Project Report
Chiorean I. and Golovco S.
Programming for bioinformatics

TEDAR validation project

The first step we had to make, was analyzing the given data from TEDAR algorithm and comparing them with the data of the other three static algorithms: *mensile, trimestrale, statico*. The datasets *notorietà 2018* and n*otorietà 2019* were used for the validation of the algorithms. In order to deal with it, we merged the given datasets to obtain the unique values and then we produced the following data table.

	Notorietà 2018			
	Chi3	**Chi5**	**CI3**	**CI5**
Mensile #elementi prima	*5456*	*3067*	*6064*	*3410*
	11493	5890	12673	6532
Trimestrale #elementi prima	*5136*	*3284*	*6115*	*3916*
	11050	6572	12999	7784
Statico #elementi prima	*3984*	*2963*	*5339*	*3996*
	8776	6087	11485	8119
TEDAR #elementi prima	*5362*	*3596*	*6429*	*4302*
	11591	7308	13698	8723

#elementi notorietà 2018 = 26847

	Notorietà 2019			
	Chi3	**Chi5**	**CI3**	**CI5**
Mensile #elementi prima	*5512*	*3094*	*6134*	*3442*
	11493	5890	12673	6532
Trimestrale #elementi prima	*5194*	*3315*	*6174*	*3942*
	11050	6572	12999	7784
Statico #elementi prima	*4042*	*3005*	*5400*	*4042*
	8776	6087	11485	8119
TEDAR #elementi prima	*5424*	*3633*	*6485*	*4338*
	11591	7308	13698	8723

#elementi notorietà 2019 = 27926

Here *CI* means confidence interval and *Chi* refers to the Chi Square test. The corresponding numbers are the thresholds used for the tests. Each cell of the table contains the number of known signals of the algorithm identified by applying the threshold.

The second step was producing the related Venn diagrams and plots to better understand and visualize the results. (More detailed pictures can be found in the GitHub repository together with the scripts.)

First of all, precision and sensitivity had to be calculated before producing the diagrams. The two formulas are the following: ***sensitivity= tp/(tp+fn)*** and ***precision=tp/(tp+fp)*** where tp=true positives, fn=false negatives, fp=false positive.

The way we obtained true positives, false negatives and false positives is described in the given script.

Sensitivity and Precision for TEDAR

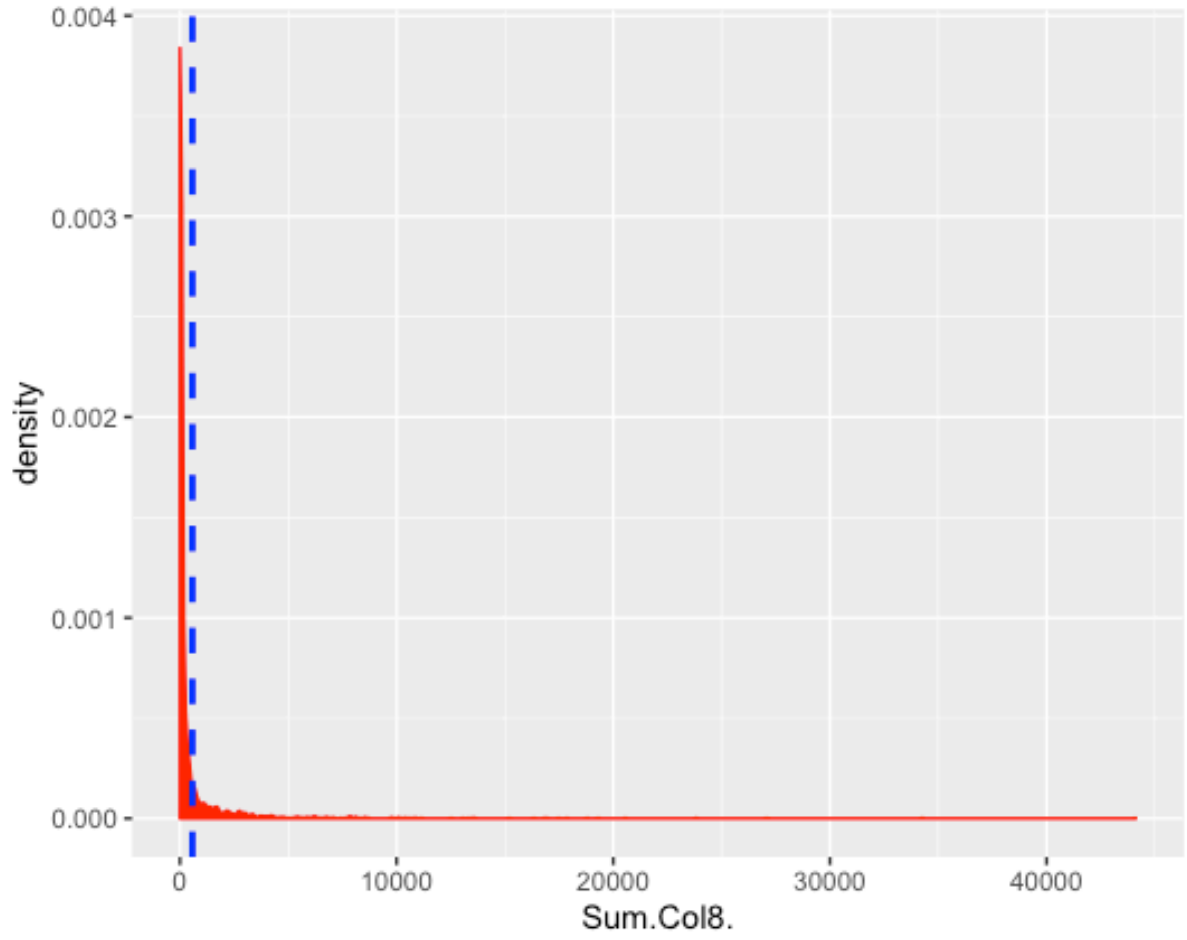
	Notorietà 2018			
	Chi3	**Chi5**	**CI3**	**CI5**
True positive	*5362*	*3596*	*6429*	*4302*
False negative	*21485*	*23251*	*20418*	*22545*
False positive	*6229*	*3712*	*7269*	*4421*
Sensitivity	*0.1997244*	*0.1339442*	*0.2394681*	*0.1602414*
Precision	*0.4626003*	*0.4920635*	*0.4693386*	*0.493179*

