | -0.085 | 0.061 | -0.054 | -0.027 | -0.076 | -0.063 | 0.056 | -0.075 | -0.045 | 0.161 | 0.078 | -0.114 | -0.099 |
| c2_pct_periods | -0.086 | -0.146 | 0.000 | -0.026 | -0.037 | -0.055 | -0.010 | -0.029 | 0.060 | -0.015 | -0.155 | 0.104 | -0.048 | -0.018 | -0.071 | -0.190 | -0.153 |
| c2_pct_apost | -0.086 | -0.146 | 0.000 | -0.026 | -0.037 | -0.055 | -0.010 | -0.029 | 0.060 | -0.015 | -0.155 | 0.104 | -0.048 | -0.018 | -0.071 | -0.190 | -0.153 |
| c2_useful_review_count | 0.142 | -0.055 | -0.011 | 0.040 | -0.025 | -0.024 | -0.086 | 0.110 | -0.108 | -0.070 | 0.082 | -0.092 | 0.035 | 0.180 | -0.077 | -0.004 | 0.019 |
| c2_log10_biz_review_count | -0.055 | 0.261 | -0.029 | 0.007 | -0.151 | -0.043 | -0.069 | 0.052 | -0.042 | -0.015 | -0.139 | 0.097 | -0.004 | -0.093 | -0.133 | -0.128 | -0.119 |
| c2_item_count | -0.202 | 0.106 | 0.200 | 0.075 | 0.034 | -0.057 | 0.042 | -0.117 | 0.158 | -0.003 | -0.042 | -0.112 | 0.095 | -0.062 | 0.233 | 0.148 | 0.158 |
| c2_good_review | -0.04537 | 0.746814 | 0.048869 | -0.1310 | 0.147354 | -0.11707 | -0.10142 | 0.128739 | 0.062661 | -0.0812 | -0.09894 | 0.042309 | -0.00127 | -0.09841 | -0.05842 | 0.167519 | 0.178853 |
| c2_cluster_similarity | -0.102 | 0.052 | 0.024 | -0.011 | -0.073 | 0.025 | -0.037 | 0.052 | -0.095 | 0.051 | 0.001 | -0.020 | -0.038 | 0.009 | 0.020 | -0.049 | -0.052 |
| c3_sentiment | -0.091 | 0.033 | 0.595 | 0.072 | 0.066 | -0.158 | -0.078 | -0.102 | 0.096 | -0.011 | -0.049 | -0.066 | -0.137 | -0.006 | -0.051 | 0.037 | 0.045 |
| c3_pct_exclam | -0.087 | -0.051 | 0.112 | -0.088 | -0.019 | 0.099 | 0.041 | -0.028 | -0.044 | -0.086 | -0.073 | 0.071 | -0.124 | -0.108 | 0.002 | -0.137 | -0.085 |
| c3_log10_usr_review_count | 0.068 | 0.011 | -0.299 | -0.062 | 0.095 | -0.052 | -0.053 | 0.097 | 0.184 | -0.036 | 0.040 | 0.041 | 0.070 | 0.154 | -0.043 | 0.076 | 0.005 |
| c3_pct_quest | -0.053 | -0.034 | -0.105 | -0.093 | -0.044 | 0.083 | 0.036 | 0.055 | -0.063 | -0.058 | -0.038 | 0.072 | -0.106 | -0.053 | -0.011 | -0.150 | -0.092 |
| c3_review_size | 0.040 | -0.065 | -0.288 | -0.079 | 0.018 | -0.126 | 0.054 | 0.014 | 0.004 | -0.014 | 0.047 | -0.143 | 0.176 | 0.219 | -0.057 | -0.079 | -0.092 |
| c3_pct_periods | -0.057 | 0.043 | -0.019 | 0.065 | 0.104 | 0.022 | -0.108 | 0.017 | -0.034 | -0.025 | 0.054 | 0.175 | -0.109 | -0.005 | 0.066 | 0.072 | 0.042 |
| c3_pct_apost | -0.057 | 0.043 | -0.019 | 0.065 | 0.104 | 0.022 | -0.108 | 0.017 | -0.034 | -0.025 | 0.054 | 0.175 | -0.109 | -0.005 | 0.066 | 0.072 | 0.042 |
| c3_useful_review_count | 0.083 | 0.002 | -0.155 | -0.079 | -0.041 | -0.074 | 0.005 | 0.014 | -0.039 | -0.058 | 0.024 | 0.033 | -0.019 | 0.150 | -0.168 | -0.128 | -0.093 |
| c3_log10_biz_review_count | -0.179 | -0.079 | 0.119 | -0.003 | 0.059 | 0.117 | 0.098 | 0.166 | -0.101 | -0.036 | 0.065 | 0.040 | -0.092 | -0.004 | -0.160 | 0.004 | -0.016 |
| c3_item_count | -0.176 | 0.040 | 0.236 | 0.027 | 0.052 | -0.029 | 0.032 | -0.179 | 0.182 | 0.034 | -0.077 | -0.156 | 0.103 | -0.010 | 0.199 | 0.122 | 0.143 |
| c3_good_review | -0.15816 | 0.046373 | 0.767709 | 0.069176 | 0.102486 | -0.13352 | -0.03231 | -0.14793 | 0.04411 | -0.0189 | -0.07832 | -0.014 | -0.15559 | -0.07567 | -0.03052 | 0.070217 | 0.067437 |
| c3_cluster_similarity | -0.056 | 0.072 | 0.041 | -0.014 | -0.012 | 0.017 | -0.061 | 0.024 | -0.040 | 0.058 | 0.016 | -0.066 | -0.001 | 0.022 | -0.013 | -0.004 | -0.008 |
| c4_sentiment | -0.017 | -0.056 | -0.086 | 0.605 | 0.039 | -0.100 | -0.116 | -0.047 | -0.097 | 0.062 | -0.007 | -0.125 | 0.109 | 0.074 | 0.009 | 0.084 | 0.049 |
| c4_pct_exclam | -0.039 | -0.025 | 0.028 | 0.367 | -0.045 | 0.033 | 0.021 | -0.040 | -0.033 | 0.131 | 0.160 | -0.029 | -0.150 | -0.117 | 0.046 | 0.111 | 0.044 |
| c4_log10_usr_review_count | 0.037 | -0.024 | -0.144 | -0.234 | 0.015 | -0.043 | -0.070 | -0.155 | 0.098 | -0.094 | -0.019 | -0.026 | 0.103 | 0.088 | -0.006 | -0.174 | -0.149 |
| c4_pct_quest | 0.002 | 0.060 | 0.084 | -0.178 | -0.058 | 0.127 | 0.013 | -0.030 | 0.029 | -0.009 | -0.055 | 0.021 | -0.046 | -0.138 | -0.024 | -0.071 | -0.042 |
| c4_review_size | 0.064 | 0.037 | -0.082 | -0.328 | 0.070 | 0.075 | -0.038 | -0.062 | 0.142 | -0.139 | -0.047 | 0.054 | -0.027 | 0.030 | 0.032 | -0.079 | -0.062 |
| c4_pct_periods | 0.053 | 0.016 | 0.058 | 0.000 | -0.002 | 0.032 | -0.083 | 0.036 | -0.005 | 0.034 | -0.035 | -0.023 | -0.005 | -0.054 | -0.098 | -0.034 | -0.010 |
| c4_pct_apost | 0.053 | 0.016 | 0.058 | 0.000 | -0.002 | 0.032 | -0.083 | 0.036 | -0.005 | 0.034 | -0.035 | -0.023 | -0.005 | -0.054 | -0.098 | -0.034 | -0.010 |
| c4_useful_review_count | 0.158 | -0.021 | -0.082 | 0.062 | -0.022 | -0.037 | 0.021 | 0.003 | -0.030 | -0.022 | 0.091 | -0.018 | -0.054 | 0.011 | -0.188 | -0.065 | -0.049 |
| c4_log10_biz_review_count | 0.114 | 0.136 | -0.016 | 0.142 | -0.058 | 0.149 | 0.010 | 0.109 | -0.051 | 0.001 | -0.032 | -0.023 | 0.052 | -0.138 | -0.145 | 0.085 | 0.037 |
| c4_item_count | -0.156 | 0.020 | 0.125 | 0.151 | 0.064 | -0.073 | 0.036 | -0.095 | -0.009 | 0.045 | -0.088 | -0.121 | 0.076 | 0.104 | 0.162 | 0.104 | 0.126 |
| c4_good_review | -0.06599 | -0.07132 | -0.03708 | 0.789765 | 0.020223 | -0.09136 | -0.05678 | -0.02096 | -0.10116 | 0.110174 | 0.044339 | -0.21417 | 0.067842 | 0.040419 | -0.02586 | 0.134869 | 0.070716 |
| c4_cluster_similarity | -0.087 | 0.031 | 0.034 | -0.047 | -0.048 | 0.062 | -0.072 | 0.027 | -0.011 | 0.044 | 0.021 | -0.041 | 0.007 | -0.001 | 0.024 | -0.017 | -0.052 |
| c5_sentiment | -0.082 | -0.052 | -0.032 | -0.038 | 0.601 | -0.052 | -0.191 | -0.058 | 0.023 | -0.004 | 0.102 | -0.016 | -0.082 | 0.134 | -0.022 | 0.081 | 0.062 |
| c5_pct_exclam | -0.084 | 0.030 | -0.011 | -0.024 | 0.225 | -0.095 | -0.064 | 0.056 | -0.071 | -0.102 | -0.006 | 0.108 | 0.068 | 0.005 | -0.033 | 0.001 | 0.001 |
| c5_log10_usr_review_count | 0.094 | -0.141 | -0.124 | -0.054 | -0.425 | -0.004 | 0.102 | -0.089 | -0.178 | 0.106 | -0.029 | 0.041 | 0.053 | 0.050 | 0.000 | -0.219 | -0.201 |
| c5_pct_quest | -0.015 | -0.066 | 0.080 | 0.120 | -0.185 | -0.068 | 0.089 | 0.059 | -0.008 | 0.120 | -0.003 | -0.072 | -0.030 | -0.030 | 0.043 | 0.014 | 0.040 |
| c5_review_size | 0.062 | 0.011 | -0.059 | -0.078 | -0.273 | 0.026 | -0.023 | -0.135 | -0.023 | 0.052 | 0.029 | 0.051 | 0.081 | -0.050 | -0.024 | -0.128 | -0.110 |
| c5_pct_periods | 0.001 | 0.036 | 0.033 | -0.024 | -0.169 | 0.026 | 0.040 | 0.065 | 0.013 | -0.009 | -0.058 | -0.064 | -0.031 | 0.062 | -0.056 | -0.051 | -0.036 |
| c5_pct_apost | 0.001 | 0.036 | 0.033 | -0.024 | -0.169 | 0.026 | 0.040 | 0.065 | 0.013 | -0.009 | -0.058 | -0.064 | -0.031 | 0.062 | -0.056 | -0.051 | -0.036 |
| c5_useful_review_count | 0.226 | -0.018 | -0.152 | -0.133 | -0.159 | 0.041 | -0.025 | -0.074 | -0.017 | 0.029 | -0.048 | 0.095 | 0.188 | -0.067 | -0.070 | -0.071 | -0.009 |
| c5_log10_biz_review_count | 0.002 | 0.131 | 0.037 | 0.032 | 0.153 | -0.055 | -0.153 | -0.053 | 0.054 | 0.098 | -0.046 | -0.048 | -0.022 | -0.036 | -0.101 | -0.006 | 0.007 |
| c5_item_count | -0.139 | 0.085 | 0.155 | -0.039 | 0.198 | -0.064 | -0.026 | -0.175 | 0.209 | -0.001 | -0.032 | -0.093 | -0.047 | 0.013 | 0.188 | 0.099 | 0.074 |
| c5_good_review | -0.12287 | 0.038659 | 0.013006 | -0.06351 | 0.783078 | -0.08402 | -0.22752 | -0.02172 | 0.035868 | -0.07291 | 0.055737 | 0.043546 | -0.05957 | 0.076938 | -0.04562 | 0.125329 | 0.128992 |



