**SAINT PETERSBURG STATE UNIVERSITY**

Faculty of Applied Mathematics and Control Processes

Mathematical Game Theory and Statistical Decisions Department

Applied Statistics in R

Exam

Professor: Parilina Elena M.

Student: Orlov Ivan M., 19.M09-пу

Saint Petersburg

2020

Contents

[1. Descriptive statistics. 2](#_Toc43579776)

[2. Statistical goodness-of-fit tests. 5](#_Toc43579777)

[3. Comparing two or more samples. 6](#_Toc43579778)

[4. Regression analysis. 7](#_Toc43579779)

[5. Cluster analysis. 9](#_Toc43579780)

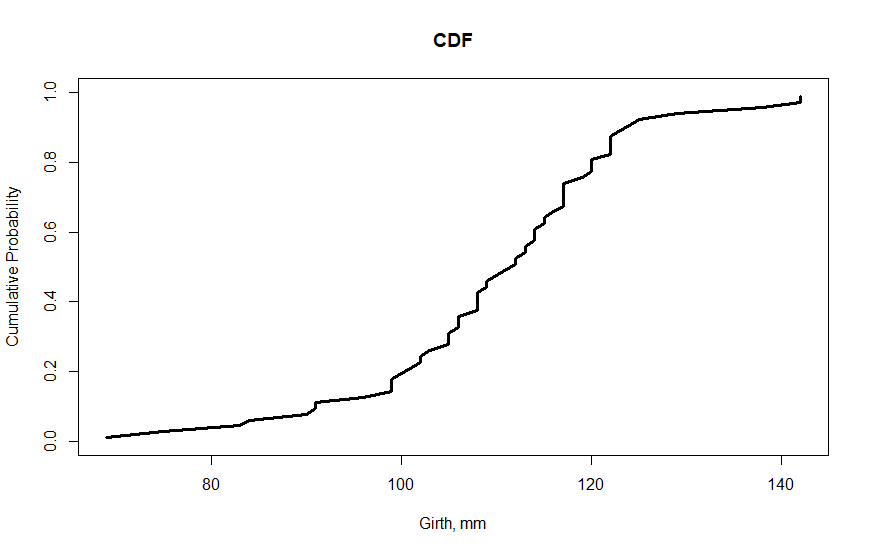
[Applications (code) 22](#_Toc43579781)

# Descriptive statistics.

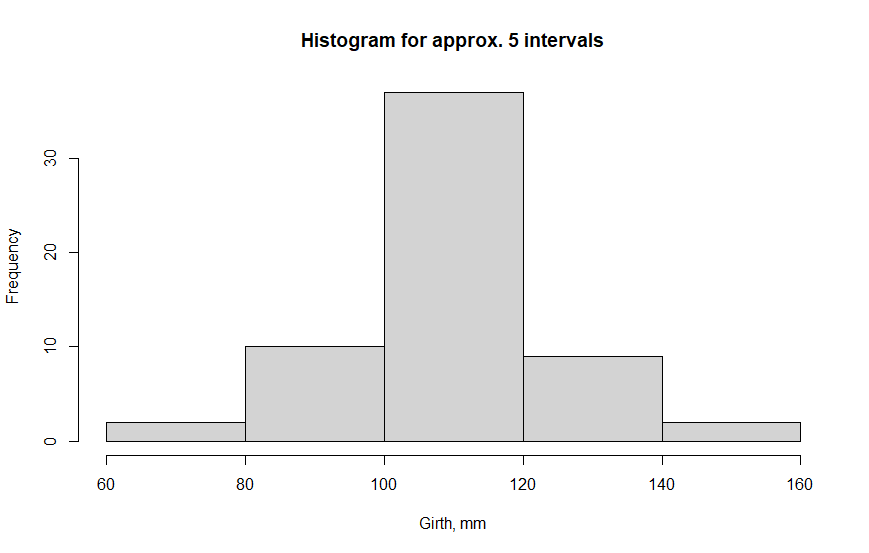
*The data represent trunk girth (mm) of a random sample of 60 four-year-old apple trees at East Malling*

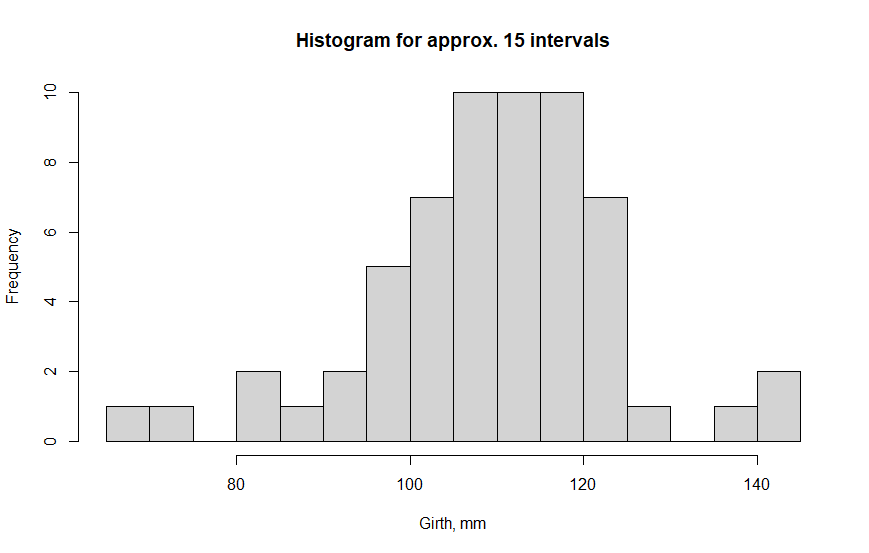
*Research Station (England). Reference: S.C. Pearce, University of Kent at Canterbury*

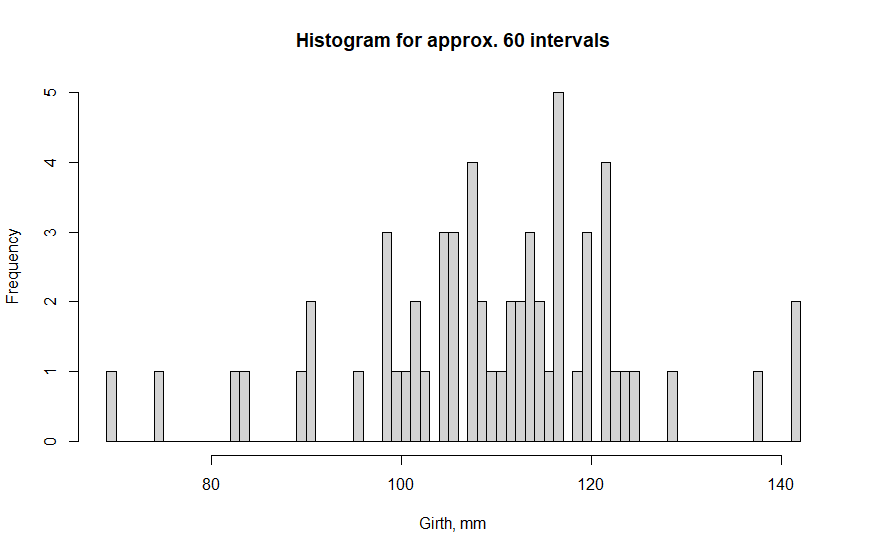
|  |  |
| --- | --- |
| **Statistics** | **Value** |
| n | 60 |
| mean | 110.0167 |
| variance | 197.237 |
| standard deviation | 14.04411 |
| median | 111.5 |
| 0.25-quantile | 102.75 |
| 0.5-quantile | 111.5 |
| 0.75-quantile | 117.5 |
| mode | 117 |
| minimal value | 69 |
| maximal value | 142 |
| skewness | -0.39736 |
| kurtosis | 0.832999 |