	Notorietà 2019			
	Chi3	**Chi5**	**CI3**	**CI5**
True positive	*5424*	*3633*	*6485*	*4338*
False negative	*22502*	*24293*	*21441*	*23588*
False positive	*6167*	*3675*	*7213*	*4385*
Sensitivity	*0.1942276*	*0.1300938*	*0.2322209*	*0.1553391*
Precision	*0.4679493*	*0.4971264*	*0.4734268*	*0.497306*

Density plots

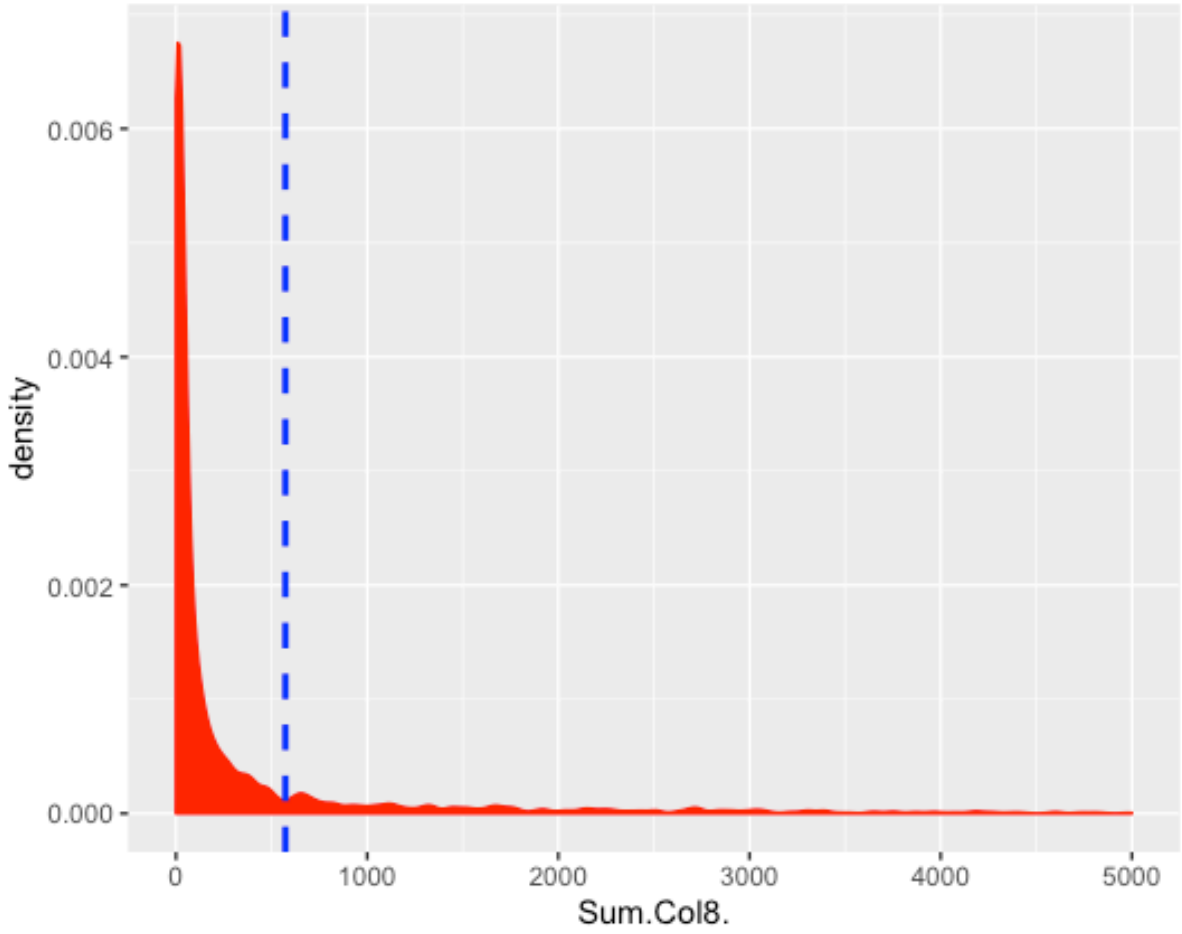