| | c1_good_cluster_flag | c2_good_cluster_flag | c3_good_cluster_flag | c4_good_cluster_flag | c5_good_cluster_flag | c6_good_cluster_flag | c7_good_cluster_flag | c8_good_cluster_flag | c9_good_cluster_flag | c10_good_cluster_flag | c11_good_cluster_flag | c12_good_cluster_flag | c13_good_cluster_flag | c14_good_cluster_flag | c15_good_cluster_flag | good_cluster_count | good_system |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c5_cluster_similarity | -0.024 | -0.012 | 0.019 | -0.015 | -0.017 | -0.033 | -0.047 | 0.036 | -0.044 | 0.101 | 0.043 | -0.071 | -0.098 | 0.042 | 0.064 | -0.023 | -0.010 |
| c6_sentiment | -0.120 | -0.044 | 0.010 | -0.103 | 0.007 | 0.428 | -0.023 | 0.090 | 0.049 | -0.185 | -0.040 | 0.094 | -0.128 | -0.023 | -0.176 | -0.056 | -0.110 |
| c6_pct_exclam | -0.102 | -0.081 | -0.013 | -0.105 | -0.067 | 0.426 | 0.102 | 0.107 | -0.117 | -0.195 | -0.098 | 0.066 | -0.063 | -0.117 | -0.081 | -0.117 | -0.115 |
| c6_log10_usr_review_count | 0.174 | 0.028 | -0.117 | -0.052 | -0.074 | -0.258 | -0.023 | 0.005 | 0.133 | 0.134 | 0.031 | -0.024 | -0.065 | 0.003 | -0.109 | -0.093 | -0.067 |
| c6_pct_quest | 0.074 | -0.053 | 0.053 | 0.071 | -0.002 | -0.143 | 0.023 | 0.061 | -0.080 | 0.008 | -0.036 | 0.084 | 0.071 | -0.010 | 0.066 | 0.067 | 0.054 |
| c6_review_size | 0.004 | 0.055 | 0.081 | -0.006 | 0.094 | -0.283 | -0.135 | -0.077 | 0.190 | 0.038 | -0.050 | -0.040 | -0.068 | 0.145 | -0.106 | -0.064 | -0.052 |
| c6_pct_periods | 0.186 | -0.076 | -0.056 | -0.056 | -0.075 | 0.075 | 0.109 | -0.076 | -0.067 | 0.119 | 0.003 | -0.033 | 0.113 | -0.074 | 0.129 | 0.084 | 0.104 |
| c6_pct_apost | 0.186 | -0.076 | -0.056 | -0.056 | -0.075 | 0.075 | 0.109 | -0.076 | -0.067 | 0.119 | 0.003 | -0.033 | 0.113 | -0.074 | 0.129 | 0.084 | 0.104 |
| c6_useful_review_count | 0.089 | 0.098 | -0.003 | -0.069 | 0.037 | -0.171 | -0.104 | -0.056 | 0.163 | -0.010 | -0.093 | -0.002 | -0.043 | 0.091 | -0.156 | -0.093 | -0.072 |
| c6_log10_biz_review_count | 0.027 | -0.034 | 0.009 | -0.061 | -0.093 | 0.223 | 0.013 | 0.039 | 0.083 | -0.061 | -0.028 | 0.079 | 0.013 | -0.117 | -0.031 | 0.027 | -0.006 |
| c6_item_count | -0.163 | 0.107 | 0.094 | 0.005 | 0.206 | 0.013 | -0.116 | -0.128 | 0.194 | 0.029 | 0.005 | -0.125 | 0.056 | 0.082 | 0.140 | 0.161 | 0.173 |
| c6_good_review | -0.04502 | -0.16743 | -0.02987 | -0.1683 | -0.10765 | 0.812731 | 0.124499 | 0.044103 | -0.01984 | -0.14973 | -0.06166 | 0.098849 | -0.0579 | -0.12714 | -0.05811 | 0.049084 | -0.01031 |
| c6_cluster_similarity | -0.060 | -0.040 | 0.011 | -0.041 | -0.102 | 0.056 | -0.047 | -0.057 | 0.002 | -0.019 | 0.109 | 0.012 | 0.016 | -0.070 | 0.071 | -0.057 | -0.071 |
| c7_sentiment | 0.026 | -0.059 | 0.015 | 0.014 | -0.287 | 0.004 | 0.636 | 0.008 | -0.190 | 0.047 | -0.024 | -0.096 | 0.106 | -0.059 | 0.021 | 0.072 | 0.085 |
| c7_pct_exclam | 0.093 | 0.081 | -0.079 | -0.011 | -0.099 | -0.060 | 0.352 | 0.035 | -0.093 | -0.084 | -0.030 | -0.015 | 0.000 | 0.010 | -0.088 | 0.002 | 0.025 |
| c7_log10_usr_review_count | 0.124 | -0.107 | -0.092 | -0.005 | 0.037 | -0.141 | -0.391 | -0.043 | 0.032 | 0.082 | 0.090 | 0.101 | -0.108 | 0.087 | 0.038 | -0.125 | -0.088 |
| c7_pct_quest | -0.042 | -0.025 | -0.073 | -0.009 | 0.139 | 0.109 | -0.099 | 0.151 | -0.069 | -0.077 | -0.054 | 0.100 | -0.077 | -0.062 | -0.091 | -0.070 | -0.088 |
| c7_review_size | -0.076 | 0.147 | 0.023 | -0.040 | 0.140 | -0.116 | -0.313 | 0.025 | 0.035 | -0.012 | -0.054 | 0.198 | -0.037 | -0.008 | 0.050 | -0.012 | 0.003 |
| c7_pct_periods | 0.021 | 0.030 | -0.081 | -0.059 | -0.022 | 0.002 | -0.089 | 0.004 | 0.068 | 0.041 | -0.005 | 0.025 | 0.010 | -0.002 | 0.029 | -0.010 | 0.031 |
| c7_pct_apost | 0.021 | 0.030 | -0.081 | -0.059 | -0.022 | 0.002 | -0.089 | 0.004 | 0.068 | 0.041 | -0.005 | 0.025 | 0.010 | -0.002 | 0.029 | -0.010 | 0.031 |
| c7_useful_review_count | 0.097 | 0.008 | -0.104 | 0.039 | 0.099 | -0.173 | -0.273 | 0.004 | 0.015 | -0.016 | 0.170 | 0.063 | -0.079 | 0.106 | 0.128 | 0.016 | 0.043 |
| c7_log10_biz_review_count | -0.054 | 0.073 | 0.039 | -0.078 | -0.004 | 0.019 | -0.008 | 0.109 | -0.116 | -0.027 | -0.101 | 0.106 | -0.131 | -0.095 | -0.051 | -0.117 | -0.081 |
| c7_item_count | -0.191 | 0.011 | 0.227 | -0.051 | 0.000 | 0.077 | 0.182 | -0.165 | 0.188 | 0.007 | -0.053 | -0.193 | 0.051 | -0.027 | 0.112 | 0.084 | 0.098 |
| c7_good_review | -0.0195 | -0.02029 | 0.018876 | 0.004322 | -0.27487 | 0.057854 | 0.803067 | 0.091409 | -0.19849 | 0.001118 | -0.02155 | -0.1051 | 0.102687 | -0.11307 | 0.022193 | 0.144643 | 0.147199 |
| c7_cluster_similarity | -0.068 | -0.018 | 0.033 | -0.029 | -0.109 | 0.066 | 0.012 | 0.007 | -0.104 | 0.053 | 0.016 | -0.043 | 0.023 | 0.026 | 0.009 | -0.042 | -0.014 |
| c8_sentiment | -0.040 | 0.106 | -0.116 | 0.030 | 0.065 | 0.120 | 0.046 | 0.665 | -0.137 | -0.030 | -0.005 | 0.024 | -0.102 | -0.141 | -0.168 | 0.104 | 0.058 |
| c8_pct_exclam | -0.102 | 0.115 | 0.008 | -0.039 | 0.058 | -0.039 | 0.122 | 0.257 | 0.014 | -0.054 | 0.080 | -0.007 | -0.139 | 0.008 | -0.017 | 0.095 | 0.041 |
| c8_log10_usr_review_count | 0.109 | -0.159 | -0.066 | 0.130 | -0.122 | 0.014 | 0.012 | -0.366 | 0.057 | 0.245 | 0.077 | -0.197 | 0.086 | -0.059 | -0.041 | -0.104 | -0.099 |