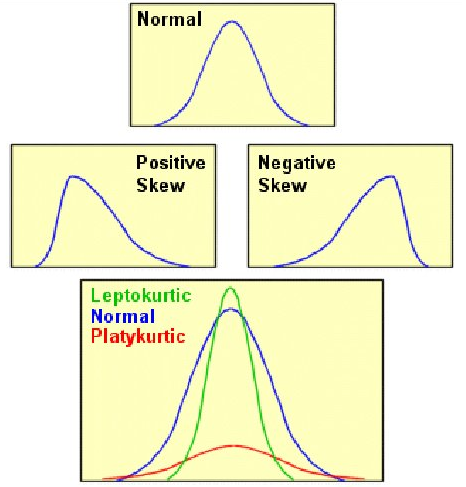
Cumulitive distribution function graph shows that the biggest grow is in the invterval between 100 and 120 mm: almost 60% of girths are in there.

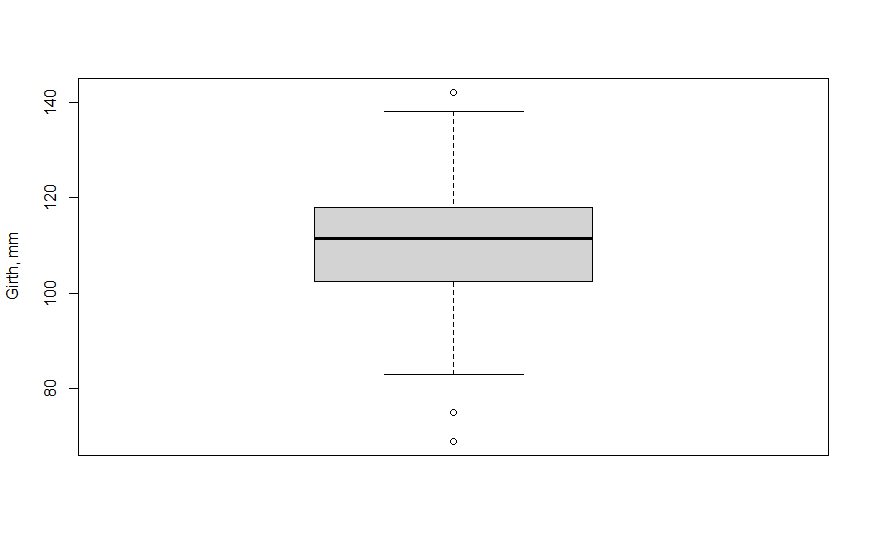




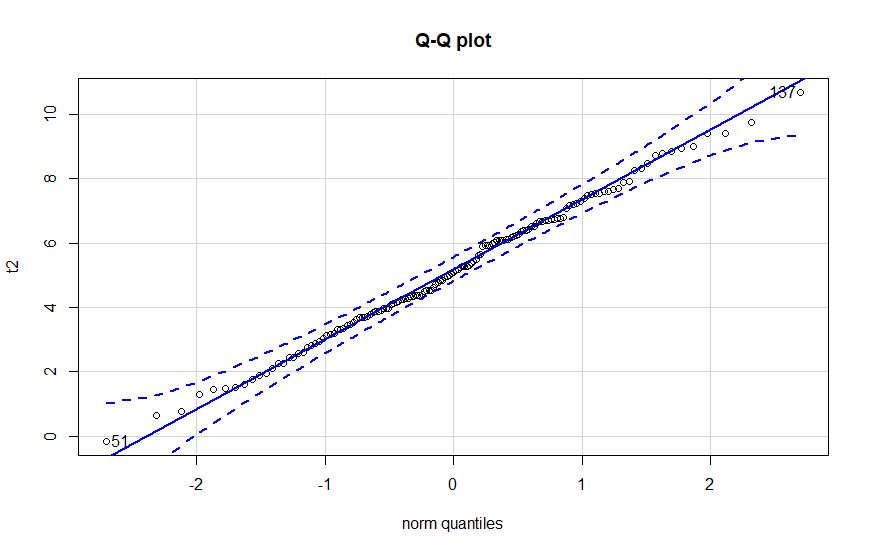


Asd





# Statistical goodness-of-fit tests.



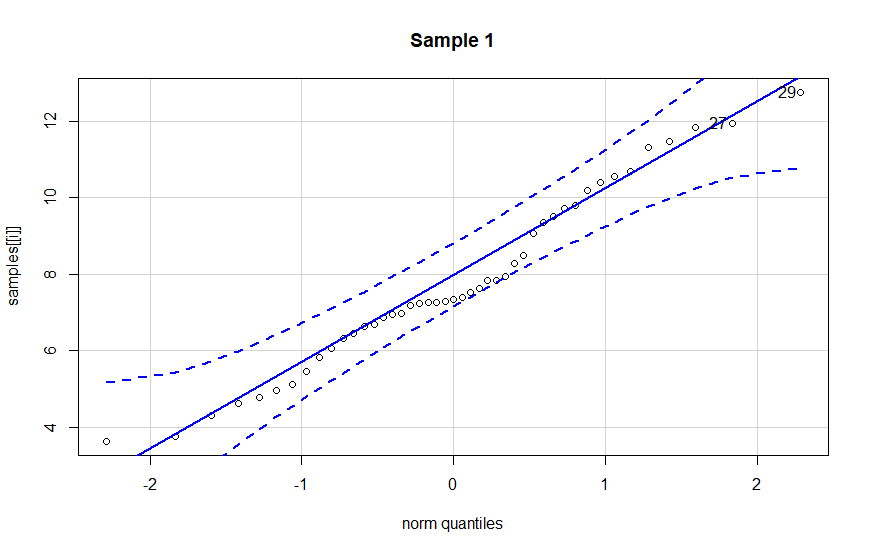
uniNorm function does no longer exist in R 4.0.1.

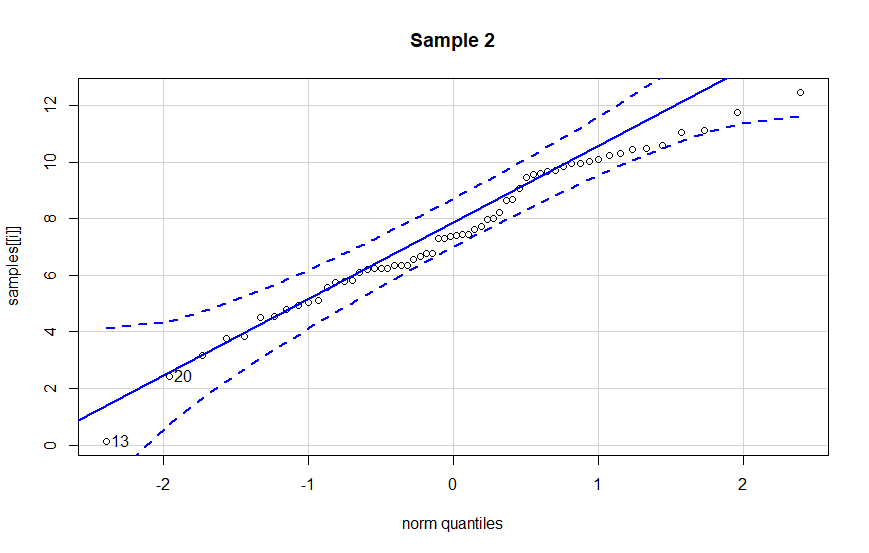
|  |  |
| --- | --- |
| **Statistics** | **Value** |
| n | 145 |
| Mean | 5.148819 |
| Std.Dev | 2.104814 |
| Median | 5.083483 |
| Min | -0.1551615 |
| Max | 10.67074 |
| 25th | 3.726937 |
| 75th | 6.656663 |
| Skew | 0.04403965 |
| Kurtosis | -0.388785 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Test** | **Statistic** | **p-value** | **Normality** |
| Pearson chi-square (adjusted) | 9.655172 | 0.6461832 | YES |
| Pearson chi-square | 9.655172 | 0.7869349 | YES |
| Lilliefors (Kolmogorov-Smirnov) | 0.0511 | 0.466 | YES |
| Cramer-von Mises | 0.0408 | 0.6648 | YES |
| Anderson-Darling | 0.2117 | 0.8542 | YES |

