**MT5758 - Group Project (Time related results)**

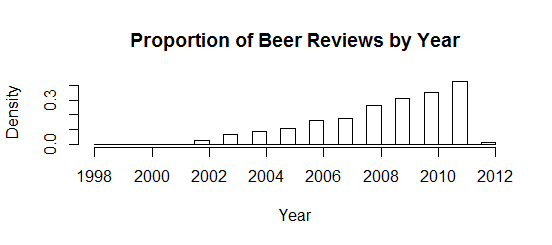
We used the 'as.POSIXlt' function in R to convert the time stamp associated with each review from UNIX based to the conventional format. We then created 5 additional variables; 'hour', 'day', 'month', 'season' and 'year' in order to explore if there was any time related variability exhibited by the data.

Exploratory Data Analysis ("EDA") indicated that there was little difference between the number of ratings per month. The lowest proportion of ratings occurred in June (7.6%) and the highest in December (9.6%). Therefore, whilst month alone does not give you much information about the number of ratings, there is a clear pattern of frequent ratings during (US) cold months, as shown by the below table.

Fig.x

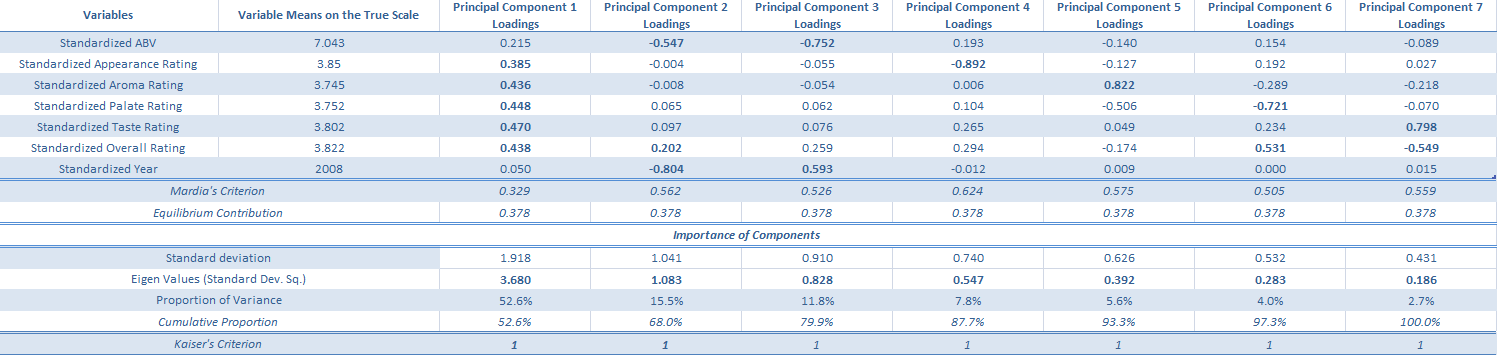
|  |  |  |
| --- | --- | --- |
| **Season** | **Months** | **Proportion of ratings** |
| Winter | Nov-Jan | 27.5% |
| Autumn | Aug-Oct | 25.2% |
| Summer | May-July | 23.7% |
| Spring | Feb-Apr | 25.0% |

Our EDA also confirmed that as year increases so too does the number of ratings (Fig x1). The observations were collected between the founding of BeerAdvocate.com in January 1998 up until January 2012. It is therefore not surprising that the number of ratings increases with year as it likely reflects the growth of both the popularity of the website and the growth of internet usage itself.

****

We had intended to investigate the 'hour' variable, however, as there is no time code specified for each observation we could not be confident about our interpretations, given the somewhat international nature of our data.

Based on the results of our aforementioned EDA we undertook two further PCA investigations. The variables included were the same as those of our previous PCAs but with the addition of the scaled 'year' and 'month' respectively. The results can be seen in Figures x2 and x3 below;