| c8_pct_quest | 0.023 | 0.055 | 0.064 | -0.020 | 0.048 | -0.071 | -0.052 | -0.255 | 0.205 | -0.126 | -0.011 | 0.000 | -0.002 | 0.011 | 0.138 | 0.006 | 0.040 |
| c8_review_size | 0.118 | -0.196 | -0.025 | 0.137 | -0.109 | -0.005 | -0.068 | -0.114 | 0.043 | 0.168 | 0.006 | -0.096 | 0.119 | -0.032 | 0.037 | -0.010 | 0.002 |
| c8_pct_periods | -0.060 | 0.124 | 0.091 | -0.118 | 0.051 | -0.007 | 0.069 | -0.040 | -0.008 | -0.069 | -0.103 | -0.041 | -0.157 | 0.012 | 0.017 | -0.084 | -0.064 |
| c8_pct_apost | -0.060 | 0.124 | 0.091 | -0.118 | 0.051 | -0.007 | 0.069 | -0.040 | -0.008 | -0.069 | -0.103 | -0.041 | -0.157 | 0.012 | 0.017 | -0.084 | -0.064 |
| c8_useful_review_count | 0.184 | -0.101 | -0.033 | 0.170 | -0.099 | -0.043 | -0.154 | -0.078 | 0.016 | 0.228 | -0.050 | -0.090 | 0.092 | -0.149 | 0.016 | -0.042 | -0.019 |
| c8_log10_biz_review_count | 0.045 | 0.026 | -0.071 | -0.074 | -0.131 | -0.016 | -0.042 | 0.048 | 0.005 | -0.034 | 0.020 | 0.014 | 0.025 | -0.100 | 0.046 | -0.092 | -0.096 |
| c8_item_count | -0.185 | 0.030 | 0.048 | 0.074 | -0.020 | 0.016 | 0.081 | 0.038 | -0.045 | -0.030 | -0.034 | -0.031 | 0.061 | -0.041 | 0.128 | 0.047 | 0.035 |
| c8_good_review | -0.01583 | 0.157056 | -0.10457 | 0.040748 | 0.073591 | 0.053424 | 0.088668 | 0.781261 | -0.1704 | -0.06935 | -0.00883 | 0.025076 | -0.09508 | -0.15188 | -0.17118 | 0.141741 | 0.073286 |
| c8_cluster_similarity | -0.040 | 0.057 | 0.050 | -0.125 | -0.047 | 0.044 | -0.013 | 0.015 | -0.010 | -0.019 | 0.033 | 0.003 | -0.073 | 0.016 | 0.063 | -0.015 | 0.001 |
| c9_sentiment | -0.064 | 0.074 | -0.006 | -0.197 | 0.020 | -0.003 | -0.060 | -0.133 | 0.667 | -0.050 | -0.002 | -0.100 | -0.092 | -0.078 | 0.094 | 0.035 | -0.004 |
| c9_pct_exclam | -0.016 | -0.026 | -0.010 | -0.114 | 0.101 | 0.021 | -0.148 | -0.177 | 0.367 | -0.107 | 0.062 | -0.003 | -0.050 | 0.071 | 0.025 | -0.001 | -0.021 |
| c9_log10_usr_review_count | 0.049 | -0.027 | 0.002 | 0.034 | -0.068 | 0.079 | 0.063 | 0.223 | -0.371 | -0.096 | 0.013 | 0.177 | -0.126 | -0.034 | -0.201 | -0.118 | -0.121 |
| c9_pct_quest | 0.052 | -0.079 | 0.025 | 0.065 | 0.037 | -0.049 | 0.055 | -0.015 | -0.189 | -0.018 | -0.035 | 0.048 | -0.019 | 0.002 | 0.076 | -0.018 | 0.035 |
| c9_review_size | 0.005 | -0.034 | 0.064 | 0.034 | -0.117 | 0.030 | 0.140 | 0.008 | -0.324 | -0.024 | 0.162 | 0.059 | -0.079 | 0.006 | -0.032 | -0.044 | 0.013 |
| c9_pct_periods | 0.021 | -0.026 | -0.096 | -0.025 | -0.062 | 0.004 | 0.080 | -0.017 | -0.088 | 0.112 | -0.098 | -0.082 | 0.255 | 0.019 | 0.104 | 0.048 | 0.058 |
| c9_pct_apost | 0.021 | -0.026 | -0.096 | -0.025 | -0.062 | 0.004 | 0.080 | -0.017 | -0.088 | 0.112 | -0.098 | -0.082 | 0.255 | 0.019 | 0.104 | 0.048 | 0.058 |
| c9_useful_review_count | 0.033 | -0.100 | -0.049 | 0.066 | -0.078 | -0.030 | -0.005 | -0.060 | -0.122 | 0.077 | 0.098 | -0.060 | 0.139 | 0.125 | -0.040 | -0.006 | 0.019 |
| c9_log10_biz_review_count | -0.037 | -0.020 | -0.079 | -0.091 | -0.058 | -0.080 | -0.006 | -0.110 | 0.265 | -0.044 | 0.038 | -0.009 | -0.028 | 0.052 | 0.149 | -0.016 | 0.054 |
| c9_item_count | -0.222 | 0.113 | 0.219 | 0.017 | 0.117 | -0.049 | -0.077 | -0.144 | 0.270 | 0.006 | -0.101 | -0.066 | -0.027 | -0.058 | 0.151 | 0.073 | 0.066 |
| c9_good_review | -0.06917 | 0.075647 | 0.018562 | -0.23513 | 0.016834 | -0.00668 | -0.08149 | -0.19019 | 0.781759 | -0.01651 | -0.02483 | -0.14752 | -0.05066 | -0.08413 | 0.12544 | 0.054588 | 0.028606 |
| c9_cluster_similarity | -0.023 | 0.089 | 0.012 | -0.006 | -0.058 | -0.013 | -0.011 | 0.018 | -0.034 | 0.003 | -0.062 | -0.053 | -0.050 | 0.008 | 0.006 | -0.063 | -0.058 |
| c10_sentiment | 0.042 | 0.032 | 0.012 | 0.109 | -0.069 | -0.129 | -0.100 | -0.005 | 0.011 | 0.695 | 0.025 | -0.166 | 0.062 | -0.114 | -0.045 | 0.138 | 0.147 |
| c10_pct_exclam | -0.051 | -0.015 | 0.135 | 0.017 | -0.046 | -0.126 | -0.063 | 0.000 | 0.087 | 0.398 | -0.038 | -0.142 | -0.021 | 0.018 | -0.114 | 0.017 | 0.001 |
| c10_log10_usr_review_count | 0.216 | 0.171 | -0.079 | -0.091 | 0.063 | -0.050 | 0.058 | 0.174 | -0.077 | -0.261 | -0.089 | 0.118 | 0.060 | -0.136 | -0.068 | -0.011 | -0.003 |
| c10_pct_quest | 0.031 | 0.108 | 0.030 | 0.013 | -0.038 | 0.056 | 0.141 | 0.104 | -0.049 | -0.209 | -0.092 | 0.068 | 0.057 | -0.124 | -0.047 | 0.017 | 0.004 |
| c10_review_size | 0.097 | 0.162 | -0.120 | 0.009 | 0.052 | 0.066 | -0.016 | 0.279 | -0.142 | -0.263 | -0.048 | 0.063 | -0.021 | 0.022 | -0.113 | -0.004 | -0.006 |
| c10_pct_periods | 0.031 | 0.002 | 0.083 | -0.027 | -0.050 | -0.032 | -0.069 | -0.071 | 0.053 | -0.146 | -0.016 | 0.025 | -0.067 | -0.005 | 0.151 | -0.053 | -0.008 |
| c10_pct_apost | 0.031 | 0.002 | 0.083 | -0.027 | -0.050 | -0.032 | -0.069 | -0.071 | 0.053 | -0.146 | -0.016 | 0.025 | -0.067 | -0.005 | 0.151 | -0.053 | -0.008 |
| c10_useful_review_count | 0.231 | 0.144 | -0.133 | 0.054 | 0.054 | -0.103 | -0.038 | 0.167 | -0.125 | 0.004 | 0.017 | 0.008 | 0.037 | -0.087 | -0.183 | -0.005 | 0.009 |
| c10_log10_biz_review_count | -0.075 | 0.006 | -0.002 | -0.020 | -0.190 | 0.022 | -0.013 | -0.044 | -0.077 | 0.149 | 0.070 | 0.046 | -0.138 | 0.073 | 0.044 | -0.050 | -0.045 |
| c10_item_count | -0.123 | 0.017 | 0.018 | 0.001 | 0.117 | -0.067 | 0.002 | -0.166 | 0.064 | 0.010 | 0.017 | -0.093 | 0.064 | 0.075 | 0.234 | 0.076 | 0.099 |



