

MT5758: Project 1

April 2014

120022775

Applied Statistics & Data Mining, MsC

Contents

[**1.** **Introduction** 3](#_Toc386538661)

[**2.** **Research Questions** 4](#_Toc386538662)

[**3.** **Implications** 4](#_Toc386538663)

[**4.** **Data Concerns** 4](#_Toc386538664)

[**5.** **Data Set Description** 5](#_Toc386538665)

[**6.** **Cleaning & Descriptive Analysis** 5](#_Toc386538666)

[**7.** **Methods** 7](#_Toc386538667)

[**7.1** **PCA** 7](#_Toc386538668)

[**7.2** **Clustering Analysis** 8](#_Toc386538669)

[**8.** **Result** 10](#_Toc386538670)

[**8.1** **PCA** 10](#_Toc386538671)

[**8.2** **Clustering** 11](#_Toc386538672)

[**8.3** **High Frequent Raters vs. Low Frequent Raters** 11](#_Toc386538673)

[**8.4** **False/Suspicious Ratings** 12](#_Toc386538674)

[**8.5** **Beer Recommendation Application** 12](#_Toc386538675)

[**9.** **Further Studies and Improvements** 13](#_Toc386538676)

[**10.** **References and Literature use** 13](#_Toc386538677)

[**11.** **Software used** 13](#_Toc386538678)

[**12.** **Tables and figures** 14](#_Toc386538679)

1. **Introduction**

The aim of this project is to explore real data using various multivariate analysis techniques. After considering a variety of datasets the group decided to use a dataset of online beer reviews which was made available by the Stanford Network Analysis Project. The reviews were collected by BeerAdvocate.com , an independent community of beer enthusiasts dedicated to supporting and promoting beer.

The data set contains approximately 1.5 million online beer reviews which were collected between 1998 and 2011. A large variety of international beers were scored on a scale of 1-5 (at intervals of 0.5) for the following categories: appearance, aroma, palate, taste and overall rating. Individual beers can be identified by their name and brewing company. The dataset also contains information about alcohol percentage of the beer, beer style, the time and date from when the review was registered (as a Unix timestamp), profile name of the reviewer and a free text field containing a personal review from the beer consumer. No other information was collected about the reviewer, such as gender, age, or nationality. Figure 1 displays a screenshot of a test review from BeerAdvocate.com.

As well as being of interesting subject matter and free availability, this dataset lends itself well to multivariate techniques as many of the beer attributes are inherently related to each other. For example, it is well known that aroma and taste are related in some way, as well as taste and palate. Including all of the attributes in our analysis will allow us to explore these complex relationships.

1. **Research Questions**

The following research questions are of interest to the group:

* 1. What characterizes a beer that has a high or low overall rating? We are interested in the variables that significantly affect the overall rating, and their relationship to both overall rating and each other.
  2. How prevalent is bias (both intentional and unintentional) in large scale online review websites? We are interested in the following sources of bias that may affect a beers rating:

• Repeated reviews of a beer from a single user

• Consistently extreme reviews from a single user (very high or very low)

• False reviews (e.g. high ratings on the individual attributes but a low overall rating)

* 1. Can we identify beers that are similar to each other? This would allow beer companies to successfully target potential customers.
  2. Can we identify time changes in beer preference? This may reflect seasonal or longer term changes associated with changing consumer preferences or brewing style.

**Jonathan**

* 1. Can we identify a gap in the market for a beer, based on consumer preferences?

**Natasha**

1. **Implications**

The results from this analysis would be useful to beer brewers, merchants and consumers. Beer brewers would be able to identify the characteristics that lead to a highly rated beer which may improve their current product. They would also be able to identify close competitors in the market and potential gaps in the market. Beer merchants would benefit as they more effectively manage the stock of preferred beer types and recommend alternative beers that are similar to each other. Thus improving their customer offering and enhancing their revenues. Beer consumers could use the results to identify highly rated beers and find beers that are similar to ones that they enjoy. Finally, interested researchers can benefit from our insights into the magnitude of bias on online review sites and the possible types of bias.

1. **Data Concerns**

As is common with online data collection open to the general public there are initial concerns about the validity of the data. Initial exploration suggests some ratings lie outside the 1 to 5 range (at 0), which may correspond to missing data or to a very low rating. Similarly, alcohol percentage is out of the ‘normal’ range for beer in some reviews and missing in others, however this should be easily resolved by investigating the beer itself. With regards to the beer reviewers, there may be abundant bias concerned with promoting beers they are affiliated with and criticizing the opposition. For example there are consistently low reviews on different beers by a reviewer with the profile name “dog” and some reviewers have rated the same beer several times. It must be decided how to deal with these reviewers such as by eliminating their review, taking the most recent or averaging them.

1. **Data Set Description**

The data set contains 1 586 614 reviews with the following variables:

1. ***name:*** categorical The name of the beer that was reviewed

2. ***beer\_id:*** numerical Numeric beer\_id of a specific beer

3. ***brewer\_id:*** numerical Brewer\_id of the beer

4. ***ABV:*** continuous Alcohol percentage in beer

5. ***beerStyle:*** categorical Beer style of rated beer, BeerAdvocat.com have classified beers into certain types on the site. There are 104 beer styles registered in the data set.

6. ***reviewAppearance*** ordinal Review of the beers appearance on a scale from 1-5 with 0.5 intervals

7. ***reviewAroma:*** ordinal Review of the beers aroma on a scale from 1-5 with 0.5 intervals

**8. reviewPalate** ordinal Review of the beers palate on a scale from 1-5 with 0.5 intervals

**9.** **reviewTaste** ordinal Review of the beers taste on a scale from 1-5 with 0.5 intervals

**10.** **rating** ordinal Overall rating of a beer on a scale from 1-5 with 0.5 intervals

**11.** **Time** datetim e UNIX time stamp

**12.** **profileName** categorical Profile Name of the reviewer

1. **Cleaning & Descriptive Analysis**

Prior to focusing on advanced multivariate methods some description of the dataset and how it was cleaned will be given in this section.

The first challenge was loading the data into R. The dataset is considerably large and time consuming to load. A common “**read.csv”** in R file was first attempted but this failed as it appears the data is stored in rows rather in the more conventionally way with rows as observations and variable as columns. There are various ways of doing things in R but useful functions for loading big data into efficiently into R were found. [[1]](#footnote-2)

Once the data was loaded into R some initial exploratory analysis was conducted in order to uncover potential cleaning schemes:

* The dataset contains 1 586 607 reviews from 2002-2011
* Number of reviews with missing alcohol content was 67785. These appeared to be non-systematic missing variables, meaning they did not seem to relate to whether the beer was alcohol free or not or to a certain type of group observations
* There were 7 reviews that had 0 values for review appearance and overall rating. 0 is a value that is not possible review value hence these values were removed.
* The number of duplicate ratings through time (same user rating the same beer) was 14294 (about 0.8% of the total reviews). It was decided to only keep latest rating a user has made on a beer an remove the other ones. The dataset had 1 571 808 rows after this cleaning the duplicate rows.
* There are 104 unique beer groups represented in the data set which the various beer are grouped under. See Table 2.
* Reviews pr. profile is 47.07 meaning that a lot of users seem to be rating lot. Interestingly 20% of the raters stand for 92.18 % of the ratings.
* Figure 2 to Figrure 7 contain frequency plots of the review categories. The reviews appear to be left skewed meaning that raters tend to use the upper bound of the scale (giving high rating) as opposed to lower ratings.
* The correlation matrix in Table 1 shows that, not so surprisingly, that taste and overall rating appear to have the highest positive correlation (0.78). Overall rating vs. the beers alcohol percentage appear to have the lowest positive correlation 0.13, meaning that for example a good overall rating and alcohol high percentage does not appear to be related.
* Not surprisingly to beer knowers the Belgian Quadrupel [[2]](#footnote-3)beer style appears to be the beer style that has the highest probability (See Figure 8) of receiving a 5 on overall rating. In this analysis only the beer styles that have been rated more than 100 times are considered.
* “Pliny the Younger”, a beer brewed in Santa Rosa, CA appears to have the highest probability of receiving a 5 (about 40% chance) from the beers that have been rated over 100 times and taking a short search on the net this appears to be the case [[3]](#footnote-4) The next on the list is a Belgium beer called “Trappist Westvleteren 12 (XII)”[[4]](#footnote-5) (See Figure 9)
* Figure 10 shows beer that have the highest porportions of ones to the total number of reviews. “Crazy Ed's Cave Creek Chili Beer “comes out as the worst beer. Some funny review examples are listed in the very back of the Table and Figures Section, Example 1.:

1. **Methods**

Two multivariate methods were used in this project; PCA and clustering. Both these methods have been covered during lectures and leans nicely towards this particular dataset for exploratory data analysis. The reason for this is that we want to explore patterns in a multidimensional space without having too many hypothesis or pre assumptions on the data set.

