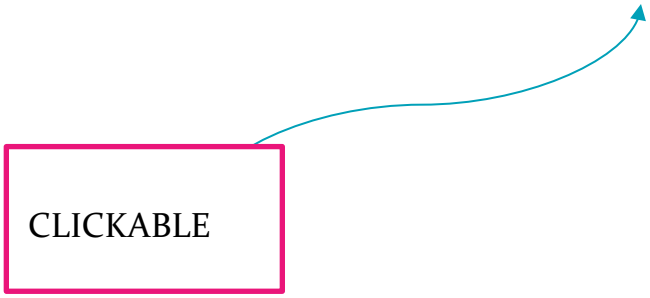
EM 623 - Data Science and Knowledge Discovery

EXERCISE 03

Rushabh Barbhaya | Exercise Report | 2/23/2018

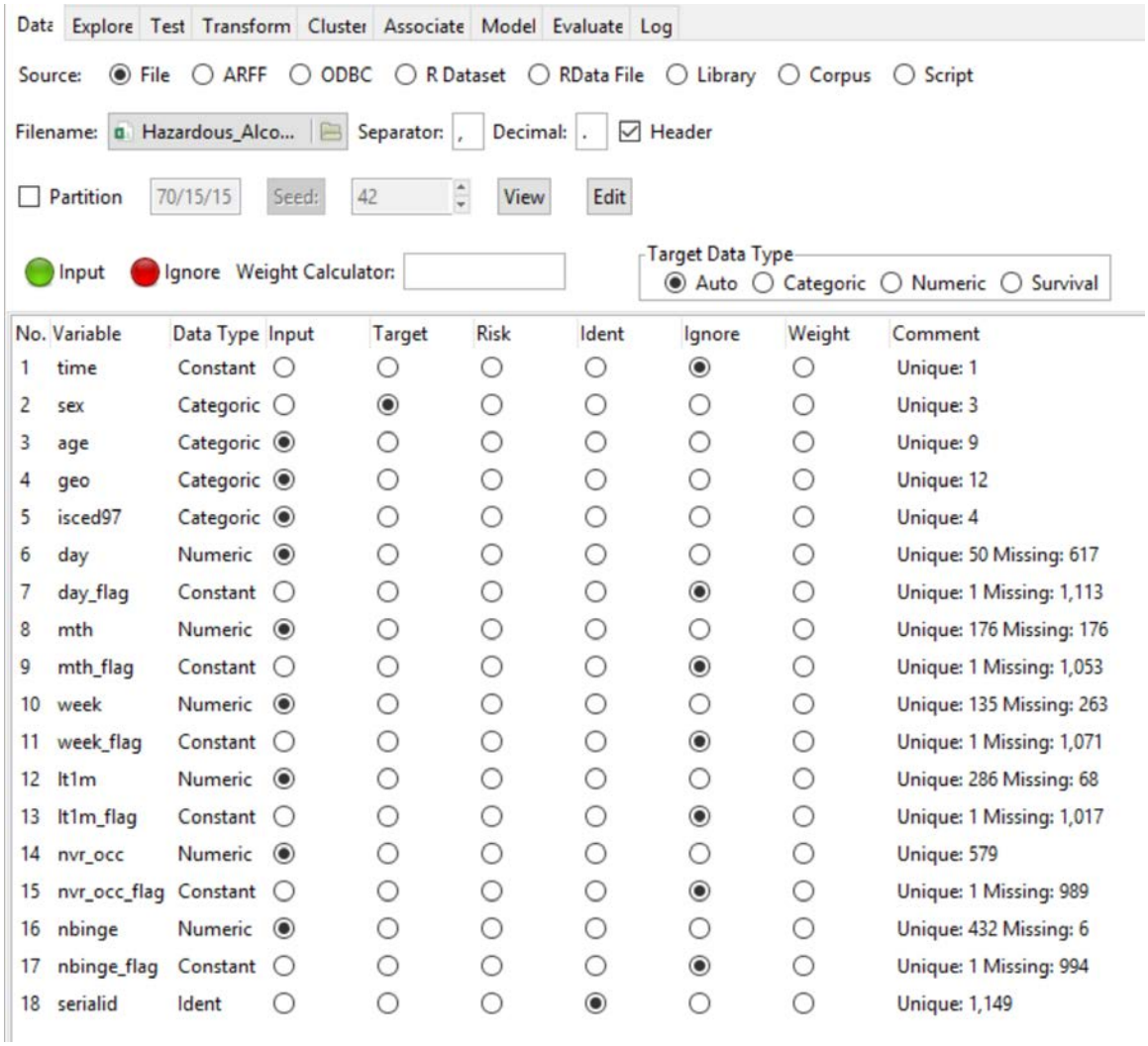
Contents

List of Figures



Missing Values

To calculate the total number of missing values from 'Hazardous Alcohol Consumption.csv', load the file in Rattle and press 'execute'. It will present the data with number of unique values in a particular column header along with missing values.



Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

Source: (•) File ( ) ARFF ( ) ODBC ( ) R Dataset ( ) RData File ( ) Library ( ) Corpus ( ) Script

Filename: Hazardous_Alco... Separator: , Decimal: . [x] Header

[ ] Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator:

Target Data Type: (•) Auto ( ) Categoric ( ) Numeric ( ) Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	time	Constant	○	○	○	○	●	○	Unique: 1
2	sex	Categoric	○	●	○	○	○	○	Unique: 3
3	age	Categoric	●	○	○	○	○	○	Unique: 9
4	geo	Categoric	●	○	○	○	○	○	Unique: 12
