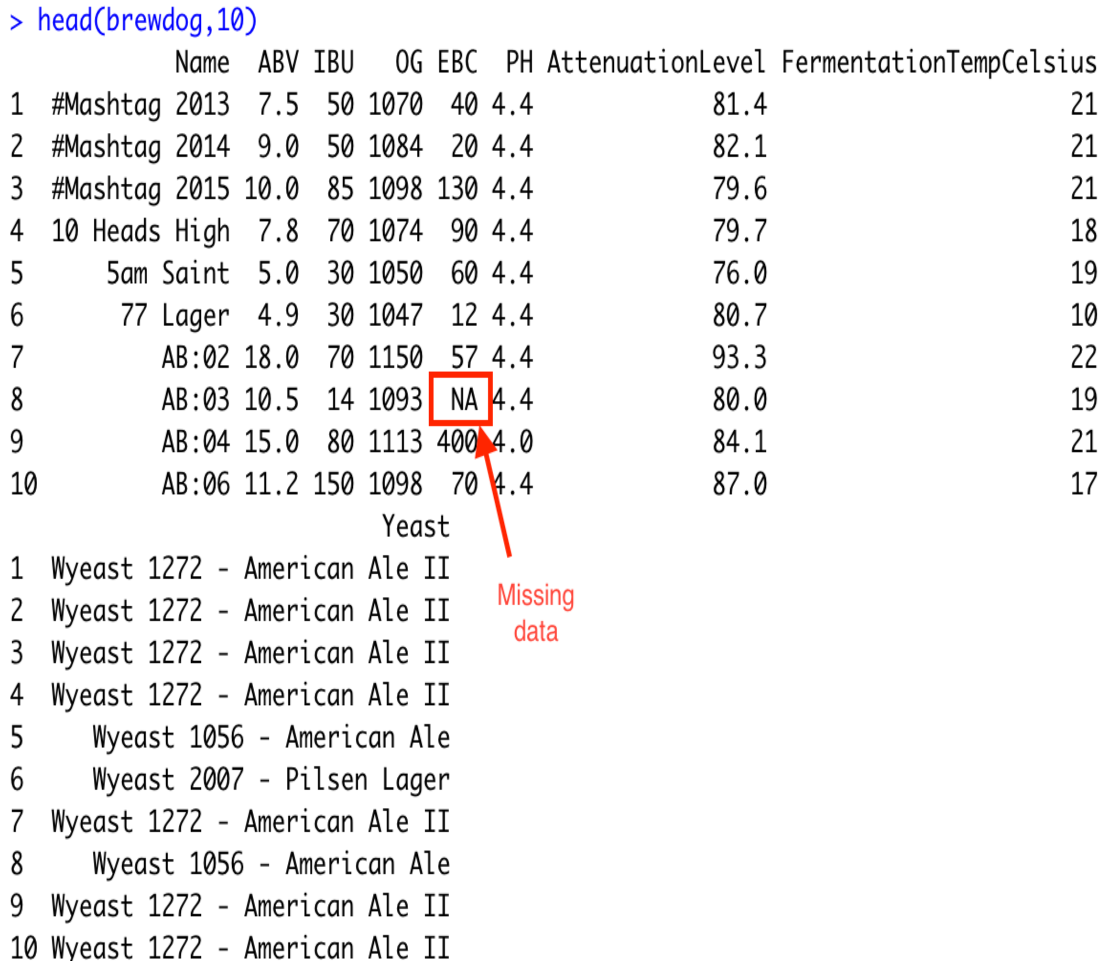
1.1. Introduction

Brewdog dataset consists of 199 rows of different beers with 9 columns including Name, Alcohol by Volume (ABV), International Bitterness Units (IBU), Original Gravity (OG), Colour Units from the European Brewery Convention (EBC), pH (Acid & Base Scale), Attenuation Level, Fermentation temperature in Celsius and Yeast type.

The first 10 rows of the dataset have been displayed. The dataset consists of some missing values in certain columns, e.g., 9th row → value for EBC is missing.

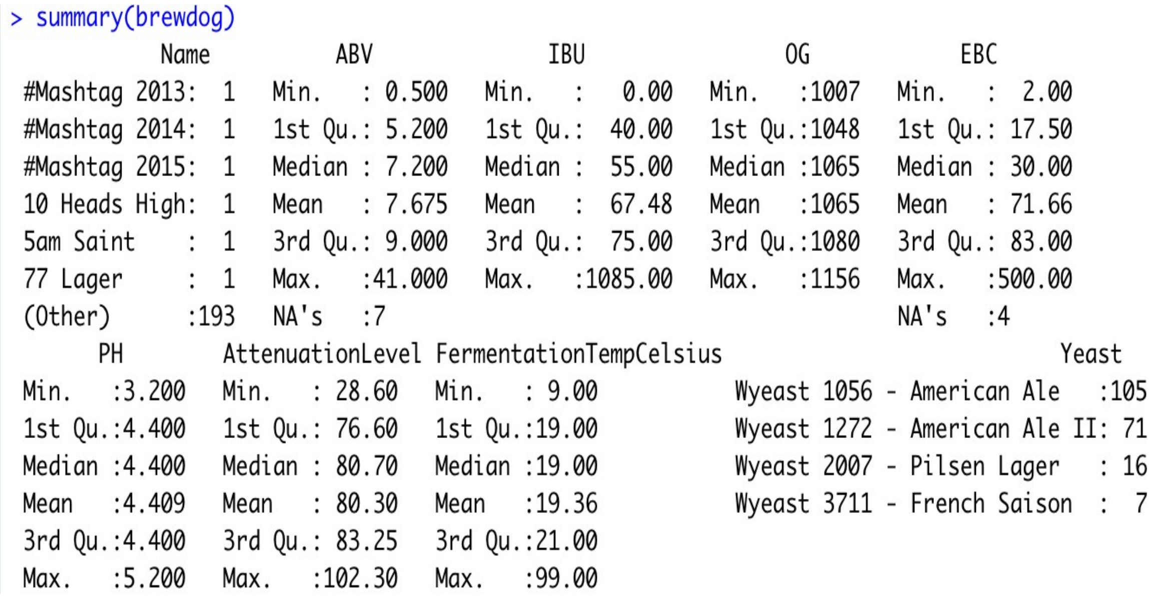


```
> head(brewdog,10)
            Name  ABV IBU   OG EBC  PH AttenuationLevel FermentationTempCelsius
1  #Mashtag 2013  7.5  50 1070  40 4.4             81.4                      21
2  #Mashtag 2014  9.0  50 1084  20 4.4             82.1                      21
3  #Mashtag 2015 10.0  85 1098 130 4.4             79.6                      21
4  10 Heads High  7.8  70 1074  90 4.4             79.7                      18
5      5am Saint  5.0  30 1050  60 4.4             76.0                      19
6       77 Lager  4.9  30 1047  12 4.4             80.7                      10
7          AB:02 18.0  70 1150  57 4.4             93.3                      22
8          AB:03 10.5  14 1093  NA 4.4             80.0                      19
9          AB:04 15.0  80 1113 400 4.0             84.1                      21
10         AB:06 11.2 150 1098  70 4.4             87.0                      17
                           Yeast
1  Wyeast 1272 - American Ale II
2  Wyeast 1272 - American Ale II
3  Wyeast 1272 - American Ale II
4  Wyeast 1272 - American Ale II
5     Wyeast 1056 - American Ale
6     Wyeast 2007 - Pilsen Lager
7  Wyeast 1272 - American Ale II
8     Wyeast 1056 - American Ale
9  Wyeast 1272 - American Ale II
10 Wyeast 1272 - American Ale II
```

Figure: Overview of Brewdog dataset

1.2. Identifying missing data

The summary of the dataset provides information about the number of missing values in each column. ABV and EBC consists of 7 and 4 missing variables respectively.



```
> summary(brewdog)
              Name          ABV              IBU                OG            EBC        
 #Mashtag 2013:  1   Min.   : 0.500   Min.   :   0.00   Min.   :1007   Min.   :  2.00  
 #Mashtag 2014:  1   1st Qu.: 5.200   1st Qu.:  40.00   1st Qu.:1048   1st Qu.: 17.50  
 #Mashtag 2015:  1   Median : 7.200   Median :  55.00   Median :1065   Median : 30.00  
 10 Heads High:  1   Mean   : 7.675   Mean   :  67.48   Mean   :1065   Mean   : 71.66  
 5am Saint    :  1   3rd Qu.: 9.000   3rd Qu.:  75.00   3rd Qu.:1080   3rd Qu.: 83.00  
 77 Lager     :  1   Max.   :41.000   Max.   :1085.00   Max.   :1156   Max.   :500.00  
 (Other)      :193   NA's   :7                                         NA's   :4       
       PH        AttenuationLevel FermentationTempCelsius                          Yeast    
 Min.   :3.200   Min.   : 28.60   Min.   : 9.00           Wyeast 1056 - American Ale   :105  
 1st Qu.:4.400   1st Qu.: 76.60   1st Qu.:19.00           Wyeast 1272 - American Ale II: 71  
 Median :4.400   Median : 80.70   Median :19.00           Wyeast 2007 - Pilsen Lager   : 16  
 Mean   :4.409   Mean   : 80.30   Mean   :19.36           Wyeast 3711 - French Saison  :  7  
 3rd Qu.:4.400   3rd Qu.: 83.25   3rd Qu.:21.00                                             
 Max.   :5.200   Max.   :102.30   Max.   :99.00                                             
```

Figure: Summary of Brewdog

The aggr() in VIM package provides a visual representation of Brewdog. The plots clearly depict the number of missing variables in each column. The graph on the left shows the number of missing variables in each column as red bars. The combination graph on the right shows where information is missing as red block and the scale right shows the number of missing records for that combination.

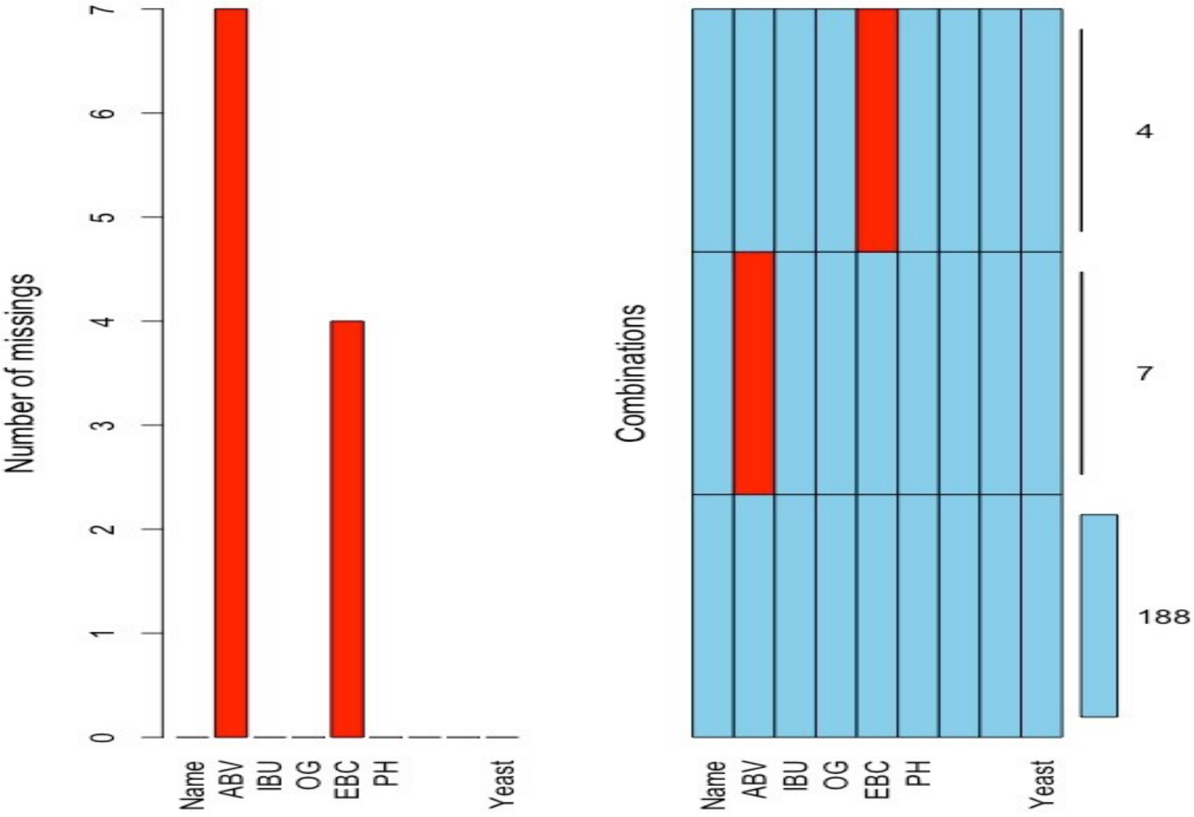


Figure: Combinations plot

1.3. Identifying relationship between missing variables and other variables

To obtain the relationship between the missing variables and other variables, a copy of Brewdog is placed in a variable named “missdata”. An additional column named missing is added to missdata, containing all the incomplete cases in BrewDog. Correlation analysis is performed between the complete cases and incomplete cases using corrgram() which accepts missdata as its parameter.