# Comparing two or more samples.

S





# Regression analysis.

*The data (X1, X2, X3) are for each patient.*

*Y = systolic blood pressure*

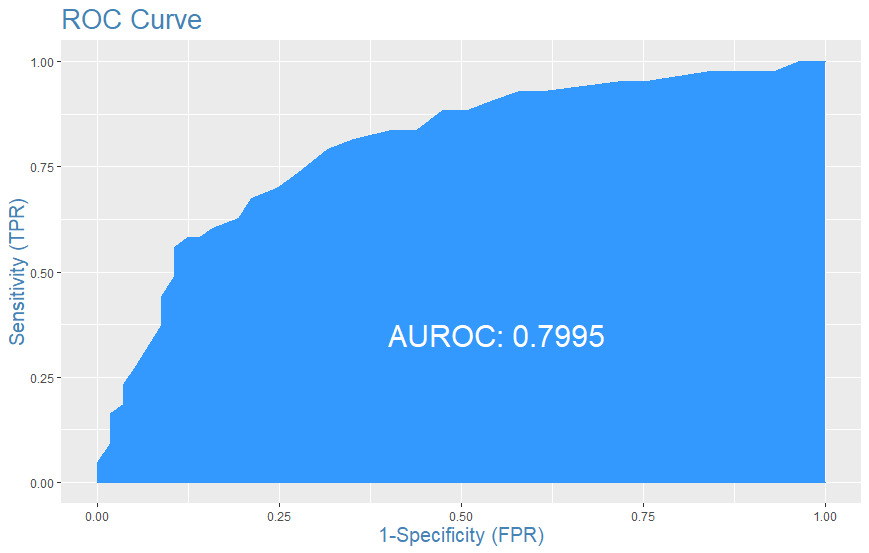
*X1 = age in years*

*X2 = weight in pounds*

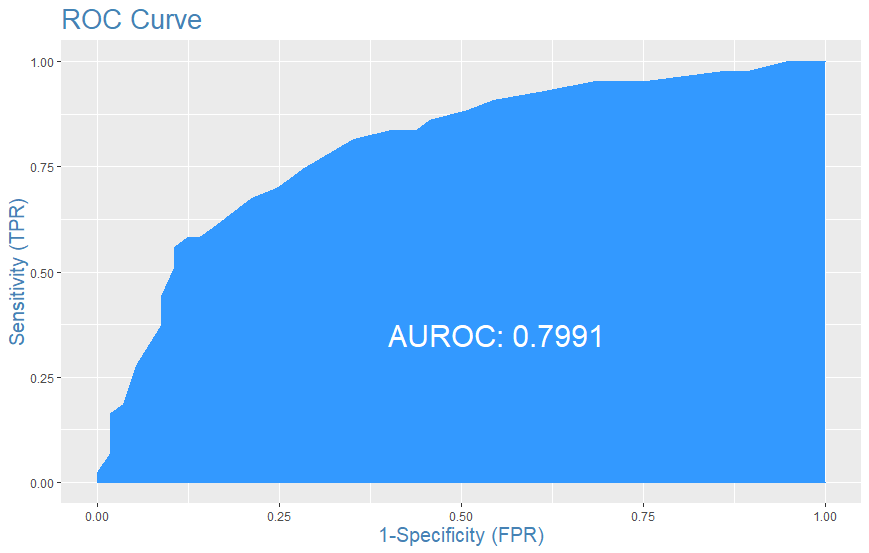
*X = Age*

*Y = "1" if person has kids and "0" if person has no kids*

Logit



Probit

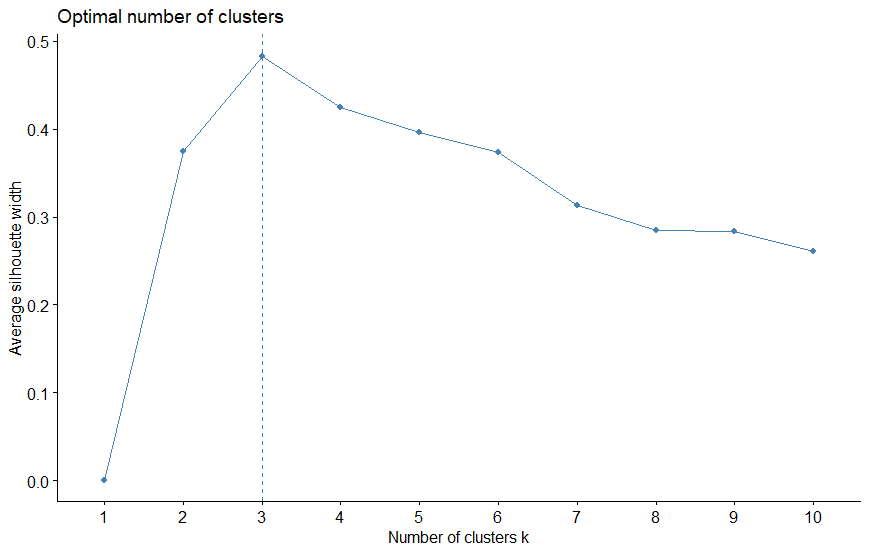


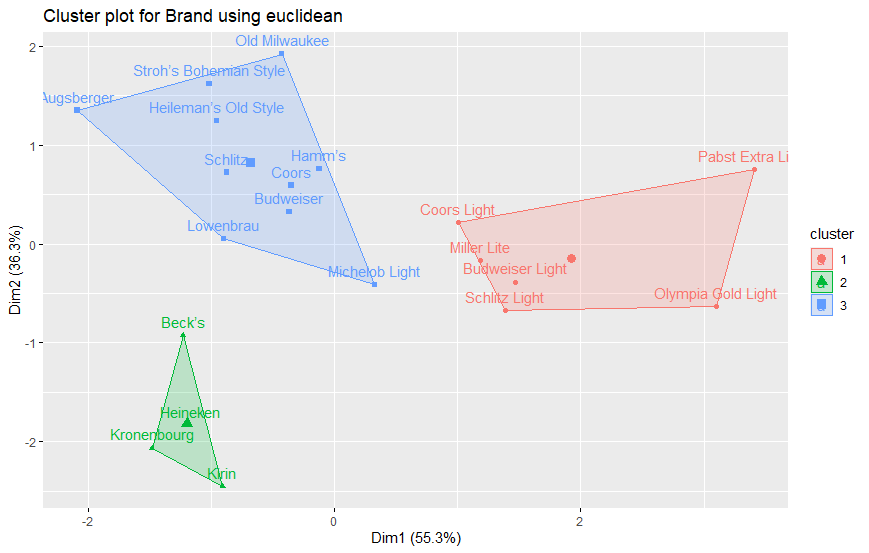
# Cluster analysis.