In R there are methods for carrying out PCA quickly but to give the reader a broader overview of the method it will be shortly described in the following section:

* 1. **PCA**

The idea of principal component analysis (PCA) is to find a small number of linear combinations of the variables so as to capture most of the variation in a dataset as a whole, in our case the review data. In our data set we had 6 attributes and if possible (without losing too much information) it would be practical to consider a small number of combinations of the review data and alcohol content rather than all six attributes. Principal components analysis finds a set of orthogonal standardized linear combinations which together explain all the variation in the original data. There are as many principal components as there are variables, but typically it is only the first few of them that explain important amounts of the total variation. In R there are few functions for carrying out PCA automatically. The general preferred method for numerical accuracy is prcomp (where the calculation is done by a singular value decomposition of the centered and scaled data matrix, not using eigenvalues on the covariance matrix, as in the alternative function princomp[[5]](#footnote-6)). The following process conceptual overview of how the PCA components are calculated:

1. Get some multivariate data (preferably more than 3 dimensions as graphing can then be used as a visual tool to a greater extent)
2. Subtract the mean. All the means of each dimensions (rating categories and alcohol) is subtracted from each observation. This produces a dataset with mean 0.
3. Calculate the covariance matrix. This will in terms calculate the variance between each of the covariates. We have 6 dimensions (that we care about) in the data set hence the squared 6x6 matrix will contain covariance metrics including the dimensions own variance.
4. Calculate the eigenvectors and eigenvalues of the covariance matrix. Since the covariance matrix is square, we can calculate the eigenvectors and eigenvalues for this matrix. These are rather important, as they tell us useful information about our data. Basically these values attempt to reduce the dimensions (using new reduced dimensions) of the data but maintaining the information contains. In our case we have a 6x6 covariance matrix and hence most likely 6 eigenvectors and associated eigenvalues. One example is for example if the relationship between taste and overall rating seems to be linear, we can then find the eigenvector and associated eigenvalue and reduce this relationship to a single line (or a new dimension)
5. Choosing components and forming a feature vector. The eigenvector with the highest eigenvalue is the principle component of the dataset. In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of signiﬁcance. Now, if you like, you can decide to ignore the components of lesser signiﬁcance. You do lose some information, but if the eigenvalues are small, you don’t lose much. If you leave out some components, the ﬁnal data set will have less dimensions than the original. To be precise, if you originally have dimensions in your data, and so you calculate eigenvectors and eigenvalues, and then you choose only the ﬁrst n eigenvectors, then the ﬁnal data set has only n dimensions.
6. This the ﬁnal step in PCA, and is also the easiest. Once we have chosen the components

(eigenvectors) that we wish to keep in our data and formed a feature vector, we simply

take the transpose of the vector and multiply it on the left of the original data set,

transposed. We then get the values mapped onto a pane.

PCA will be used to explore the relationships between the various rating categories.

* 1. **Clustering Analysis**

The aim of clustering analysis or classification has the simple aim to finding natural clusters or partitioning of the data set based on the sampling units without having prior information to what cluster a point belongs to.

The are two main methods in clustering analysis; partitioning and hierarchical clustering. The hierarchical methods are restricted to a nested structure where each level is constrained to the previous. For portioning or segmenting methods the solution is at any level is independent.

The k-means method usually performs better than the hierarchical methods of clustering. [[6]](#footnote-7) The idea with this method is that set of initial seeds or centers are set as starting points. For this method you have to give your initial idea of how many clusters there are. The algorithm can be described short as follows:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.[[7]](#footnote-8)

The metric that is usually minimized is called trace(W) where W is the within cluster variance co-variance matrix pooled over all cluster.

It is important to note that the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect.

Hierarchal clustering is another method that will be attempted in this project and compared with the k-means method to see whether they display similar patterns. The main difference to partition is that these methods assume that the groupings in the data have a hierarchical structure. We will be using an agglomerative method [[8]](#footnote-9) (using group average linkage for distance measure when fusing) which starts from the individual sampling units forming them into groups and fusing the groups till there is only one that includes all the points (bottom-up).

The algorithm for hierarchical clustering may be shortly summed up by the following steps:

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar), based on group average linkage, pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. Of course there is no point in having all the N items grouped in a single cluster but, once you have got the complete hierarchical tree, if you want k clusters you just have to cut the k-1 longest links.[[9]](#footnote-10)

There are several packages for producing dendograms and cluster analysis in R and these will be utilized for analysis.

1. **Result**
   1. **PCA**

We attempted a simple PCA on the whole data set at first using the scaled ratings and alcohol content as inputs. This was to identify if we could reduce the ratings overview and alcohol to a single or fewer dimensions/entities that contain almost the same information. Specifically one of the purposes for doing this was to see whether we could use any of the PCA components for a beer recommendation system. If for example one PCA is apparent in describing a good beer and explains a high proportion of the variance in the data then this would leans itself nicely to be used in a recommendation system.

As seen in Table1 there are as many PCA components as there are variables and the table also states the amount of variance that explaine by each component. PC1 and PC2 represents 0.7649 of the total variance in the data cloud (See Table 1).

In Table 2 the first PC may overall seem to explain what the reviewer think of as a good beer. All the review loadings are postive meaning that the higher postive difference from the scaled mean the higher this score/loading will be. Taste seems to be (not surprisingly) the strongest contributor. This component captures 61% of the dataset. The second PC may give some indication what a bad beer is. The score for this compoent comes out higher if the lower the overall rating is and if the alchohol content is very high/meaning strong beer. The other variables have minor contributions. In other words by this we can say that a segment of the reviweres think a beer is in general bad if the alchohol conent is very high. This component captures 15% of the variance in the data. The most notable compoent following this one is PC6. It only accounts for about 3% in the data cloud, but the main take away is that the overall rating and taste rating goes in oposite directions. The score for this component will be high if the taste rating is high but the overall rating is low. This does not make much sense as these are the highest correlated values (78.8%) and we suspect that these ratings may be biased somehow, either the user makes mistakes or dilabratly makes these kind of reviews. Please refer to section 8.4 False/Supiscious reviews.

* 1. **Clustering**

Based on the two first PCA components we wanted to see how well clustering methods would be able to discriminate good and bad beers. Since the two first PCs explain about 75% of the variance we would expect clustering methods to do fairly well in discriminating between good and bad beers depending on the definition we use for good and bad beers.

We decided for illustrative purposes to do select the top 3 beers that have highest proportion fives in overall rating and over 1000(ratings) and the top 3 beers that have the highest proportion of 1s on overall rating. First k-means using 2 seed/centers was used and the result is displayed in Figure 15. As seen the clustering algorithm is discriminating fairly well between the good and bad beers based on the compressed rating information contained in PC1 and PC2.

Figure 16 shows a simple dendogram of a random sample of the top3 and bottom3 beers. A red line indicating a separation on 2 clusters is also highlighted. As with k-means the hierarchical clustering methods (using group average distances), based on the sample, also seem to performing well in discriminating between good bad rated beers only using PC1 and PC2. Figure 17 shows all the top 3 beers and bottom 3 beers reviews clustered based on hierarchical clusters. It gives almost the same picture as the k-means algorithm.

Naturally there are some beers that can both have a high proportion of 1s of reviews and also a high proportion of 5s, as seen in Figure 15 and Figure17 these it is difficult to say which clusters these should belong to.

The intention of this section is to show that even if we have compressed rating from the data cloud in PC1 an PC2 we are still able to discriminate fairly well between good beer and beers. The clustering algorithms may be used for more extensive purposes such as e.g. identifying similar beers or users in a recommendation system. We have made a content based filtering recommendation(CBFR) [[10]](#footnote-11)(See section 8.5) system where the algorithms tries to recommend items that are similar to those that a user liked in the past. We used PCs when cread our CBFR recommendation system.