| | c1_good_cluster_flag | c2_good_cluster_flag | c3_good_cluster_flag | c4_good_cluster_flag | c5_good_cluster_flag | c6_good_cluster_flag | c7_good_cluster_flag | c8_good_cluster_flag | c9_good_cluster_flag | c10_good_cluster_flag | c11_good_cluster_flag | c12_good_cluster_flag | c13_good_cluster_flag | c14_good_cluster_flag | c15_good_cluster_flag | good_cluster_count | good_system |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c10_good_review | 0.028217 | -0.01837 | 0.019999 | 0.115168 | -0.08558 | -0.14352 | -0.13411 | -0.04515 | 0.019208 | 0.809294 | 0.061894 | -0.17876 | 0.061851 | -0.08535 | -0.03297 | 0.151273 | 0.158706 |
| c10_cluster_similarity | -0.099 | -0.036 | 0.045 | -0.048 | -0.065 | 0.048 | -0.036 | -0.009 | -0.048 | -0.003 | 0.000 | -0.020 | -0.043 | 0.061 | 0.067 | -0.064 | -0.040 |
| c11_sentiment | -0.014 | -0.224 | -0.159 | 0.196 | -0.072 | 0.096 | -0.040 | -0.087 | -0.083 | -0.080 | 0.367 | -0.030 | -0.085 | 0.064 | -0.010 | -0.075 | -0.074 |
| c11_pct_exclam | -0.047 | -0.155 | -0.064 | 0.072 | 0.044 | -0.082 | -0.173 | -0.070 | -0.044 | 0.012 | 0.270 | 0.052 | -0.099 | 0.108 | 0.153 | -0.015 | -0.006 |
| c11_log10_usr_review_count | 0.127 | -0.113 | -0.105 | 0.089 | -0.052 | 0.129 | -0.037 | 0.021 | -0.067 | -0.065 | -0.163 | 0.139 | -0.026 | -0.091 | -0.079 | -0.111 | -0.106 |
| c11_pct_quest | 0.042 | -0.066 | -0.064 | -0.027 | -0.068 | 0.023 | 0.001 | 0.173 | -0.010 | 0.088 | -0.061 | 0.000 | -0.002 | -0.151 | 0.030 | -0.036 | -0.017 |
| c11_review_size | 0.070 | -0.051 | 0.043 | 0.085 | -0.009 | -0.022 | -0.049 | -0.038 | -0.009 | -0.042 | -0.214 | 0.003 | -0.065 | 0.043 | -0.095 | -0.134 | -0.068 |
| c11_pct_periods | 0.102 | -0.056 | 0.002 | -0.130 | -0.081 | 0.042 | -0.004 | 0.016 | 0.010 | 0.072 | -0.012 | -0.017 | -0.010 | -0.016 | 0.038 | -0.018 | -0.042 |
| c11_pct_apost | 0.102 | -0.056 | 0.002 | -0.130 | -0.081 | 0.042 | -0.004 | 0.016 | 0.010 | 0.072 | -0.012 | -0.017 | -0.010 | -0.016 | 0.038 | -0.018 | -0.042 |
| c11_useful_review_count | 0.226 | -0.004 | -0.086 | 0.065 | -0.032 | 0.015 | -0.096 | 0.092 | -0.079 | 0.121 | 0.000 | 0.059 | -0.032 | -0.082 | -0.114 | 0.005 | 0.001 |
| c11_log10_biz_review_count | 0.026 | -0.086 | -0.016 | -0.117 | 0.045 | 0.040 | 0.088 | 0.026 | 0.026 | -0.061 | 0.102 | -0.020 | -0.097 | 0.042 | 0.043 | 0.011 | -0.013 |
| c11_item_count | -0.176 | 0.117 | 0.146 | -0.072 | 0.065 | -0.007 | 0.041 | -0.167 | 0.197 | -0.050 | 0.037 | -0.096 | 0.041 | 0.065 | 0.151 | 0.123 | 0.132 |
| c11_good_review | -0.07874 | -0.16045 | -0.1386 | 0.231426 | -0.01057 | 0.020147 | 0.054097 | 0.03436 | -0.10583 | 0.035654 | 0.746206 | -0.09699 | -0.14653 | 0.055718 | -0.03891 | 0.12325 | 0.040582 |
| c11_cluster_similarity | -0.113 | 0.056 | 0.041 | -0.026 | -0.070 | 0.056 | -0.048 | -0.018 | -0.044 | 0.010 | -0.004 | -0.028 | -0.065 | 0.027 | 0.077 | -0.049 | -0.079 |
| c12_sentiment | -0.015 | 0.102 | -0.041 | -0.256 | 0.070 | 0.096 | -0.126 | 0.035 | -0.161 | -0.238 | -0.051 | 0.699 | -0.068 | -0.091 | -0.104 | -0.047 | -0.059 |
| c12_pct_exclam | -0.082 | 0.005 | 0.006 | -0.044 | 0.049 | 0.020 | -0.020 | -0.007 | -0.058 | -0.057 | -0.023 | 0.249 | -0.081 | -0.088 | -0.039 | -0.058 | -0.062 |
| c12_log10_usr_review_count | 0.061 | -0.029 | 0.151 | -0.094 | -0.011 | -0.014 | 0.074 | -0.150 | 0.183 | 0.016 | -0.008 | -0.338 | 0.121 | 0.120 | -0.035 | 0.014 | 0.043 |
| c12_pct_quest | 0.053 | -0.006 | 0.019 | 0.005 | 0.066 | 0.014 | -0.039 | -0.017 | -0.032 | -0.008 | -0.028 | -0.216 | 0.081 | 0.012 | -0.015 | -0.049 | -0.068 |
| c12_review_size | 0.127 | -0.071 | -0.016 | -0.007 | -0.026 | -0.044 | 0.076 | -0.008 | 0.104 | 0.030 | 0.115 | -0.152 | 0.068 | 0.128 | -0.037 | 0.097 | 0.040 |
| c12_pct_periods | 0.004 | -0.011 | -0.117 | 0.093 | -0.023 | 0.048 | -0.010 | -0.058 | 0.008 | 0.041 | 0.128 | -0.274 | -0.067 | 0.132 | -0.094 | -0.085 | -0.056 |
| c12_pct_apost | 0.004 | -0.011 | -0.117 | 0.093 | -0.023 | 0.048 | -0.010 | -0.058 | 0.008 | 0.041 | 0.128 | -0.274 | -0.067 | 0.132 | -0.094 | -0.085 | -0.056 |
| c12_useful_review_count | 0.217 | 0.014 | -0.082 | -0.034 | -0.031 | -0.139 | 0.018 | 0.134 | 0.020 | 0.054 | 0.073 | -0.184 | 0.139 | 0.097 | -0.031 | 0.081 | 0.099 |
| c12_log10_biz_review_count | 0.026 | 0.003 | 0.045 | -0.138 | 0.063 | 0.052 | 0.069 | 0.000 | -0.139 | -0.134 | 0.026 | 0.289 | -0.067 | 0.066 | -0.191 | -0.012 | 0.015 |
| c12_item_count | -0.180 | 0.147 | 0.192 | 0.035 | 0.044 | -0.089 | 0.005 | -0.110 | 0.115 | -0.011 | -0.135 | -0.032 | 0.036 | -0.008 | 0.206 | 0.099 | 0.113 |
| c12_good_review | -0.06009 | 0.116567 | -0.04358 | -0.26562 | 0.0845 | 0.087773 | -0.13115 | 0.062131 | -0.17175 | -0.23081 | -0.04371 | 0.788796 | -0.08571 | -0.121 | -0.07369 | -0.01932 | -0.0251 |
| c12_cluster_similarity | -0.113 | 0.040 | 0.068 | -0.012 | -0.051 | -0.012 | -0.060 | 0.014 | -0.084 | 0.054 | -0.033 | 0.020 | -0.072 | 0.016 | 0.003 | -0.078 | -0.030 |
| c13_sentiment | 0.156 | -0.068 | -0.063 | -0.121 | -0.075 | -0.015 | 0.114 | -0.034 | -0.003 | 0.106 | -0.090 | -0.123 | 0.672 | -0.006 | -0.078 | 0.146 | 0.122 |
| c13_pct_exclam | -0.030 | 0.066 | -0.027 | 0.038 | -0.100 | -0.011 | 0.019 | -0.052 | 0.099 | 0.034 | -0.072 | 0.029 | 0.302 | -0.103 | 0.028 | 0.095 | 0.041 |