5	isced97	Categoric	●	○	○	○	○	○	Unique: 4
6	day	Numeric	●	○	○	○	○	○	Unique: 50 Missing: 617
7	day_flag	Constant	○	○	○	○	●	○	Unique: 1 Missing: 1,113
8	mth	Numeric	●	○	○	○	○	○	Unique: 176 Missing: 176
9	mth_flag	Constant	○	○	○	○	●	○	Unique: 1 Missing: 1,053
10	week	Numeric	●	○	○	○	○	○	Unique: 135 Missing: 263
11	week_flag	Constant	○	○	○	○	●	○	Unique: 1 Missing: 1,071
12	lt1m	Numeric	●	○	○	○	○	○	Unique: 286 Missing: 68
13	lt1m_flag	Constant	○	○	○	○	●	○	Unique: 1 Missing: 1,017
14	nvr_occ	Numeric	●	○	○	○	○	○	Unique: 579
15	nvr_occ_flag	Constant	○	○	○	○	●	○	Unique: 1 Missing: 989
16	nbinge	Numeric	●	○	○	○	○	○	Unique: 432 Missing: 6
17	nbinge_flag	Constant	○	○	○	○	●	○	Unique: 1 Missing: 994
18	serialid	Ident	○	○	○	●	○	○	Unique: 1,149

Figure 1 Missing Values from Rattle

You can add them all up coming to a total of 7367 missing values. This operation can also be performed in excel using the following function:

```
=COUNTBLANK(A1:R1150)
```

Figure 2 Missing Values using Excel

It amounts to the same value.

