Analysis of the Clinical Data of the Universal Diagnostics experiments

Antonio Adsuar

The current document exposes the results of the different analysis made to the data provided by Universal Diagnostics, and gives some conclusions about such data.

04/07/2014

Contenido

Data preparation

First of all, we needed to preprocess data to perform any study.

We got the following problems to solve:

a) Both data files weren't compatible between them at two points:
    a. The description of each sample was differently done at each file; in the raw data file, each sample had an ID described as follows: UDXYYY_neg.mzdata, being YYY the number of the sample. In the description, each sample was described as follows: UDXYYY, being YYY the number of the simple.
    b. In the description of each sample, each row of the file was a new sample; however, in the raw data file, each column of the file was a new simple.
b) In the description of each sample file, some data needed to be normalized. For example, the cancer stage wasn't normalized (it was case sensitive, there where different ways of specifying the **Stage IIIA**).

So the first part of the analysis of the clinical data was to preprocess the different files to get an homogenous file.

The description file of the clinical data had 195 samples with 8 features, including the ID.

The raw data file had 194 samples with 5514 features (including the ID too).

As a result of the preprocessing and the merging of data into one single file, we got a new data file with 193 samples and 5521 features.

The number of features can be calculated with an obvious formula:

$$description_{features} + rawdata_{features} - 1$$

In other words, the sum of both features minus the common one (that is, ID).

However, the number of samples is not as simple. The reason is that the raw data file didn't had any information about the sample UDX6. On the

other hand, the description file didn’t had any information about the samples UDX25 and UDX286. Both samples were removed in order to have samples with every feature. That’s the reason why the common list of samples had only 193 of them (instead of the 196 that could be expected).

Discrimination between the different values of the feature Label

Analysis of Sensibility, Specificity, PPV, NPV and AUC

The data was splitted in two groups:

- Train data set: the 70% of the samples (136 samples)
- Test data set: the 30% of the samples (57 samples)

Once such splitting was executed, we trained a classifier with the train data set. The classifier I chose was the SVM because of its robustness and quality of performance.

The result of the prediction of the test data set allowed us to generate a confusion matrix from which we could get the following factors:

		Reference		
		Class 0	Class 1	Class 2
Prediction	Class 0	14	0	0
	Class 1	6	30	3
	Class 2	0	0	4

	Class 0	Class 1	Class 2
Sensitivity	0.7000	1.0000	0.5714
Specificity	1.0000	0.6667	1.0000
Positive Predictive Value	1.0000	0.7692	1.0000
Negative Predictive Value	0.8605	1.0000	0.9434

And an **accuracy** of the prediction model: **0.8421053**.

This study shows a very good sensitivity value (1.0000) for Class 1 (and thus negative predictive value), and very good specificity value (1.0000) for both classes 0 y 2 (and so it is the positive predictive value). In summary, the accuracy of the prediction model is not very good (some studies consider as a good prediction model accuracies >0.85, and even >0.90).

These data are of interest when we're trying to work in a screening test. In this sense, a screening test should be high sensitive (although it has low specificity), in order to avoid false negative patients that could lead to a late diagnosis of the disease. Once we've selected these patients (which would be true and false positive), we would apply a test with a high specificity (which, on the far, would be 1), to select patients to whom apply correct treatment.

ROC curve

- Area under curve (AUC) for the prediction model of this variable is 0.9615. Nevertheless, I haven't been able to generate the chart for this curve.
- Otherwise, if we consider the variable as a 2 categories one (which would be, "0" and "no 0"), then the AUC for the prediction model would be 0.9231 and the graph would be as follows:

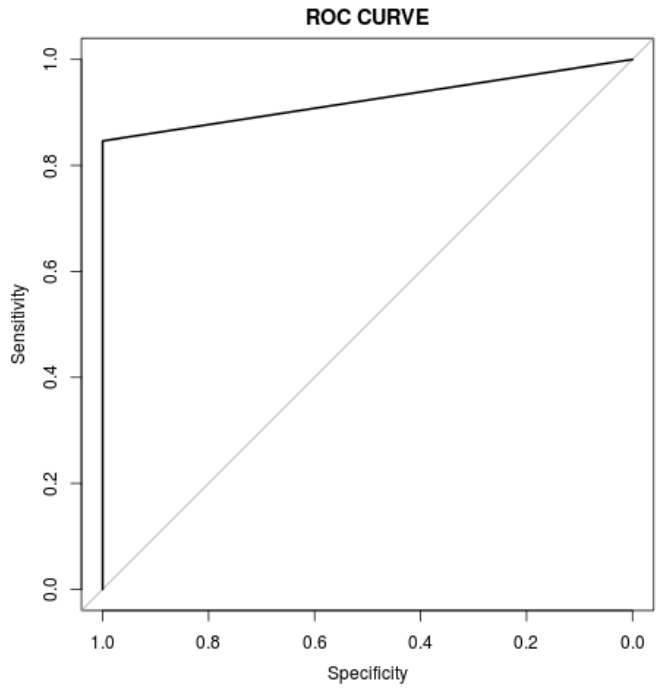

Figure 1. ROC for the variable **Label**

This is a very good AUC (values > 0.75 are considered as significant ones).

RANGE	INTERPRETATION
0.9 < AUC < 1.0	Excellent
0.8 < AUC < 0.9	Good
0.7 < AUC < 0.8	Worthless
0.6 < AUC < 0.7	Not Good
0.5 < AUC < 0.6	Failed

Analyzing accuracy for different amounts of measurements

	NUMBER OF FEATURES			
	5	10	20	50
ACCURACY	0.71930	0.77193	0.89474	0.89474

As we can see in the table above, accuracy improves as the number of features increases. Nonetheless, it's important to note that the accuracy obtained with 20 and 50 features is even better than the one got with the full number of features (that is, 0.8421 versus 0.8947). The reason for this is the selection of the features involved in the procedure. As the reference is more defined by selected features (which is the situation, because we're selecting the best 5 to 50 features in terms of the **Label** feature), accuray will be better.

Discrimination between the different values of the features Gender, Age, Histology and Stage

Gender

As we did before, we create the cross tabulation in order to show the data.

		Reference	
		Male	**Female**
Prediction	**Male**	29	21
	Female	1	6

Sensitivity	0.9667
Specificity	0.2222
Positive Predictive Value	0.5800
Negative Predictive Value	0.8571

Finally, it has an accuracy of **0.614**.

As we can see from data above, considering gender male as the reference value, this test has a high sensitivity to detect them (>95%), with a negative predictive value >85%. Nonetheless, the specificity remains poor, what means that the test isn't able to identify subjects without the condition (non-male subjects).

It is important to remember that either sensitivity and specificity are related to the test. On the other hand, positive and negative predictive values also depend on the prevalence of every feature of the variable.

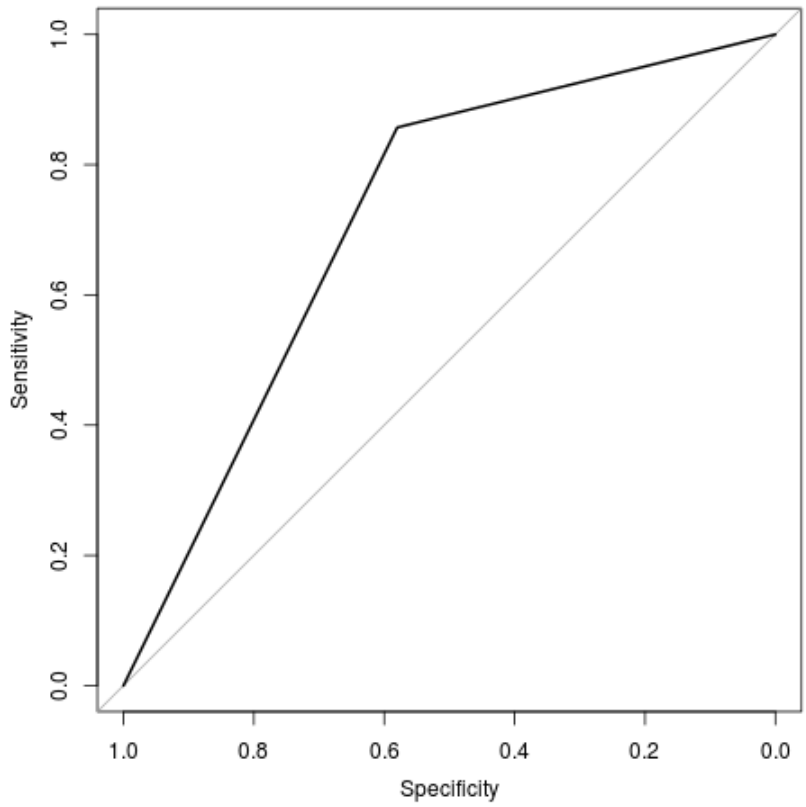

Figure 2. ROC for the variable **Gender**

Area under curve: 0.7186

In order to this AUC value, we can state that this test is **worthless**. In other words, it won't help us at all.

