Intro and assumptions

Introduction

Brewdog provided us with the data from 196 types of beers and eight different features(variables). There are some missing data records, so we are supposed to impute the missing data using a suitable method. Then we will discuss products clustering that will be used for marketing purposes.

Preprocessing and assumptions

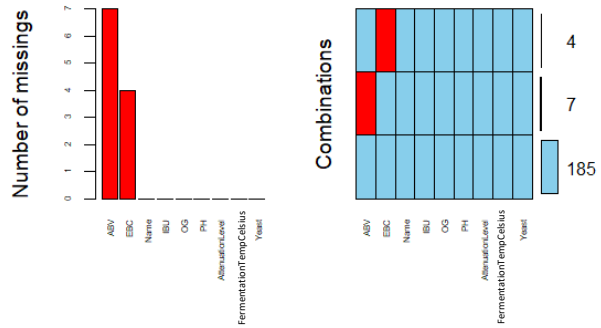
We will define dummy variables (Yi) to transform categorical data to numerical. We define four different "Yeast" categories. If a beer belongs to this category, it takes 1; Otherwise 0. This way, we will check the relation between variables. These dummy variables are not essential for the imputation process as our chosen method will automatically make these dummy variables.

First, we check if the missing data are in the form of MCAR, MAR or NMAR. In the case of MCAR, there is no logical correlation among the features, and we can use simple imputation methods. However, it may cause biases if there is a high number of missing data. (Jamshidian and Bentler, 1999)

On the other hand, if we can prove a dependency between missing data and other features, we need multiple imputation methods in which regression will consider the correlation of variables. (Buck, 1960) multiple imputation process is effective for both MAR and MCAR (under regularity conditions). (Jamshidian and Mata, 2007) If we can prove a correlation between missing data and variables, the NMAR scenario is a possible class. We assume there is no intention in the missing data, as the data belongs to well-known products with food and health-related certificates.

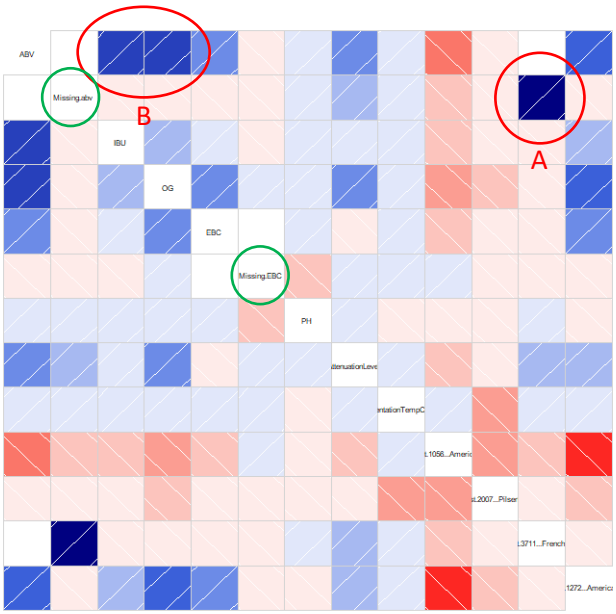
There are 11 missing data in total, seven missed records belong to ABV, and four belong to EBC. To check the relation between missing data and other features, we filtered data for missing fields and looked for possible patterns. Also, we aggregate the data to plot aggregated missing data as chart 1.

Chart 1 – Aggregated missing data



To find a possible relationship among variables, also between missing data and other variables, we have calculated Pearson correlation(Table 1) which demonstrates a strong relation between missing data in ABV and Wyeast 3711.(Chart2 – A) Seven missing data in the ABV column are related to only seven Wyeast 3711. Table 1 shows no other meaningful correlation between ABV's missing data and the other features.

Chart 2 – Corrgram- variables and missing data



Also, significant correlations (threshold set to 0.6) between ABV and IBU, OG exist. (Chart2- B). It will help us accurately impute missing data. On the contrary, this correlation can cause collinearity and mislead the model during the clustering process.