Collaborative filtering methods [[11]](#footnote-12) [[12]](#footnote-13)are based on collecting and analyzing a large amount of information on users’ behaviors, activities or preferences and predicting what users will like based on their similarity to other users. In this case e k-nearest neighbor may be used to identify the closeness to a cluster of users with same rating characterstics and make recommendations based on this rather on specific historical actions. A practical application of this would be to for example try to seperate the data into 10 clusters based on PC1 and PC2. For each of the clusters take the beers that have the highest ratings inside these cluster. So based on a new rating or/and including your previous ratings which cluster do the specific user rating /user rating history belong to? The fit function in R for k-means may be used for this. The top rated beers are then retrieved for the clusters you belong to and recommended as the beers you may like (excluding the ones you have already time)

* 1. **High Frequent Raters vs. Low Frequent Raters**

One of the tasks assigned internally in the group was to see whether high frequent raters (> 10 rating) have a different rating pattern than low frequent raters (less than or equal to10 ratings). For this task we carried out a PCA first on the profiles that had more than 10 ratings and the ones that had less than 10. The intention of doing this was to see whether different PCA components will emerge from the different segments indicating differences in rating patterns between the two groups. 1 441 901 of the ratings were made by profiles that made more than 10 ratings in the data set and 62468 rating were made by profiles having a total rating frequency less than 10.

The two most important PCA components seem to portray the same picture for both segments, meaning the two segments seem to have more or less the same overall opinion of what is a good beer and what is a bad beer. There are minor differences in the smaller components but they do not explain a lot of the total variance. The standard deviations of the two first components seem to be higher for the low raters group and this may be driven by lower sample size. Also it may well be that low raters use the width of the rating scales in larger sense.

Figure 11 shows that low rates may seem to be less critical as they have a higher proportion of higher overall ratings compared to the high frequent raters. This may be the case as those who have tried many beer types through the years may be more prone of being more critical towards various beer types.

* 1. **False/Suspicious Ratings**

One of the PCA components/eigenvectors that was found in the data set seem to display a potentially suspicious relationship between taste and overall. See Table 8, PCA 5. As seen the scores or loading go in opposite direction meaning that high rating for taste would imply a low overall rating. This does not make a lot of sense as taste rating and overall rating is the relationship that has the highest correlation in the dataset (See Table 1). One hypothesis was that this PC describes user error ratings or false ratings.

A quick look at the top 10 000 observations with highest score for this component reveals that the taste ratings indeed are quite different from the overall ratings. See Figure 12. Even though the component does not describe a lot of the variance in the dataset on a total basis (probably because there are not that many observations) it still provides some useful insights and the relationship should be investigated further to find out if this is a systematic error or random. With systematic error it is meant that for example there might be specific profiles or periods in time that stands for these types of suspicious reviews.

* 1. **Beer Recommendation Application**

One of the research questions we set out to investigate was to identify which beer is similar to one another. We took a practical approach on this and decided to attempt to make a simple beer recommendation system. The book “Programming Collective Intelligence: Building Smart Web 2.0 Applications” By Toby Segaran, outlined some nice ideas on how do this on reduced dimensions/PCA.

We decided to take one of the ideas from which is set up as follows:

* Providing the application of a beer you like, in return the application lists top n (5 in our case) other beer you may like based on experience of other people who have had the beer you liked.

Based on the idea from the book we develop an application that did the following in Shiny R: (Please refer to the code section (13.Code) for details on how this was implemented)

1. PCA1 seem to explain over 60% of the variance in the dataset and to some extent describe what a good beer means in terms rating, reading the loading. (See table x.x). Hence we score all the reviews based on PCA so each review had a PCA1 score.
2. The top 50 rated beers were selected (for simplicity and code execution time) for this application and a 50x 50 matrix was created with all the beer names in the rows and columns.
3. For each of the beer combinations, profiles that had made a reviews on the two beers were identified. The associated PCA1 score was also found.
4. Then the correlation between each of the beer pairs were made based on PCA1 and stored in the 50x50 matrix. This matrix was then stored on the hard drive for use in the beer recommendation application. See illustrative example in Table 3:
5. A simple GUI was made were you can select between the 50 beer types and then the top 5 correlations to other beers based on PCA1 score pops up. (See Figure 13)

This code can be applied on all beer combinations as long as review combinations exist between a particular beer and another one (even though if there only exists 1 or 2 reviews between a pair the recommendation does not become very reliable). This application may be enhanced by e.g. including the PCA2 scores as well and create a second correlation matrix and e.g weight the correlations of PCA1 and PCA2. By doing so one account for more variance in the data when recommendations are made.

1. **Further Studies and Improvements**
2. **References and Literature**

[1] (Paper reference) J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

[2] (Web-Book, 2011) Collaborative Filtering Recommender Systems

By Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan <http://files.grouplens.org/papers/FnT%20CF%20Recsys%20Survey.pdf>

[3] (Book) “Programming Collective Intelligence: Building Smart Web 2.0 Applications” By Toby Segaran

1. **Software used**

R, Microsoft Excel 2010

1. **Tables and figures**

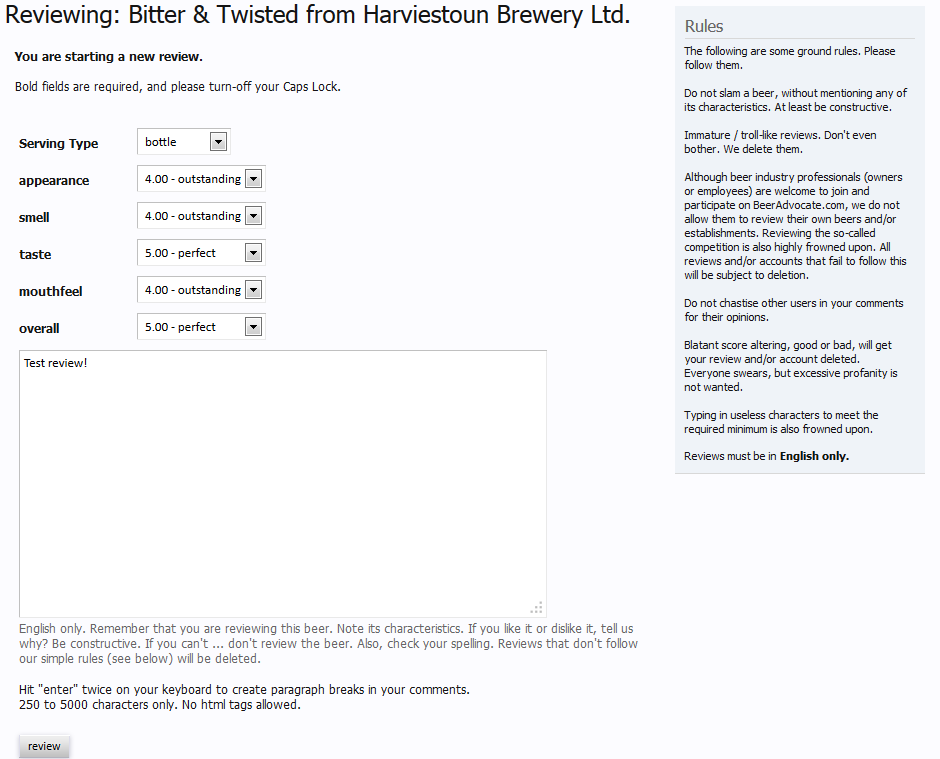


Figure 1: Screenshot of a test review http://www.beeradvocate.com

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** |
| Standard deviation | 1.9161 | 0.9581 | 0.74097 | 0.62615 | 0.53263 | 0.43097 |
| Proportion of Variance | 0.6119 | 0.153 | 0.09151 | 0.06534 | 0.04728 | 0.03096 |
| Cumulative Proportion | 0.6119 | 0.7649 | 0.85642 | 0.92176 | 0.96904 | 1 |

Table 1: Importance of componant. PCA Table from R using all review variables and alcohol content

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Scores / Loadings** |  |  |  |  |  |  |
| **Variable** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** |
| AlcoholContent | 0.2141756 | 0.9312366 | 0.18914893 | -0.1395482 | 0.1550273 | -0.0874225 |
| reviewAppearance | 0.3847739 | 0.04591445 | -0.8934419 | -0.1246567 | 0.1874952 | 0.0301445 |
| reviewAroma | 0.4367273 | 0.04914228 | 0.00769209 | 0.82189892 | -0.2903324 | -0.2167587 |
| reviewPalate | 0.448397 | -0.0872295 | 0.10587183 | -0.5067474 | -0.7200605 | -0.0695856 |
| reviewTaste | 0.4706472 | -0.119537 | 0.26622112 | 0.04809865 | 0.2358234 | 0.7971212 |
| rating | 0.4390654 | -0.3261552 | 0.28955573 | -0.174249 | 0.5314314 | -0.5515611 |