Another task was producing the density plots for the drugs and for the couples drug-ADR given other input datasets: *drugfreq.csv* and *drugsCoupled.csv*. The latter was produced as well as the first one by keeping only the first two columns of the given csv file.

These are the obtained results:



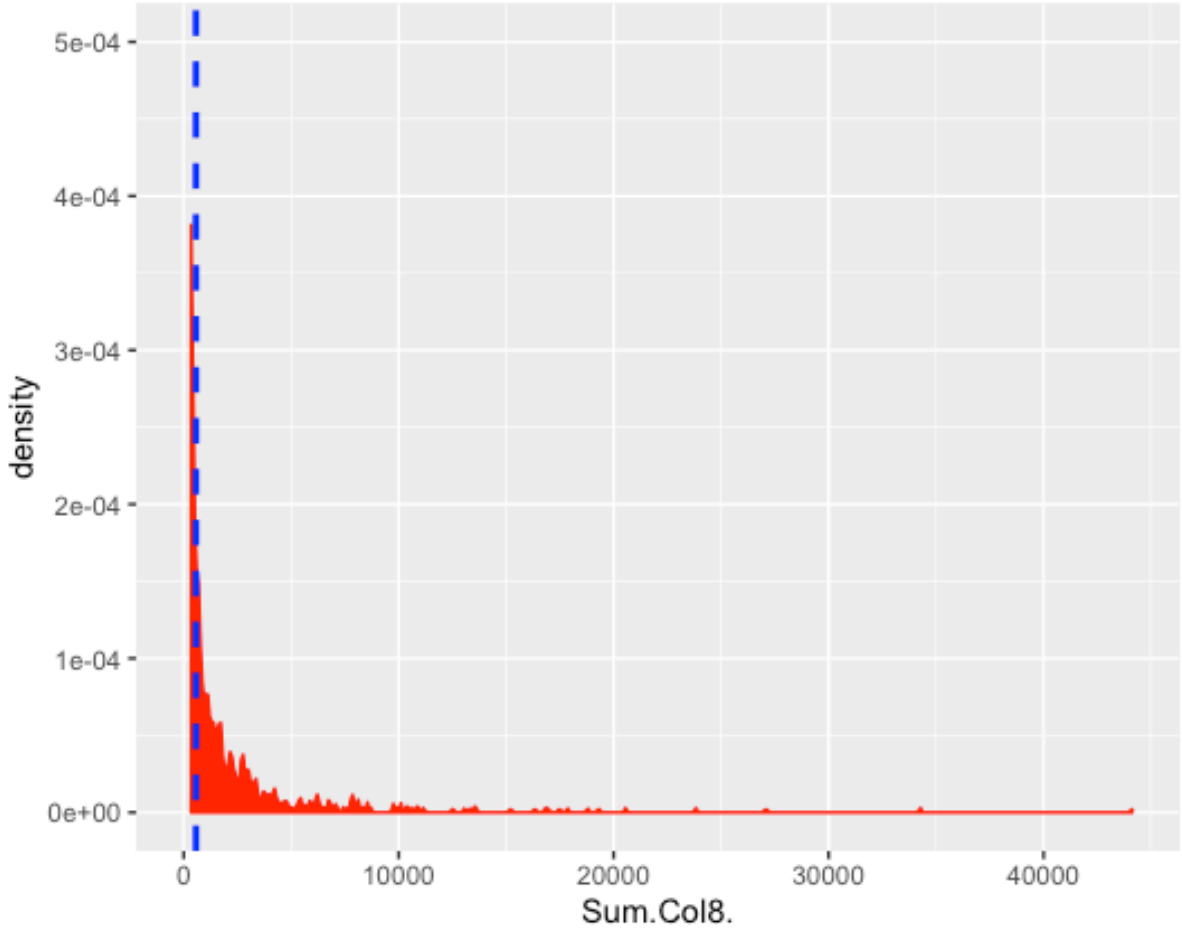
Density per drug with the dashed line representing the mean