| c13_log10_usr_review_count | 0.032 | -0.076 | 0.015 | -0.033 | 0.011 | -0.029 | 0.059 | 0.042 | 0.020 | 0.080 | 0.066 | -0.079 | -0.325 | 0.044 | -0.107 | -0.116 | -0.100 |
| c13_pct_quest | -0.092 | 0.118 | 0.024 | 0.022 | 0.092 | 0.019 | -0.108 | -0.018 | -0.015 | 0.079 | 0.090 | 0.056 | -0.006 | 0.011 | 0.021 | 0.113 | 0.087 |
| c13_review_size | -0.055 | -0.130 | 0.082 | -0.021 | -0.027 | 0.135 | 0.166 | 0.034 | 0.025 | 0.006 | 0.002 | -0.058 | -0.309 | -0.050 | -0.034 | -0.088 | -0.060 |
| c13_pct_periods | 0.060 | -0.047 | -0.037 | -0.015 | 0.010 | 0.058 | -0.043 | 0.036 | -0.071 | 0.016 | -0.076 | -0.029 | -0.078 | 0.023 | -0.006 | -0.078 | -0.063 |
| c13_pct_apost | 0.060 | -0.047 | -0.037 | -0.015 | 0.010 | 0.058 | -0.043 | 0.036 | -0.071 | 0.016 | -0.076 | -0.029 | -0.078 | 0.023 | -0.006 | -0.078 | -0.063 |
| c13_useful_review_count | -0.050 | -0.089 | 0.093 | 0.082 | -0.149 | 0.052 | 0.100 | 0.010 | -0.031 | 0.044 | -0.080 | -0.017 | -0.239 | -0.037 | 0.075 | -0.084 | -0.075 |
| c13_log10_biz_review_count | -0.010 | -0.121 | 0.027 | -0.041 | -0.023 | 0.153 | 0.074 | -0.053 | -0.156 | -0.060 | -0.189 | 0.069 | 0.106 | -0.113 | 0.008 | -0.111 | -0.096 |
| c13_item_count | -0.041 | -0.012 | 0.071 | 0.021 | 0.040 | -0.101 | 0.002 | -0.189 | 0.053 | 0.004 | 0.063 | -0.025 | 0.128 | 0.010 | 0.167 | 0.079 | 0.104 |
| c13_good_review | 0.164344 | -0.05258 | -0.07518 | -0.05312 | -0.11164 | -0.02027 | 0.102989 | -0.03716 | -0.00498 | 0.111892 | -0.08758 | -0.10358 | 0.813514 | -0.07662 | -0.02921 | 0.211411 | 0.169417 |
| c13_cluster_similarity | -0.069 | 0.112 | 0.069 | -0.106 | 0.023 | 0.087 | -0.039 | 0.026 | -0.063 | -0.063 | 0.008 | 0.039 | -0.049 | -0.057 | -0.031 | -0.040 | -0.036 |
| c14_sentiment | 0.027 | -0.128 | -0.045 | 0.112 | 0.068 | -0.091 | 0.006 | -0.095 | 0.018 | -0.017 | 0.044 | -0.148 | 0.023 | 0.564 | -0.036 | 0.107 | 0.055 |
| c14_pct_exclam | -0.091 | -0.096 | -0.055 | 0.071 | -0.076 | -0.024 | -0.006 | -0.031 | -0.090 | 0.050 | 0.068 | -0.079 | -0.056 | 0.261 | -0.012 | -0.061 | -0.054 |
| c14_log10_usr_review_count | 0.030 | 0.185 | 0.113 | 0.011 | 0.023 | -0.056 | -0.117 | 0.069 | -0.001 | 0.105 | -0.032 | 0.016 | -0.106 | -0.341 | 0.045 | -0.026 | 0.014 |
| c14_pct_quest | -0.065 | -0.042 | -0.016 | 0.129 | -0.132 | 0.106 | -0.005 | 0.029 | 0.037 | 0.097 | 0.005 | -0.039 | -0.095 | -0.223 | 0.050 | -0.058 | -0.062 |
| c14_review_size | 0.097 | 0.081 | 0.067 | 0.064 | -0.008 | 0.033 | 0.047 | 0.007 | -0.058 | 0.045 | 0.018 | -0.022 | 0.137 | -0.289 | -0.034 | 0.064 | 0.027 |
| c14_pct_periods | -0.033 | 0.054 | -0.033 | -0.030 | -0.058 | 0.008 | -0.073 | -0.060 | 0.045 | -0.031 | 0.164 | 0.050 | -0.097 | -0.063 | -0.082 | -0.095 | -0.072 |
| c14_pct_apost | -0.033 | 0.054 | -0.033 | -0.030 | -0.058 | 0.008 | -0.073 | -0.060 | 0.045 | -0.031 | 0.164 | 0.050 | -0.097 | -0.063 | -0.082 | -0.095 | -0.072 |
| c14_useful_review_count | 0.215 | 0.053 | 0.104 | -0.024 | -0.144 | 0.092 | 0.071 | -0.008 | -0.027 | 0.056 | -0.095 | -0.031 | 0.134 | -0.226 | -0.014 | 0.054 | 0.118 |
| c14_log10_biz_review_count | 0.096 | -0.112 | -0.116 | -0.021 | -0.064 | 0.010 | -0.006 | -0.053 | -0.075 | -0.181 | 0.065 | 0.066 | 0.038 | 0.176 | -0.055 | -0.095 | -0.094 |
| c14_item_count | -0.128 | 0.046 | 0.120 | 0.006 | 0.195 | -0.054 | -0.040 | -0.179 | 0.207 | 0.003 | -0.051 | -0.155 | 0.087 | 0.040 | 0.251 | 0.144 | 0.179 |
| c14_good_review | -0.02969 | -0.19215 | -0.05596 | 0.097297 | 0.024125 | -0.11396 | -0.04198 | -0.16839 | 0.052164 | -0.02826 | 0.061733 | -0.15007 | -0.03042 | 0.770227 | -0.024 | 0.061421 | 0.019657 |
| c14_cluster_similarity | -0.085 | 0.037 | 0.050 | 0.014 | -0.015 | 0.023 | -0.048 | 0.008 | -0.063 | 0.054 | 0.023 | -0.044 | -0.032 | 0.009 | 0.014 | -0.017 | -0.019 |
| c15_sentiment | -0.139 | -0.055 | -0.020 | -0.054 | -0.073 | -0.098 | 0.019 | -0.149 | 0.034 | -0.084 | -0.049 | 0.019 | -0.047 | -0.054 | 0.674 | -0.005 | 0.003 |
| c15_pct_exclam | -0.047 | 0.082 | -0.084 | -0.039 | -0.079 | 0.086 | -0.069 | -0.088 | 0.095 | 0.022 | -0.061 | -0.005 | 0.055 | -0.031 | 0.198 | 0.028 | 0.008 |
| c15_log10_usr_review_count | 0.135 | 0.060 | -0.055 | -0.093 | -0.017 | 0.045 | -0.040 | 0.234 | -0.171 | 0.032 | 0.049 | 0.069 | -0.062 | -0.054 | -0.381 | -0.114 | -0.091 |
| c15_pct_quest | -0.025 | -0.040 | 0.069 | 0.124 | -0.003 | -0.018 | 0.027 | 0.007 | -0.014 | 0.062 | 0.018 | -0.022 | -0.054 | 0.054 | -0.271 | -0.039 | -0.100 |
| c15_review_size | 0.150 | 0.033 | -0.028 | -0.047 | 0.096 | 0.011 | 0.034 | 0.137 | -0.115 | -0.127 | -0.052 | 0.119 | 0.022 | -0.007 | -0.301 | -0.043 | -0.052 |
| c15_pct_periods | -0.007 | 0.035 | 0.028 | 0.057 | 0.042 | 0.092 | -0.051 | 0.050 | -0.036 | -0.040 | 0.074 | 0.015 | 0.024 | -0.021 | -0.232 | 0.003 | -0.004 |
| c15_pct_apost | -0.007 | 0.035 | 0.028 | 0.057 | 0.042 | 0.092 | -0.051 | 0.050 | -0.036 | -0.040 | 0.074 | 0.015 | 0.024 | -0.021 | -0.232 | 0.003 | -0.004 |
| c15_useful_review_count | 0.232 | -0.001 | 0.017 | 0.037 | 0.057 | -0.016 | -0.022 | 0.127 | -0.223 | -0.029 | -0.051 | 0.108 | 0.083 | -0.022 | -0.209 | 0.015 | 0.030 |
| c15_log10_biz_review_count | 0.056 | -0.008 | 0.054 | -0.157 | -0.013 | 0.044 | 0.044 | -0.090 | 0.020 | -0.112 | 0.000 | 0.052 | 0.071 | 0.074 | 0.130 | 0.066 | 0.012 |