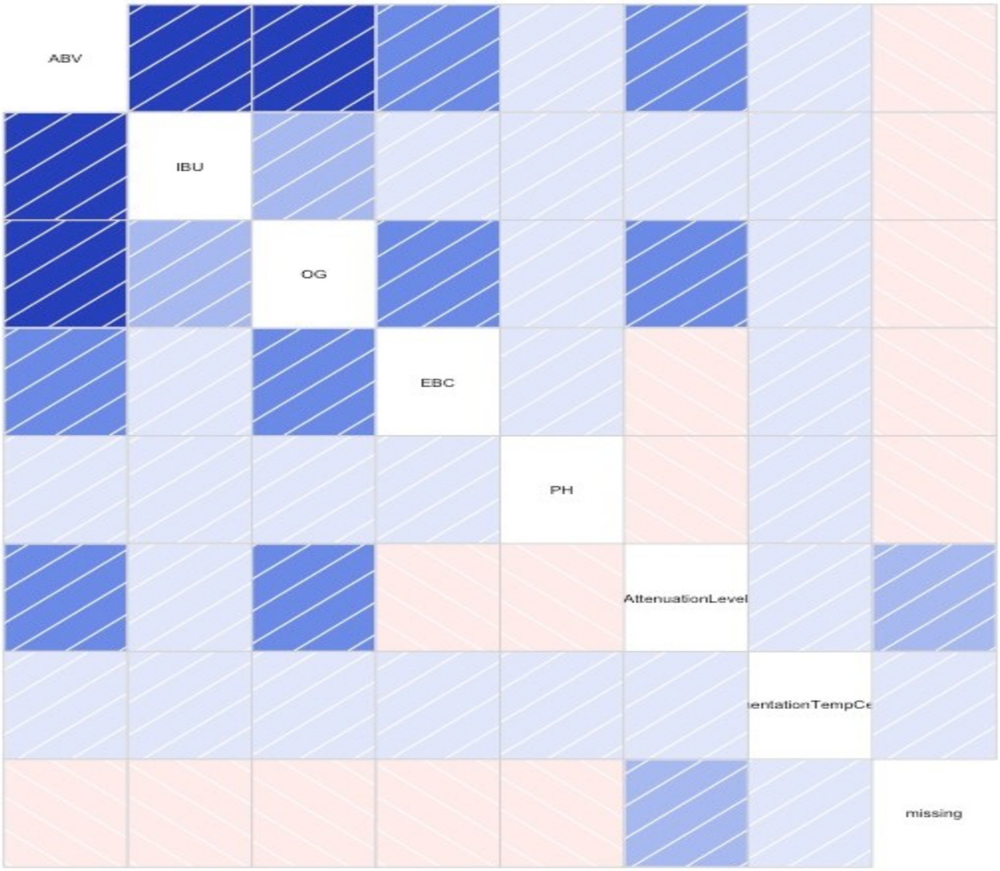


Figure: Correlation Analysis

The correlation analysis proves that there is positive relationship between the missing variables and the complete variables. Dark blue indicates strong positive relationship. ABV has strong correlation with IBU, OG, EBC and attenuation variables. EBC has strong blue relationship with OG and ABV variables.
The missing data have been observed to be missing at random (MAR) since

- there is strong relationship between ABV and EBC
- the missing variables can be predicted using the other variables due to their strong relationship with them.
- ABV is missing for the yeast type – Wyeast 3711 French Saison
- EBC is missing for few records of yeast type – Wyeast 1056 American Ale

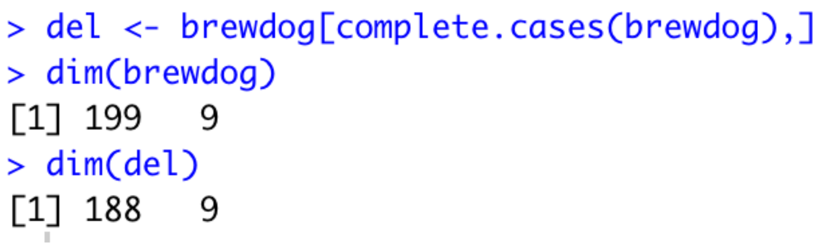
Hence, there is a clear pattern of relationship.

1.4. **Handling missing data**

Missing data can cause distortions in variable distribution in a dataset leading to biased analysis. Therefore, it is important to handle missing data effectively. Missing data can be handled, either by deletion or imputation.

2.4.1. **Deletion**

Deletion removes the missing data entirely from the dataset. In listwise deletion, all the rows in which ABV and EBC values are missing will be removed completely, resulting in loss of data including the complete values. Deletion is preferred when you have less than 5% missing data, yet since Brewdog consists of only 199 rows of data, losing data is not considerable.



```
> del <- brewdog[complete.cases(brewdog),]
> dim(brewdog)
[1] 199   9
> dim(del)
[1] 188   9
```

Figure: Listwise deletion

2.4.2. **Imputation**

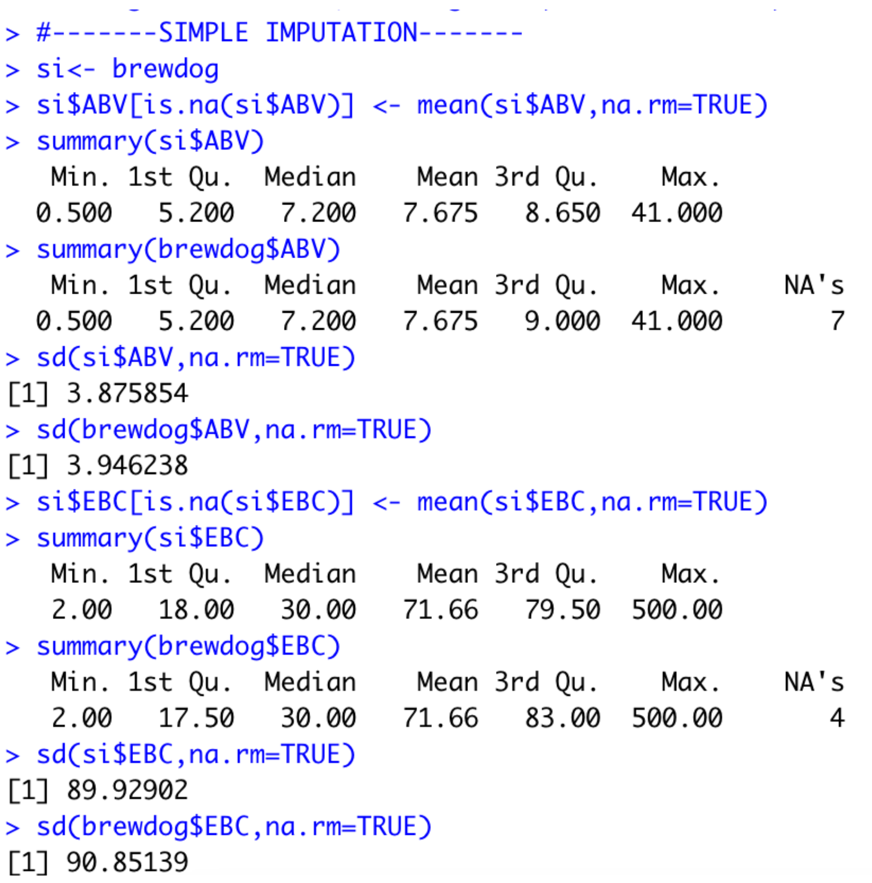
Imputation is a technique to replace the missing variable with substitute values to retain the information in the dataset. Imputation is preferred over deletionbecause deletion results in loss of information thereby reducing the size of the dataset.

There are two types of imputation:

- Simple imputation: Missing values are replaced with mean, median or mode.
- Multiple imputation: Missing data is replaced with multiple acceptedvalues obtained from predictions using methods like ANOVA, regression.

Simple Imputation

Simple imputation is performed by replacing the missing values with the mean value. There is no significant change in the distribution data after imputation.



```
> #-------SIMPLE IMPUTATION-------
> si<- brewdog
> si$ABV[is.na(si$ABV)] <- mean(si$ABV,na.rm=TRUE)
> summary(si$ABV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.500   5.200   7.200   7.675   8.650  41.000 
> summary(brewdog$ABV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.500   5.200   7.200   7.675   9.000  41.000       7 
> sd(si$ABV,na.rm=TRUE)
[1] 3.875854
> sd(brewdog$ABV,na.rm=TRUE)
[1] 3.946238
> si$EBC[is.na(si$EBC)] <- mean(si$EBC,na.rm=TRUE)
> summary(si$EBC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00   18.00   30.00   71.66   79.50  500.00 
> summary(brewdog$EBC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   2.00   17.50   30.00   71.66   83.00  500.00       4 
> sd(si$EBC,na.rm=TRUE)
[1] 89.92902
> sd(brewdog$EBC,na.rm=TRUE)
[1] 90.85139
```