Age

Because age is a continuous variable, is not possible to get the values of sensitivity, specificity and so, due to the wide range of data. In order to solve this problem, we need to categorize this variable. We will use the median (calculated as 63) to create 2 categories: ≤63 (class 0) and >63 (class 1).

	AGE (YEARS)
MEDIAN	63
MINIMUM	23
MAXIMUM	89
INTERQUARTILERANGE	21

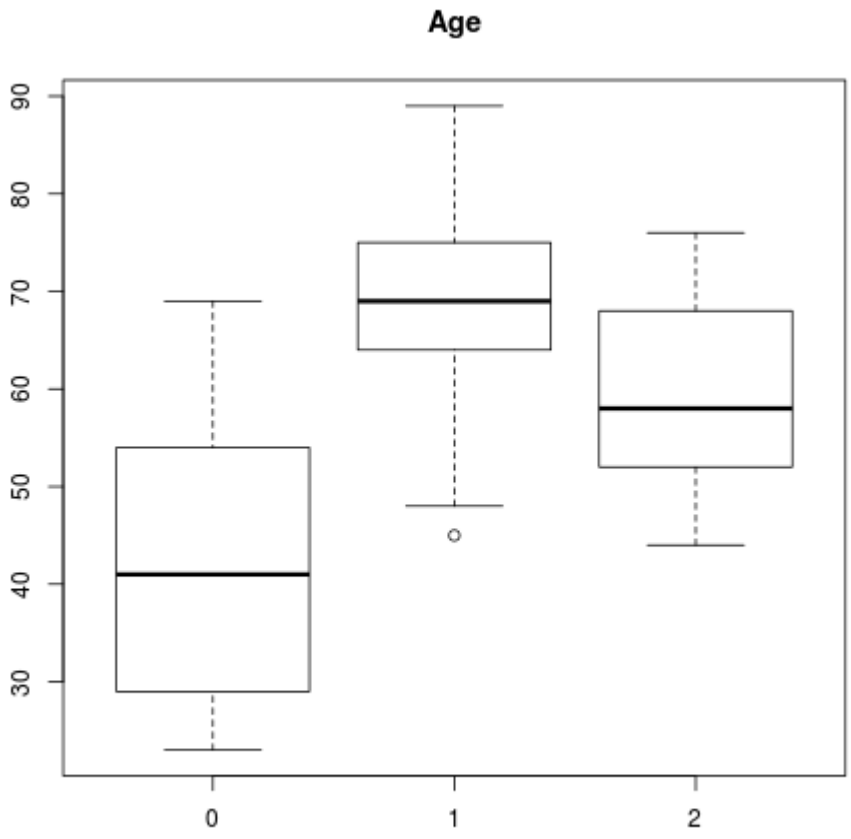

Figure 3. Distribution of age for the whole set of samples for the variable **Label**.

		Reference	
		Age ≤ 63	**Age > 63**
Prediction	**Age ≤ 63**	19	2
	Age > 63	10	26

Sensitivity	0.6552
Specificity	0.9286
Positive Predictive Value	0.9048
Negative Predictive Value	0.7222

Finally, it has an accuracy of **0.7895**.

