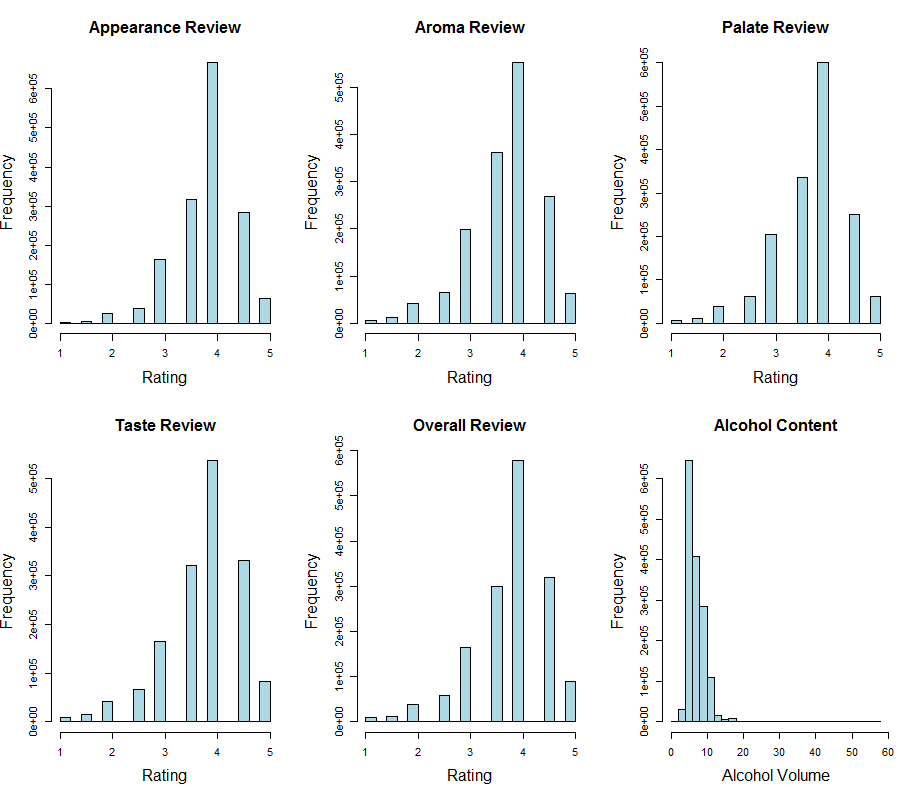
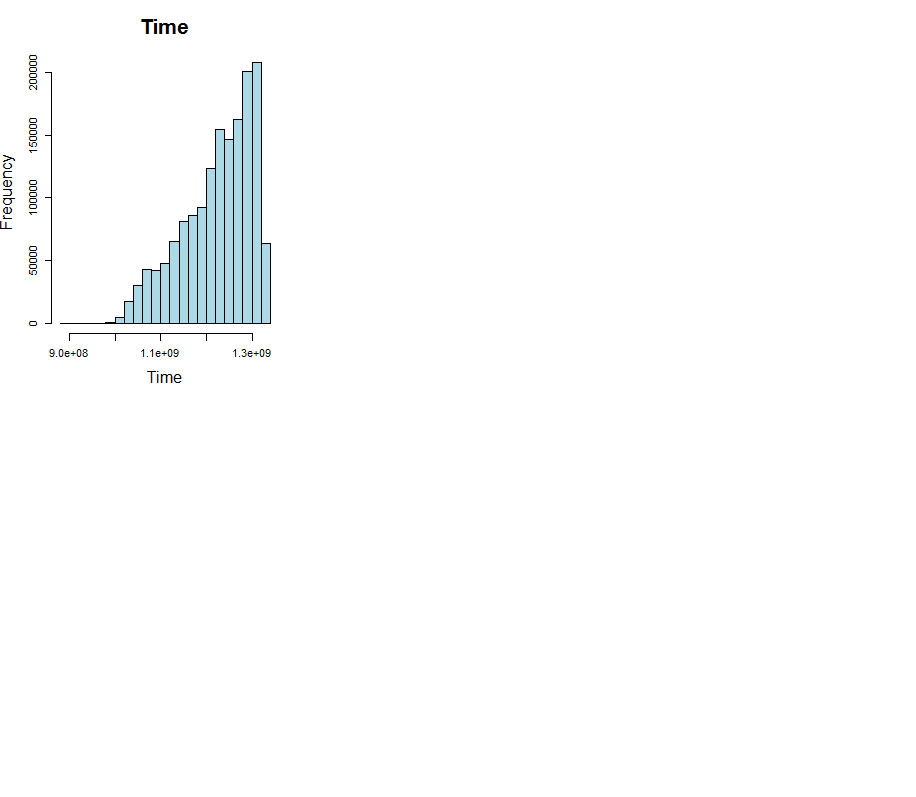