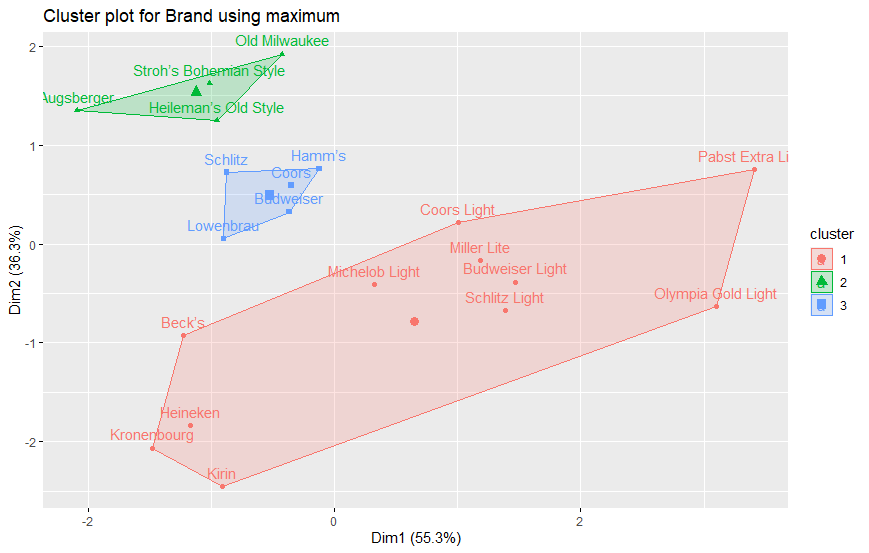
*Table 10.12 Attributes and Prices of Beer Brands for Cluster Analysis*

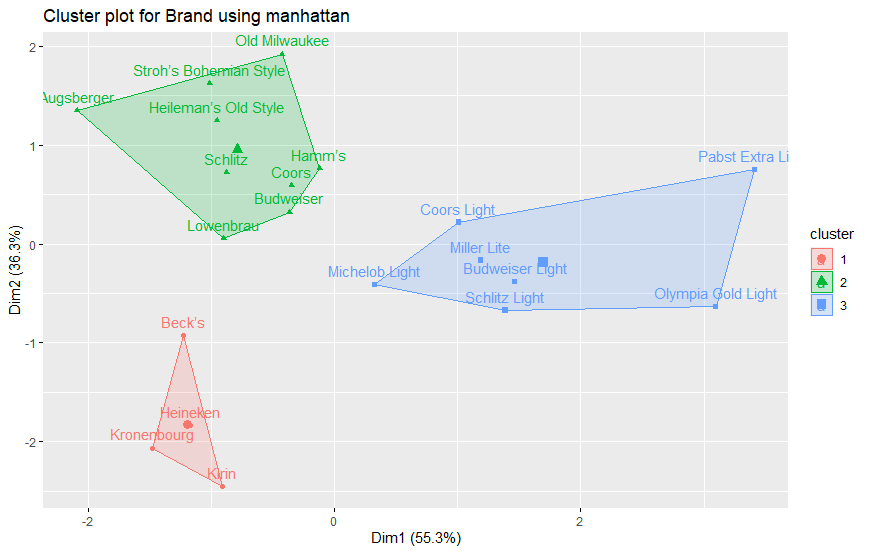
What parameters to use in cluster analysis?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Brand** | **Calories  (12 oz.)** | **Sodium**  **(mg/12 oz)** | **Alcohol  (%)** | **Price (Wholesale)** |



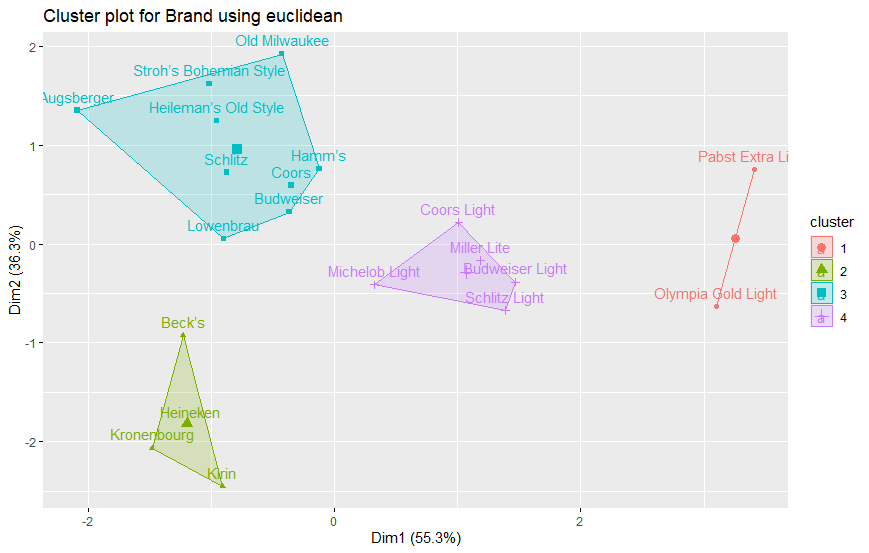


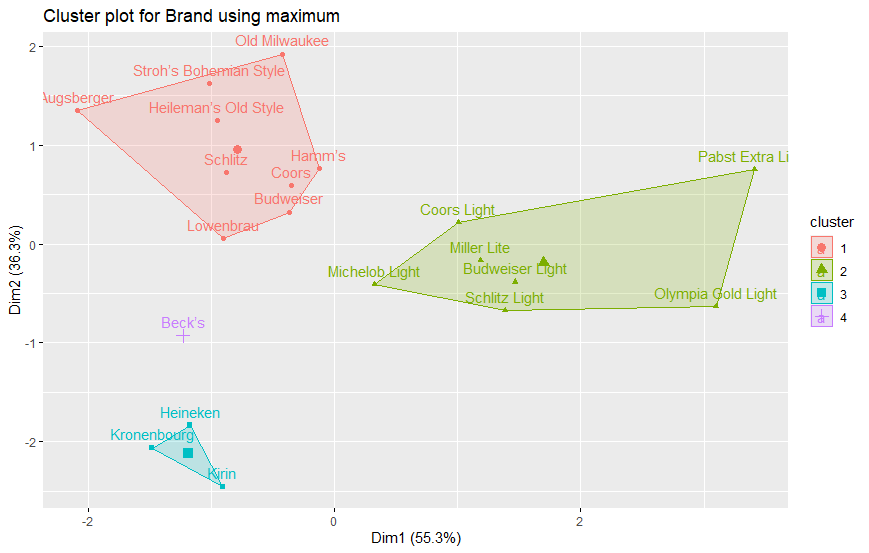


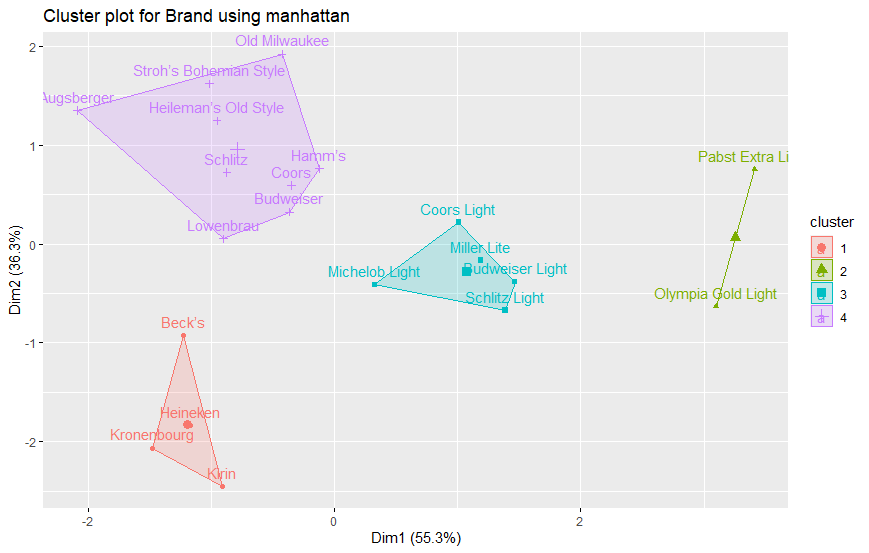


Number of clusters should be anywhere between the extreme 1 and 21 (in both of these situations cluster analysis is meaningless since either everything or nothing is the same), so better use numbers between 2 and 20.