| | c1_good_cluster_flag | c2_good_cluster_flag | c3_good_cluster_flag | c4_good_cluster_flag | c5_good_cluster_flag | c6_good_cluster_flag | c7_good_cluster_flag | c8_good_cluster_flag | c9_good_cluster_flag | c10_good_cluster_flag | c11_good_cluster_flag | c12_good_cluster_flag | c13_good_cluster_flag | c14_good_cluster_flag | c15_good_cluster_flag | good_cluster_count | good_system |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c15_item_count | -0.131 | 0.061 | 0.170 | -0.039 | 0.064 | -0.098 | 0.036 | -0.193 | 0.137 | -0.027 | -0.041 | -0.069 | 0.056 | -0.037 | 0.336 | 0.102 | 0.113 |
| c15_good_review | -0.14789 | -0.09875 | 0.001565 | -0.02926 | -0.05652 | -0.0654 | -0.02851 | -0.19789 | 0.079225 | -0.07645 | -0.04093 | -0.0171 | 0.003631 | -0.02489 | 0.82586 | 0.075322 | 0.071996 |
| c15_cluster_similarity | -0.173 | -0.052 | 0.072 | -0.079 | -0.092 | 0.027 | 0.019 | -0.045 | -0.017 | 0.046 | 0.013 | -0.098 | -0.059 | 0.037 | 0.057 | -0.118 | -0.103 |
| c1_good_cluster_flag | 1.000 | -0.039 | -0.148 | -0.039 | -0.145 | -0.104 | -0.004 | -0.003 | -0.054 | 0.084 | -0.120 | -0.092 | 0.160 | -0.010 | -0.134 | 0.082 | 0.091 |
| c2_good_cluster_flag | -0.039 | 1.000 | 0.042 | -0.122 | 0.105 | -0.112 | -0.047 | 0.188 | 0.035 | -0.097 | -0.103 | 0.065 | -0.039 | -0.162 | -0.097 | 0.232 | 0.217 |
| c3_good_cluster_flag | -0.148 | 0.042 | 1.000 | 0.001 | -0.001 | -0.032 | 0.039 | -0.117 | 0.007 | 0.001 | -0.124 | -0.022 | -0.108 | -0.162 | -0.044 | 0.129 | 0.131 |
| c4_good_cluster_flag | -0.039 | -0.122 | 0.001 | 1.000 | 0.009 | -0.114 | 0.004 | -0.015 | -0.147 | 0.098 | 0.067 | -0.204 | 0.007 | 0.033 | -0.013 | 0.196 | 0.144 |
| c5_good_cluster_flag | -0.145 | 0.105 | -0.001 | 0.009 | 1.000 | -0.065 | -0.255 | 0.079 | 0.060 | -0.107 | 0.101 | 0.071 | -0.103 | 0.066 | -0.079 | 0.265 | 0.224 |
| c6_good_cluster_flag | -0.104 | -0.112 | -0.032 | -0.114 | -0.065 | 1.000 | 0.133 | 0.078 | -0.065 | -0.167 | -0.011 | 0.074 | -0.048 | -0.156 | -0.073 | 0.145 | 0.053 |
| c7_good_cluster_flag | -0.004 | -0.047 | 0.039 | 0.004 | -0.255 | 0.133 | 1.000 | 0.076 | -0.162 | -0.076 | 0.000 | -0.106 | 0.083 | -0.099 | -0.027 | 0.223 | 0.177 |
| c8_good_cluster_flag | -0.003 | 0.188 | -0.117 | -0.015 | 0.079 | 0.078 | 0.076 | 1.000 | -0.188 | -0.103 | -0.024 | 0.050 | -0.109 | -0.115 | -0.207 | 0.196 | 0.127 |
| c9_good_cluster_flag | -0.054 | 0.035 | 0.007 | -0.147 | 0.060 | -0.065 | -0.162 | -0.188 | 1.000 | -0.007 | 0.016 | -0.175 | -0.046 | 0.027 | 0.108 | 0.162 | 0.137 |
| c10_good_cluster_flag | 0.084 | -0.097 | 0.001 | 0.098 | -0.107 | -0.167 | -0.076 | -0.103 | -0.007 | 1.000 | 0.057 | -0.200 | 0.054 | -0.076 | -0.052 | 0.159 | 0.160 |
| c11_good_cluster_flag | -0.120 | -0.103 | -0.124 | 0.067 | 0.101 | -0.011 | 0.000 | -0.024 | 0.016 | 0.057 | 1.000 | -0.091 | -0.064 | 0.091 | -0.020 | 0.261 | 0.145 |
| c12_good_cluster_flag | -0.092 | 0.065 | -0.022 | -0.204 | 0.071 | 0.074 | -0.106 | 0.050 | -0.175 | -0.200 | -0.091 | 1.000 | -0.070 | -0.110 | -0.032 | 0.082 | 0.061 |
| c13_good_cluster_flag | 0.160 | -0.039 | -0.108 | 0.007 | -0.103 | -0.048 | 0.083 | -0.109 | -0.046 | 0.054 | -0.064 | -0.070 | 1.000 | -0.019 | 0.034 | 0.287 | 0.239 |
| c14_good_cluster_flag | -0.010 | -0.162 | -0.162 | 0.033 | 0.066 | -0.156 | -0.099 | -0.115 | 0.027 | -0.076 | 0.091 | -0.110 | -0.019 | 1.000 | -0.032 | 0.098 | 0.086 |
| c15_good_cluster_flag | -0.134 | -0.097 | -0.044 | -0.013 | -0.079 | -0.073 | -0.027 | -0.207 | 0.108 | -0.052 | -0.020 | -0.032 | 0.034 | -0.032 | 1.000 | 0.156 | 0.157 |
| good_cluster_count | 0.082 | 0.232 | 0.129 | 0.196 | 0.265 | 0.145 | 0.223 | 0.196 | 0.162 | 0.159 | 0.261 | 0.082 | 0.287 | 0.098 | 0.156 | 1.000 | 0.806 |
| good_system | 0.091 | 0.217 | 0.131 | 0.144 | 0.224 | 0.053 | 0.177 | 0.127 | 0.137 | 0.160 | 0.145 | 0.061 | 0.239 | 0.086 | 0.157 | 0.806 | 1.000 |