Metadata file

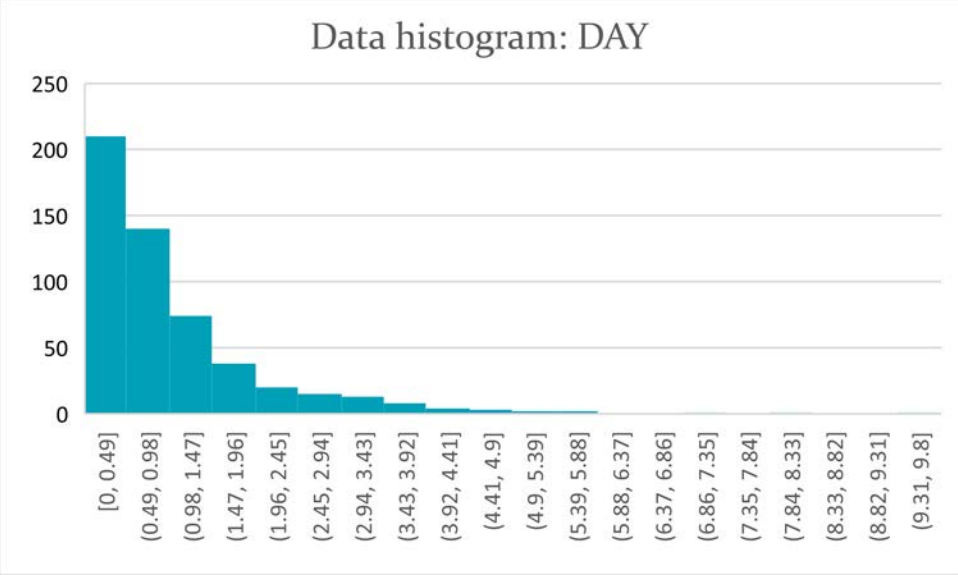
From the file 'Alcohol Consumption-Metadata.csv' we get a little description about the file 'Hazardous_Alcohol_Consumption.csv' we get the whole table name → **Hazardous Alcohol Consumption (Binge Drinking) by Sex Age and Educational Attainment Level**. We also get details about the source of the file, and its datapath.

We also get information about column headers. Like the data was cumulated from January, ISCED97 stands for 'International Standard Classification of Education 1997.

Outliers

In order to determine the outliers I used a histogram for each data point.

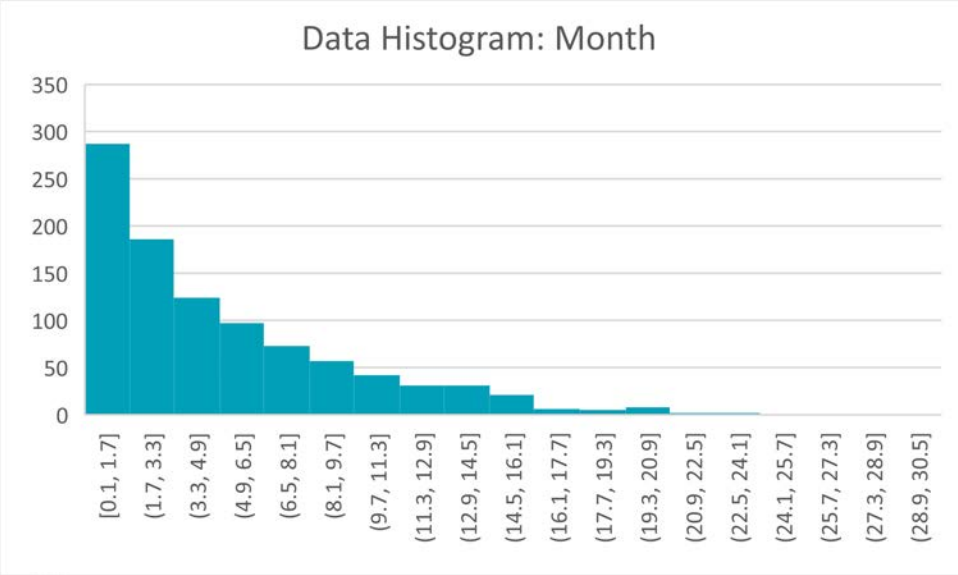
A. Day



Graph 1 Data Histogram: DAY

From the data represented in this histogram the data above the range of 3.92 seem to be outliers. As they fall outside the normal curve of the graph.