Table 2: Loadings/Scores for each PC on the scaled variables in the dataset

|  |  |  |
| --- | --- | --- |
| ***profileName*** | ***PCA1 Score Beer 1*** | ***PCA1 Score Beer 2*** |
| beerQueen | 3.2 | 5.4 |
| beerorama | 3.2 | 5.2 |
| TheDude | 3.1 | 5.1 |
| Correlation/Similarity Score Beer 1 and Beer 2 |  | 0.755928946 |
|  |  |  |
| ***profileName*** | ***PCA1 Score Beer 3*** | ***PCA1 Score Beer 4*** |
| beerQueen | 5 | 3 |
| beerorama | 6 | 2 |
| TheDude | 6 | 1 |
| Correlation/Similarity Score Beer 3 and Beer 4 |  | -0.866025404 |

Table 3: Illustrative example of correlation between PCA1 for beer1 and beer2 and beer1 and beer 3.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Correlation Matrix** |  |  |  |  |  |  |
| **Variable** | **ABV** | **reviewAppearance** | **reviewAroma** | **reviewPalate** | **reviewTaste** | **rating** |
| ABV | 1 | 0.2636079 | 0.3320096 | 0.2862067 | 0.2903481 | 0.1383751 |
| reviewAppearance | 0.2636079 | 1 | 0.5584348 | 0.563913 | 0.5439085 | 0.4981777 |
| reviewAroma | 0.3320096 | 0.5584348 | 1 | 0.6143096 | 0.7143693 | 0.61281 |
| reviewPalate | 0.2862067 | 0.563913 | 0.6143096 | 1 | 0.7318339 | 0.6989623 |
| reviewTaste | 0.2903481 | 0.5439085 | 0.7143693 | 0.7318339 | 1 | 0.7874138 |
| rating | 0.1383751 | 0.4981777 | 0.61281 | 0.6989623 | 0.7874138 | 1 |

Table 4: Correlation Matrix between the rating variables

|  |  |  |
| --- | --- | --- |
|  | **Name** |  |
| [1] | "Altbier" | "American Adjunct Lager" |
| [3] | "American Amber / Red Ale" | "American Amber / Red Lager" |
| [5] | "American Barleywine" | "American Black Ale" |
| [7] | "American Blonde Ale" | "American Brown Ale" |
| [9] | "American Dark Wheat Ale" | "American Double / Imperial IPA" |
| [11] | "American Double / Imperial Pilsner" | "American Double / Imperial Stout" |
| [13] | "American IPA" | "American Malt Liquor" |
| [15] | "American Pale Ale (APA)" | "American Pale Lager" |
| [17] | "American Pale Wheat Ale" | "American Porter" |
| [19] | "American Stout" | "American Strong Ale" |
| [21] | "American Wild Ale" | "Baltic Porter" |
| [23] | "Belgian Dark Ale" | "Belgian IPA" |
| [25] | "Belgian Pale Ale" | "Belgian Strong Dark Ale" |
| [27] | "Belgian Strong Pale Ale" | "Berliner Weissbier" |
| [29] | "BiÃ¨re de Champagne / BiÃ¨re Brut" | "BiÃ¨re de Garde" |
| [31] | "Black & Tan" | "Bock" |
| [33] | "Braggot" | "California Common / Steam Beer" |
| [35] | "Chile Beer" | "Cream Ale" |
| [37] | "Czech Pilsener" | "Doppelbock" |
| [39] | "Dortmunder / Export Lager" | "Dubbel" |
| [41] | "Dunkelweizen" | "Eisbock" |
| [43] | "English Barleywine" | "English Bitter" |
| [45] | "English Brown Ale" | "English Dark Mild Ale" |
| [47] | "English India Pale Ale (IPA)" | "English Pale Ale" |
| [49] | "English Pale Mild Ale" | "English Porter" |
| [51] | "English Stout" | "English Strong Ale" |
| [53] | "Euro Dark Lager" | "Euro Pale Lager" |
| [55] | "Euro Strong Lager" | "Extra Special / Strong Bitter (ESB)" |
| [57] | "Faro" | "Flanders Oud Bruin" |
| [59] | "Flanders Red Ale" | "Foreign / Export Stout" |
| [61] | "Fruit / Vegetable Beer" | "German Pilsener" |
| [63] | "Gose" | "Gueuze" |
| [65] | "Happoshu" | "Hefeweizen" |
| [67] | "Herbed / Spiced Beer" | "Irish Dry Stout" |
| [69] | "Irish Red Ale" | "Japanese Rice Lager" |
| [71] | "KÃ¶lsch" | "Keller Bier / Zwickel Bier" |
| [73] | "Kristalweizen" | "Kvass" |
| [75] | "Lambic - Fruit" | "Lambic - Unblended" |
| [77] | "Light Lager" | "Low Alcohol Beer" |
| [79] | "MÃ¤rzen / Oktoberfest" | "Maibock / Helles Bock" |
| [81] | "Milk / Sweet Stout" | "Munich Dunkel Lager" |
| [83] | "Munich Helles Lager" | "Oatmeal Stout" |
| [85] | "Old Ale" | "Pumpkin Ale" |
| [87] | "Quadrupel (Quad)" | "Rauchbier" |
| [89] | "Roggenbier" | "Russian Imperial Stout" |
| [91] | "Rye Beer" | "Sahti" |
| [93] | "Saison / Farmhouse Ale" | "Schwarzbier" |
| [95] | "Scotch Ale / Wee Heavy" | "Scottish Ale" |
| [97] | "Scottish Gruit / Ancient Herbed Ale" | "Smoked Beer" |
| [99] | "Tripel" | "Vienna Lager" |
| [101] | "Weizenbock" | "Wheatwine" |
| [103] | "Winter Warmer" | "Witbier" |

Table 5: Unique beer groups represented in the dataset. Some styles have unique unicoding