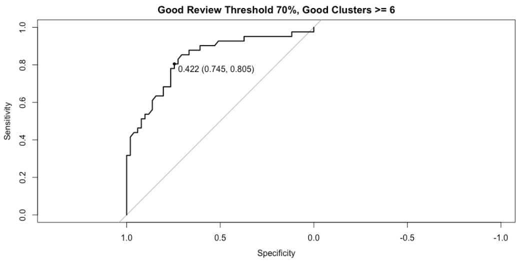
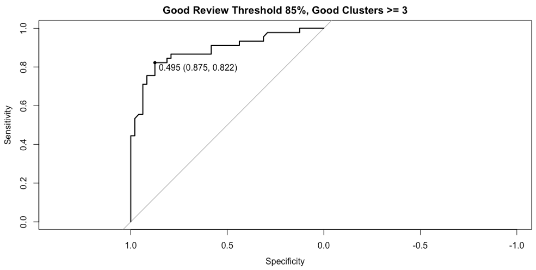
Table 6-1. Correlation matrix for selected attributes for 70% good review threshold. Blue are negative numbers, yellow are near zero, and orange are positive values. Please note how each of the columns represents a “local target attribute” which shows what attributes are strong within a cluster over time. The final column shows the “global target attribute” named “good system”. Attributes that correlate well with this attribute tended to be selected by the random forest variable selection process.

```
good_review_threshold    0.7      (70%)
cluster_count_threshold  6


  0   1
259 210


Good system
c9_item_count                  100
c14_item_count               60.29
c1_item_count                  51.59
c1_log10_biz_review_count        45.14
c15_item_count               29.75
c3_pct_exclam                  28.75
c13_item_count               27.64
c5_item_count                  22.79
c8_item_count                  22.05
c5_pct_exclam                  18.88
c2_log10_biz_review_count         15.36
c2_useful_review_count       15.36
c2_item_count                  15.24
c12_cluster_similarity       15
c7_review_size               13.1
c10_cluster_similarity       12.57
c14_log10_biz_review_count 11.99
c8_sentiment                   11.32
c5_log10_usr_review_count         10.92
c7_item_count                  10.38


mtry      Accuracy            Kappa
2             0.683925       0.352371
83          0.686506         0.362803
165         0.693469         0.377503




            Reference
Prediction       bad         good
bad                40.7      16
good           14.6        28.7


Accuracy(average)        0.6936
      bad        good
bad        42     13
good     11       26
```

```
cluster_count_threshold  0.85    (85%)
cluster_count_threshold  3


  0   1
242 227


Good system
c14_item_count            100
c5_sentiment                85.757
c10_item_count            40.521
c2_item_count               31.352
c15_item_count            27.711
c5_item_count               27.34
c12_item_count            26.232
c9_item_count               19.408
c1_item_count               19.318
c11_item_count            13.89
c11_useful_review_count   12.297
c1_pct_quest                11.75
c5_review_size            10.736
c9_useful_review_count    9.577
c7_pct_quest                8.749
c14_pct_periods           8.51
c5_useful_review_count    8.371
c7_item_count               8.13
c13_pct_periods           8.088
c14_pct_apost               8.071


mtry      Accuracy            Kappa
2             0.772907       0.5443271
83          0.778774         0.5562298
165         0.770837         0.540325


            Reference
Prediction       bad         good
bad                42.3      13
good           9.2           35.6


Accuracy(average)        0.7787


            bad   good
bad       42         9
good     5          37
```