ROC curve

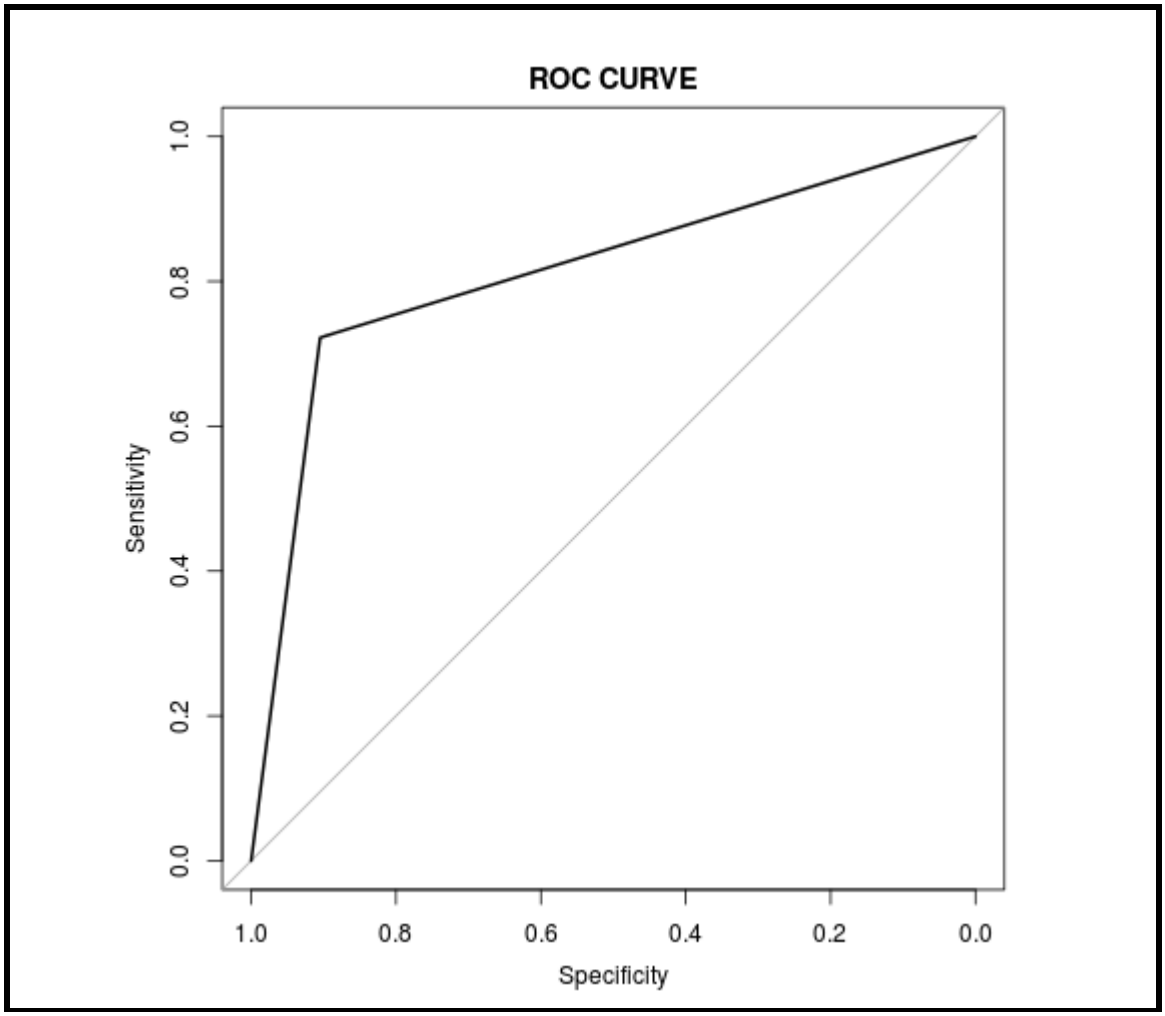

Figure 4. ROC for the variable **Age**

Area under curve: 0.8135

In order to this AUC value, we can state that this test is **good**. But remember that it's good since we change the categorization. It's not feasible if we work with age as a continuous variable.

Histology

Regarding this variable, the problem relies on the high number of empty values (58/193, 30.1%), what makes the prediction very poor. For solving this situation, we should consider for the analysis only those samples in which this variable was instantiated.

Anyway, since the number of samples with its histology instantiated and with a X value is the 0.9124% of the whole space, it makes no use to work with this classification.

As an example, my different tests with the proposal from above led me to the next situation:

		REFERENCE	
		X	**Y**
PREDICTION	**X**	37	3
	Y	0	0

SENSITIVITY	1.0000
SPECIFICITY	0.0000
POSITIVE PREDICTIVE VALUE	0.9250
NEGATIVE PREDICTIVE VALUE	NaN

Finally, it has an accuracy of **0.925**.

As we can see, the balance between sensitivity and specificity is unaffordable.