Figure: Performing simple imputation

Multiple Imputation

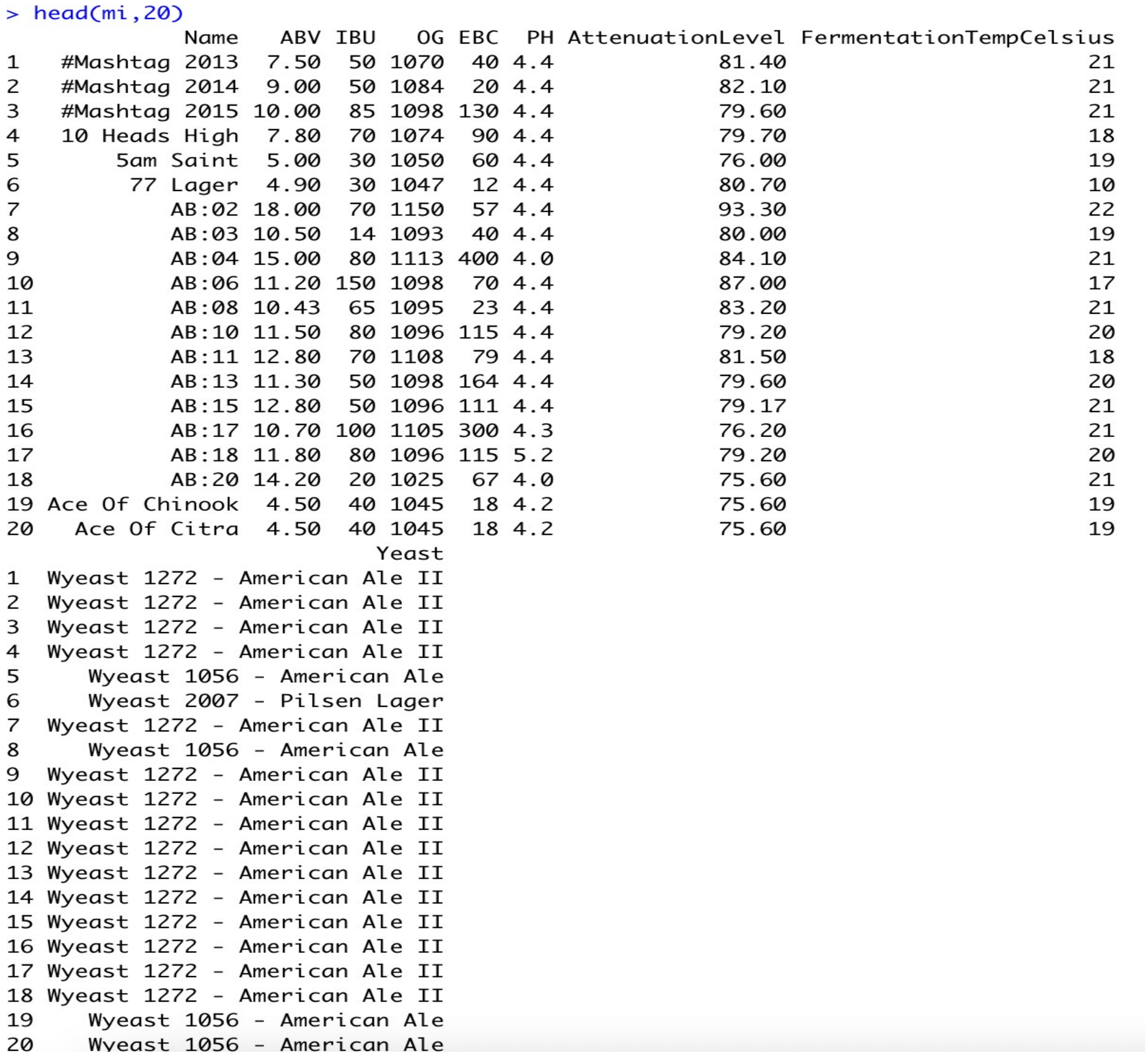
Multiple imputation involves replacing the missing values with substituted values using chained equation approach, resulting in complete cases.

Multiple imputation is preferred over simple imputation (Dziura et al., 2013, p.350) for Brewdog dataset, since,

- The validity of simple imputation does not consider whether the data is missing at random (MAR), but rather depends on assumptions about the missing values for example, are identical to the last observed value.
- Simple imputation results in underestimation of variability of unseen data by imputing a constant for all missing values regardless of other variable characteristics.

Multiple imputation forms 'M' complete datasets by imputing each missing value 'M' times where the multiple values are obtained from a distribution of possibilities. The 'M' complete datasets are combined into a valid statistical inference that properly reflect the uncertainty due to missing values. It produces unbiased estimates and handles the missing covariate information along with the missing outcomes (Dziura et al, 2013, p.351).

Multiple imputation is performed using mice package in R which accepts Brewdog, the number of imputations(m) and the maximum number of iterations(maxit) as its parameters. The complete dataset is stored in variable “mi”.



```
> head(mi,20)
             Name   ABV IBU   OG EBC  PH AttenuationLevel FermentationTempCelsius
1   #Mashtag 2013  7.50  50 1070  40 4.4            81.40                      21
2   #Mashtag 2014  9.00  50 1084  20 4.4            82.10                      21
3   #Mashtag 2015 10.00  85 1098 130 4.4            79.60                      21
4   10 Heads High  7.80  70 1074  90 4.4            79.70                      18
5       5am Saint  5.00  30 1050  60 4.4            76.00                      19
6        77 Lager  4.90  30 1047  12 4.4            80.70                      10
7           AB:02 18.00  70 1150  57 4.4            93.30                      22
8           AB:03 10.50  14 1093  40 4.4            80.00                      19
9           AB:04 15.00  80 1113 400 4.0            84.10                      21
10          AB:06 11.20 150 1098  70 4.4            87.00                      17
11          AB:08 10.43  65 1095  23 4.4            83.20                      21
12          AB:10 11.50  80 1096 115 4.4            79.20                      20
13          AB:11 12.80  70 1108  79 4.4            81.50                      18
14          AB:13 11.30  50 1098 164 4.4            79.60                      20
15          AB:15 12.80  50 1096 111 4.4            79.17                      21
16          AB:17 10.70 100 1105 300 4.3            76.20                      21
17          AB:18 11.80  80 1096 115 5.2            79.20                      20
18          AB:20 14.20  20 1025  67 4.0            75.60                      21
19 Ace Of Chinook  4.50  40 1045  18 4.2            75.60                      19
20   Ace Of Citra  4.50  40 1045  18 4.2            75.60                      19
                           Yeast
1  Wyeast 1272 - American Ale II
2  Wyeast 1272 - American Ale II
3  Wyeast 1272 - American Ale II
4  Wyeast 1272 - American Ale II
5     Wyeast 1056 - American Ale
6     Wyeast 2007 - Pilsen Lager
7  Wyeast 1272 - American Ale II
8     Wyeast 1056 - American Ale
9  Wyeast 1272 - American Ale II
10 Wyeast 1272 - American Ale II
11 Wyeast 1272 - American Ale II
12 Wyeast 1272 - American Ale II
13 Wyeast 1272 - American Ale II
14 Wyeast 1272 - American Ale II
15 Wyeast 1272 - American Ale II
16 Wyeast 1272 - American Ale II
17 Wyeast 1272 - American Ale II
18 Wyeast 1272 - American Ale II
19    Wyeast 1056 - American Ale
20    Wyeast 1056 - American Ale
```

Figure: Overview of Brewdog after multiple imputation

Visualising the dataset after multiple imputation shows that all the missing values have been replaced with the possible values without losing any information.

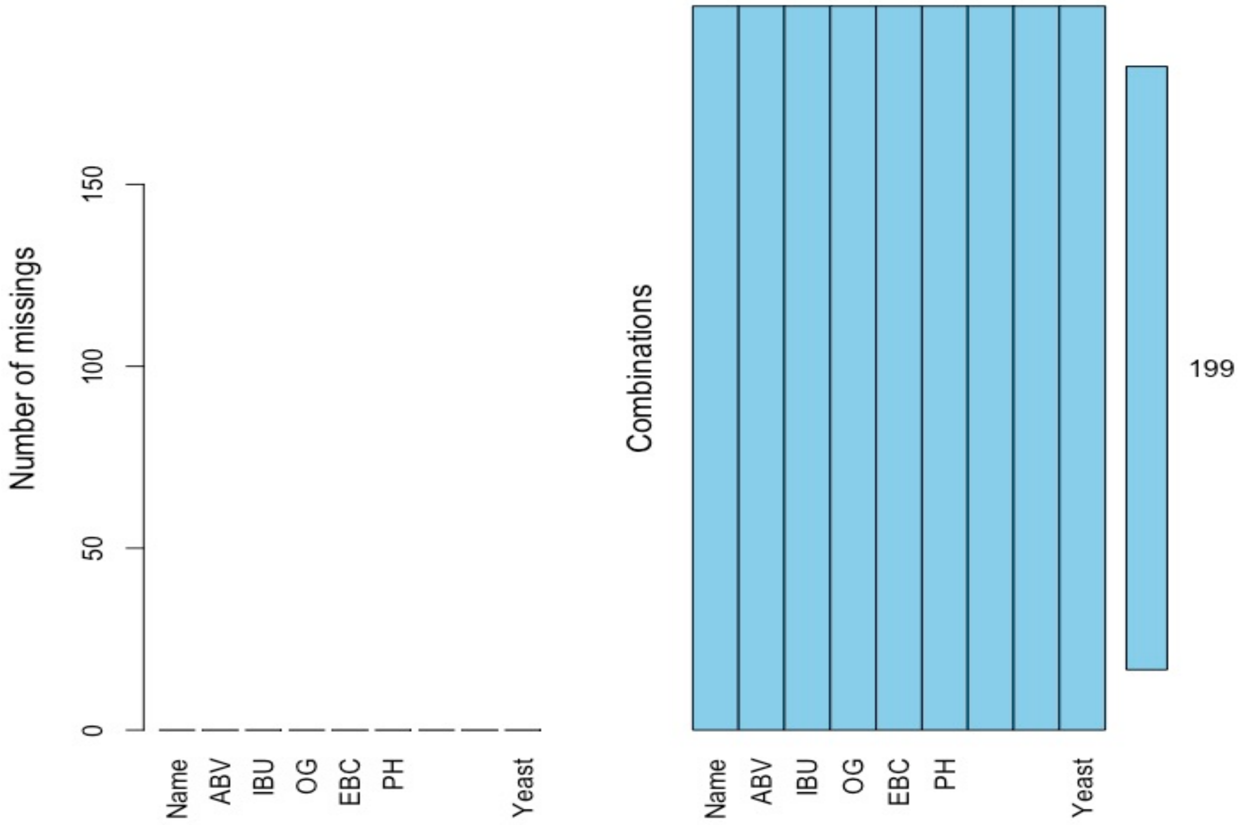
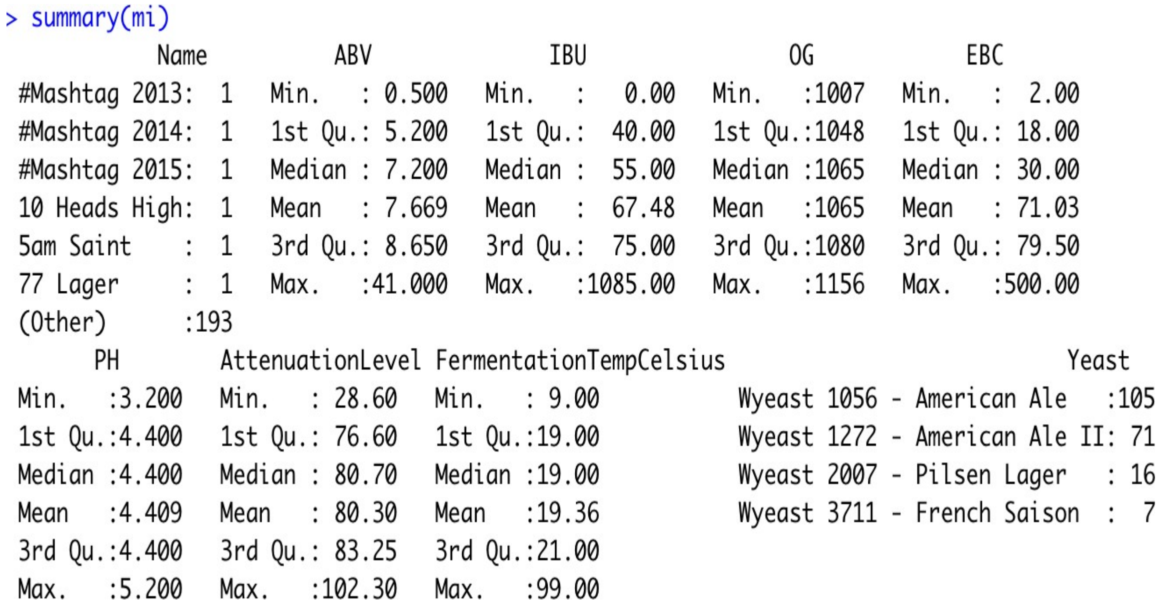