Sum.Col8. refers to the frequency of each drug. This holds for all the other following plots.



Density per drug with xlim and the dashed line representing the mean

Here x axis is limited between the range 0 and 5000.

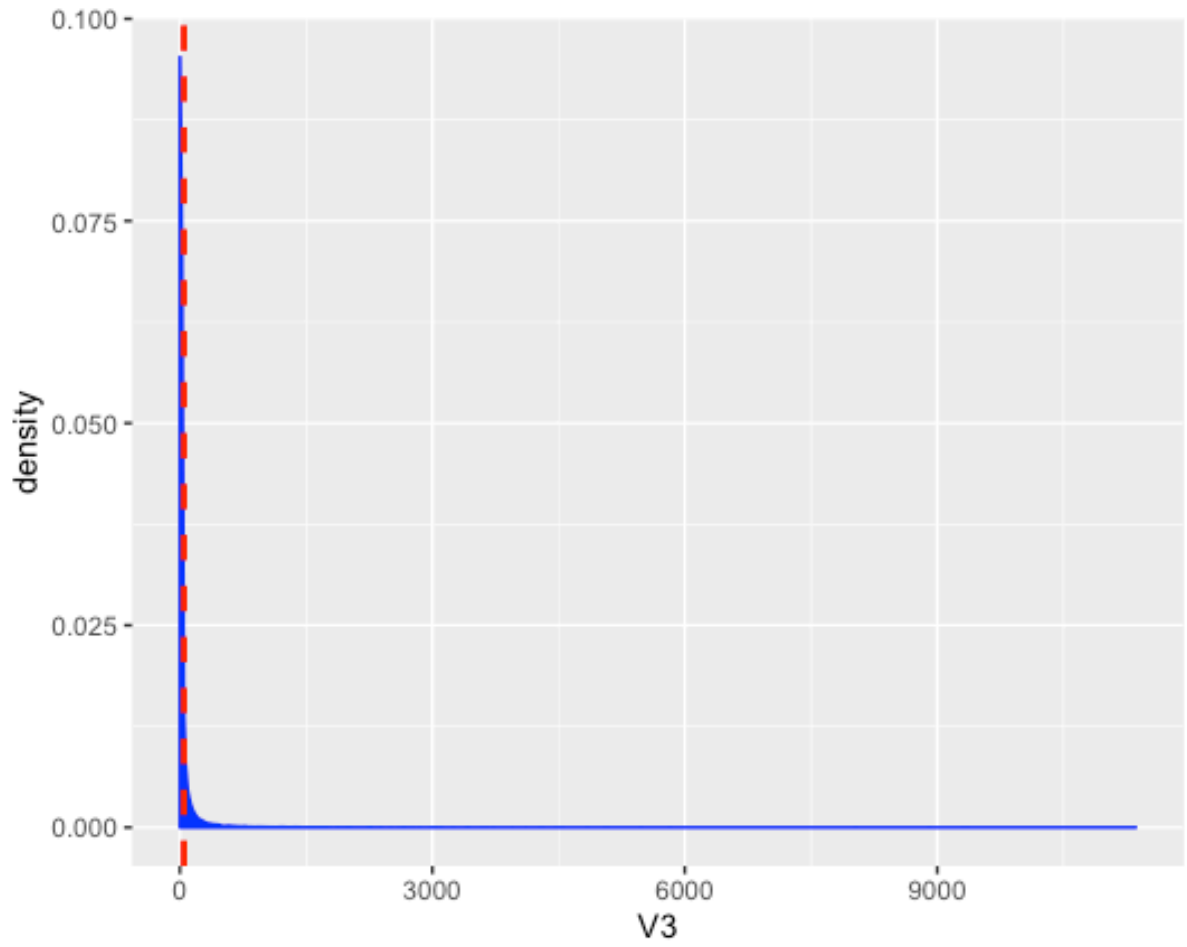


Density per drug with ylim and the dashed line representing the mean

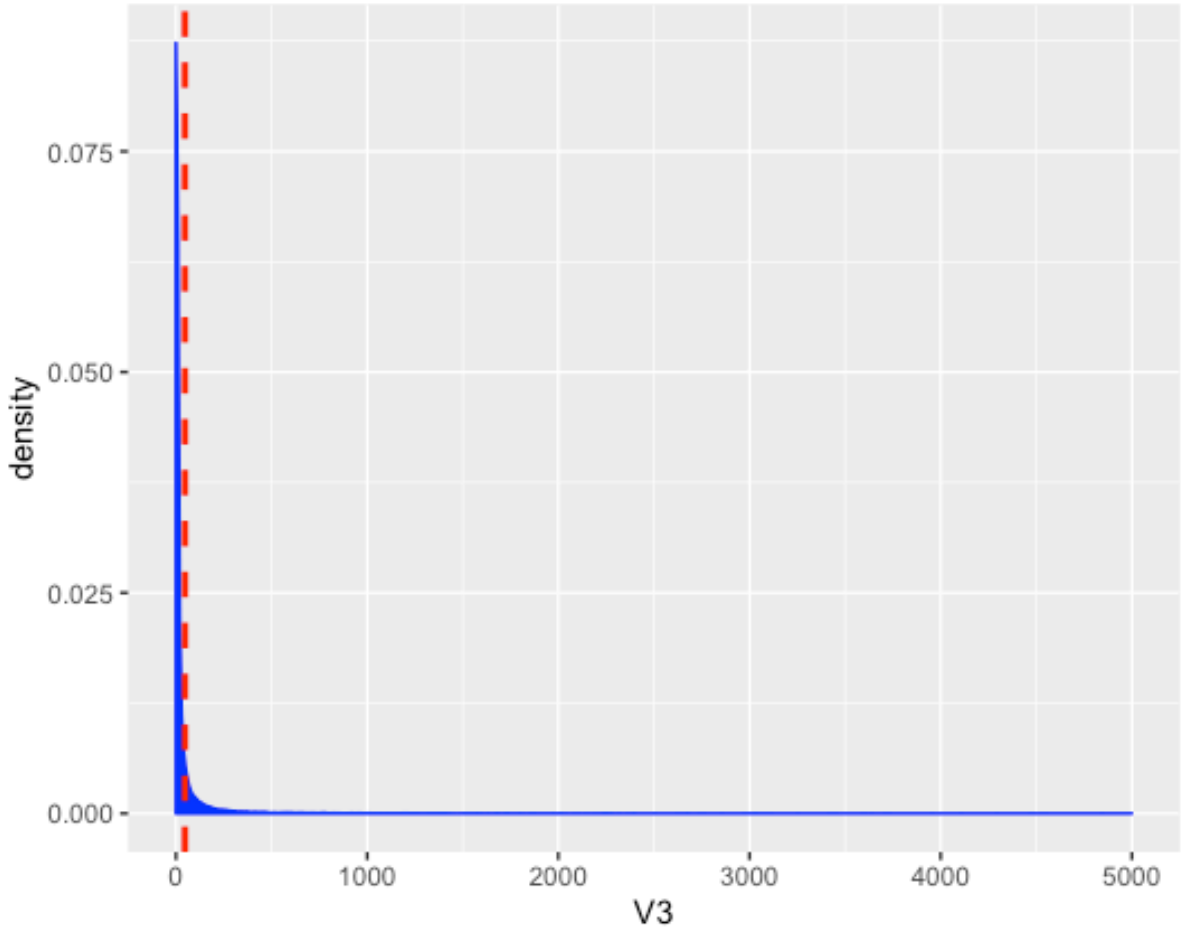
Here y axis is limited between the range 0.000 and 0.0005.