Table 1 – Correlation between ABV and missing data in ABV and other variables

ABV cells have missing data	ABV vs IBU	ABV VS OG	ABV vs EBC	ABV vs PH	ABV vs Att	ABV vs Fer	Y1	Y2	Y3	Y4
Electric India	38	1045	15	4.4	88.9	22	0	0	1	0
TM10	20	1048	14	4.2	89.6	22	0	0	1	0
Magic Stone Dog (w/Magic Rock & Stone Brewing Co.)	30	1043	15	4.4	81.4	23	0	0	1	0
Baby Saison - B-Sides	9	1032	2	4.4	88	21	0	0	1	0
Everday Anarchy	45	1081	15	4.4	93.8	23	0	0	1	0
Rhubarb Saison - B-Sides	25	1052	10	5.2	87	20	0	0	1	0
Black Jacques	45	1089	15	4.4	94.4	23	0	0	1	0
Pearson correlation ABV's missing data	-0.09	-0.07	-0.13	0.05	0.23	0.08	-0.21	-0.06	1.00	-0.14
Pearson correlation ABV (after IMPU)	0.65	0.63	0.32	0.04	0.38	0.09	-0.43	-0.14	-0.04	0.54

We have tested the idea of the relationship between missing data in EBC and other features similarly. But we can not find any meaningful correlation neither for EBC nor missing data. So, it is logical to assume that missing data in ABV are in MAR class, and we can classify missing data in EBC as MCAR class.

As (Jamshidian and Mata, 2007) suggest, we will use multiple imputation for ABV to guarantee unbiased and valid associations. The missing ratio in the EBC column is negligible, and we could not define a meaningful correlation. We use a simple imputation for this variable, but testing the possible differences will also calculate the multiple imputations.

Another essential assumption is that all features, including ABV(taste and pleasure), IBU(taste), OG(taste), EBC(color), pH(taste), attenuation(taste and smell), fermentation temperature(taste and smell) and yeast type(taste) affect the customer's satisfaction. The purpose of clustering is marketing, and we need further proof of the effectiveness of these features on customer satisfaction.

We will assume we cannot eliminate correlated variables (as a decision criteria) for clustering purposes. It will increase the weights of correlated variables in distance calculation. Then, we will eliminate ABV as a strongly correlated variable to IBU and OG to examine the changes. Though, there are other options like factor analysis to handle this issue. Also, we assume the scale of different features are sensible, and we do not need normalization.

We will discuss the unsupervised clustering method as a part of data are categorical, and we can not use k-means. Using the agglomerative hierarchical clustering will let us analyze the possible clusters. To decide the optimal number of clusters, we will examine the inconsistency level of links(difference between the height of a link and sub-level links) and NbClust function results.

Calculation and processing

Imputation

Referring to our assumptions about EBC, we have used the MEAN of EBC column for current values as 70.61 to impute missing data. Knowing that the means are equal, the variances are different as 8098.166 and 7932.05 for before and after imputation, respectively.

We can not assume normality for data to test variances equality. But we used the Kolmogorov-Smirnov to test distribution equality. (Abayomi et al., 2008) According to the test, we can not reject the hypothesis of equality of distributions. We also ran Multivariate Imputation by Chained Equations (MICE) processes for EBC, and the result was the same as 70.61615.

Picture 1 – Kolmogorov-Smirnov test result for EBC

```
        Two-sample Kolmogorov-Smirnov test

data:  Beer$EBC and BeerImputSimp$EBC
D = 0.014987, p-value = 1
alternative hypothesis: two-sided
```

We have conducted a MICE imputation process to impute missing data in ABV. We ran the model for m=20 and maxit=100 and imputed the missing data with the results. We assumed a normal

distribution (according to the histogram) and performed statistical tests on equality of means(t-test), variances (f-test) and distributions (Kolmogorov-Smirnov test).

Picture 2 – Two population summaries -ABV

```
summary(ABVCSV2COM$ABV)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.50    5.20    7.15    7.62    8.85   41.00
summary(ABVCSV2$ABV)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.500   5.200   7.200   7.644   9.000  41.000       7
```

The result demonstrates that we can not reject the hypothesis of equality in the means at 0.9524, variances at 0.889 and distribution at the approximate 1 p-values level.