Figure: Combinations plot after multiple imputation

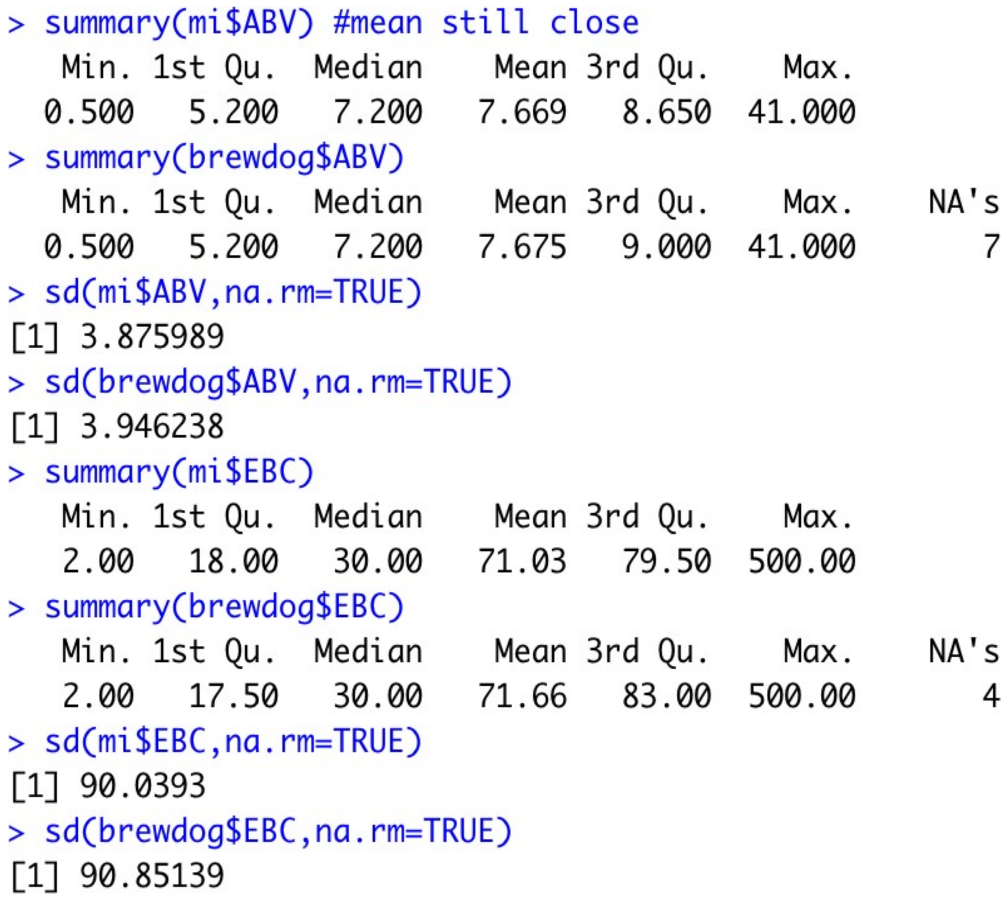
The below figure shows the summary of the mi dataset. There are no NA values.



```
> summary(mi)
            Name          ABV             IBU               OG            EBC
 #Mashtag 2013:  1   Min.   : 0.500   Min.   :   0.00   Min.   :1007   Min.   :  2.00
 #Mashtag 2014:  1   1st Qu.: 5.200   1st Qu.:  40.00   1st Qu.:1048   1st Qu.: 18.00
 #Mashtag 2015:  1   Median : 7.200   Median :  55.00   Median :1065   Median : 30.00
 10 Heads High:  1   Mean   : 7.669   Mean   :  67.48   Mean   :1065   Mean   : 71.03
 5am Saint    :  1   3rd Qu.: 8.650   3rd Qu.:  75.00   3rd Qu.:1080   3rd Qu.: 79.50
 77 Lager     :  1   Max.   :41.000   Max.   :1085.00   Max.   :1156   Max.   :500.00
 (Other)      :193
       PH        AttenuationLevel FermentationTempCelsius                          Yeast
 Min.   :3.200   Min.   : 28.60   Min.   : 9.00          Wyeast 1056 - American Ale   :105
 1st Qu.:4.400   1st Qu.: 76.60   1st Qu.:19.00          Wyeast 1272 - American Ale II: 71
 Median :4.400   Median : 80.70   Median :19.00          Wyeast 2007 - Pilsen Lager   : 16
 Mean   :4.409   Mean   : 80.30   Mean   :19.36          Wyeast 3711 - French Saison  :  7
 3rd Qu.:4.400   3rd Qu.: 83.25   3rd Qu.:21.00
 Max.   :5.200   Max.   :102.30   Max.   :99.00
```

Figure: Summary of mi

1.5. Checking ABV and EBC variables before and after multiple imputation



```
> summary(mi$ABV) #mean still close
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.500   5.200   7.200   7.669   8.650  41.000 
> summary(brewdog$ABV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.500   5.200   7.200   7.675   9.000  41.000       7 
> sd(mi$ABV,na.rm=TRUE)
[1] 3.875989
> sd(brewdog$ABV,na.rm=TRUE)
[1] 3.946238
> summary(mi$EBC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00   18.00   30.00   71.03   79.50  500.00 
> summary(brewdog$EBC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   2.00   17.50   30.00   71.66   83.00  500.00       4 
> sd(mi$EBC,na.rm=TRUE)
[1] 90.0393
> sd(brewdog$EBC,na.rm=TRUE)
[1] 90.85139
```

Figure: Checking variable distribution

The mean values for ABV and EBC after imputation is very close to that of the original dataset. Similarly, the standard deviation is also close enough which implies that the distribution of variables about the mean have not changed significantly after imputation.

The histogram obtained after multiple imputation is evenly distributed, with no significant change from the original data distribution for both the variables.