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Frequency Appearance | |  |  |  | Frequency Aroma | |  |  |
| **Interval** | **Frequency** | **Cum Freq** | **relative** |  | **Interval** | **Frequency** | **Cum Freq** | **relative** |
| (0.5,1] | 3291 | 3291 | 0.21% |  | (0.5,1] | 6815 | 6815 | 0.43% |
| (1,1.5] | 6071 | 9362 | 0.39% |  | (1,1.5] | 12384 | 19199 | 0.79% |
| (1.5,2] | 25187 | 34549 | 1.60% |  | (1.5,2] | 42221 | 61420 | 2.69% |
| (2,2.5] | 39178 | 73727 | 2.49% |  | (2,2.5] | 65858 | 127278 | 4.19% |
| (2.5,3] | 164834 | 238561 | 10.49% |  | (2.5,3] | 198599 | 325877 | 12.64% |
| (3,3.5] | 316147 | 554708 | 20.11% |  | (3,3.5] | 362513 | 688390 | 23.06% |
| (3.5,4] | 667809 | 1222517 | 42.49% |  | (3.5,4] | 552086 | 1240476 | 35.12% |
| (4,4.5] | 284812 | 1507329 | 18.12% |  | (4,4.5] | 268182 | 1508658 | 17.06% |
| (4.5,5] | 64479 | 1571808 | 4.10% |  | (4.5,5] | 63150 | 1571808 | 4.02% |
|  |  |  |  |  |  |  |  |  |
| Frequency Palate | |  |  |  | Frequency Taste | |  |  |
| **Interval** | **Frequency** | **Cum Freq** | **relative** |  | **Interval** | **Frequency** | **Cum Freq** | **relative** |
| (0.5,1] | 6818 | 6818 | 0.43% |  | (0.5,1] | 9916 | 9916 | 0.63% |
| (1,1.5] | 10941 | 17759 | 0.70% |  | (1,1.5] | 15022 | 24938 | 0.96% |
| (1.5,2] | 38044 | 55803 | 2.42% |  | (1.5,2] | 41695 | 66633 | 2.65% |
| (2,2.5] | 62392 | 118195 | 3.97% |  | (2,2.5] | 66125 | 132758 | 4.21% |
| (2.5,3] | 205574 | 323769 | 13.08% |  | (2.5,3] | 165756 | 298514 | 10.55% |
| (3,3.5] | 336030 | 659799 | 21.38% |  | (3,3.5] | 322264 | 620778 | 20.50% |
| (3.5,4] | 600939 | 1260738 | 38.23% |  | (3.5,4] | 536435 | 1157213 | 34.13% |
| (4,4.5] | 249798 | 1510536 | 15.89% |  | (4,4.5] | 331942 | 1489155 | 21.12% |
| (4.5,5] | 61272 | 1571808 | 3.90% |  | (4.5,5] | 82653 | 1571808 | 5.26% |
|  |  |  |  |  |  |  |  |  |
| Frequency Rating | |  |  |  | Frequency Alcohol Percentage % | | |  |
| **Interval** | **Frequency** | **Cum Freq** | **relative** |  | **Interval** | **Frequency** | **Cum Freq** | **relative** |
| (0.5,1] | 10891 | 10891 | 0.69% |  | (0,2] | 1339 | 1339 | 0.09% |
| (1,1.5] | 12885 | 23776 | 0.82% |  | (2,4] | 29748 | 31087 | 1.98% |
| (1.5,2] | 37998 | 61774 | 2.42% |  | (4,6] | 643369 | 674456 | 42.77% |
| (2,2.5] | 58140 | 119914 | 3.70% |  | (6,8] | 408246 | 1082702 | 27.14% |
| (2.5,3] | 164571 | 284485 | 10.47% |  | (8,10] | 283390 | 1366092 | 18.84% |
| (3,3.5] | 299644 | 584129 | 19.06% |  | (10,12] | 108706 | 1474798 | 7.23% |
| (3.5,4] | 577378 | 1161507 | 36.73% |  | (12,14] | 14968 | 1489766 | 1.00% |
| (4,4.5] | 320463 | 1481970 | 20.39% |  | (14,16] | 5645 | 1495411 | 0.38% |
| (4.5,5] | 89838 | 1571808 | 5.72% |  | (16,18] | 7302 | 1502713 | 0.49% |
|  |  |  |  |  | (18,20] | 914 | 1503627 | 0.06% |
|  |  |  |  |  | (20,22] | 66 | 1503693 | 0.00% |
|  |  |  |  |  | (22,24] | 20 | 1503713 | 0.00% |
|  |  |  |  |  | (24,26] | 105 | 1503818 | 0.01% |
|  |  |  |  |  | (26,28] | 357 | 1504175 | 0.02% |
|  |  |  |  |  | (28,30] | 16 | 1504191 | 0.00% |
|  |  |  |  |  | (30,32] | 89 | 1504280 | 0.01% |
|  |  |  |  |  | (38,40] | 10 | 1504290 | 0.00% |

Table 6: Frequnecy Distributions for various rating categories.