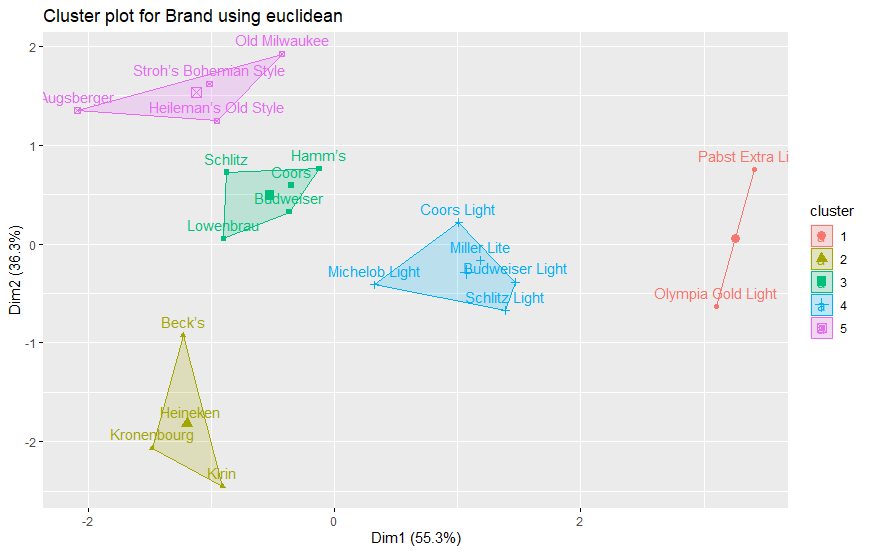
With *n=4:*

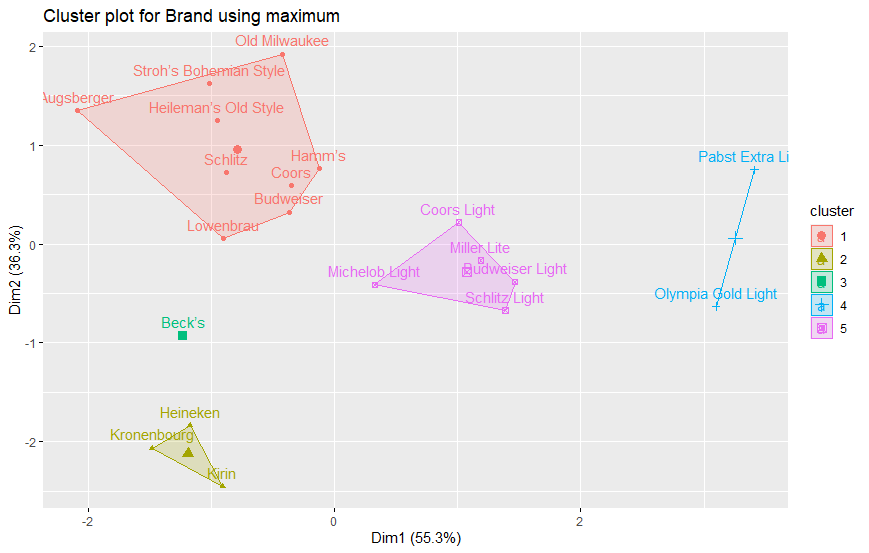


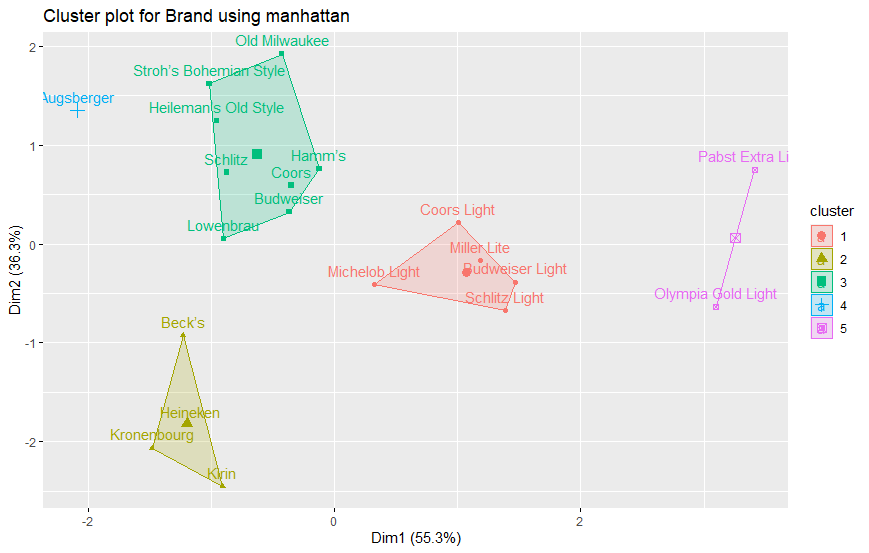




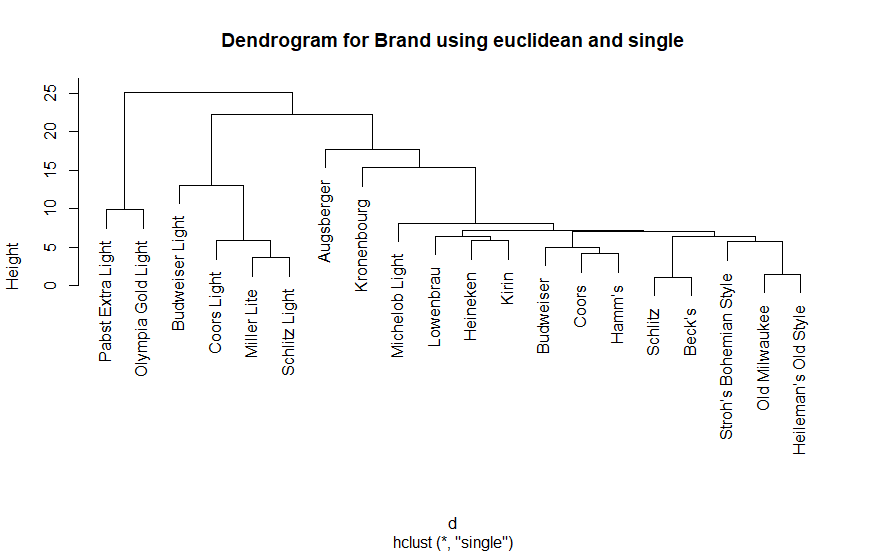
With *n=5:*

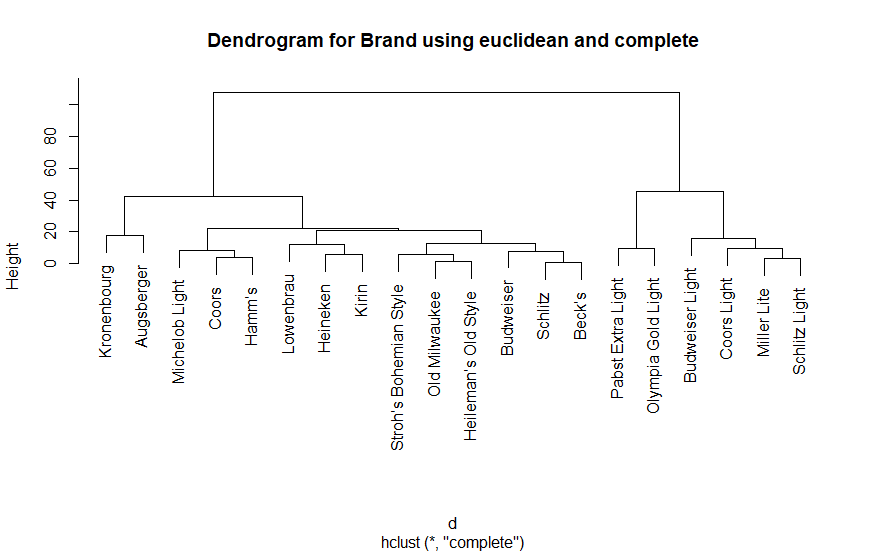