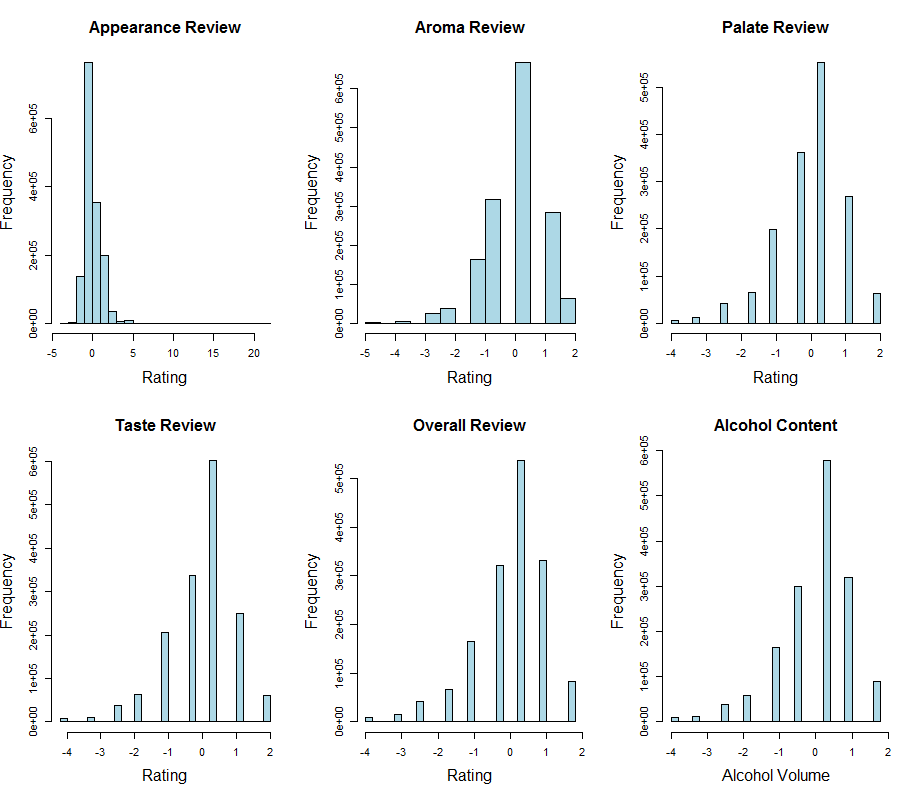
Distribution of raw variables:





Distribution of scaled variables:



-once these have been scaled the non-normality in alcohol volume is dramatically reduced.

All Janines notes are in blue…

**Why/how were they scaled?**

Centered by column mean ie column mean was subtracted

often, differences between the absolute values of variables are not of interest

differences between the averages for each variable obscure the interesting patterns

often the (co-)variation among variables is particularlyinteresting

**AND** Scaled by dividing by standard deviation of the column

Two main reasons for standardising variables

to put them on similar, standard, scales

or to make the scale more interpretable

if all the variables have different units, convert them to the same, standard, scale

- the variation within the variables is of comparable magnitude

- geometrically this is equalising the variation on each axis so

the data cloud tends to be more spherical

Assumptions:

From Janines Slides: Assumptions of PCA

PCA is normally an exploratory, hypothesis generating technique it doesn't really make any assumptions

however, PCA is more useful and the results easier to interpret (and more relevant) if certain assumptions can be made

in particular, random or at least representative sampling, and linearity of the relationships between variables

PCA assumes that the covariances (or correlations) adequately describe the relationships between variables

this is only true when the relationships are linear, or at least monotonic

**All quotes below comes from the book:**

Principal Component Anlaysis, Second Edition by Jolliffe, I.T. (2002) Springer: New York.

**Available at** [**http://f3.tiera.ru/2/M\_Mathematics/MV\_Probability/MVsa\_Statistics%20and%20applications/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)\_MVsa\_.pdf**](http://f3.tiera.ru/2/M_Mathematics/MV_Probability/MVsa_Statistics%20and%20applications/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf)

* Multivariate normality

“It will become apparent that many of the properties and applications of PCA…. have no need for explicit distributional assumptions. It cannot be disputed that linearity and covariances/correlations, both of which play a central role in PCA, have especial relevance when distributions are multivariate normal, but this does not detract from the usefulness of PCA when data have otherforms”. **Pg 19.**

“Admittedly, correlations or covariances, on which PCs are based, have particular relevance for normal random variables, but they are still valid for discrete variables provided that the possible values of the discrete variables have a genuine interpretation.” **Pg 69.**

**-ours definitely have an interpretation, this is a rating scale**

**Alternative methods to deal with non-(multivariate)-normality:**

“As a complete contrast to the strict assumptions made in most work

on the distributions of PCs, Efron and Tibshirani (1993, Section 7.2) look

at the use of the ‘bootstrap’ in this context. The idea is, for a particular

sample of *n* observations **x**1, **x**2,. . . , **x***n*, to take repeated random samples

of size *n* from the distribution, calculate the PCs for each sample, and build up empirical distributions for PC coefficients and variances.” **Pg 49**

techniques are available for testing the null hypothesis of MVN, but generally the relatively informal graphical methods are best (univariate and bivariate)

also allow checking for outliers

Linearity:

in real data the relationship between two variables may be markedly non-linear ; or the covariances may fail to describe the situation for other reasons, so the resulting PCA may not

give good results

many multivariate techniques assume that the covariance or correlation matrix adequately describes the relationship amongst the variables, i.e. that relationships are **linear**

* Independence between observations- not of huge concern

- by the very nature of the data we have to accept this. Multiple reviews from the same people, bias introduced by only a certain type of people involving in the reviews

“However,when the main objective of PCA is descriptive, not inferential, complicationssuch as non-independence do not seriously affect this objective.” **Pg 299**

* **In clustering we are actually looking for dependence!**
* **Lack of independence over time**
* **-there are alternative multivariate methods that can be used when we can’t assume independence, many of them designed to incorporate time dependence**

“we can use a principal component-like technique, called *singular spectrum analysis* (SSA), to analyse the autocorrelation in the series.” **Pg 302**

* > cor(beer.data.clean[,c(3,6)])
* Time ABV
* Time 1.000000 0.143855
* ABV 0.143855 1.000000

**-given that alcohol content increases with time we have non independent data**

**- this drastic non-normal distribution in time may be a good reason that it was not included in the overall, primary PCA**

* Outliers

the presence of outliers may distort the components

two major ways in which outliers can affect the results :

by distorting the covariances between the variables

by adding spurious dimensions or obscuring the cut off point for choosing components

one of the less important components may be in place solely due to the presence of a single outlying data point

-We do have extreme values in alcohol content but we have seen that these types of beers do have very high alcohol content and so we do not want to get rid of them