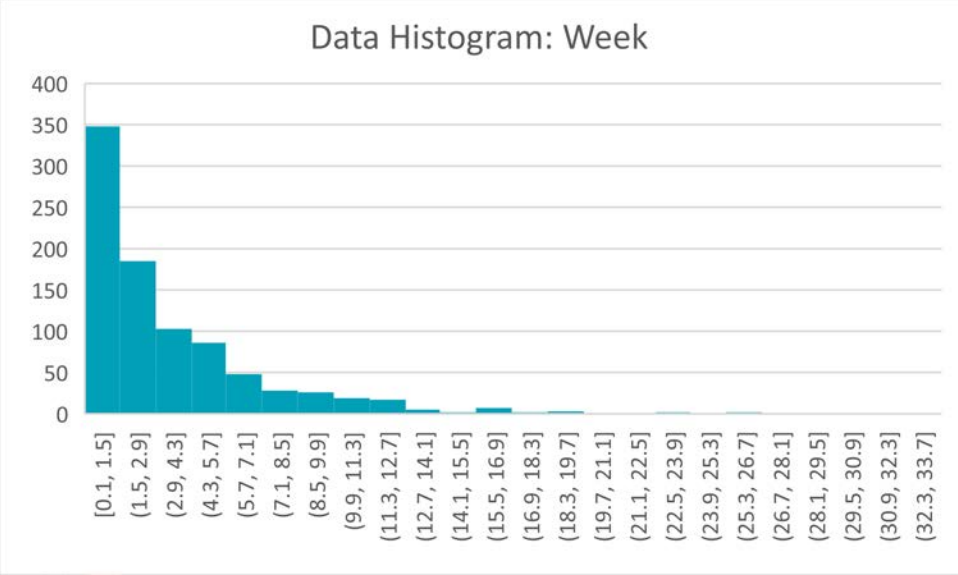
B. Month



Graph 2 Data Histogram: Month

From the data of month represented in the histogram above it can be assumed that the data above 19.3 seem to be outliers. As they fall outside the normal flow of graph.

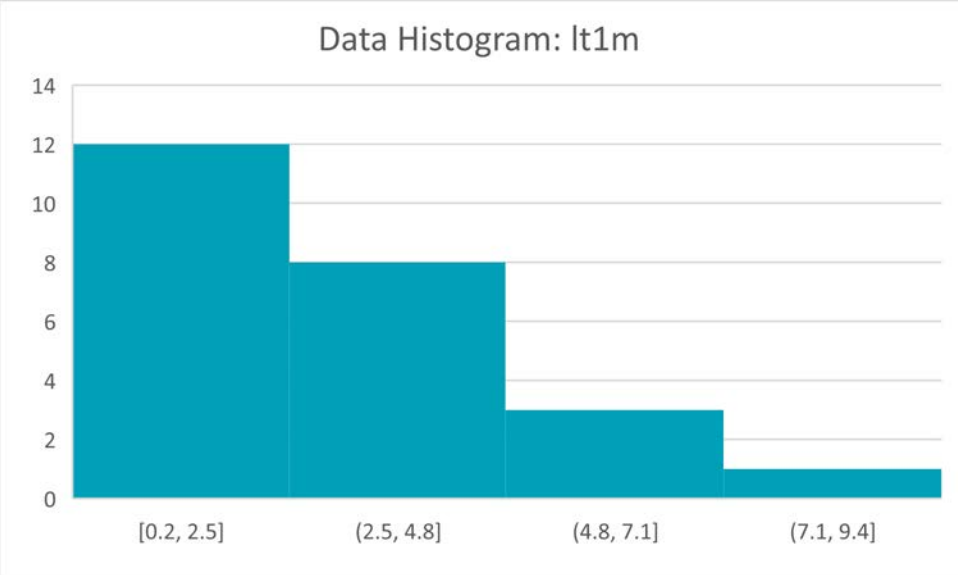
C. Week



Graph 3 Data Histogram: Week

From the data shown above it can be assumed that the data above 15.5 can be considered as outliers as they fall outside the normal curve of the graph.