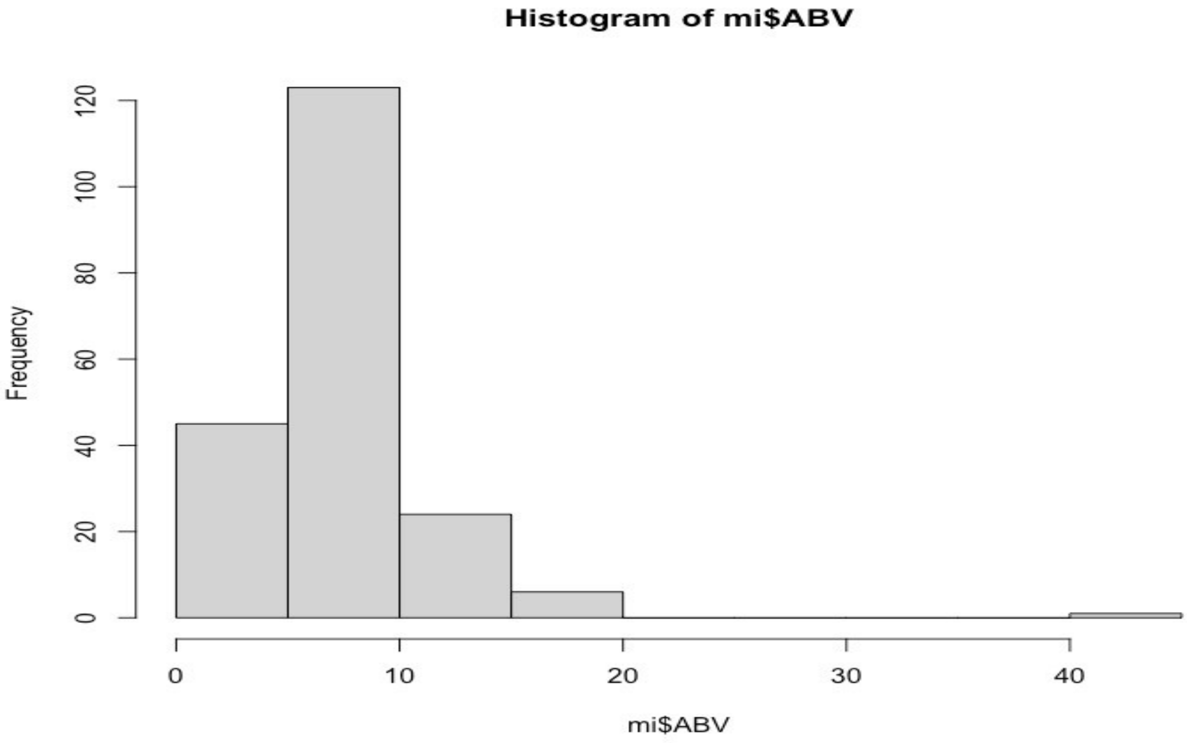


Figure: Histogram of ABV in mi

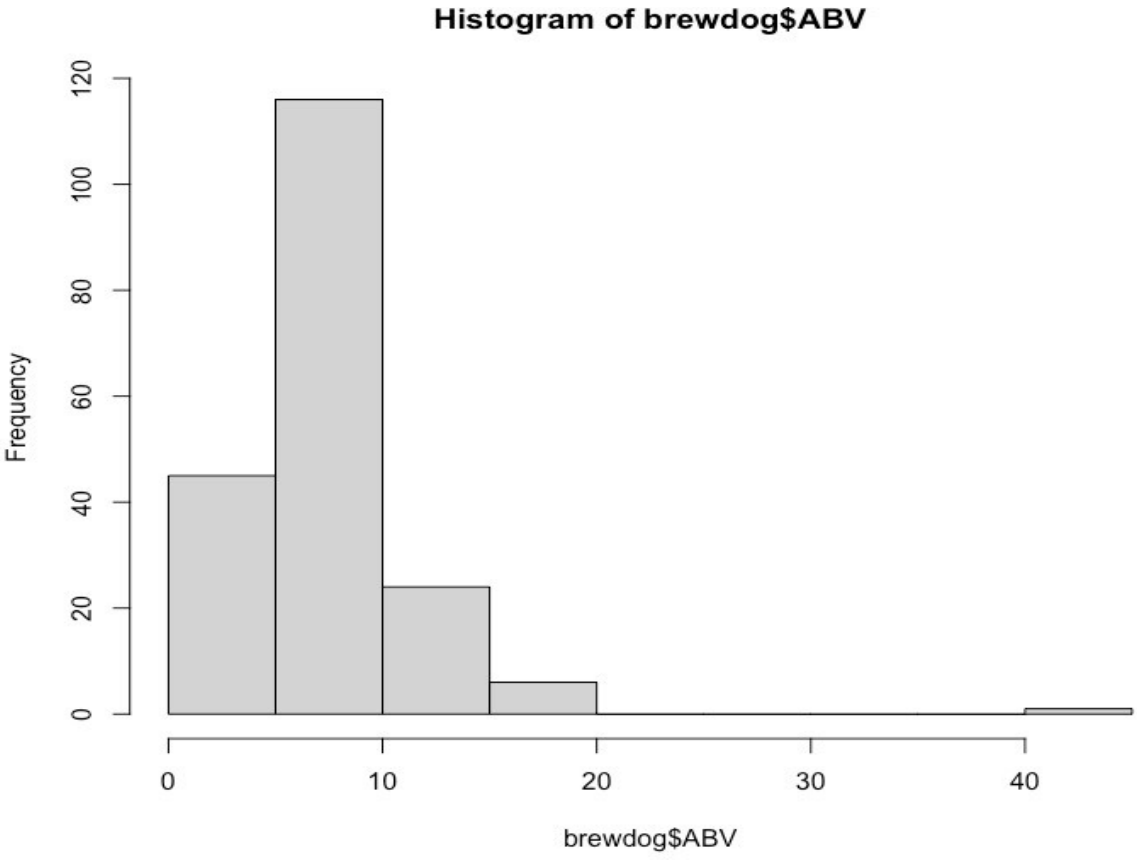


Figure: Histogram of ABV in Brewdog

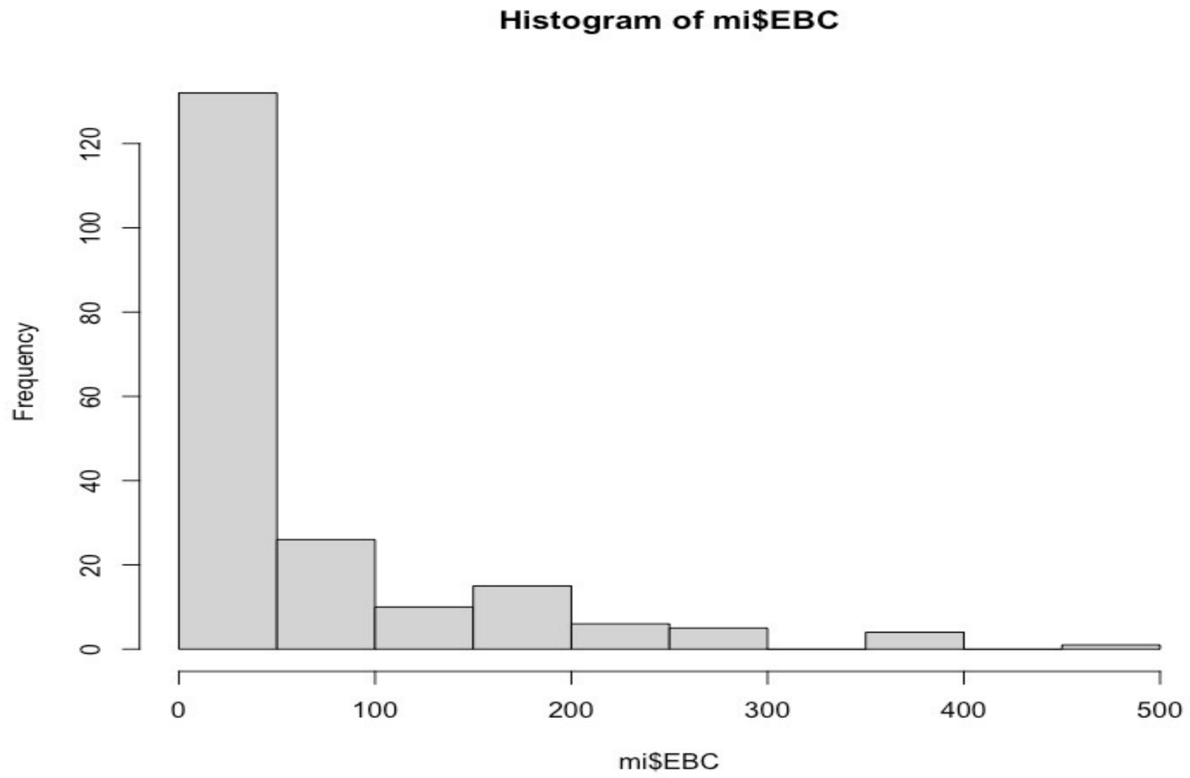


Figure: Histogram of EBC in mi

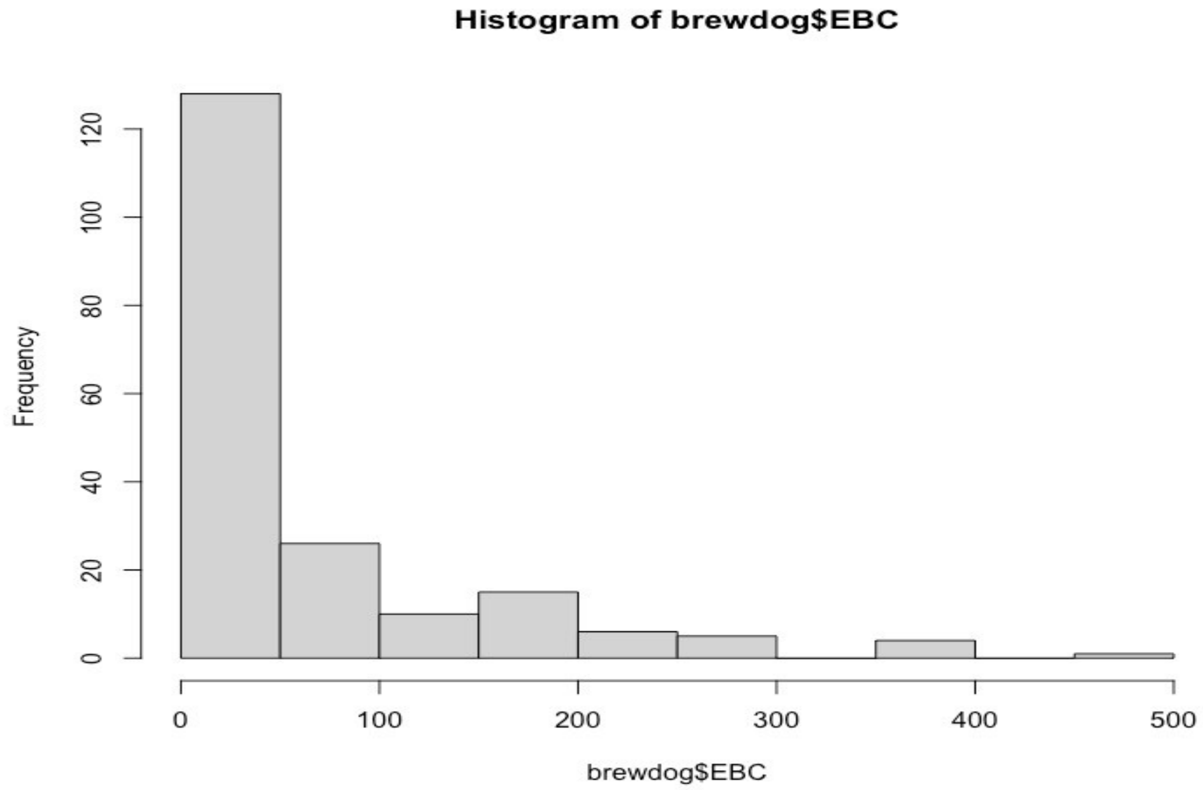


Figure: Histogram of EBC in Brewdog

1.6. Scaling the complete dataset

Brewdog consists of mixed numerical data with different units. Hence, the data have been scaled to maintain a normalised distribution. The dataframe is sliced by including only the numerical data.

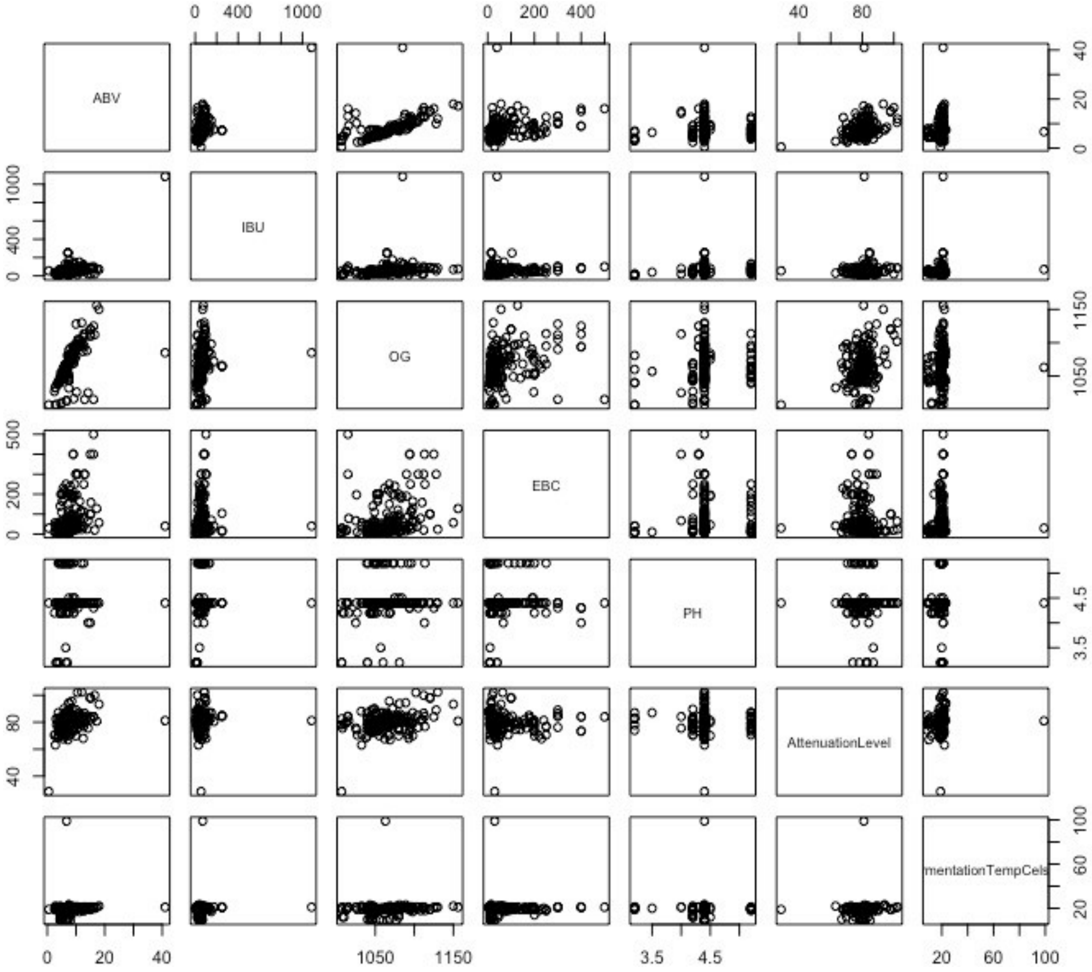


Figure 2.6.1: Numerical data before scaling

The location of the data points after scaling have not changed compared to its original location, whereas the scale on the x and y axis are aligned.