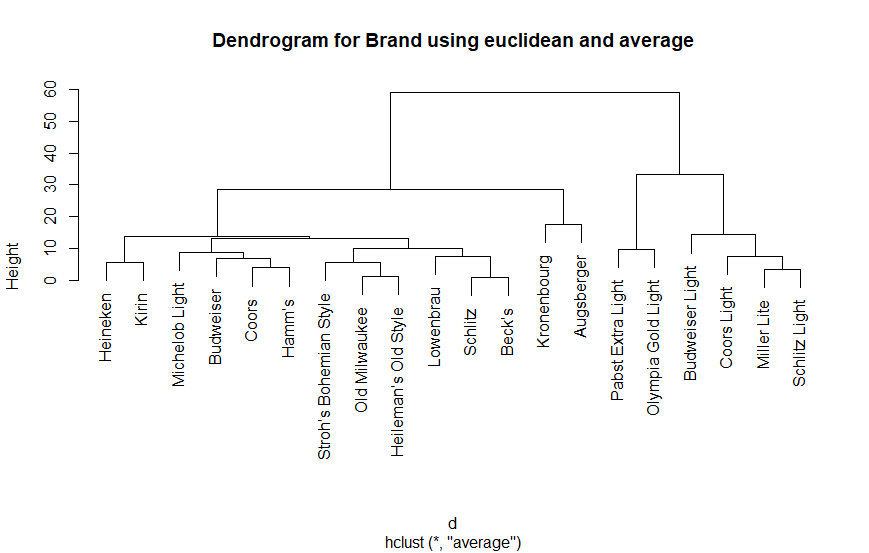


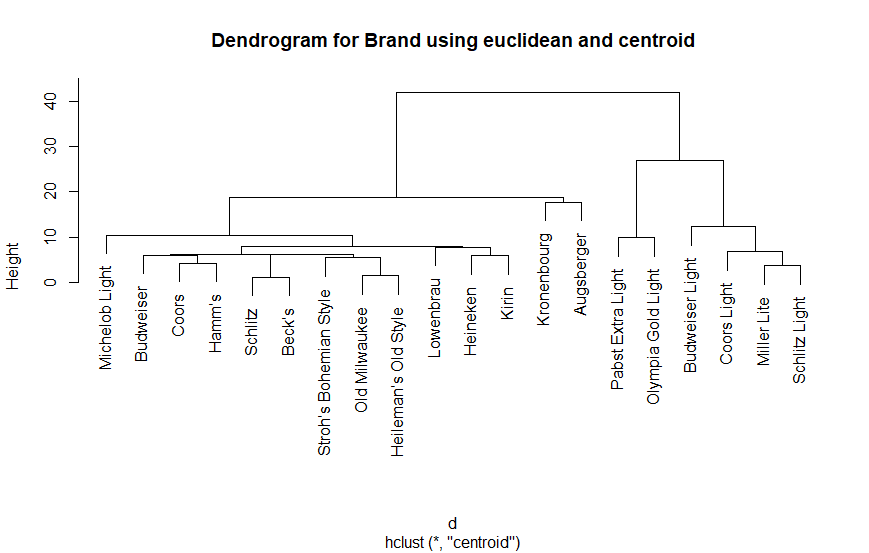


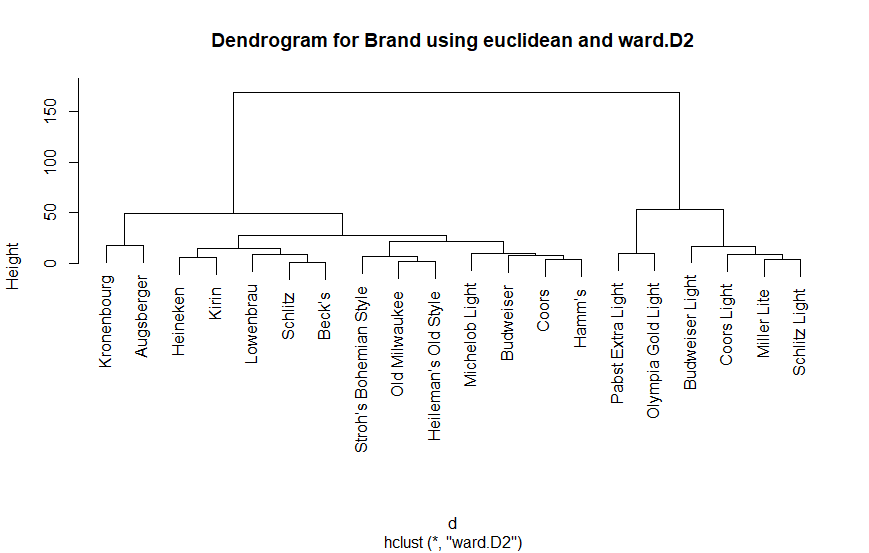
Hierarchical clusters for **euclidean/maximum/manhattan** distance methods and **single/complete/average/centroid/ward.D2** clusterization methods.