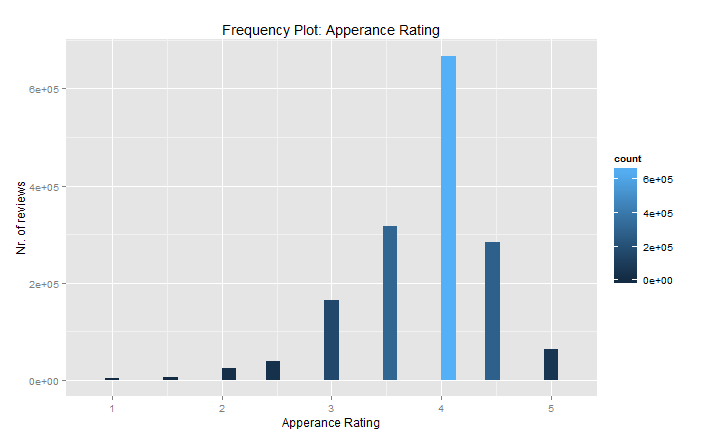


Figure 2: Apperance Rating Frequency Plot

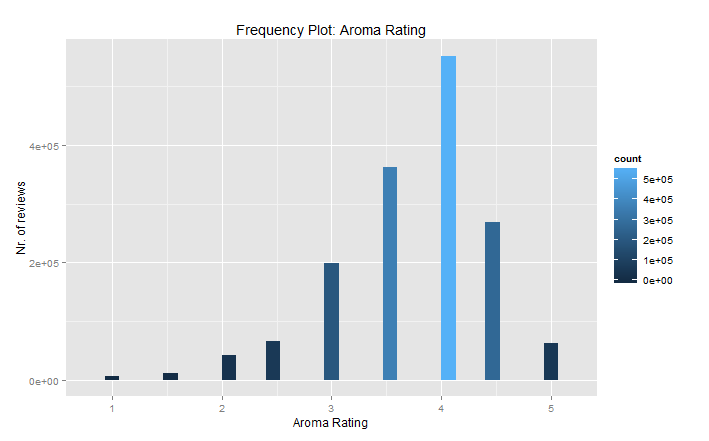


Figure 3: Aroma Rating Frequency Plot

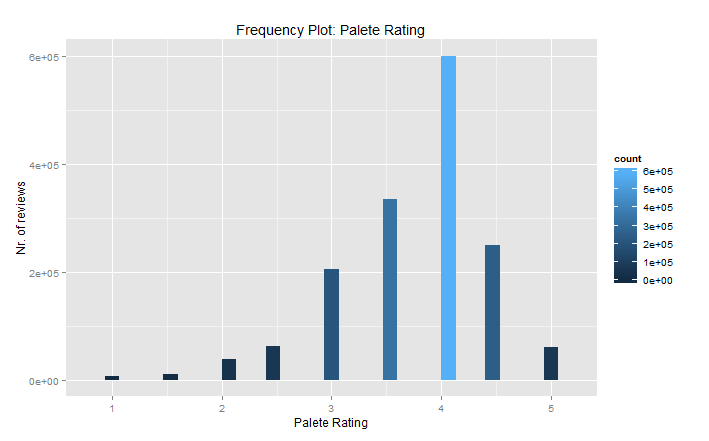


Figure 4: Palete Rating Frequency Plot

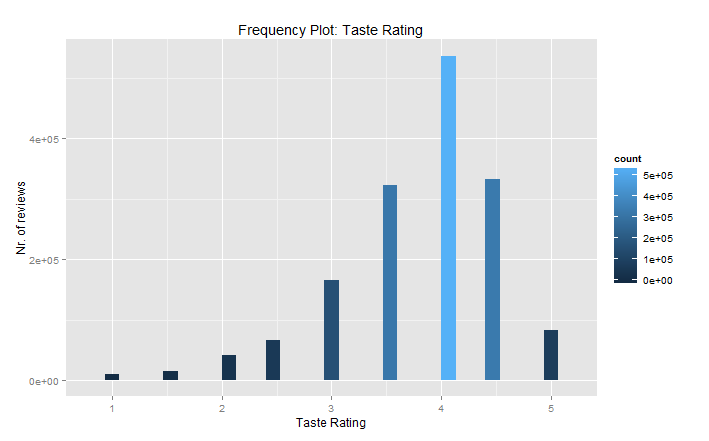


Figure 5:Taste Rating Frequency Plot

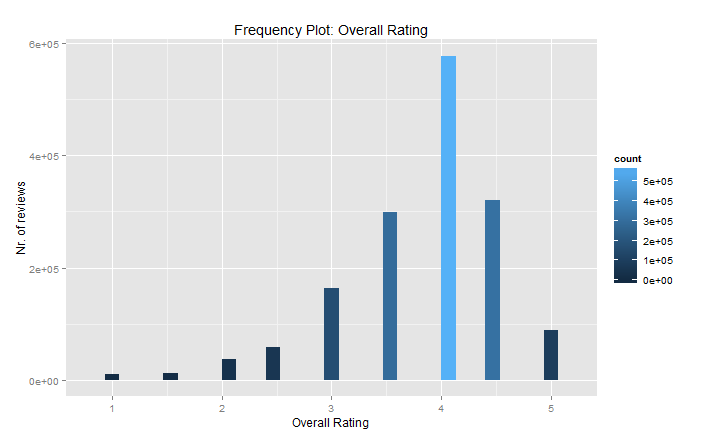


Figure 6: Overall Rating Frequency Plot

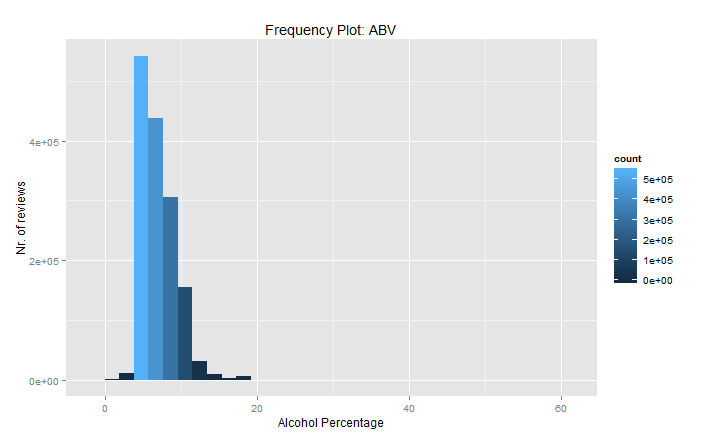


Figure 7: ABV Rating Frequency Plot

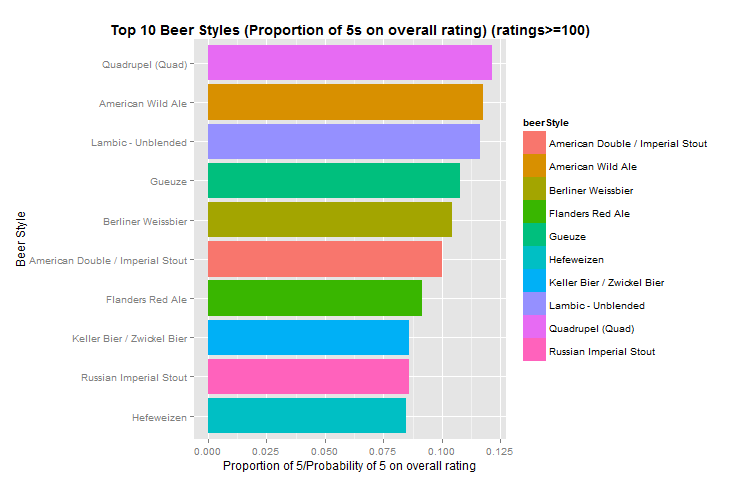


Figure 8: Probability of receiving 5 for beer styles that have been rated more than 99 times.

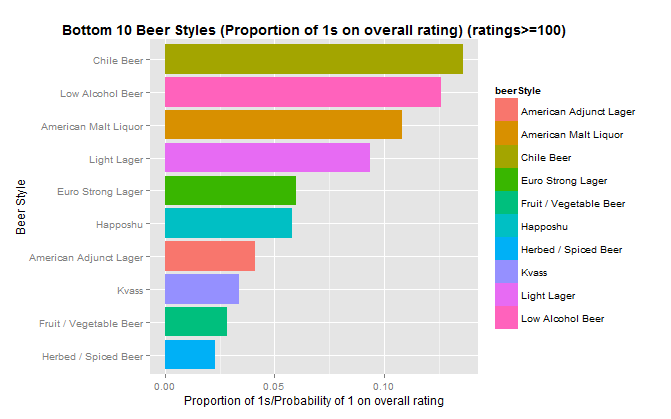


Figure 9: Probability of receiving 1 as rating for beer groups that have been rated more than 99 times

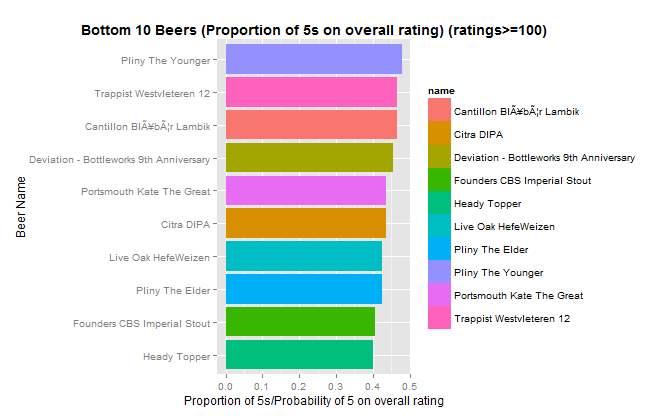


Figure 10: Probability of receiving 5 for beers that have been rated more than 99 times

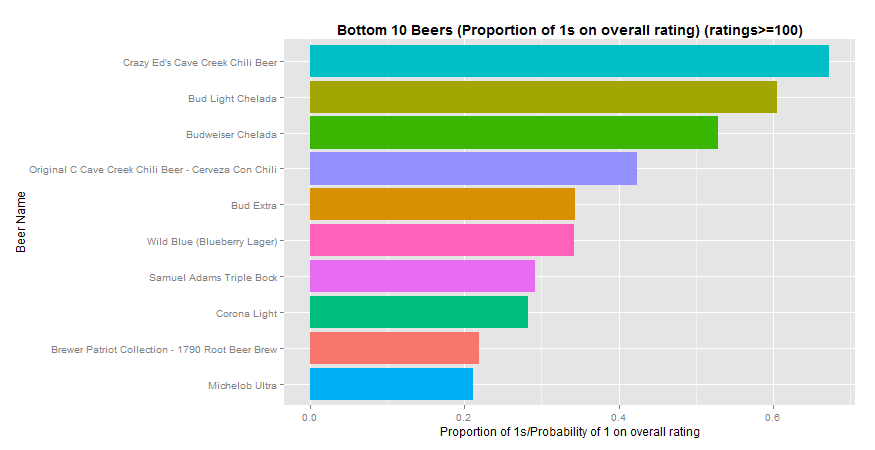


Figure 11: Probability of receiving 1 as rating for beers that have been rated more than 99 times

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Importance of components: |  |  |  |  |  |  |
| **Metric** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** |
| Standard deviation | 1.9136 | 0.9569 | 0.74489 | 0.6283 | 0.5339 | 0.4333 |
| Proportion of Variance | 0.6103 | 0.1526 | 0.09248 | 0.06579 | 0.0475 | 0.03129 |
| Cumulative Proportion | 0.6103 | 0.7629 | 0.85541 | 0.92121 | 0.9687 | 1 |

Table 7: Importance of high raters PCA component. High raters are defined as profiles making more than 10 reviews.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Importance of components: |  |  |  |  |  |  |
| **Metric** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** |
| Standard deviation | 1.9516 | 0.9767 | 0.6876 | 0.59029 | 0.508 | 0.39748 |
| Proportion of Variance | 0.6348 | 0.159 | 0.0788 | 0.05807 | 0.043 | 0.02633 |
| Cumulative Proportion | 0.6348 | 0.7938 | 0.8726 | 0.93066 | 0.9737 | 1 |

Table 8: Importance of low raters PCA component. Low raters are defined as profiles making less or equal than 10 reviews.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Frequent Raters (ratings > 10) | |  |  |  |  |  |
| **Variable** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** |
| ABV | 0.2164616 | 0.9306484 | 0.1845804 | -0.1418958 | 0.1569754 | -0.090511 |
| reviewAppearance | 0.3830577 | 0.04358 | -0.8966743 | -0.11295 | 0.1834087 | 0.0307618 |
| reviewAroma | 0.4368328 | 0.0492114 | 0.0172449 | 0.8185249 | -0.2993401 | -0.2164957 |
| reviewPalate | 0.4482227 | -0.0888569 | 0.1034866 | -0.5146121 | -0.7152144 | -0.0642977 |
| reviewTaste | 0.47077 | -0.1180963 | 0.2658745 | 0.0489844 | 0.2413978 | 0.7956553 |
| rating | 0.4393862 | -0.3282229 | 0.2832122 | -0.1729156 | 0.5313294 | -0.5538859 |

Table 9: PCA component high raters. High raters are defined as profiles making more than 10 reviews.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Non Frequent Raters (ratings =< 10) | |  |  |  |  |  |
| **Variable** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** |
| ABV | 0.1828697 | 0.9344017 | 0.2683055 | 0.0722666 | -0.1178234 | -0.0485752 |
| reviewAppearance | 0.4076751 | 0.0884648 | -0.7939294 | 0.3496989 | -0.2696483 | 0.0255244 |
| reviewAroma | 0.4355474 | 0.0487709 | -0.2037554 | -0.8303762 | 0.1773495 | -0.2131339 |
| reviewPalate | 0.4510375 | -0.0694815 | 0.1355569 | 0.4021451 | 0.7671165 | -0.1522278 |
| reviewTaste | 0.4684608 | -0.1386586 | 0.2686713 | -0.0672799 | -0.1383223 | 0.8157661 |
| rating | 0.4333217 | -0.30436 | 0.4069546 | 0.1292914 | -0.5237879 | -0.5127526 |

