EM 623 - Data Science and Knowledge Discovery

Exercise 03

Rushabh Barbhaya | Exercise Report | 2/23/2018

Contents

[Missing Values 2](#_Toc507177954)

[Metadata file 3](#_Toc507177955)

[Outliers 3](#_Toc507177956)

[Normalization of nbinge 6](#_Toc507177957)

[Relation Matrix 8](#_Toc507177958)

List of Figures

[Figure 1 Missing Values from Rattle 2](#_Toc507178173)

[Figure 2 Missing Values using Excel 2](#_Toc507178174)

[Figure 3 Normalizing option 6](#_Toc507178175)

[Figure 4 Column selection 7](#_Toc507178176)

[Figure 5 Normalized Values 7](#_Toc507178177)

[Figure 6 Transformation options 8](#_Toc507178178)

[Figure 7 Correlation 8](#_Toc507178179)

[Figure 8 Relation Matrix 9](#_Toc507178180)

CLICKABLE

# 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.

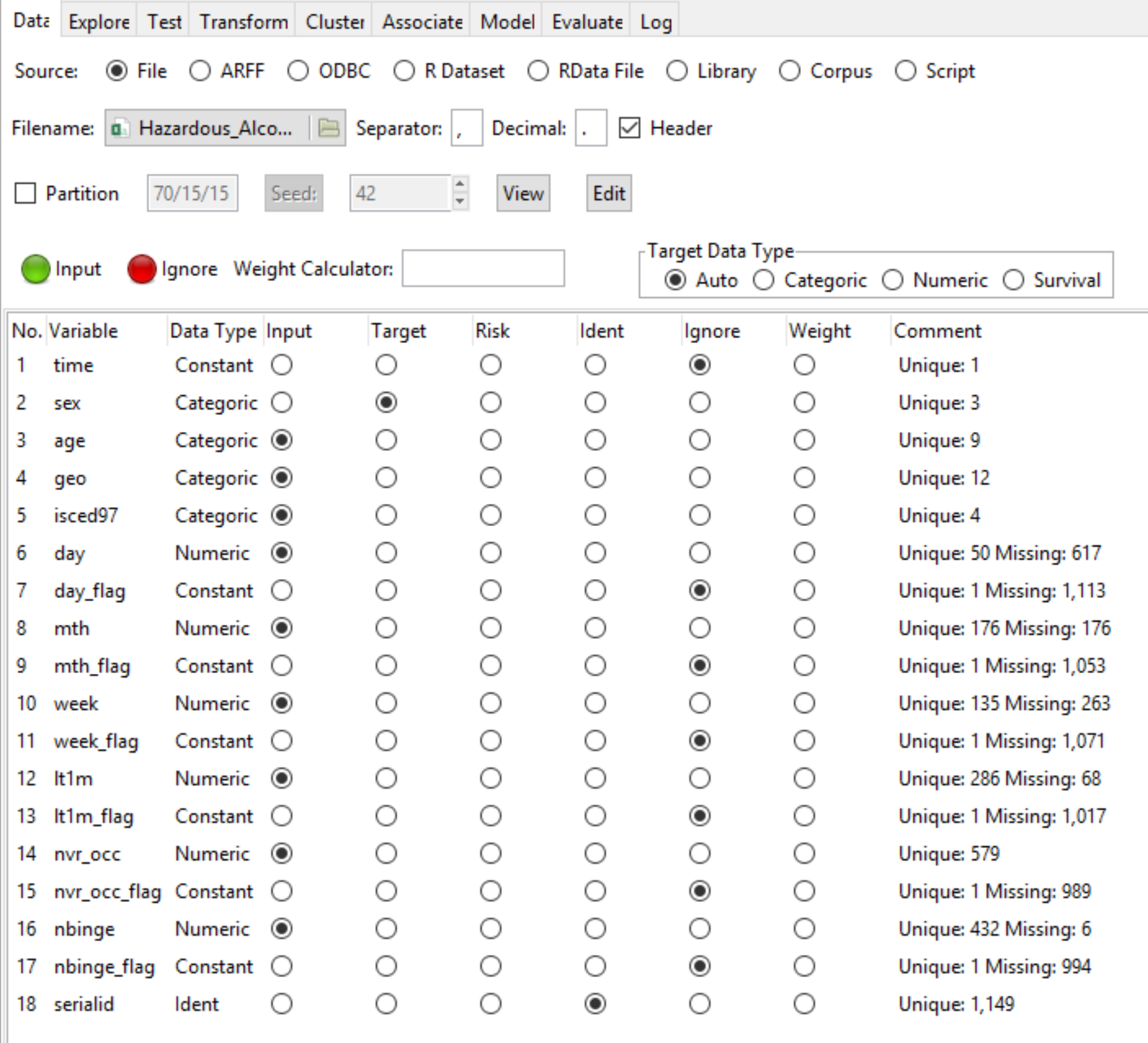


Figure 1 Missing Values from Rattle

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:



Figure 2 Missing Values using Excel

It amounts to the same value.