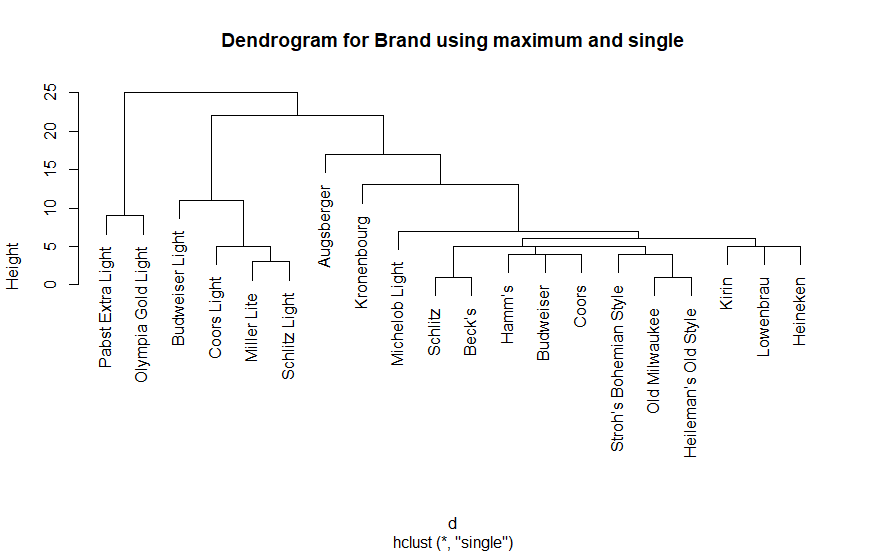


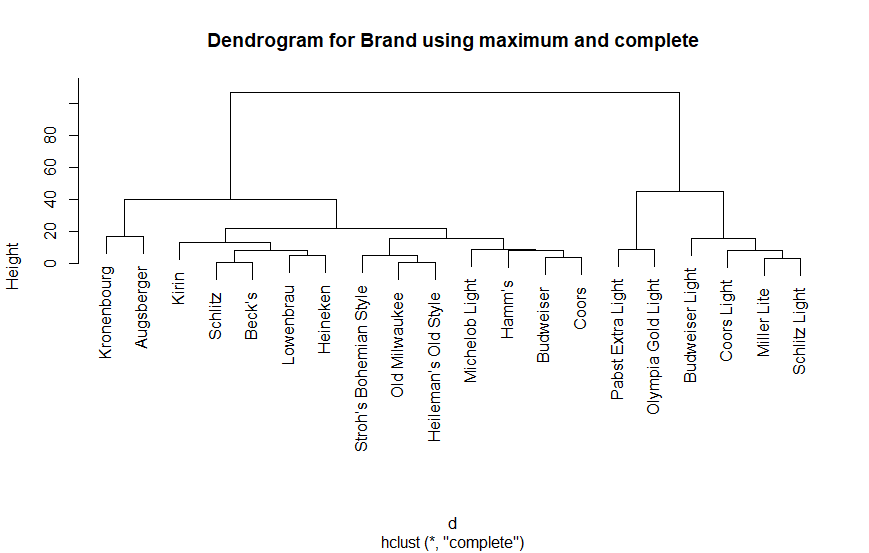


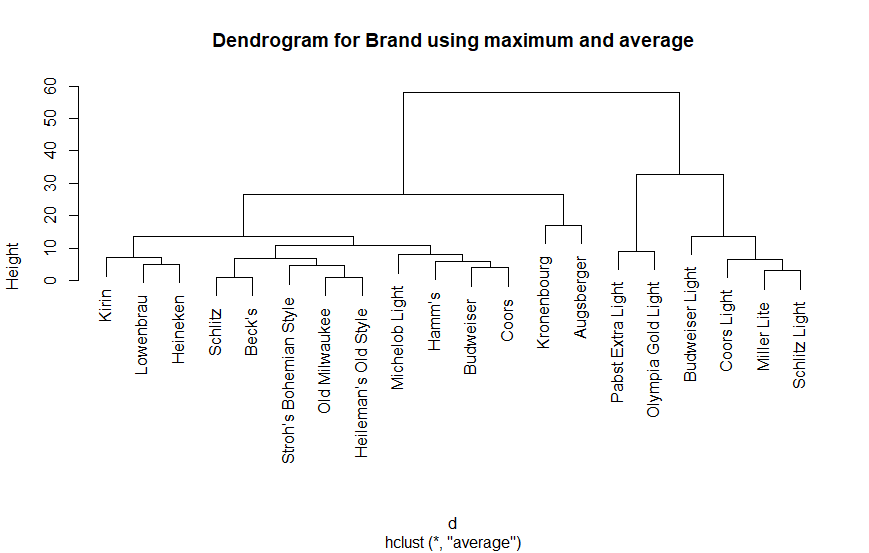


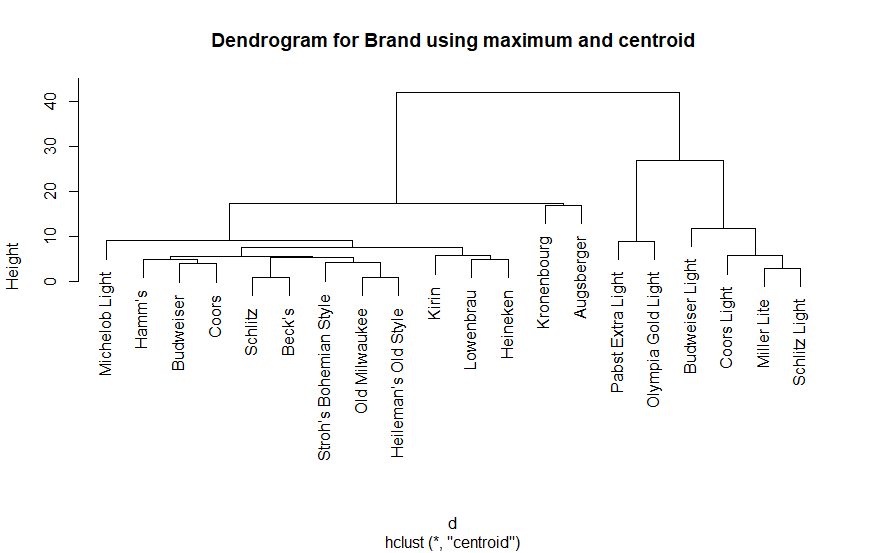


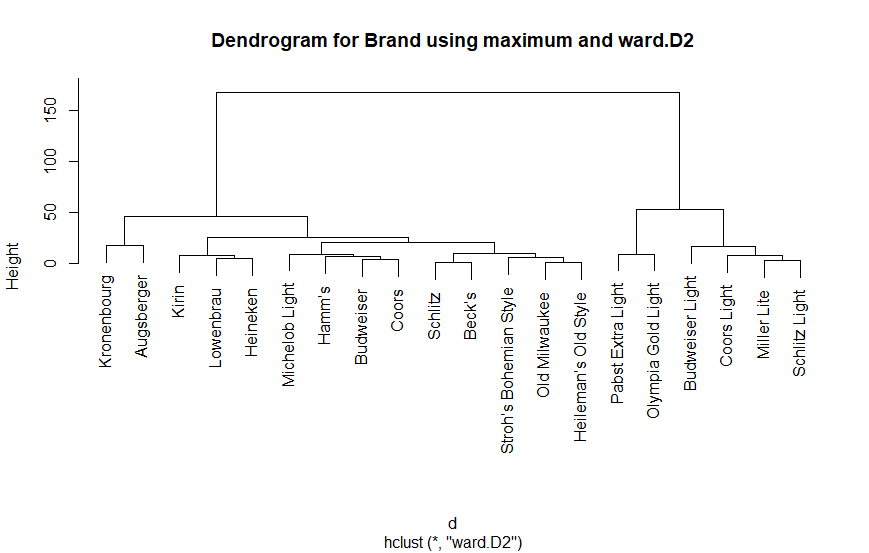


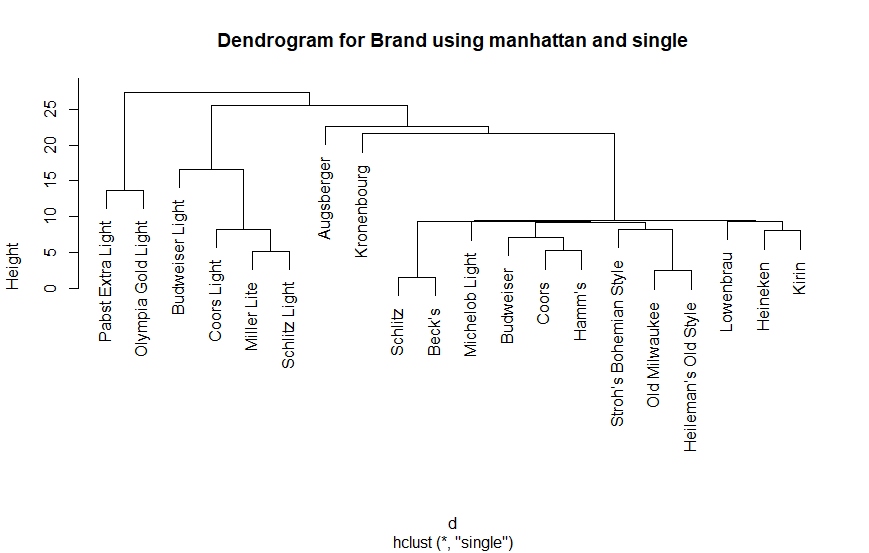


