Cleaned Beer Data Principal Component Analysis:

We did Principal Component Analysis (PCA) over the alcohol content, individual ratings and overall ratings of different beers in our dataset. We used R <reference if required> for our analysis. It has 2 functions for doing PCA, ‘princomp’ and ‘prcomp’. ‘prcomp’ is considered to be more robust as it uses ‘single value decomposition’ technique (using correlation matrix) to find eigen values and so, we’ll just use it.

Also, the variances for alcohol content and ratings were significantly different and so, we standardized them so as to have a variance of 1 for all the variables. Using these scaled variables we extracted the loadings (eigen vectors) for alcohol content and ratings that are shown below





We can see here that PCA gives 6 principal components for out data, which is the same as the number of variables we used for PCA. In general, PCA compresses the information of all the variables in 1st few principal components while the remaining components explains the left over small bit of variance in the data. This way we would capture maximum information of the data in fewer variables (principal components) and can analyze it on the reduced scale.

Important thing here would be to interpret these principal components, as not always these components measure the real variables. Mostly these components just reflect some pattern of covariance (correlation) in the data that can have different reasons.

We can see that 1st Principal Component explains maximum, 61.2%, of variation in our data and 2nd Principal Component explains 15.3% of variation. Collectively, they both explain 76.5% variation in the data and so, contain maximum information of the data in them. Remaining Principal Components 3rd, 4th, 5th and 6th explains 9.2%, 6.5%, 4.7% and 3.1% variance in data respectively.

The1st Principal Component says that all the variables have positive projections on it. This would mean that the scores of all the variables for 1st Principal Component would be positive if their values are above their specific means, whereas, it would be negative if the values are below mean[[1]](#footnote-1). This suggests that all the variables have positive correlation with each other.

Though many times not all the variables have a significant influence on the position of the each Principal Component. This means that not all variables contribute to the major trend in the data captured by the particular Principal Component. The two approaches used to determine the variables having significant influence on the Principal Component are Mardia’s Criterion[[2]](#footnote-2) and Equilibrium Contribution[[3]](#footnote-3). While Mardia’s Criterion is argued to work quite well, we would be using lesser conservative of both the approaches to select significant variables for our Principal Components, as we do not have very large number of variables in our data.

Based on this, it can be seen that both Mardia’s and Equilibrium criterion suggests that the individual and overall ratings contribute to 1st Principal Component while alcohol content does not contribute much to it. This would mean that higher score on 1st Principal component would reflect a higher individual and overall rating in the data. It would tell us if we are buying the best beer with great taste, palate, appearance and aroma or a bad beer with not so good taste, palate, appearance and aroma.

2nd Principal component tell something a little more interesting. It shows a contrast between the Alcohol Content and overall rating for the beer. Individual rating hardly projects on to it at all. Though Mardia’s and Equilibrium criterion only suggests alcohol content to have significant influence on 2nd Principal Component, but it is interesting to consider overall rating important as well[[4]](#footnote-4) as its loading is not significantly small. The component suggests that if the alcohol content were high, then the beer would have a low overall rating and vice-versa. It separates less alcoholic good beers from more alcoholic bad beers.

We can also use PCA **biplot** to provide meaningful visual representation to our interpretations. It would plot the data (Principal Component scores) on the first two components along with the projections of original features (loadings of each variable) as well (Figure 1).

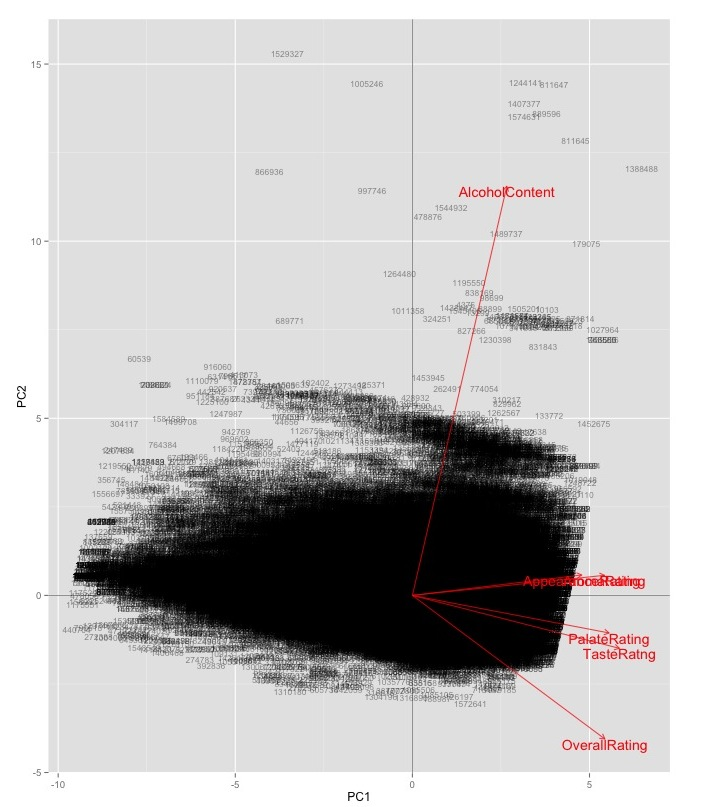


Figure 1: Biplot of Beer Data with horizontal axis representing 1st Principal Component and Vertical Axis representing 2nd. The cluttered black points are scores of data on the 2 principal components and red arrows show projections of original variables.

It is clearly visible that individual and overall ratings are close together in the space while alcohol content and overall rating are completely apart suggesting they are highly negatively correlated. While individual and overall ratings define the 1st Principal Component, alcohol content majorly define 2nd Principal Component. Also, the longer arrow length of alcohol content suggests that it has a large variance in data. We can also see that taste and palate rating go in the same direction suggesting that they are highly correlated. Similar is the case with appearance and aroma rating.

While most of the variance in data is explained by the 1st 2 components, we can ignore the remaining components and use these 1st 2 components for analyzing the data in reduced space. Though the Kaiser’s Criterion[[5]](#footnote-5) suggests that only 1st Principal Component should be used for further analysis of data. One of the best uses of this Principal Component would be to build a recommender system using it (discussed later in the report).

Though we might not be using the remaining components, but they do give some more insights to the data. While 3rd and 4th Principal Components does not provide some interesting information, it is worthwhile to look at 5th and 6th Component. 5th Principal Component suggests a polarity between Palate Rating and Overall Rating. This would mean that a high score on 5th Principal Component would suggest low palate rating and high overall rating. This could be either by chance or mistake or people actually rate some beers low on palate when they are actually good. Similarly, 6th Principal Component suggests contrast between Taste Rating and Overall Rating even though we saw before that they are highly correlated. Further in the report it is argued that these are false reviews.

1. It is to be noted that the Principal Components are calculated over the scaled variables and so the scaled values would be negative if the values of the variables were below the true variable mean and positive otherwise. [↑](#footnote-ref-1)
2. 0.7 times the largest coefficient (absolute value) in the eigen vector [↑](#footnote-ref-2)
3. 1 divided by square root of number of variables used for PCA [↑](#footnote-ref-3)
4. Mardia’s and equilibrium criterion are just few rules that can be extended, if required. [↑](#footnote-ref-4)
5. Kaiser’s Criterion tend to retain very few components when number of features are less than 20 and so should be used with caution. [↑](#footnote-ref-5)