Cancer Stage

Due to the high number of values of this variable (12, with a very low prevalence of some of them), as well a high number of empty cases (67/193, 34.7%), we have both the problem explained for *age* as well as for *histology*.

In order to solve this problem, we could only consider those samples in which this value is written, and grouping the values as it follows:

STAGE 0	1	82	**Class 0**
STAGE 1	41		
STAGE 2	40		
STAGE 3	37	44	**Class 1**
STAGE 4	7		

As stage III of cancer uses to determine infiltration to neighbor structures or even metastasis, we consider groups 0 to 2 as the same level of grouping.

		REFERENCE	
		0	**1**
PREDICTION	**0**	23	13
	1	1	0

SENSITIVITY	0.9583
SPECIFICITY	0.0000
POSITIVE PREDICTIVE VALUE	0.6389
NEGATIVE PREDICTIVE VALUE	0.0000

Finally, it has an accuracy of **0.6216**.

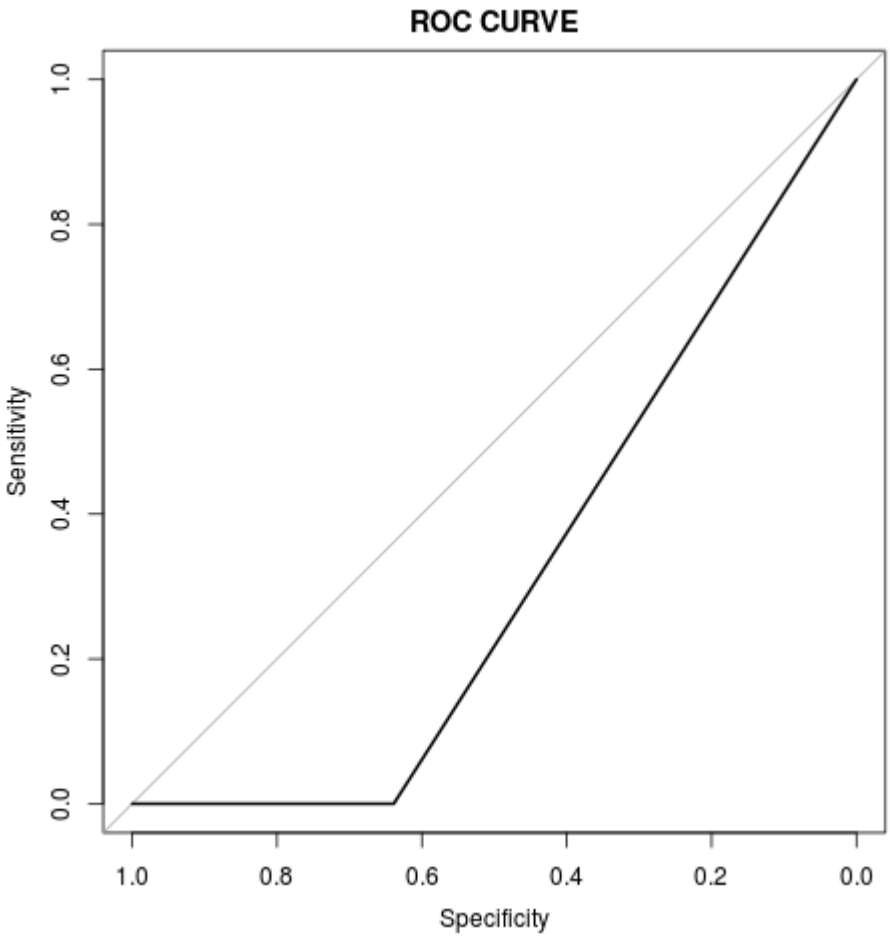

Figure 5. ROC for the variable **Stage**

Area under curve: 0.3194

As a conclusion, we can see that this test isn't valid.The specificity is 0, what means that the test isn't able to identify subjects without the condition (stages 3 and 4 subjects).

Similarly to what happened with the histology group, the balance between sensitivity and specificity is unaffordable.

Analysis of the top 20 features

Below we describe the 20 best features in terms of the best accuracy of the feature **Label**.

Such selection has been done using the RFE[1] function with the whole set of data (not the train nor the test one), a simple backwards feature selection routine that executes the classical SBS to identify the best features using a removal action.

This type of sequential algorithm has the disadvantage that is unable to reevaluate the usefulness of a feature after it has been discarded, so future dependencies between features that makes them stronger won't be identified.