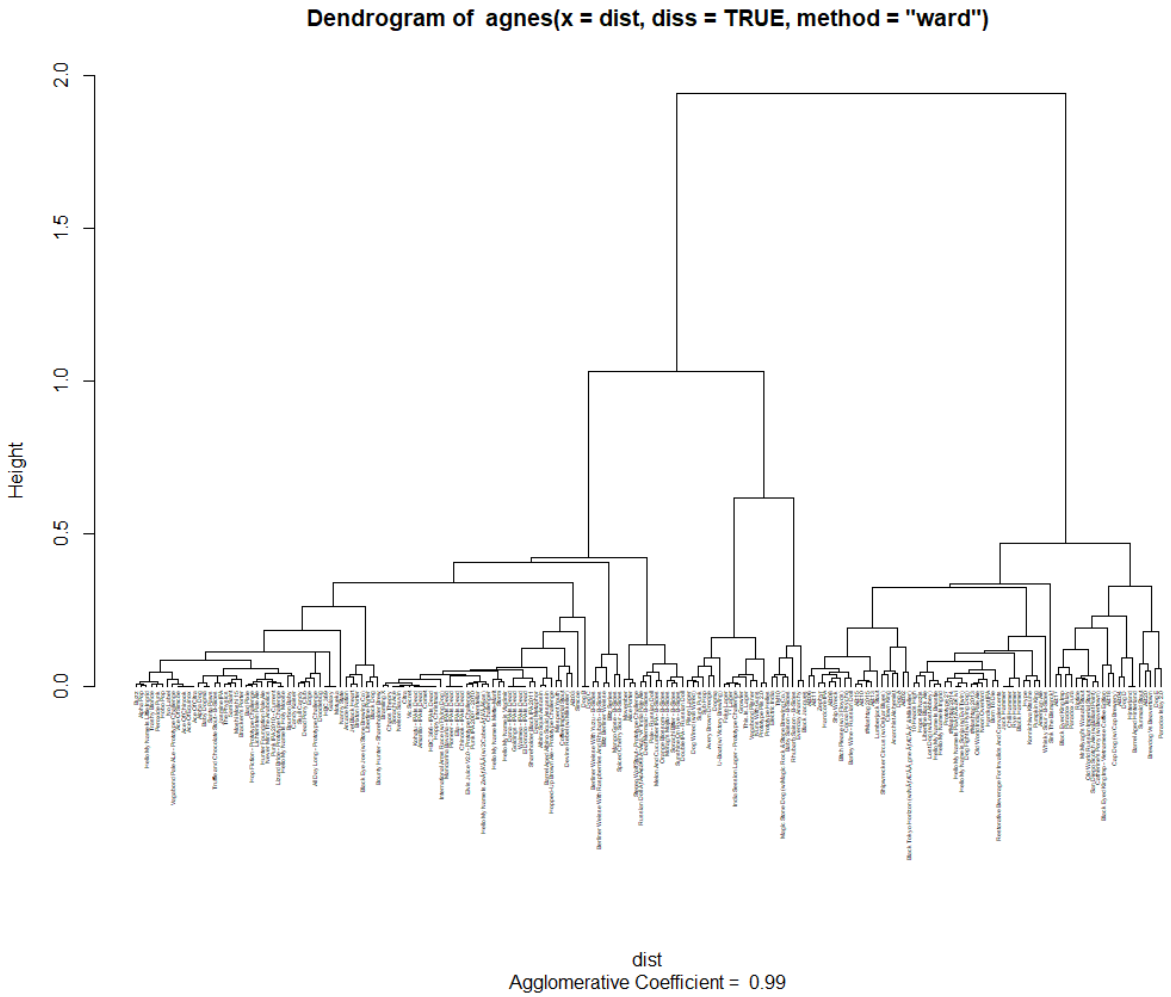
Picture 3 – t-test, f-test and Kolmogorov-Smirnov test results for ABV

```
        F test to compare two variances

data:  ABVCSV2$ABV and ABVCSV2COM$ABV
F = 1.0203, num df = 188, denom df = 195, p-value = 0.889
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7681941 1.3560880
sample estimates:
ratio of variances
          1.020275

        Welch Two Sample t-test

data:  ABVCSV2$ABV and ABVCSV2COM$ABV
t = 0.059777, df = 382.17, p-value = 0.9524
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7655465  0.8135548
sample estimates:
mean of x mean of y
 7.643545  7.619541

        Two-sample Kolmogorov-Smirnov test

data:  ABVCSV2COM$ABV and ABVCSV2$ABV
D = 0.012661, p-value = 1
alternative hypothesis: two-sided
```