| | |
|---|---|
| Accuracy 0.7391<br>95% CI (0.6371, 0.8252)<br>No Information Rate 0.5761<br>P-Value [Acc > NIR] 0.0008656<br>Kappa 0.4623<br>Mcnemar's Test P-Value 0.838256<br>Precision 0.7027<br>Recall 0.6667<br>F1 0.6842<br>Prevalence 0.4239<br>Detection Rate 0.2826<br>Detection Prevalence 0.4022<br>Balanced Accuracy 0.7296<br>'Positive' Class good | Accuracy 0.8495<br>95% CI (0.7603, 0.9152)<br>No Information Rate 0.5054<br>P-Value [Acc > NIR] 3.64E-12<br>Kappa 0.6986<br>Mcnemar's Test P-Value 0.4227<br>Precision 0.881<br>Recall 0.8043<br>F1 0.8409<br>Prevalence 0.4946<br>Detection Rate 0.3978<br>Detection Prevalence 0.4516<br>Balanced Accuracy 0.849<br>'Positive' Class good |

Table 6-2. Top attribute selection for each model with corresponding model statistics. The top twenty attributes listed here will be used for the deployment models.

| TRAINING FOR 70% GOOD REVIEW THRESHOLD<br>GOOD CLUSTER THRESHOLD = 6, 20 ATTRIBUTES | TRAINING FOR 85% GOOD REVIEW THRESHOLD<br>GOOD CLUSTER THRESHOLD = 3, 20 ATTRIBUTES |
|---|---|
| c13_item_count 100.000<br>c9_item_count 94.189<br>c1_log10_biz_review_count 86.300<br>c14_item_count 82.136<br>c1_item_count 79.746<br>c2_item_count 56.482<br>c7_item_count 43.115<br>c2_useful_review_count 43.038<br>c8_item_count 42.909<br>c5_pct_exclam 39.799<br>c3_pct_exclam 34.946<br>c15_item_count 33.381<br>c5_item_count 31.854<br>c2_log10_biz_review_count 30.807<br>c5_log10_usr_review_count 13.179<br>c7_review_size 11.422<br>c14_log10_biz_review_count 8.781<br>c8_sentiment 5.916<br>c12_cluster_similarity 2.553<br>c10_cluster_similarity 0.000 | c2_item_count 100.00<br>c5_item_count 90.51<br>c1_item_count 88.05<br>c9_item_count 84.20<br>c14_item_count 76.11<br>c10_item_count 73.16<br>c12_item_count 63.23<br>c7_item_count 58.73<br>c5_sentiment 56.73<br>c15_item_count 55.07<br>c11_item_count 44.02<br>c9_useful_review_count 43.11<br>c1_pct_quest 35.83<br>c11_useful_review_count 34.05<br>c13_pct_periods 27.94<br>c5_useful_review_count 27.63<br>c5_review_size 24.63<br>c14_pct_apost 21.26<br>c14_pct_periods 20.79<br>c7_pct_quest 0.00 |
| mtry Accuracy Kappa<br>2 0.6835384 0.3556203<br>11 0.6573741 0.3056336<br>20 0.6530648 0.2961032<br><br>Reference<br>Prediction bad good<br>bad 41.1 17.4<br>good 14.3 27.3<br><br>Accuracy (average) : 0.6835 | mtry Accuracy Kappa<br>2 0.7716560 0.5421977<br>11 0.7695507 0.5382014<br>20 0.7626240 0.5246029<br><br>Reference<br>Prediction bad good<br>bad 41.4 12.7<br>good 10.1 35.8<br><br>Accuracy (average) : 0.7717 |
| VALIDATION FOR 70% GOOD REVIEW THRESHOLD<br> <br>threshold accuracy<br>0.4220000 0.7717391<br><br>Confusion Matrix and Statistics<br><br>bad good<br>bad 38 8<br>good 13 33<br><br>Accuracy : 0.7717<br>95% CI : (0.6725, 0.8528)<br>No Information Rate : 0.5543<br>P-Value [Acc > NIR] : 1.209e-05 | VALIDATION FOR 85% GOOD REVIEW THRESHOLD<br> <br>threshold accuracy<br>0.4950000 0.8494624<br><br>Confusion Matrix and Statistics<br><br>bad good<br>bad 42 8<br>good 6 37<br>Accuracy : 0.8495<br>95% CI : (0.7603, 0.9152)<br>No Information Rate : 0.5161<br>P-Value [Acc > NIR] : 1.423e-11 |

```
                  Kappa : 0.5435
 Mcnemar's Test P-Value : 0.3827

              Precision : 0.7174
                 Recall : 0.8049
                     F1 : 0.7586
             Prevalence : 0.4457
         Detection Rate : 0.3587
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.7750

       'Positive' Class : good

        Pearson's Chi-squared test with
Yates' continuity correction

data:  my_table
X-squared = 25.343, df = 1, p-value = 4.799e-
07
```

```
                  Kappa : 0.6982
 Mcnemar's Test P-Value : 0.7893

              Precision : 0.8605
                 Recall : 0.8222
                     F1 : 0.8409
             Prevalence : 0.4839
         Detection Rate : 0.3978
   Detection Prevalence : 0.4624
      Balanced Accuracy : 0.8486

       'Positive' Class : good

        Pearson's Chi-squared test with Yates'
continuity correction

data:  my_table
X-squared = 42.658, df = 1, p-value = 6.52e-11
```

Table 6-3. Model training and validation for 70% and 85% good review models with a prediction of "good system" one time frame in advance.

## 11. References

[1] Andersen, I. How Much Are Online Reviews Actually Worth? *RevLocal*. Retrieved Jan. 31, 2019. https://www.revlocal.com/blog/reviews/how-much-are-online-reviews-actually-worth-

[2] Bialik, C. Restaurant Ratings On Yelp Are Remarkably Consistent, No Matter Who's Writing Them, When, And Where. *Yelp Official Blog*. Sept. 14, 2018. https://www.yelpblog.com/2018/09/restaurant-ratings-on-yelp-are-remarkably-consistent-no-matter-whos-writing-them-when-and-where

[3] Business Wire. New Research Quantifies the Monetary Value of an Online Share for the First Time. *Business Wire.* April 29, 2014. https://www.businesswire.com/news/home/20140429006733/en/Research-Quantifies-Monetary-Online-Share-Time#.U3JeCa1dVuC

[4] Dubey, A. Feature Selection Using Random forest: the Wisdom of Crowds. *Towards Data Science.* Dec. 14, 2018. https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

[5] International Council of Shopping Centers (2017). *The Successful Integration of Food & Beverage Within Retail Real Estate.* https://www.icsc.org/uploads/research/general/Food__Beverage_Study_US.pdf

**[6] Magma. (n.d.). In *Dictionary.com*. Retrieved from https://www.dictionary.com/browse/magma**

[7] The New York Times (2018 April 10). Mark Zuckerberg Testimony: Senators Question Facebook's Commitment to Privacy. *The New York Times.* Retrieved from https://www.nytimes.com/2018/04/10/us/politics/mark-zuckerberg-testimony.html

[8] RUser (username). How to compute ROC and AUC under ROC after training using caret in R? *Stack Overflow.* Sept. 3, 2015. https://stackoverflow.com/questions/30366143/how-to-compute-roc-and-auc-under-roc-after-training-using-caret-in-r

[9] Sekaran, Uma; Bougie, Roger (2016). *Research Methods For Business: A Skill Building Approach, 7th Edition*. Wiley. Kindle Edition.

[10] Yelp Dataset Challenge. https://www.yelp.com/dataset/challenge (Requires filling out a form to obtain the dataset).

[11] Zheng, A. (2015). *Evaluating Machine Learning Models.* Sebastopol, CA: O'Reilly Media Inc.

[12] Zheng, A. (2017). *Mastering Feature Engineering.* Sebastopol, CA: O'Reilly Media Inc.