Table 10: PCA component low raters. Low raters are defined as profiles making less or equal than 10 reviews.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loadings |  |  |  |  |  |
| **Variable** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| reviewAppearance | -0.391527 | -0.8662987 | 0.2214856 | 0.2151579 | 0.0297378 |
| reviewAroma | -0.4420859 | -0.0662847 | -0.8371547 | -0.1731853 | -0.2633349 |
| reviewPalate | -0.4578346 | 0.1222926 | 0.4251907 | -0.7629379 | -0.1121185 |
| reviewTaste | -0.4812392 | 0.2748121 | -0.0598643 | 0.1803813 | 0.810411 |
| rating | -0.4583262 | 0.3932629 | 0.2564084 | 0.5559697 | -0.5103272 |

Table 11: PCA components for all observations (Alcohol content excluded)

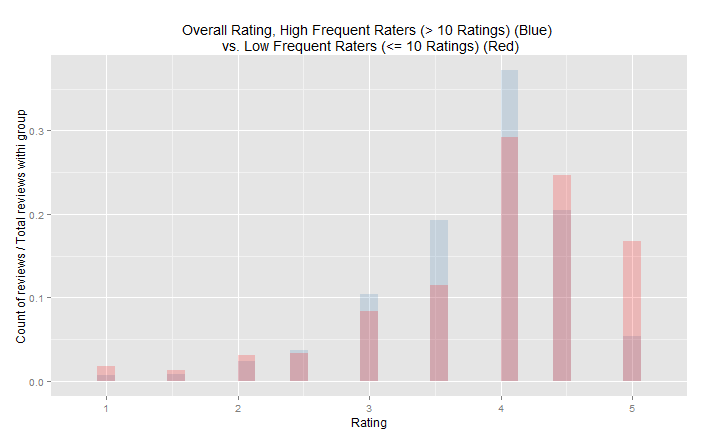


Figure 12: Overall Ratings, high frequent raters vs. low frequent raters

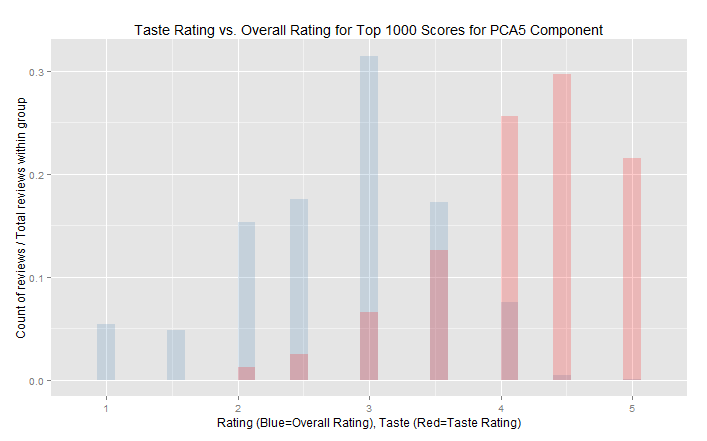


Figure 13: Taste Ratings vs. Overall Ratings for Top 1000 observations sorted by dereasing scores for PCA5

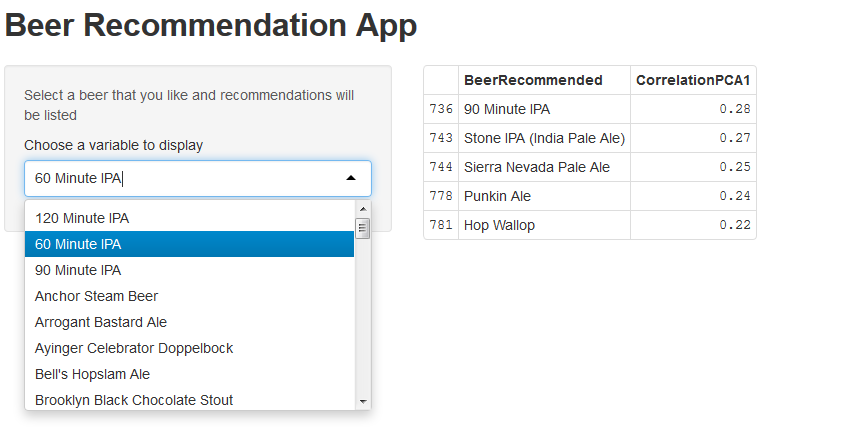


Figure 14: Our self made Beer Recommendation App of the 50 top rated beers in the dataset made in Shiny R

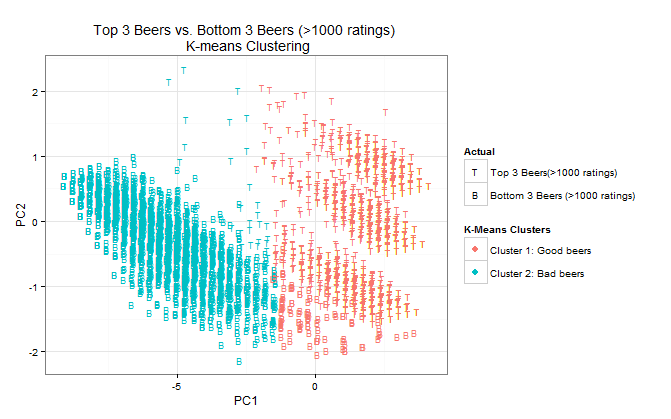


Figure 15: K-Means clustering using 2 seeds on top 3 good and bad beers. The definition of good beer is the porportion of 5s over number of ratings (for that beer). The defintion of a bad beer is the porportion of 1 over number of ratings(for that beer)

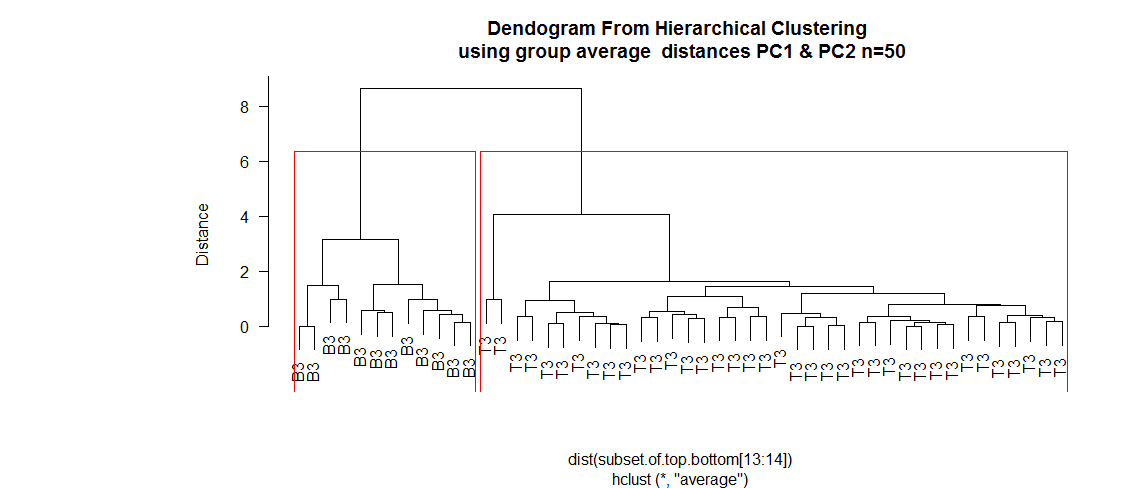


Figure 16: Dendogram for hierarchical clustering(using group average distance linking) using only a small radom sample of 50 reviews of the Top 3 beers and Bottom 3 beers. The red lines is the cluster cutoff, set to 2 in this case. B3 are bottom 3 beers sample units and T3 are top 3 beers sample units.