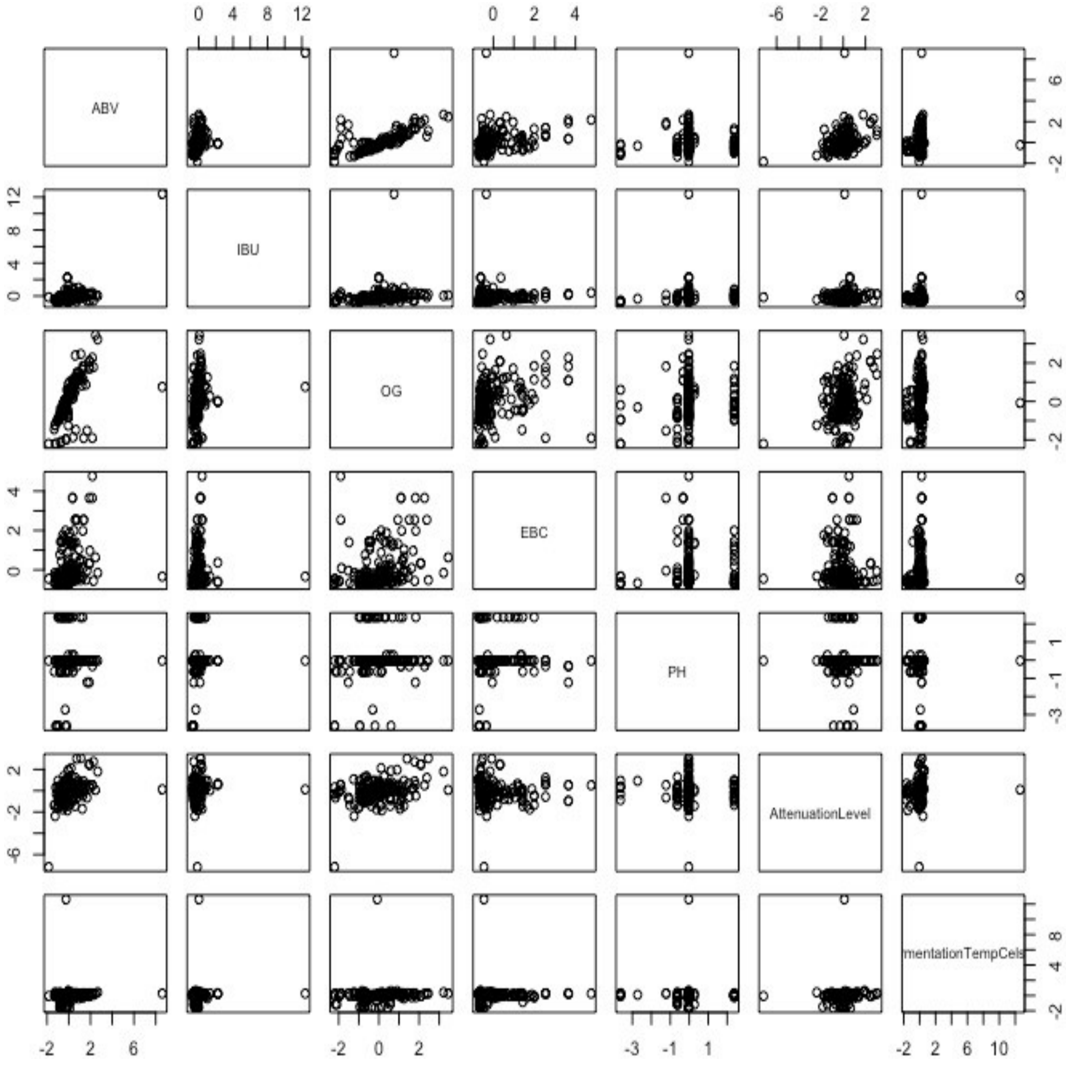


Figure: Numerical data after scaling

1.7. **Hierarchical clustering**

Clustering is the process of grouping data points into different clusters such that the objects inside the clusters are highly similar while the similarity between clusters is low. There are two types of clustering:

- Hierarchical Clustering
- Non-Hierarchical Clustering

Non-hierarchical clustering using k-means method is not considered for Brewdog dataset, since k-means algorithm isn’t directly applicable to categorical data (yeast column in Brewdog), and it requires a previous knowledge about the number of clusters.

The scaled dataset is then clustered using hierarchical clustering algorithm which groups the beers into a hierarchical series of nested clusters represented as a dendrogram. Agglomerative clustering approach is followed which initially places each data point in a single cluster and then finds the clusters which are closest to each other to merge them into a single cluster.

Hierarchical clustering is performed using fastcluster package in R. Dissimilarity matrix calculates the distance between clusters. Since, Brewdog consists of one categorical column named Yeast, the dissimilarity matrix is calculated using daisy() in cluster package which uses Euclidean distance for numerical data and Gower’s distance for categorical data.

The dissimilarity matrix produced by daisy() along with the agglomerative method wards is provided as input to agnes() to perform hierarchical clustering. The beers are clustered into 4 different clusters using hclust().

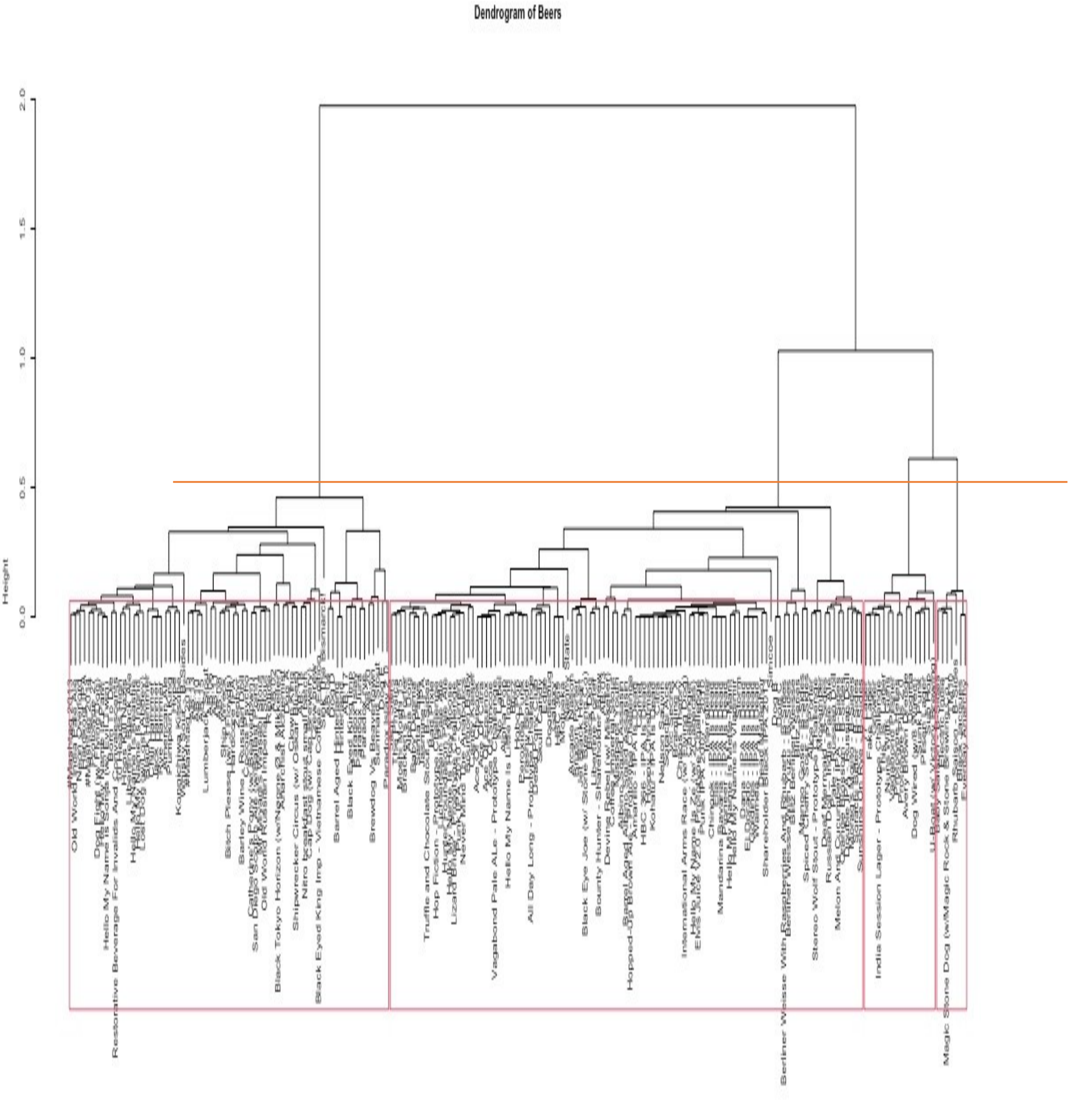
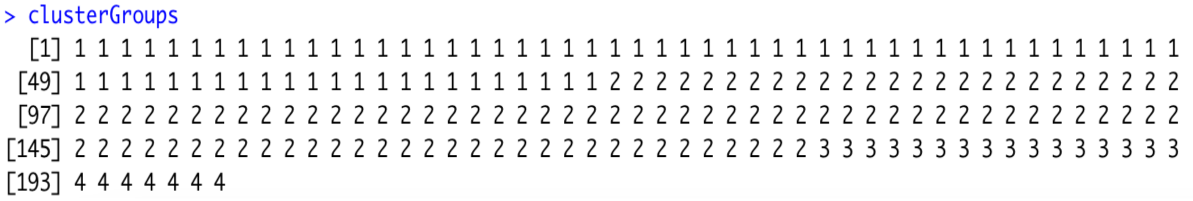


Figure: Dendrogram of Beers

1.8. Analysing the clustered information

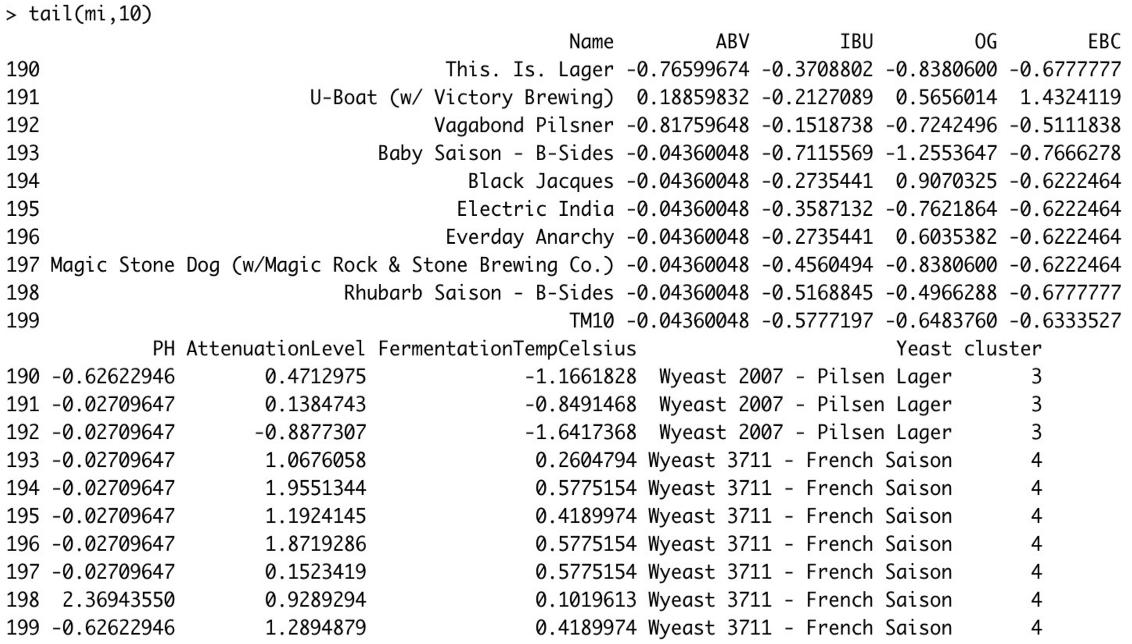
The clustered information is stored in variable “clusterGroups”.



```
> clusterGroups
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [49] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [97] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[145] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[193] 4 4 4 4 4 4 4
```

Figure 2.8.1: Cluster data

A new column named “cluster” is added to the dataframe which stores the cluster numbers of each of the beer. The data is then arranged in ascending order of the cluster numbers.