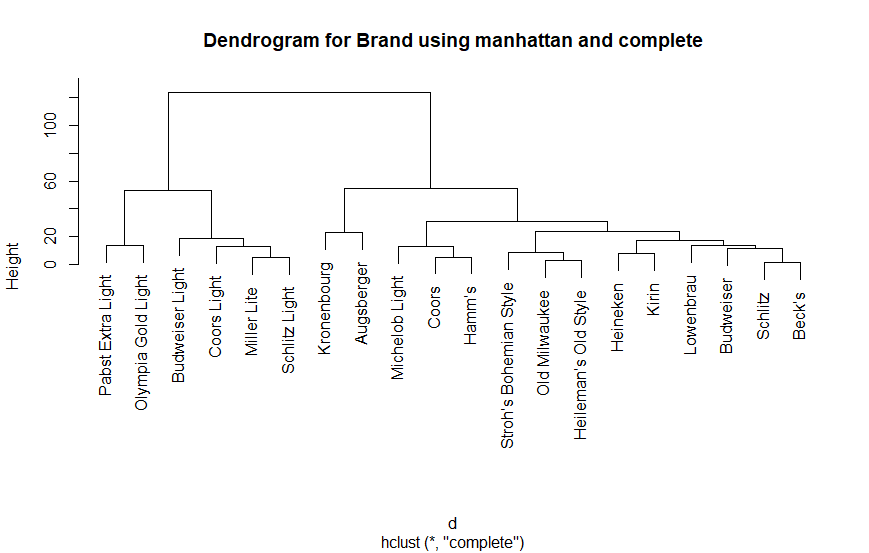


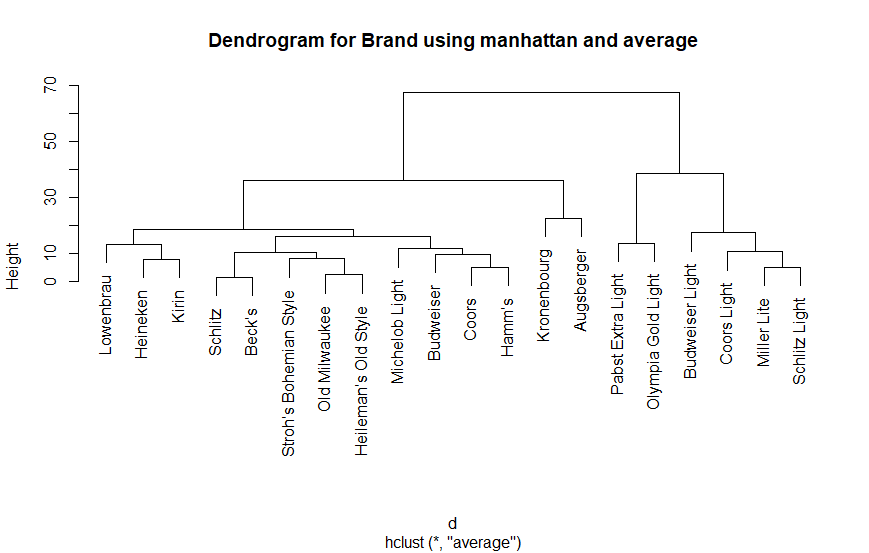


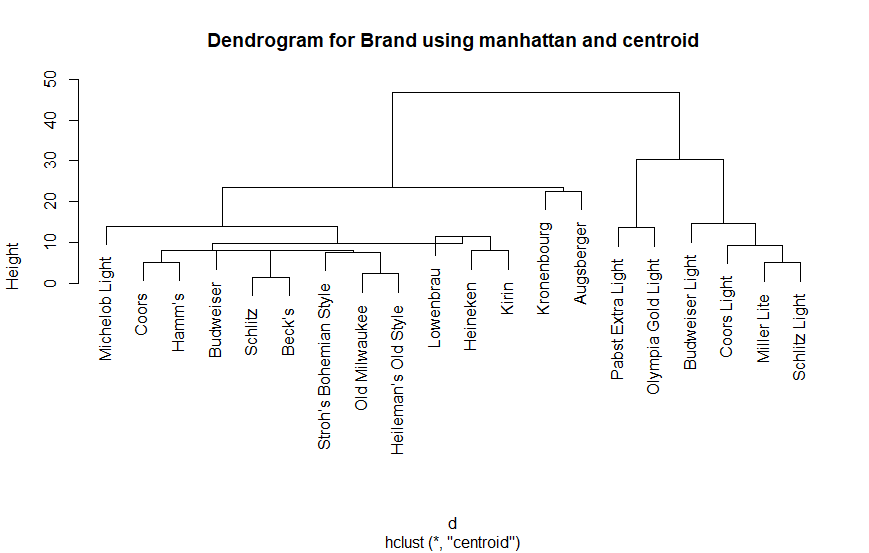


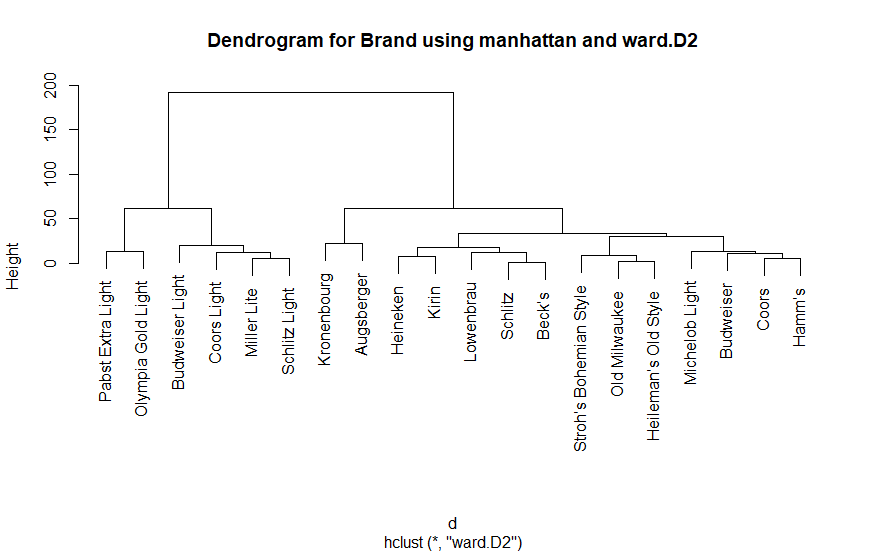












# Applications (code)

library(ggplot2)