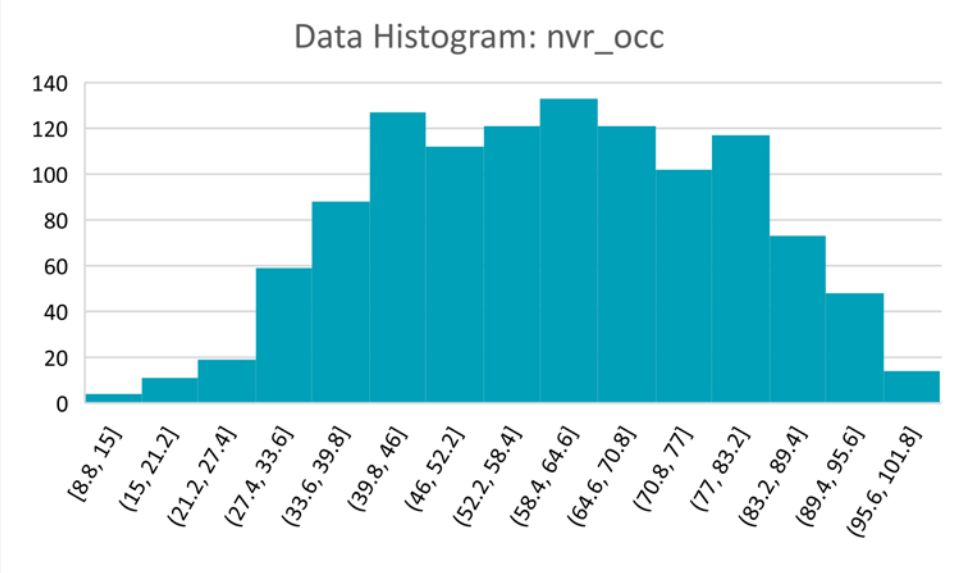
D. Lt1m



Graph 4 Data Histogram: lt1m

From the data in the graph above it can be assumed that there are no outliers as all the data fall under the normal curve of the graph.

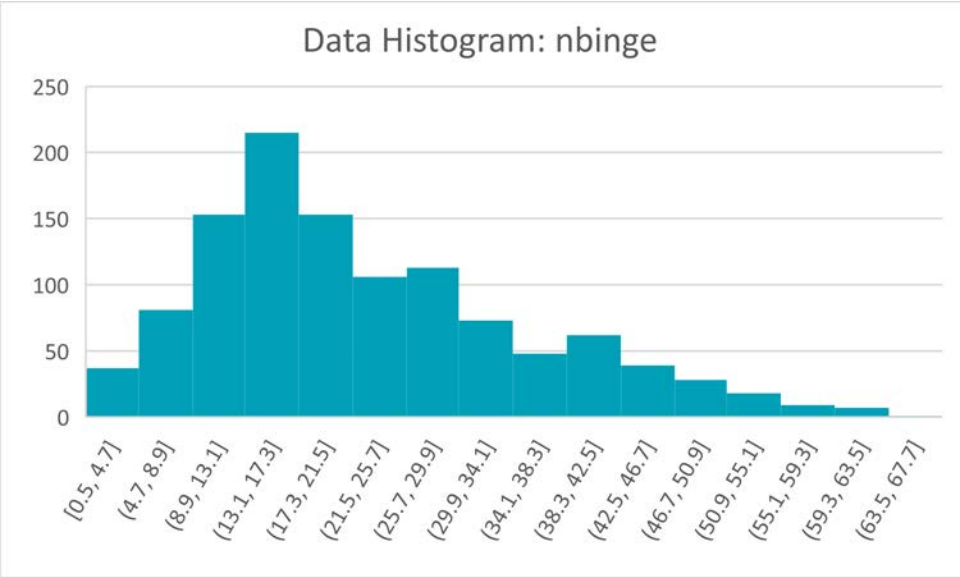
E. nvr_occ



Graph 5 Data Histogram: nvr_occ

From the data in the graph above it can be assumed that there are no outliers as all the data fall under the normal curve of the graph.