Thus, if we confront this results (where we can see high-numbered features) with the ones obtained at the previous chapter, which where the first 5 (or 10, or 20, or 50) features, we can deduce that it would be necessary an stronger analysis of the relation between features to make a better prediction about the need of discarding one feature or not.

In the following subchapters you will be able to see a close analysis of the different top 20 features. Each analysis include an statistical analysis of the different values of the measurement for each group (or class), and a graph of the distribution of such values for each group.

Measurement 1. M1432

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	3598.28	240.33	36444.61
Minimum	314.13	0E-8	69.53
Maximum	18021.32	45388.51	55473.42
Interquartile Range	7087.29	543.43	25757.13

[1] Recursive Feature Selection

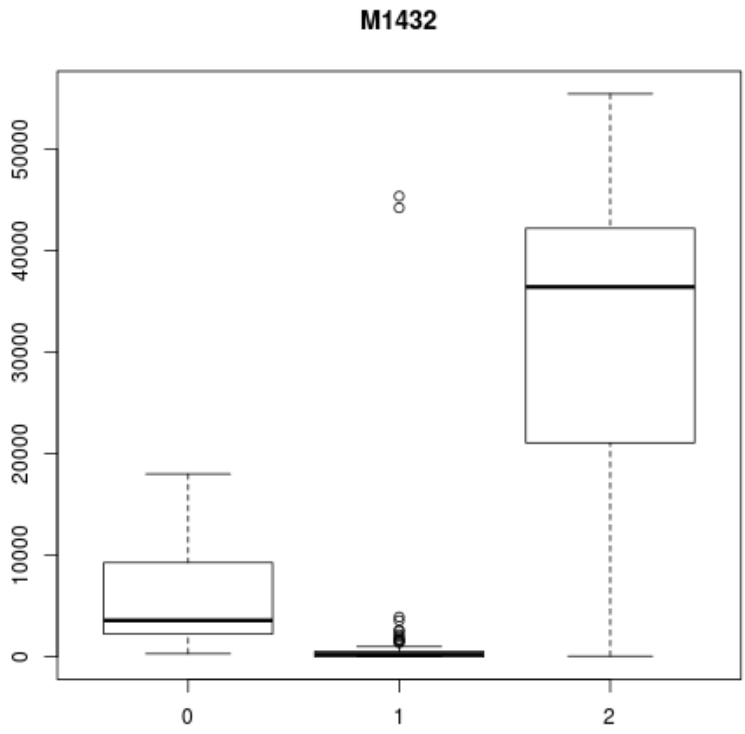

Figure 6. Distribution of the values of M1432 in the variable **Label**

Measurement 2. M299

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	16164.31	36627.89	74667.35
Minimum	1315.18	7673.98	59381.67
Maximum	41281.15	72281.32	182805.63
Interquartile Range	8845.43	14847.70	8739.89