# Metadata file

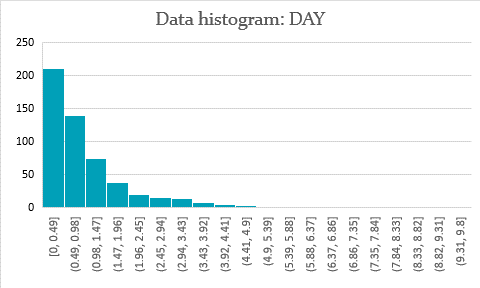
From the file ‘Alcohol Consumption-Metadata.csv’ we get a little description about the file ‘Hazardous\_Alcohol\_Consumption.csv’ we get the whole table name 🡪 **Hazardous Alcohol Consumption (Binge Drinking) by Sex Age and Educational Attainment Level**. We also get details about the source of the file, 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.

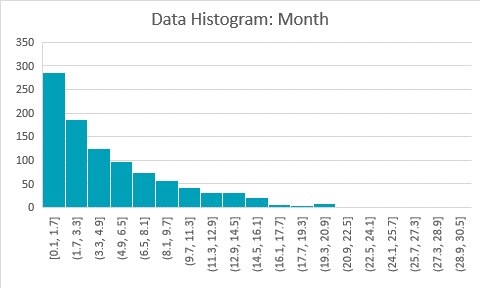
1. **Day**



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.

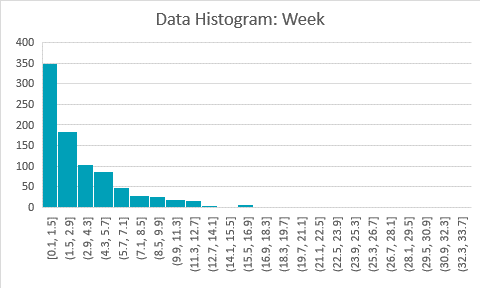
1. **Month**



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.

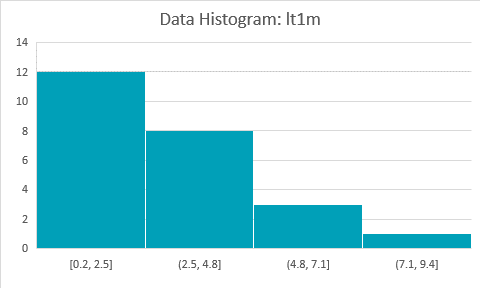
1. **Week**



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.

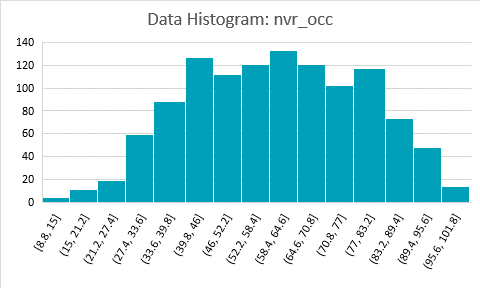
1. **Lt1m**



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.

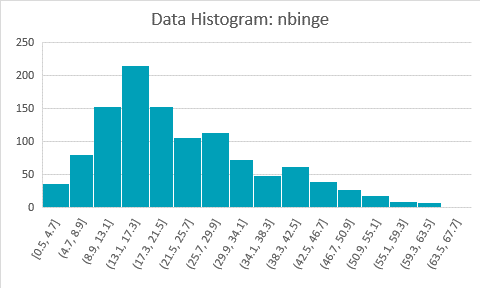
1. **nvr\_occ**



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.