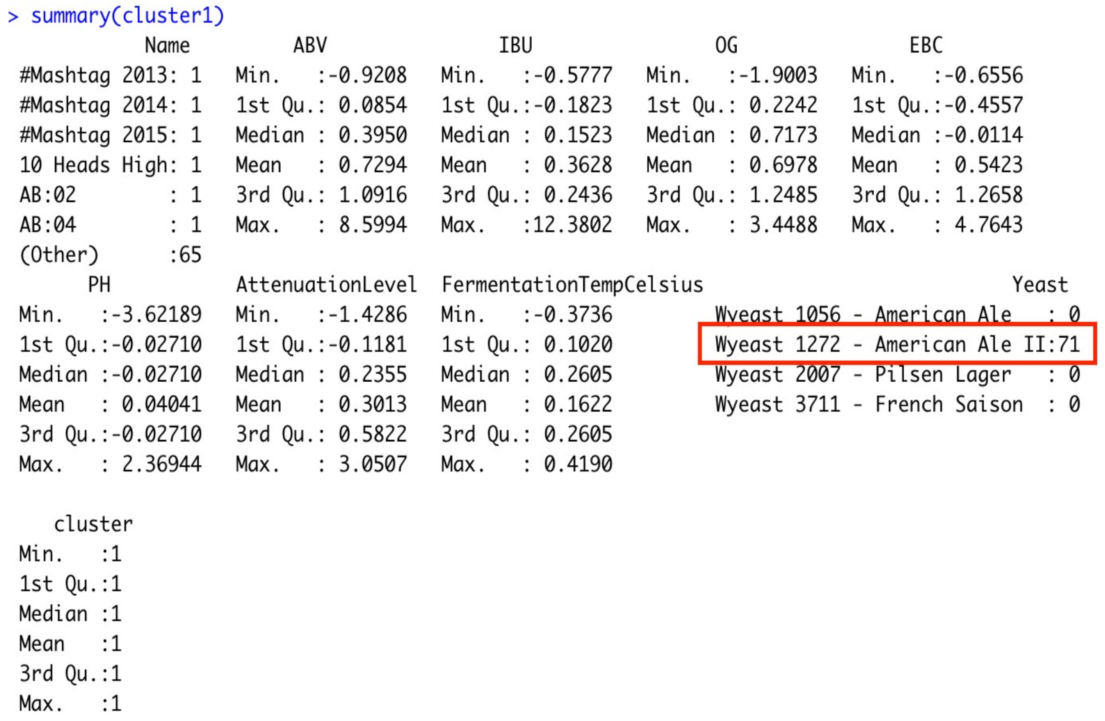


```
> tail(mi,10)
                                              Name         ABV        IBU         OG        EBC
190                                This. Is. Lager -0.76599674 -0.3708802 -0.8380600 -0.6777777
191                       U-Boat (w/ Victory Brewing)  0.18859832 -0.2127089  0.5656014  1.4324119
192                                Vagabond Pilsner -0.81759648 -0.1518738 -0.7242496 -0.5111838
193                            Baby Saison - B-Sides -0.04360048 -0.7115569 -1.2553647 -0.7666278
194                                   Black Jacques -0.04360048 -0.2735441  0.9070325 -0.6222464
195                                  Electric India -0.04360048 -0.3587132 -0.7621864 -0.6222464
196                                 Everday Anarchy -0.04360048 -0.2735441  0.6035382 -0.6222464
197 Magic Stone Dog (w/Magic Rock & Stone Brewing Co.) -0.04360048 -0.4560494 -0.8380600 -0.6222464
198                           Rhubarb Saison - B-Sides -0.04360048 -0.5168845 -0.4966288 -0.6777777
199                                            TM10 -0.04360048 -0.5777197 -0.6483760 -0.6333527
             PH AttenuationLevel FermentationTempCelsius                      Yeast cluster
190 -0.62622946        0.4712975              -1.1661828  Wyeast 2007 - Pilsen Lager       3
191 -0.02709647        0.1384743              -0.8491468  Wyeast 2007 - Pilsen Lager       3
192 -0.02709647       -0.8877307              -1.6417368  Wyeast 2007 - Pilsen Lager       3
193 -0.02709647        1.0676058               0.2604794 Wyeast 3711 - French Saison       4
194 -0.02709647        1.9551344               0.5775154 Wyeast 3711 - French Saison       4
195 -0.02709647        1.1924145               0.4189974 Wyeast 3711 - French Saison       4
196 -0.02709647        1.8719286               0.5775154 Wyeast 3711 - French Saison       4
197 -0.02709647        0.1523419               0.5775154 Wyeast 3711 - French Saison       4
198  2.36943550        0.9289294               0.1019613 Wyeast 3711 - French Saison       4
199 -0.62622946        1.2894879               0.4189974 Wyeast 3711 - French Saison       4
```

Figure: Overview of mi after clustering

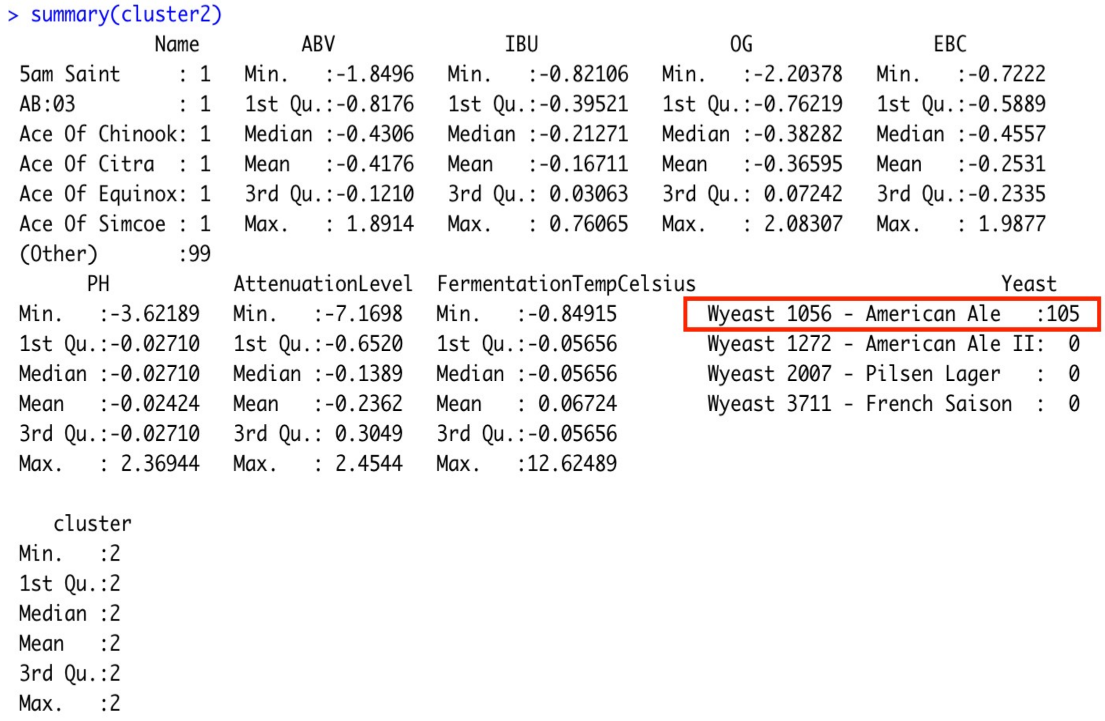
The 199 different beers have been grouped based on the yeast type because it makes the biggest difference in terms of dissimilarity between clusters and highest similarity within the cluster.

- Cluster 1 consists of 71 beers of yeast type Wyeast 1271 - American Ale II
- Cluster 2 consists of 105 beers of yeast type Wyeast 1056 - American Ale
- Cluster 3 consists of 16 beers of yeast type Wyeast 2007 - Pilsen Lager
- Cluster 4 consists of 7 beers of yeast type Wyeast 3711 - French Saison



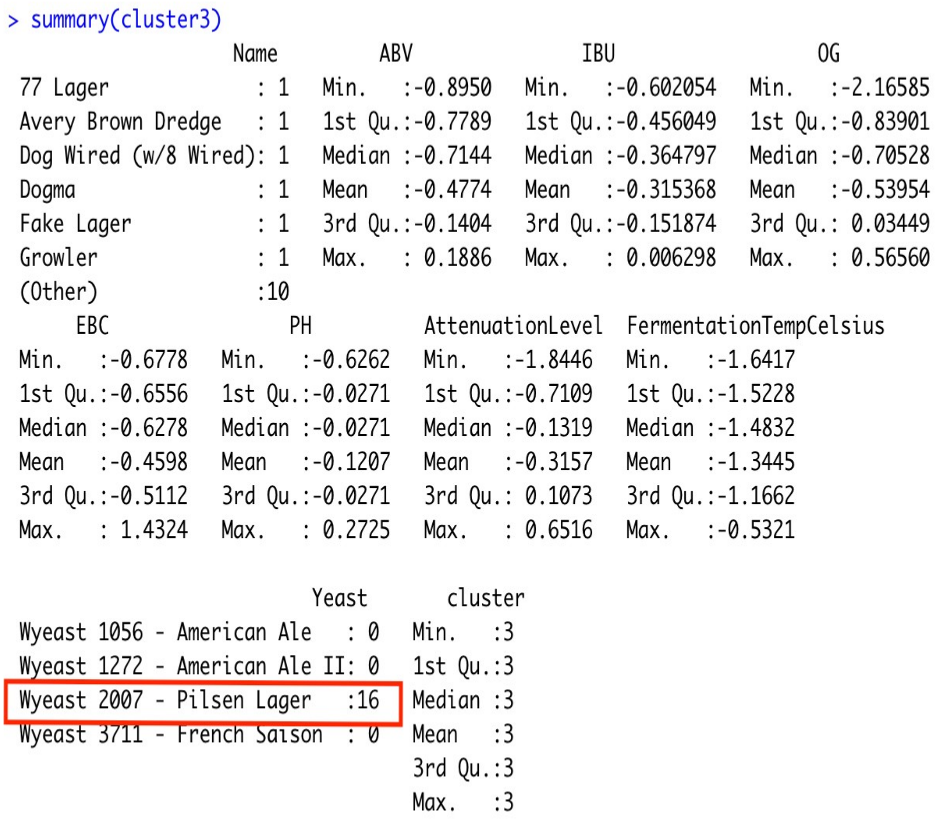
```
> summary(cluster1)
            Name           ABV                IBU                 OG                EBC         
 #Mashtag 2013: 1   Min.   :-0.9208   Min.   :-0.5777   Min.   :-1.9003   Min.   :-0.6556  
 #Mashtag 2014: 1   1st Qu.: 0.0854   1st Qu.:-0.1823   1st Qu.: 0.2242   1st Qu.:-0.4557  
 #Mashtag 2015: 1   Median : 0.3950   Median : 0.1523   Median : 0.7173   Median :-0.0114  
 10 Heads High: 1   Mean   : 0.7294   Mean   : 0.3628   Mean   : 0.6978   Mean   : 0.5423  
 AB:02        : 1   3rd Qu.: 1.0916   3rd Qu.: 0.2436   3rd Qu.: 1.2485   3rd Qu.: 1.2658  
 AB:04        : 1   Max.   : 8.5994   Max.   :12.3802   Max.   : 3.4488   Max.   : 4.7643  
 (Other)      :65                                                                           
       PH             AttenuationLevel  FermentationTempCelsius                            Yeast   
 Min.   :-3.62189   Min.   :-1.4286   Min.   :-0.3736         Wyeast 1056 - American Ale   : 0  
 1st Qu.:-0.02710   1st Qu.:-0.1181   1st Qu.: 0.1020         Wyeast 1272 - American Ale II:71  
 Median :-0.02710   Median : 0.2355   Median : 0.2605         Wyeast 2007 - Pilsen Lager   : 0  
 Mean   : 0.04041   Mean   : 0.3013   Mean   : 0.1622         Wyeast 3711 - French Saison  : 0  
 3rd Qu.:-0.02710   3rd Qu.: 0.5822   3rd Qu.: 0.2605                                          
 Max.   : 2.36944   Max.   : 3.0507   Max.   : 0.4190                                          
                                                                                               
    cluster 
 Min.   :1  
 1st Qu.:1  
 Median :1  
 Mean   :1  
 3rd Qu.:1  
 Max.   :1  
```