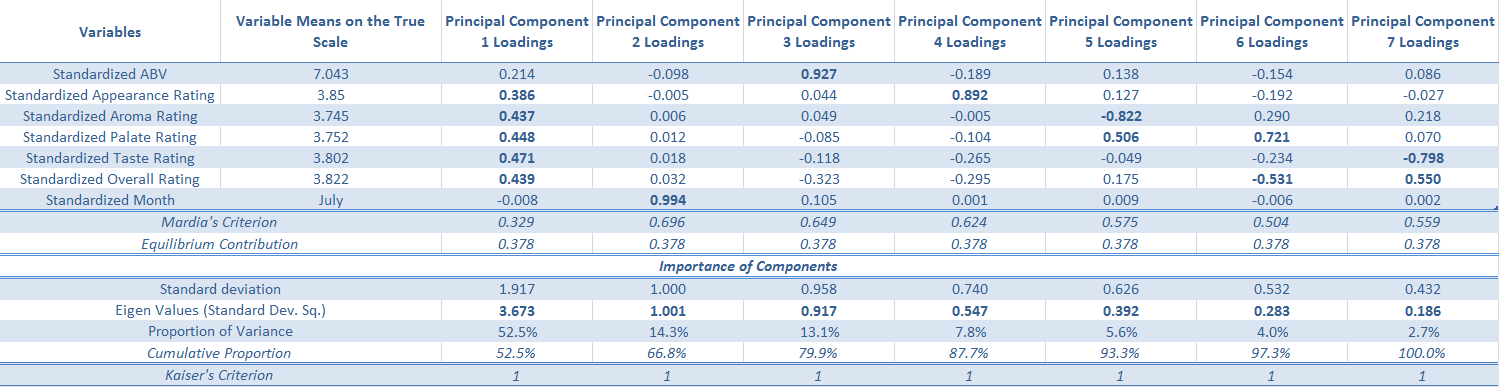
Just as was the case for the PCA carried out earlier, the first principal component explains the majority of variation in the data (52.6%) and relates to determinants of good beer rating. For example, a high score on PC1 equates to an observation having above average ratings across all of the review criteria. PC2 relates to a segment of the data that can be described ABV and more importantly year. The large negative loadings indicate that prior to the mean year (2008) people were rating (and thus consuming) beers with higher than average ABV. Therefore a very high score in PC2 would relate to the rating of a strong beer, such as an 'American Double', near the start of our data collection period (1998). PC3 effectively describes the inverse of PC2. An observations' score on this component is pulled in opposite directions by the ABV and year, meaning that a high score is indicative of a rating from nearer the end of the data collection period (2012) of a beer with a relatively low ABV, such as a 'Low Alcohol Beer'. It could be the case that PC1 and PC2 capture a change in consumer preference regarding ABV during the period. Perhaps, people are more aware of the health implications of consuming large amounts of alcohol or it could simply be a change in consumer tastes.

Both PC6 and PC7 fall well short of being regarded as important under the Kaiser Criterion. Nonetheless, they appear to confirm the widely cited[[1]](#footnote-1) tendency for online review data to have a component describing the variability resulting from *false reviews*. False reviews relate to observations whereby the overall rating and the indicators that are highly positively correlated with overall rating (in this case taste or palate) are at odds i.e. the relationship is negative rather than positive. There are two obvious hypotheses that one reaches. Firstly, that some reviewers are making honest mistakes when inputting their review scores, simple human error. However, the underlying reason could be more sinister and people are deliberately giving perfectly good beers negative reviews on purpose. There are clear incentives for the latter. For example, many of the beers rated on beeradvocate.com are made by microbreweries for whom a handful of positive (or negative) reviews can mean the life (or death) of a fledgling enterprise. As such, some unscrupulous stakeholders (business owners, friends, family etc) may take it upon themselves to write negative reviews about their competitors in order to advance their own goals. These two hypotheses seem likely to explain the variability described by PC7, however there is a third hypothesis for PC6, that is to say, some reviewers may have a different idea of what palate actually relates to. In this context, the Oxford dictionary definition of palate refers to 'The flavour of wine or beer'. However, it also refers palate as, 'A person’s ability to distinguish between and appreciate different flavours'. From the two definitions one could come to the conclusion that their palate and taste reviews should be very similar or beers with flavours that are not particularly complex (or sophisticated) should get a low review rating. The prior line of reasoning seems to be the majority view as exemplified by the strong positive correlation between palate and taste across all observations (0.73). However, the latter also appears to explain some of the variance in the data. Fig x4 contains information relating to the 10 brands of beer that appear most commonly in the subsection of the dataset with below average palate ratings and above average overall ratings;

|  |  |  |
| --- | --- | --- |
| **Beer Names** | **Proportion of Ratings in All the Ratings with Low Palate Rating and High Overall Rating** | **Proportion of Ratings in Complete Data for Analysis** |
| 120 Minute IPA | 0.49% | 0.11% |
| Bud Light | 0.49% | 0.08% |
| Budweiser | 0.54% | 0.09% |
| Coors Light | 0.46% | 0.08% |
| Guinness Draught | 0.55% | 0.13% |
| Miller High Life | 0.41% | 0.07% |
| Miller Lite | 0.39% | 0.07% |
| Pabst Blue Ribbon (PBR) | 0.49% | 0.09% |
| Samuel Adams Utopias | 0.39% | 0.02% |
| World Wide Stout | 0.40% | 0.10% |

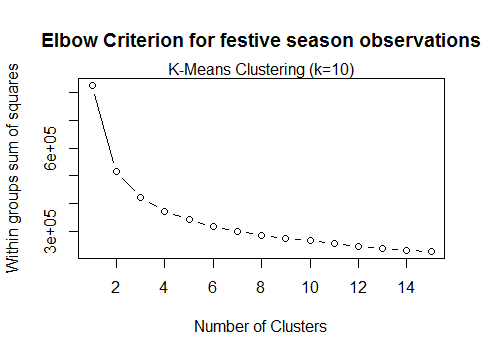
Of the beers named above, 7 out of 10 are household names and hold a large proportion of the market share in the low to mid-market beer segment and it is this position that may be causing a negative bias in the minds of reviewers.