1. **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]

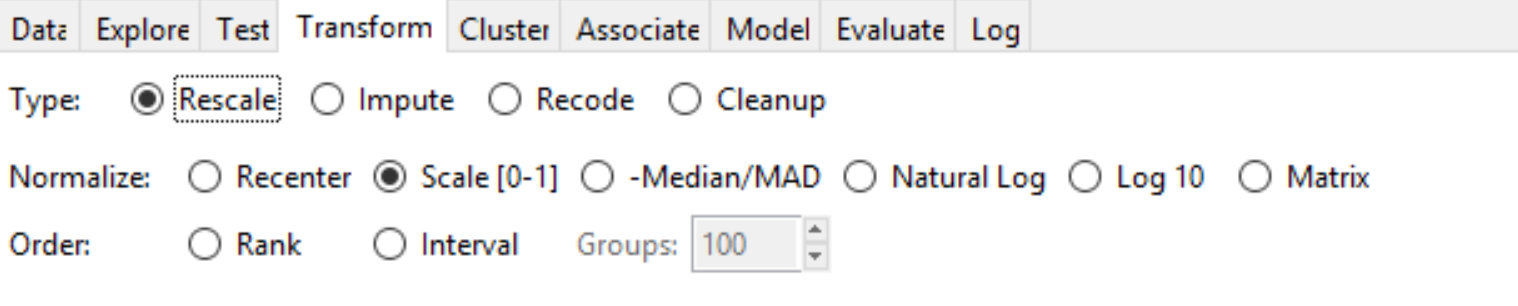


Figure 3 Normalizing option

Then select the column to be rescaled.

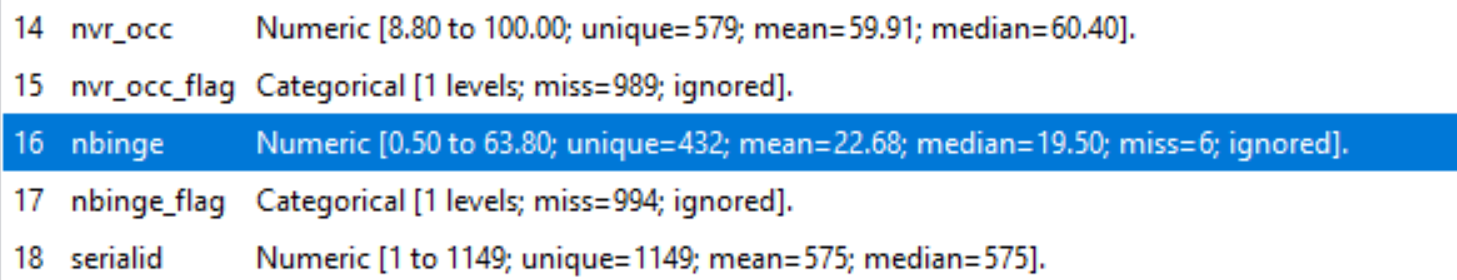


Figure 4 Column selection

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]