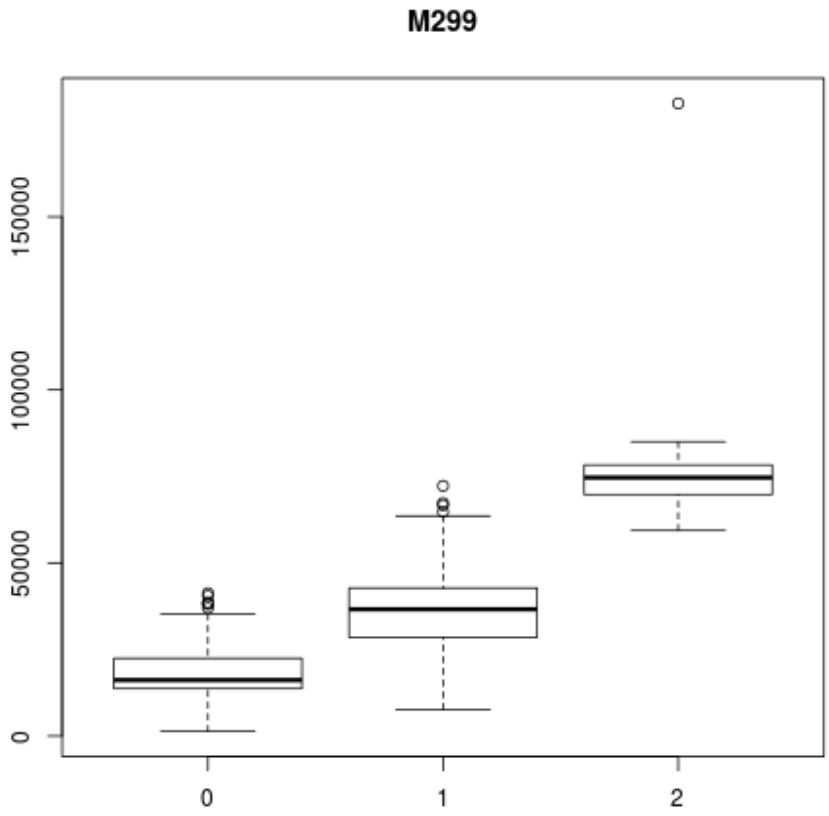

Figure 7. Distribution of the values of M299 in the variable **Label**

Measurement 3. M300

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	5696.55	12036.71	25782.98
Minimum	2504.78	2713.65	20733.72
Maximum	14843.60	24057.16	40600.48
Interquartile Range	3046.17	4904.38	3327.13

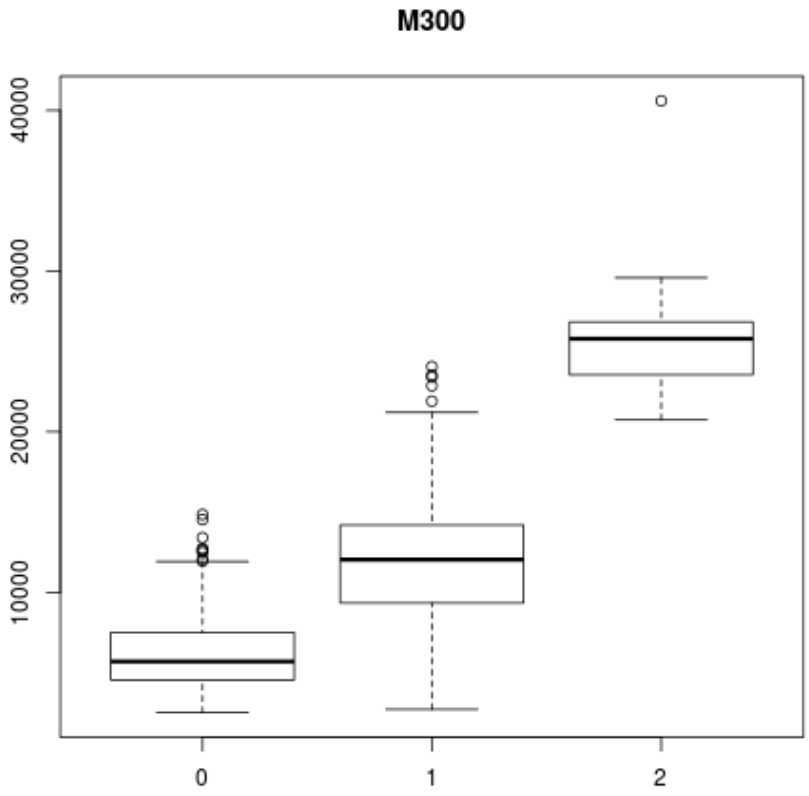

Figure 8. Distribution of the values of M300 in the variable **Label**

Measurement 4. M3581

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	20828.98	0E-10	783.43
Minimum	91.36	0E-8	219.50
Maximum	34467.41	65185.43	2672.21
Interquartile Range	6278.71	148.78	525.03