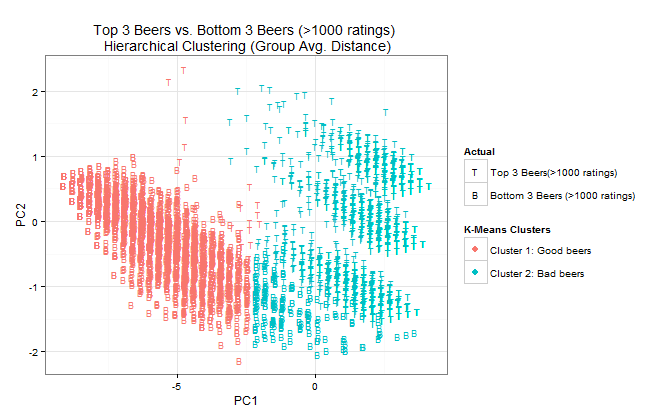


Figure 17: Hierarchical Clustinering using group avg. distance calculation to fuse cluster performed on top 3 and bottom 3 beeers. The definition of good beer is the porportion of 5s over number of ratings (for that beer). The defintion of a bad beer is the porportion of 1 over number of ratings(for that beer)

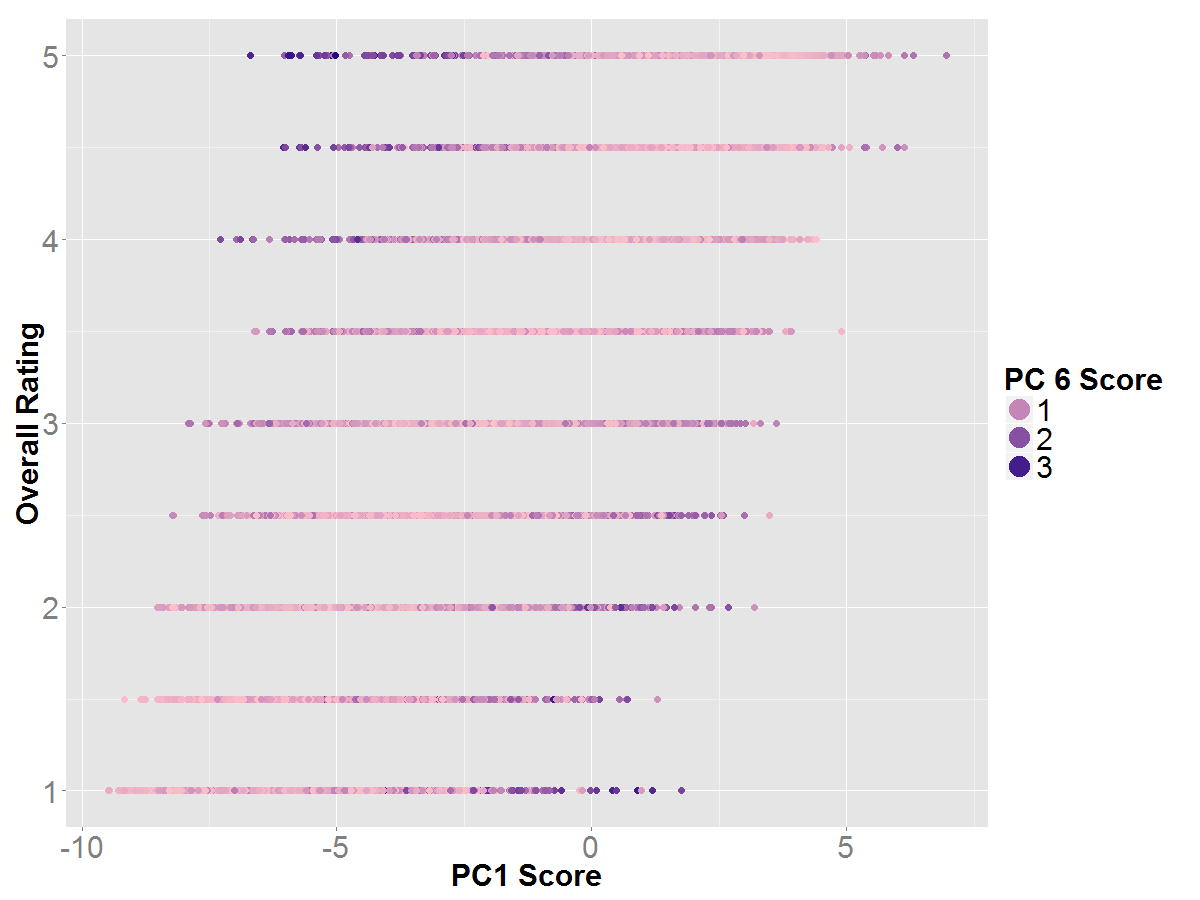


Figure 18: This shows the Overall rating against the PC1 score and is coloured by the ABSOLUTE value of PC6 score. The high absolute values (ie the dark points) are the false/weird reviews – or the darker it is the more likely to be a weird review. I plotted this with the other PCs etc but could see nothing.

*«I cannot remember who sent me this, who ever you were.... screw you.   
  
This looks like beer in the most generic sense. Yellow and yellow. There is nothing but yellow. There is no head, no lacing, no appeal. I am not sure why I did not stop right here.  
  
The aroma is atrocious. All adjunct lager and hot sauce in the nose.   
  
This beer seriously tastes like how Taco Bell feels coming out. I apologize for churching it up a bit, it is really much worse than that. Unobstructed chili flavor and heat. There is absolutely nothing desirable about this beer.  
  
All my mouth feels is hot. I can say that it is light bodied, that is about as detailed I can get.  
  
I will readily admit that this review was written based off a very small sample. I feel comfortable writing this review because I am positive it was not going to get any better, in fact I am sure it would have only gotten worse. This is not a beer, this is torture. Screw water boarding, make me drink this and I will tell you everything. Since it is the night before Thanksgiving I feel like I should end on a positive note. I am thankful I never have to drink this again.»[[13]](#footnote-14)*Here is another funny beer reviewer that had a bad experience with this one:

*«I first had this beer probably 13 years ago...I still wake in a cold sweat from the occasional nightmares that accompany the PTSD from that fateful night.  
  
I remember it all so clearly. A weak, flat, pale beer...an eye-watering fermented pepper smell...  
  
And then the taste. Oh, that terrible, terrible taste. I've never had a beer that appeared to have been intentionally made this bad. And that includes that terrible concoction of Bud Light & Clamato juice.  
  
And it doesn't go away, the taste. It burns, and not in a good way. Try as many other beers as yomoodu want, nothing will wash your palate of this awful sin.  
  
Over a decade later, and I regularly suffer from Class 3 heart-burn. I'm relatively certain I can trace it back to this beer.»*

Example 1: Example of funny reviews of the worst overall rated beer in the data set ‘Crazy Ed's Cave Creek Chili Beer’

1. **Code**

1. (Web Link, 15.04.2014, «Loading Data Into R» <http://www.r-statistics.com/2013/09/a-speed-test-comparison-of-plyr-data-table-and-dplyr/> [↑](#footnote-ref-2)
2. (Web-Link, 15.04.14) Explination of the Quadrupel Beer Style : <http://www.beeradvocate.com/beer/style/142/> [↑](#footnote-ref-3)
3. (Web-Link, 15.04.14) Description of the beer “Pliny the Younger“: <http://www.beeradvocate.com/beer/profile/863/21690/> [↑](#footnote-ref-4)
4. (Web-Link, 15.04.14) Description of the beer «Trappist Westvleteren 12»

   <http://www.beeradvocate.com/beer/profile/313/1545/> [↑](#footnote-ref-5)
5. R-help file,”prcomp” (run “??prcomp” to display the details of this function in R) [↑](#footnote-ref-6)
6. (Lecture Notes), Janine Illian (2013-2014) Cluster Analysis [↑](#footnote-ref-7)
7. (Paper reference) J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297 [↑](#footnote-ref-8)
8. (Lecture Notes), Janine Illian (2013-2014) Cluster Analysis, Hierarchical Methods [↑](#footnote-ref-9)
9. (Book reference) S. C. Johnson (1967): "Hierarchical Clustering Schemes" Psychometrika, 2:241-254 [↑](#footnote-ref-10)
10. (Book) “Programming Collective Intelligence: Building Smart Web 2.0 Applications” By Toby Segaran [↑](#footnote-ref-11)
11. (Web-Book, 2011) Collaborative Filtering Recommender Systems

    By Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan <http://files.grouplens.org/papers/FnT%20CF%20Recsys%20Survey.pdf> [↑](#footnote-ref-12)
12. (Book) “Programming Collective Intelligence: Building Smart Web 2.0 Applications” By Toby Segaran [↑](#footnote-ref-13)
13. (Web-Link 15.04.2014, Review of “Crazy Ed's Cave Creek Chili Beer “ http://www.beeradvocate.com/beer/profile/677/2213/ [↑](#footnote-ref-14)