Figure: Summary of cluster1



```
> summary(cluster2)
             Name          ABV                IBU                  OG                  EBC         
 5am Saint     : 1   Min.   :-1.8496   Min.   :-0.82106   Min.   :-2.20378   Min.   :-0.7222  
 AB:03         : 1   1st Qu.:-0.8176   1st Qu.:-0.39521   1st Qu.:-0.76219   1st Qu.:-0.5889  
 Ace Of Chinook: 1   Median :-0.4306   Median :-0.21271   Median :-0.38282   Median :-0.4557  
 Ace Of Citra  : 1   Mean   :-0.4176   Mean   :-0.16711   Mean   :-0.36595   Mean   :-0.2531  
 Ace Of Equinox: 1   3rd Qu.:-0.1210   3rd Qu.: 0.03063   3rd Qu.: 0.07242   3rd Qu.:-0.2335  
 Ace Of Simcoe : 1   Max.   : 1.8914   Max.   : 0.76065   Max.   : 2.08307   Max.   : 1.9877  
 (Other)       :99                                                                             
       PH             AttenuationLevel FermentationTempCelsius                            Yeast    
 Min.   :-3.62189   Min.   :-7.1698   Min.   :-0.84915        Wyeast 1056 - American Ale   :105  
 1st Qu.:-0.02710   1st Qu.:-0.6520   1st Qu.:-0.05656        Wyeast 1272 - American Ale II:  0  
 Median :-0.02710   Median :-0.1389   Median :-0.05656        Wyeast 2007 - Pilsen Lager   :  0  
 Mean   :-0.02424   Mean   :-0.2362   Mean   : 0.06724        Wyeast 3711 - French Saison  :  0  
 3rd Qu.:-0.02710   3rd Qu.: 0.3049   3rd Qu.:-0.05656                                         
 Max.   : 2.36944   Max.   : 2.4544   Max.   :12.62489                                         
                                                                                               
    cluster 
 Min.   :2  
 1st Qu.:2  
 Median :2  
 Mean   :2  
 3rd Qu.:2  
 Max.   :2  
```

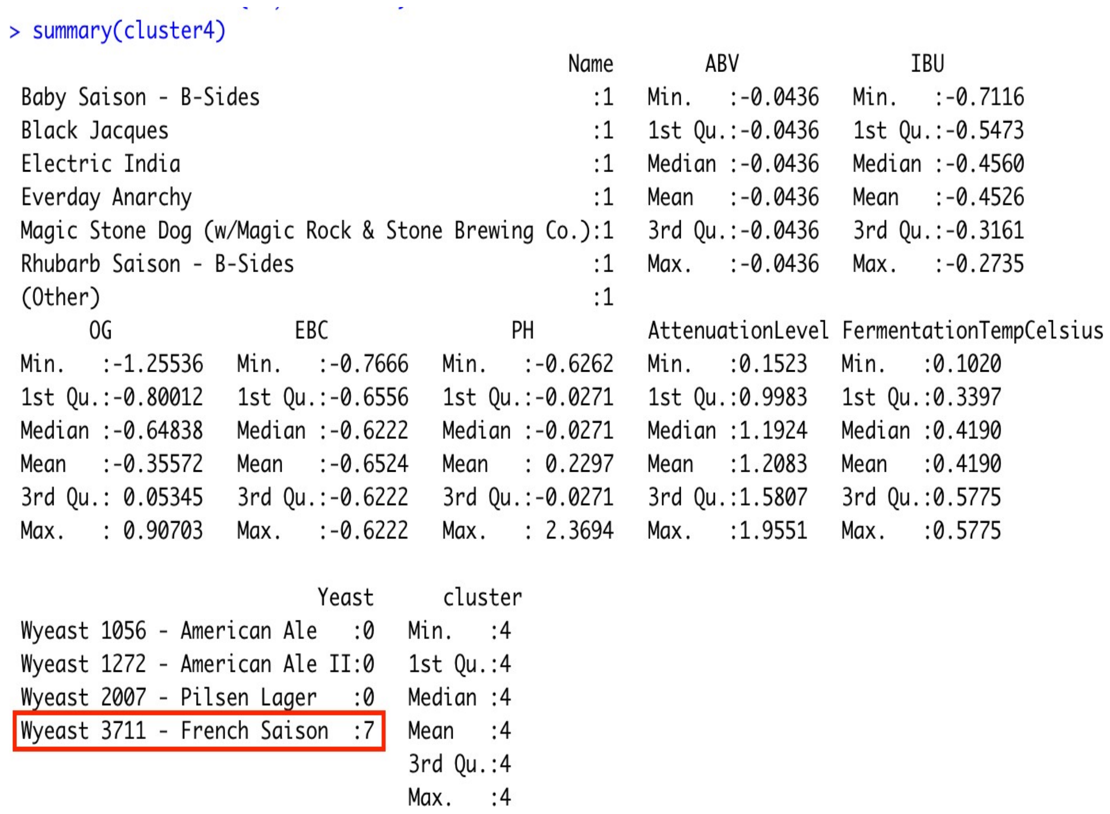
Figure: Summary of cluster 2



```
> summary(cluster3)
                     Name          ABV                IBU                    OG          
 77 Lager              : 1   Min.   :-0.8950   Min.   :-0.602054   Min.   :-2.16585  
 Avery Brown Dredge    : 1   1st Qu.:-0.7789   1st Qu.:-0.456049   1st Qu.:-0.83901  
 Dog Wired (w/8 Wired): 1   Median :-0.7144   Median :-0.364797   Median :-0.70528  
 Dogma                 : 1   Mean   :-0.4774   Mean   :-0.315368   Mean   :-0.53954  
 Fake Lager            : 1   3rd Qu.:-0.1404   3rd Qu.:-0.151874   3rd Qu.: 0.03449  
 Growler               : 1   Max.   : 0.1886   Max.   : 0.006298   Max.   : 0.56560  
 (Other)              :10                                                         
      EBC               PH           AttenuationLevel  FermentationTempCelsius
 Min.   :-0.6778   Min.   :-0.6262   Min.   :-1.8446   Min.   :-1.6417        
 1st Qu.:-0.6556   1st Qu.:-0.0271   1st Qu.:-0.7109   1st Qu.:-1.5228        
 Median :-0.6278   Median :-0.0271   Median :-0.1319   Median :-1.4832        
 Mean   :-0.4598   Mean   :-0.1207   Mean   :-0.3157   Mean   :-1.3445        
 3rd Qu.:-0.5112   3rd Qu.:-0.0271   3rd Qu.: 0.1073   3rd Qu.:-1.1662        
 Max.   : 1.4324   Max.   : 0.2725   Max.   : 0.6516   Max.   :-0.5321        

                          Yeast       cluster 
 Wyeast 1056 - American Ale    : 0   Min.   :3  
 Wyeast 1272 - American Ale II: 0   1st Qu.:3  
 Wyeast 2007 - Pilsen Lager    :16   Median :3  
 Wyeast 3711 - French Saison   : 0   Mean   :3  
                                     3rd Qu.:3  
                                     Max.   :3  
```

Figure: Summary of cluster 3



```
> summary(cluster4)
                                                      Name        ABV                IBU         
 Baby Saison - B-Sides                                  :1   Min.   :-0.0436   Min.   :-0.7116  
 Black Jacques                                          :1   1st Qu.:-0.0436   1st Qu.:-0.5473  
 Electric India                                         :1   Median :-0.0436   Median :-0.4560  
 Everday Anarchy                                        :1   Mean   :-0.0436   Mean   :-0.4526  
 Magic Stone Dog (w/Magic Rock & Stone Brewing Co.):1   3rd Qu.:-0.0436   3rd Qu.:-0.3161  
 Rhubarb Saison - B-Sides                               :1   Max.   :-0.0436   Max.   :-0.2735  
 (Other)                                                :1                                     
       OG                EBC                PH           AttenuationLevel FermentationTempCelsius
 Min.   :-1.25536   Min.   :-0.7666   Min.   :-0.6262   Min.   :0.1523   Min.   :0.1020         
 1st Qu.:-0.80012   1st Qu.:-0.6556   1st Qu.:-0.0271   1st Qu.:0.9983   1st Qu.:0.3397         
 Median :-0.64838   Median :-0.6222   Median :-0.0271   Median :1.1924   Median :0.4190         
 Mean   :-0.35572   Mean   :-0.6524   Mean   : 0.2297   Mean   :1.2083   Mean   :0.4190         
 3rd Qu.: 0.05345   3rd Qu.:-0.6222   3rd Qu.:-0.0271   3rd Qu.:1.5807   3rd Qu.:0.5775         
 Max.   : 0.90703   Max.   :-0.6222   Max.   : 2.3694   Max.   :1.9551   Max.   :0.5775         

                          Yeast      cluster 
 Wyeast 1056 - American Ale    :0   Min.   :4  
 Wyeast 1272 - American Ale II:0   1st Qu.:4  
 Wyeast 2007 - Pilsen Lager    :0   Median :4  
 Wyeast 3711 - French Saison   :7   Mean   :4  
                                    3rd Qu.:4  
                                    Max.   :4  
```

Figure: Summary of cluster 4

APPENDIX

This section consists of the R code used for the implementation of Part-2.

```
library("dplyr")
library("VIM")
library("mice")
library("corrgram")

brewdog <- read.csv("Brewdog.csv", header = TRUE, stringsAsFactors = T)
brewdog

#printing first 10 rows of Brewdog
head(brewdog,10)

#checking missing data
summary(brewdog)
aggr(brewdog, numbers=TRUE, prop= FALSE)

#creating missing data column
missdata <- brewdog
missdata$missing <- as.numeric(!complete.cases(brewdog))
corrgram(missdata)
#clear correlation between ABV and other variables. Similar positive correlation
between EBC and other variables. Possibly MAR

#------DELETION------

del <- brewdog[complete.cases(brewdog),]
dim(brewdog)
dim(del)


#-------SIMPLE IMPUTATION-------
si<- brewdog
si$ABV[is.na(si$ABV)] <- mean(si$ABV,na.rm=TRUE)
summary(si$ABV)
summary(brewdog$ABV)
sd(si$ABV,na.rm=TRUE)
sd(brewdog$ABV,na.rm=TRUE)

si$EBC[is.na(si$EBC)] <- mean(si$EBC,na.rm=TRUE)
summary(si$EBC)
summary(brewdog$EBC)
sd(si$EBC,na.rm=TRUE)
sd(brewdog$EBC,na.rm=TRUE)
```

```
#------MULTIPLE IMPUTATION------
imi<- mice(brewdog, m=10,maxit = 20)
mi<-complete(imi)
mi
head(mi,20)
aggr(mi, numbers=TRUE, prop=FALSE)
summary(mi)

# Check ABV variable results
hist(mi$ABV)
hist(brewdog$ABV)
summary(mi$ABV) #mean still close
summary(brewdog$ABV)
sd(mi$ABV,na.rm=TRUE)
sd(brewdog$ABV,na.rm=TRUE)

# Check EBC variable results.
hist(mi$EBC)
hist(brewdog$EBC)
summary(mi$EBC)
summary(brewdog$EBC)
sd(mi$EBC,na.rm=TRUE)
sd(brewdog$EBC,na.rm=TRUE)



#------SCALING THE DATASET------
mi[,2:8] <- scale(mi[,2:8],center=TRUE, scale=TRUE)
plot(mi[,2:8])
plot(brewdog[,2:8])

#------HIERARCHICAL CLUSTERING-------
library("fastcluster")
library("cluster")

#creating dissimilarity matrix
dm <- daisy(mi[2:9])

#clustering
clust <- agnes(dm,diss=TRUE, method="ward")

#plotting the dendrogram
par(cex=0.5, mar=c(6,6,6,6))
plot(clust,labels=mi$Name,main="Dendrogram of Beers", which.plots=2)
rect.hclust(clust,4)


#analysing cluster numbers
clusterGroups <- cutree(clust, k=4)
clusterGroups
```

```
#adding the clustered information
to the dataframemi$cluster <-
clusterGroups
mi<-
arran
ge(mi,
cluste
r)
tail(mi
,10)

#printing
information of
each cluster
cluster1 <-
filter(mi,cluster==1
)
summary(cluster1)

cluster2 <-
filter(mi,clust
er==2)
summary(clu
ster2)

cluster3 <-
filter(mi,clust
er==3)
summary(clu
ster3)

cluster4 <-
filter(mi,clust
er==4)
summary(clu
ster4)
```