Having completed the PCA with year as an additional variable we moved onto month, which also helps to capture the seasonal element of variation. The results are shown in Fig x3;



The proportion of variance explained (and the underlying trend captured) by each component is very similar between Fig 3 and Fig 4. For example, the variance explained by PC1 differs by only 0.1% and the variables considered to be of relevance are the same. The only new insights that we can garner come from PC2. It now explains slightly more variability (+1.2%) and is completely dominated by the month loading (0.994) with no other variable meeting either the Mardia Criterion or Equilibrium Contribution cut-offs. A high score on PC2 is representative of ratings that take place during the latter half of the year. This led us to the hypothesis that ratings are more likely to occur during winter, which can be confirmed by the Fig x. Further, this coincides with the Christmas period which lead us to investigate the patterns in ratings during this time of year to ascertain whether any of the presupposed relationships between variables differ. If successful, the information could have real commercial value, as beer companies gain insights into what beers people rate (and thus consume) over this period, thus helping them to better target their customer base with directed marketing and product launches.

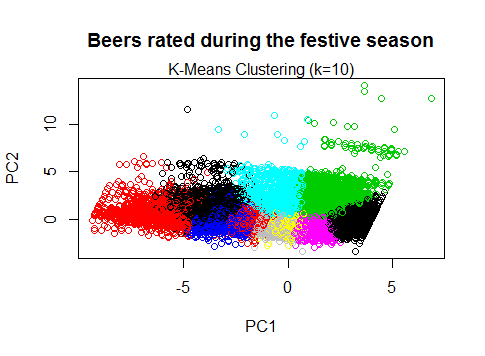
Given the fact that we are trying to gain insights into reviewer behaviour during a particular time of year, the most appropriate method at our disposal was to subset the data to just the reviews made during the Christmas period (November, December and January) and see if we could find any interesting clusters. However, we first performed some basic EDA and found that the beers that are frequently rated during the Christmas period tend to have stronger than average ABV [a barplot of ABV vs top 10 most rated beers would help this point but the top 10 xmas period beers are below along with those of cluster 3], which is surprising given our previous PCA results. We then moved onto cluster analysis and chose to use K-Means clustering as it works well with large datasets, the algorithm is simple to implement (just specify the number of clusters you want) and it is computationally inexpensive. We used the elbow criterion (Fig x4) to help us in determining an appropriate number of clusters to use.



[Explain Elbow criterion...... Also tried to use the Gap Statistic (a numerical equivalent used to interpret the above plot) but got an error message 'Error: cannot allocate vector of size 70.3gb' therefore we just interpreted the plot visually i.e. when the gradient of the curve levels off (say after 3) you take that as the number of clusters to use]

However, we decided to go for 10 clusters (arbitrarily) as we were looking to identify specific segments of consumers to target.

The below is the resulting clustering plot:



Based on the above we decided to target the green cluster (cluster 3) as it encapsulates beers with high overall ratings but typically above average ABV. A brewer equipped with the insights of our cluster analysis could add greater seasonality to their beer launches i.e. selling high ABV beer styles found to be popular in this cluster over the Christmas period. Whereas , their competitors who may only have insights surrounding the year as a whole would be less effective in targeting their customers.

Delving deeper into the characteristics of the green cluster, we found the top 10 most frequently rated beers as outlined in the below table.

(sorry for the Raw output, just trying to get you all the results faster! will amend later)

top10.cluster3

Russian Imperial Stout American Double / Imperial IPA American Double / Imperial Stout

0.13133913 0.12541832 0.11594501

Belgian Strong Dark Ale American Barleywine American Strong Ale

0.09210730 0.07609535 0.06605571

Quadrupel (Quad) Tripel Belgian Strong Pale Ale

0.04649127 0.03712094 0.03444370

English Barleywine

0.03341399

> xmas.top10

American IPA American Double / Imperial IPA Russian Imperial Stout

0.06959256 0.05023622 0.03946248

American Double / Imperial Stout American Porter American Pale Ale (APA)

0.03622307 0.03454878 0.03409016

Winter Warmer Belgian Strong Dark Ale American Amber / Red Ale

0.03268521 0.03213196 0.02681061

American Strong Ale

0.02433556

The above results, along with our EDA, PCA and clustering have enabled us to both generate and confirm our hypotheses that people rate (and thus consume) more beer during the Christmas period (PC2 of the PCA with month included) and that despite the general preference for lower than average ABV beers, over the Christmas period people drink stronger beers. Not only that, we can now provide brewers with the specific beer styles that people are consuming the most as well as being highly regarded in terms of overall rating. It is this cross section of beers that we believe brewers should either launch or produce more of, during the period.

1. INSERT REFERENCE TO CHRIS' ACADEMIC PAPER ON GITHUB [↑](#footnote-ref-1)