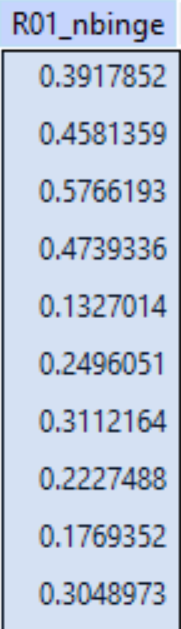


Figure 5 Normalized Values

# 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.

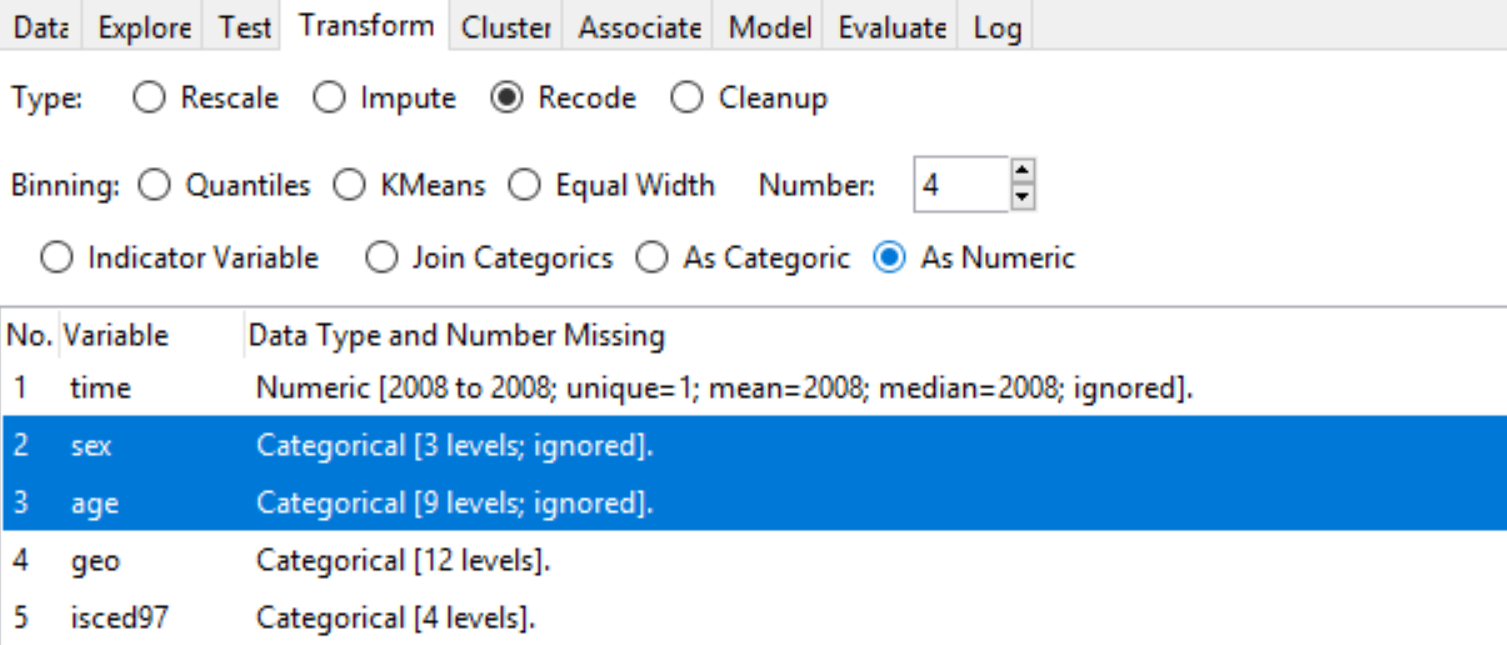


Figure 6 Transformation options

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

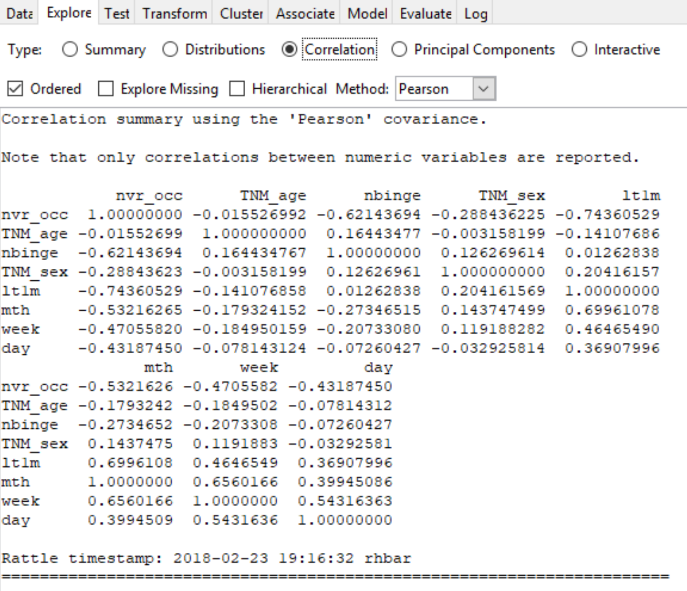


Figure 7 Correlation

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

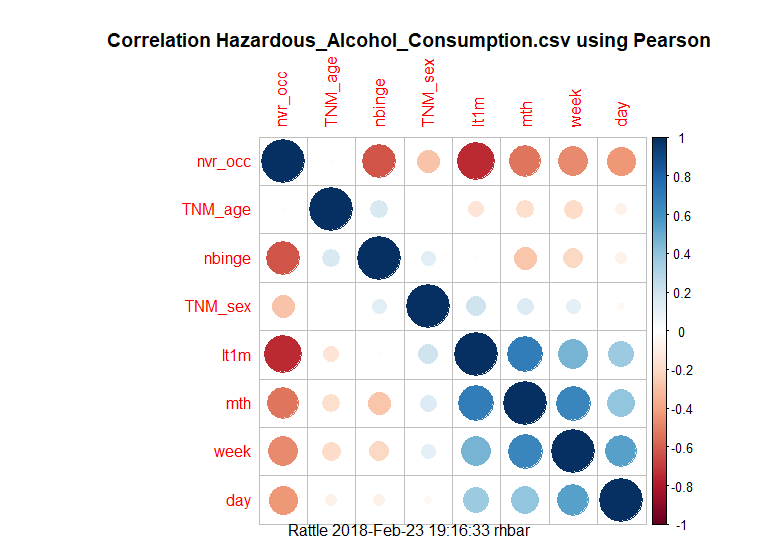


Figure 8 Relation Matrix

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.