F. nbinge



Graph 6 Data Histogram: nbinge

From the data in the graph above it can be assumed that there are no outliers as all the data fall under the normal curve of the graph.

Normalization of nbinge

First load the file in data tab and execute it. Then select the Scale [0-1] option in 'Rescale' type from 'Transform' tab [shown in figure 3]

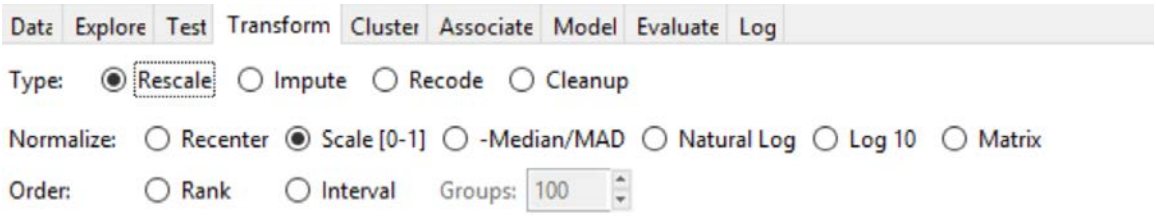


Figure 3 Normalizing option

Then select the column to be rescaled.

14	nvr_occ	Numeric [8.80 to 100.00; unique=579; mean=59.91; median=60.40].
15	nvr_occ_flag	Categorical [1 levels; miss=989; ignored].
16	nbinge	Numeric [0.50 to 63.80; unique=432; mean=22.68; median=19.50; miss=6; ignored].
17	nbinge_flag	Categorical [1 levels; miss=994; ignored].
18	serialid	Numeric [1 to 1149; unique=1149; mean=575; median=575].

Figure 4 Column selection

After pressing execute a new column will be created with a suffix of 'R01_' followed by the name of the column selected for rescaling. After that select view option from data tab and we'll be able to see the normalized table. [shown in figure 5]

R01_nbinge
0.3917852
0.4581359
0.5766193
0.4739336
0.1327014
0.2496051
0.3112164
0.2227488
0.1769352
0.3048973

Figure 5 Normalized Values

Relation Matrix

Rattle cannot display sex and age parameters as they are not numeric values. To convert those to numeric values. To do that we must select ‘as numeric’ under ‘recode’ type/option from ‘transform’ tab and select the rows we need to transform.

Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

Type: ○ Rescale ○ Impute ◉ Recode ○ Cleanup

Binning: ○ Quantiles ○ KMeans ○ Equal Width Number: 4

○ Indicator Variable ○ Join Categorics ○ As Categoric ◉ As Numeric

No.	Variable	Data Type and Number Missing
1	time	Numeric [2008 to 2008; unique=1; mean=2008; median=2008; ignored].
2	sex	Categorical [3 levels; ignored].
3	age	Categorical [9 levels; ignored].
4	geo	Categorical [12 levels].
5	isced97	Categorical [4 levels].

Figure 6 Transformation options

After that we can plot the ‘correlation’ plot from ‘explore’ tab. It will show something like this.

Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log

Type: ○ Summary ○ Distributions ◉ Correlation ○ Principal Components ○ Interactive