Clustering

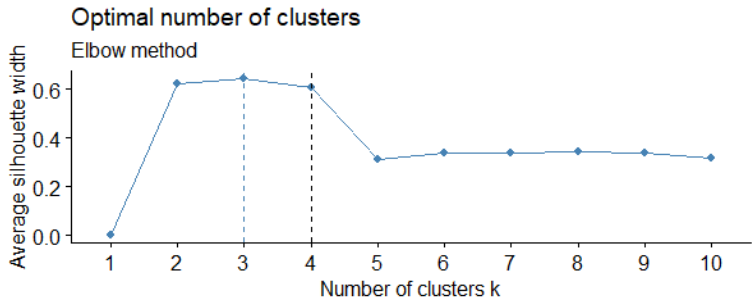
There are some categorical data in the dataset, so we calculate the dissimilarity matrix using the daisy function in R with the parameter metric set to "Gowers." Then, the "Agnes" function is used to make agglomerative hierarchical clustering for the data.

Chart 3 – Dendrogram



To define the suitable number of clusters, we have used the NbClust function in R with appropriate indices for categorical data. The result shows that "Dunn" and "Mcclain" indices recommend two, "silhouette" four (Chart 4), and "Cindex" 50 clusters, the max allowed clusters we set in the function.

Chart 4– optimal number of clusters silhouette - elbow



analyzing and conclusion

The missing data in EBC and ABV columns are in the MCAR and MAR classes. However, there is a chance for NMAR class of data for ABV; there is no evidence confirming this idea. We conducted correlation analysis and demonstrated a correlation between the missing data in ABV and the Wyeast 3711. Also, there is a correlation between ABV and OG, IBU.

We conducted both simple and multiple imputations for EBC and multiple imputation for ABV. We tested the hypothesis of equality between means, variances and distributions for both imputations. The result and histograms confirmed the imputation was successful. (Chart 3 and 4)

Chart 3 – EBC histogram before (A) and after (B) imputation

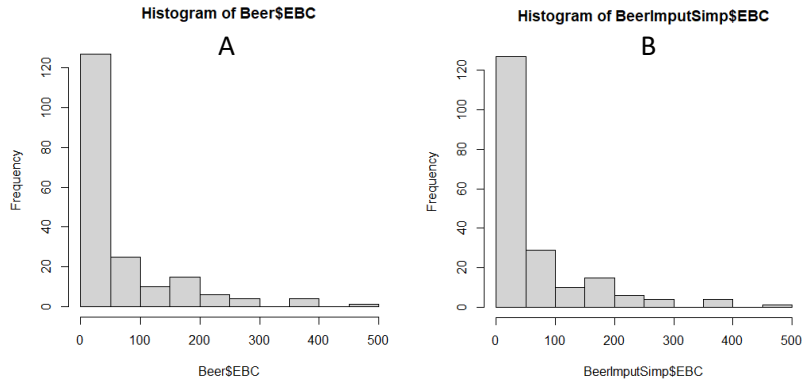
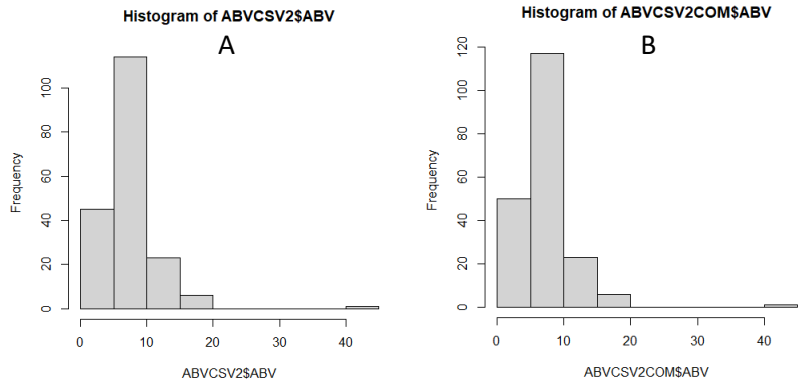
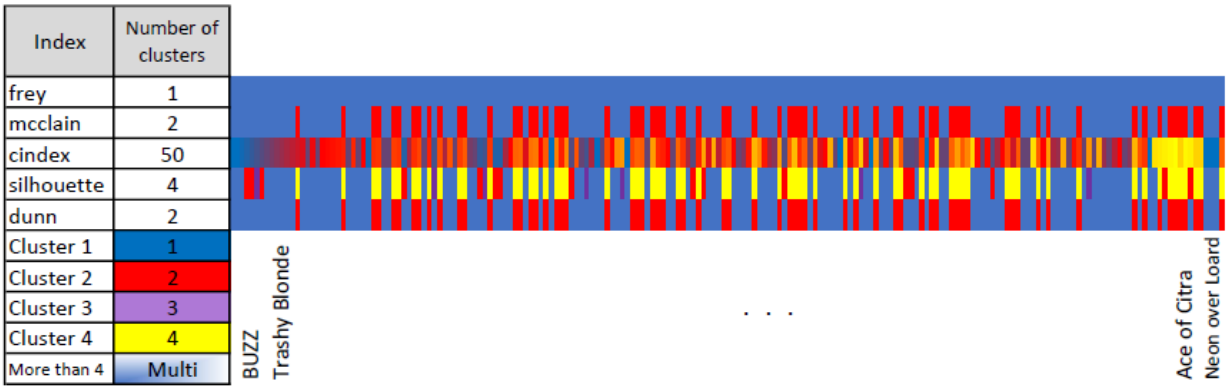