The same work was made also for the couple drug-ADR, as already said. So here there the results:

V3 refers to the frequency of each couple drug-ADR. This holds for all the other following plots.

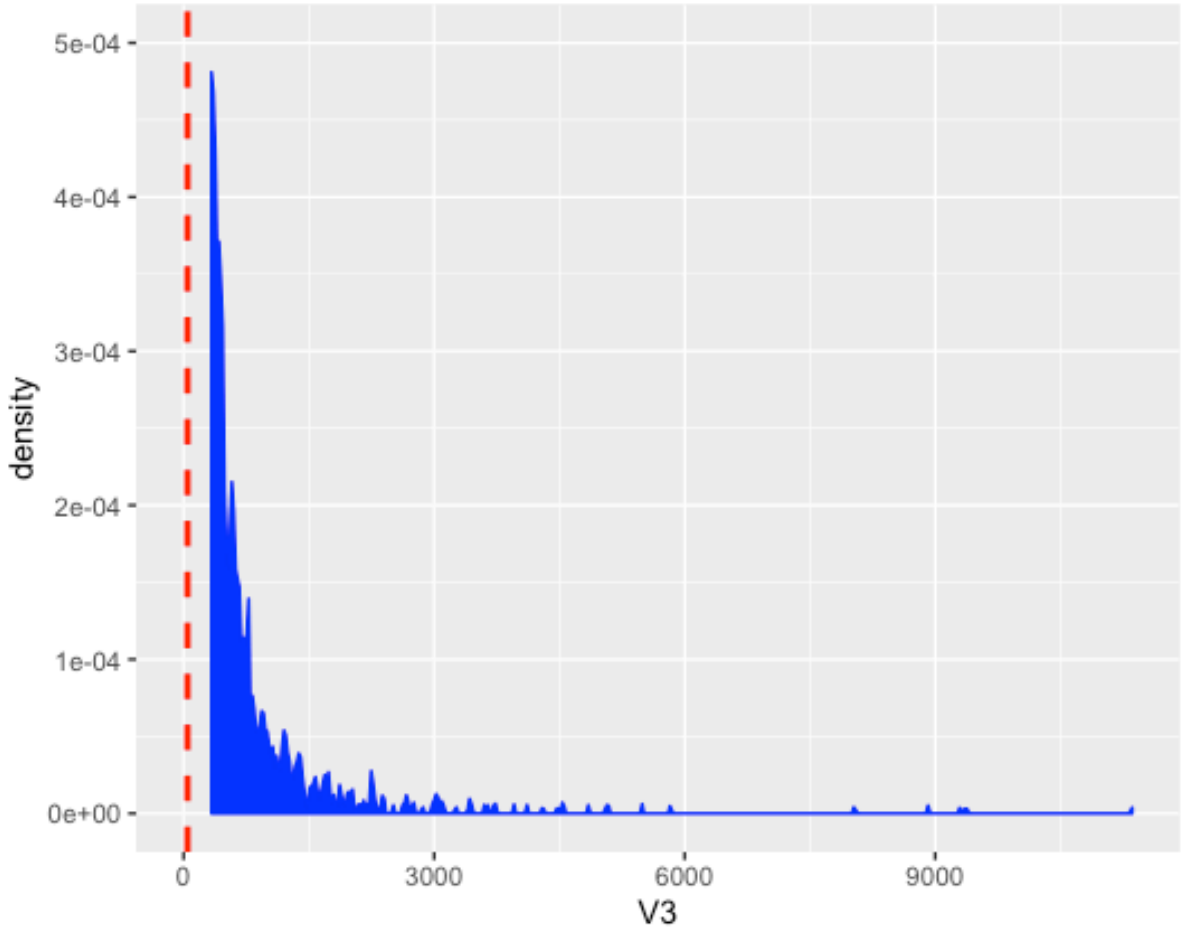


Density per couple with the dashed line representing the mean.



Density per couple with xlim and the dashed line representing the mean

Here x axis is limited between the range 0 and 5000.



Density per drug with ylim and the dashed line representing the mean

Here y axis is limited between the range 0.000 and 0.0005.