☑ Ordered ☐ Explore Missing ☐ Hierarchical Method: Pearson

```
Correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are reported.

            nvr_occ      TNM_age      nbinge      TNM_sex        ltlm
nvr_occ  1.00000000 -0.015526992 -0.62143694 -0.288436225 -0.74360529
TNM_age -0.01552699  1.000000000  0.16443477 -0.003158199 -0.14107686
nbinge  -0.62143694  0.164434767  1.00000000  0.126269614  0.01262838
TNM_sex -0.28843623 -0.003158199  0.12626961  1.000000000  0.20416157
ltlm    -0.74360529 -0.141076858  0.01262838  0.204161569  1.00000000
mth     -0.53216265 -0.179324152 -0.27346515  0.143747499  0.69961078
week    -0.47055820 -0.184950159 -0.20733080  0.119188282  0.46465490
day     -0.43187450 -0.078143124 -0.07260427 -0.032925814  0.36907996
               mth       week         day
nvr_occ -0.5321626 -0.4705582 -0.43187450
TNM_age -0.1793242 -0.1849502 -0.07814312
nbinge  -0.2734652 -0.2073308 -0.07260427
TNM_sex  0.1437475  0.1191883 -0.03292581
ltlm     0.6996108  0.4646549  0.36907996
mth      1.0000000  0.6560166  0.39945086
week     0.6560166  1.0000000  0.54316363
day      0.3994509  0.5431636  1.00000000

Rattle timestamp: 2018-02-23 19:16:32 rhbar
======================================================================
```

Figure 7 Correlation

To save the plot just open to Rstudio's plot workgroup. The matrix plot can be saved from that option.

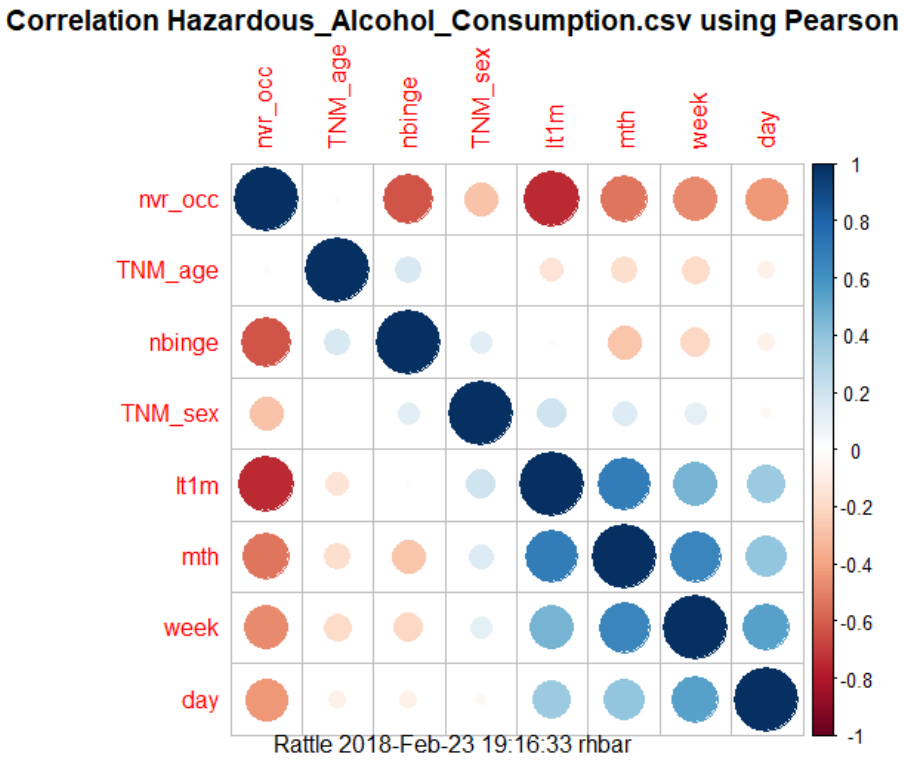


Figure 8 Relation Matrix

It is clear that the 'age' and 'sex' have no correlation at all. But 'age' is so slightly inversely correlated to 'lt1m', 'month', 'week' and 'day' and slightly directly correlated with 'nbinge'.

Similarly 'sex' is directly correlated with 'nbinge', 'lt1m', 'month' and 'week' but inversely correlated with 'nvr_occ' ever so slightly.