Chart 4 – ABV histogram before (A) and after (B) imputation



There are categorical data in the dataset, so we decided to use a hierarchical clustering algorithm. Before cutting the hierarchical tree, we used the NbClust to determine the best number of clusters. (Chart 5)

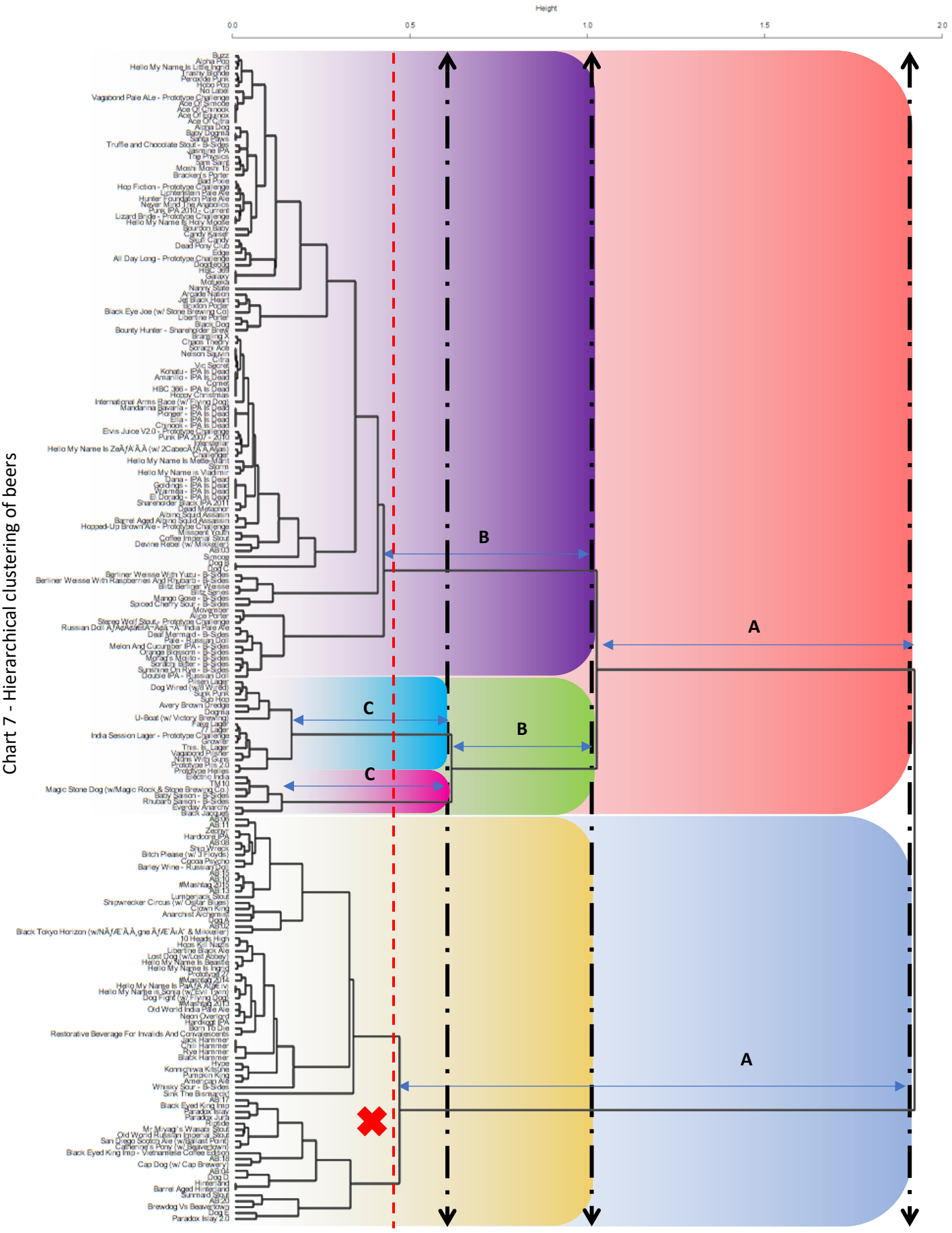
Chart 5 – NbClust result



To start, we defined two clusters according to the majority role. We cut the tree approximately at level 1.7, showing a significant inconsistency between the highest link level and two clusters' links. (Chart 7-A) We tested another cut at a 1.05 height level, demonstrating three clusters in the data. The height difference of neighbour links is still meaningful, so we can continue cutting. (Chart 7-B) Adding another cut at 0.6 level makes two new clusters with high distance to their lower links. (Chart 7-C)

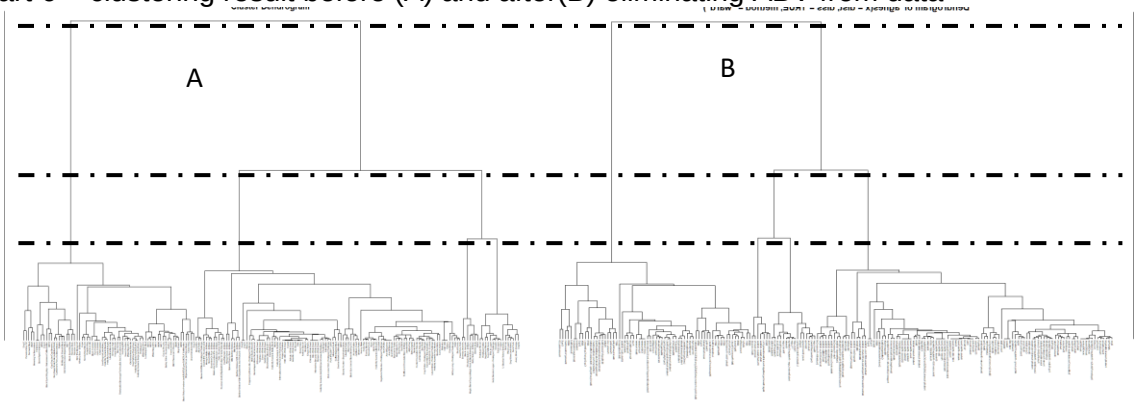
Cutting the tree to a lower height will not make an acceptable distance between upper links and their lower neighbours. We can suppose the optimal number of clusters can be four or less. Please note that the purpose of clustering is marketing. Analyzing other decision criteria, such as the total marketing budget, product/market segments, and available marketing resources guides us to a better-defined number of clusters.

Chart 7 - Hierarchical clustering of beers



To complete our analysis, we ran the model removing ABV to study the possible result of elimination. There was no change in the hierarchical clustering result, but according to NbClust, the best number of clusters differed. (1,2,3,4,50)

Chart 6 – clustering result before (A) and after(B) eliminating ABV from data



References

ABAYOMI, K., GELMAN, A. & LEVY, M. 2008. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 57**,** 273-291.

BUCK, S. F. 1960. A Method of Estimation of Missing Values in Multivariate Data Suitable for use with an Electronic Computer. *Journal of the Royal Statistical Society. Series B (Methodological),* 22**,** 302-306.

JAMSHIDIAN, M. & BENTLER, P. M. 1999. ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and behavioral Statistics,* 24**,** 21-24.

JAMSHIDIAN, M. & MATA, M. 2007. Advances in analysis of mean and covariance structure when data are incomplete. *Handbook of latent variable and related models.* Elsevier.