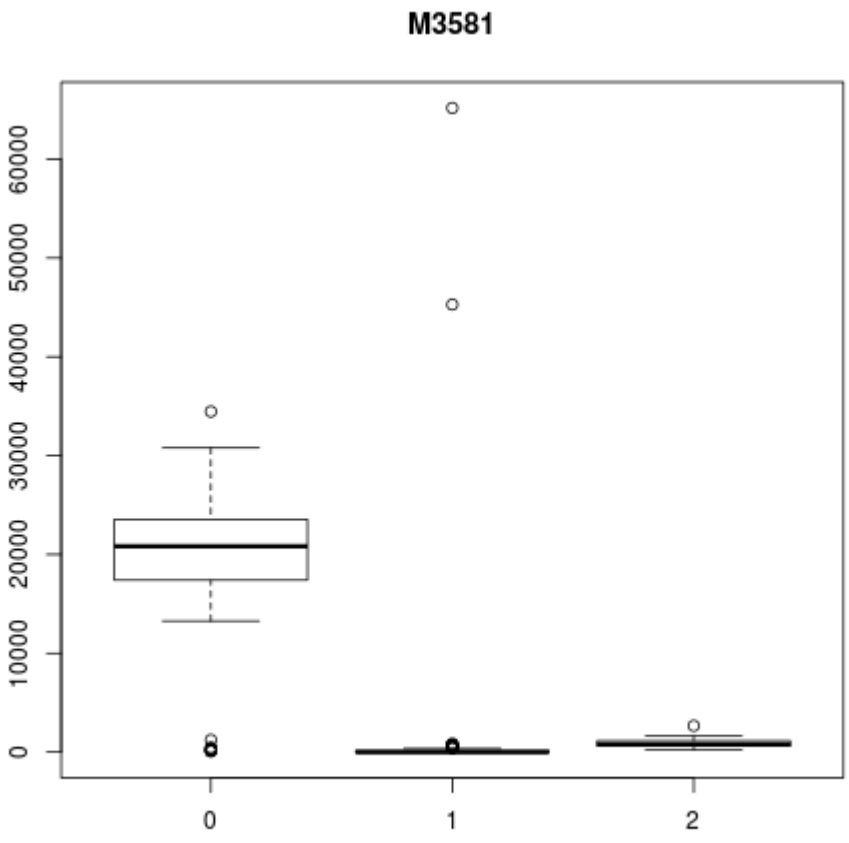

Figure 9. Distribution of the values of M3581 in the variable **Label**

Measurement 5. M3836

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	205.41	2416.22	6397.11
Minimum	0E-8	0E-8	540.29
Maximum	3438.39	5456.81	7374.77
Interquartile Range	450.18	2797.63	1401.86

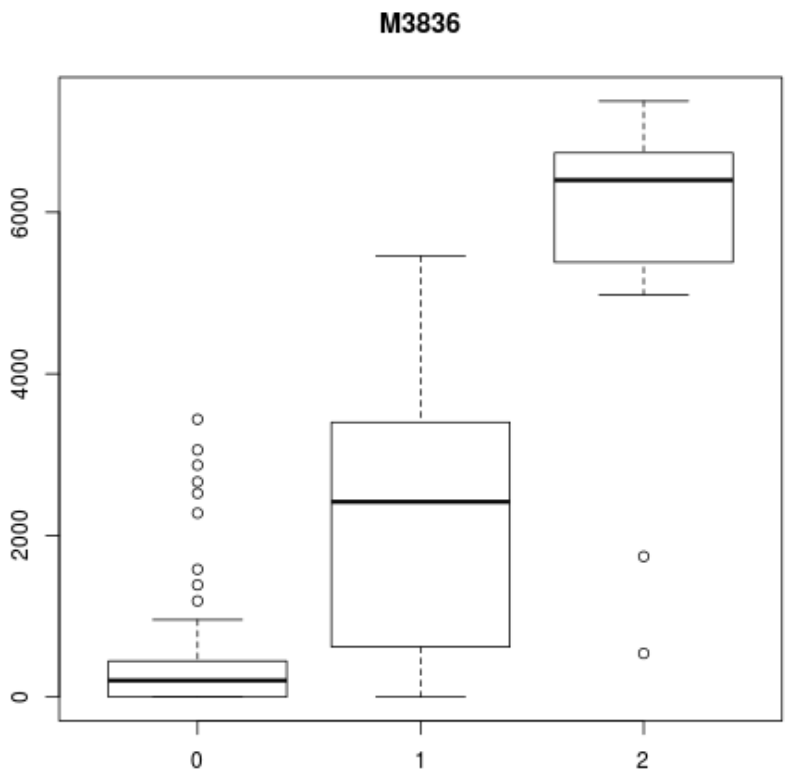

Figure 10. Distribution of the values of M3836 in the variable **Label**

Measurement 6. M3178

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	28792.97	5505.72	5933.59
Minimum	983.86	580.92	767.09
Maximum	53659.84	7325.46	7502.88
Interquartile Range	11596.34	1892.71	5251.73

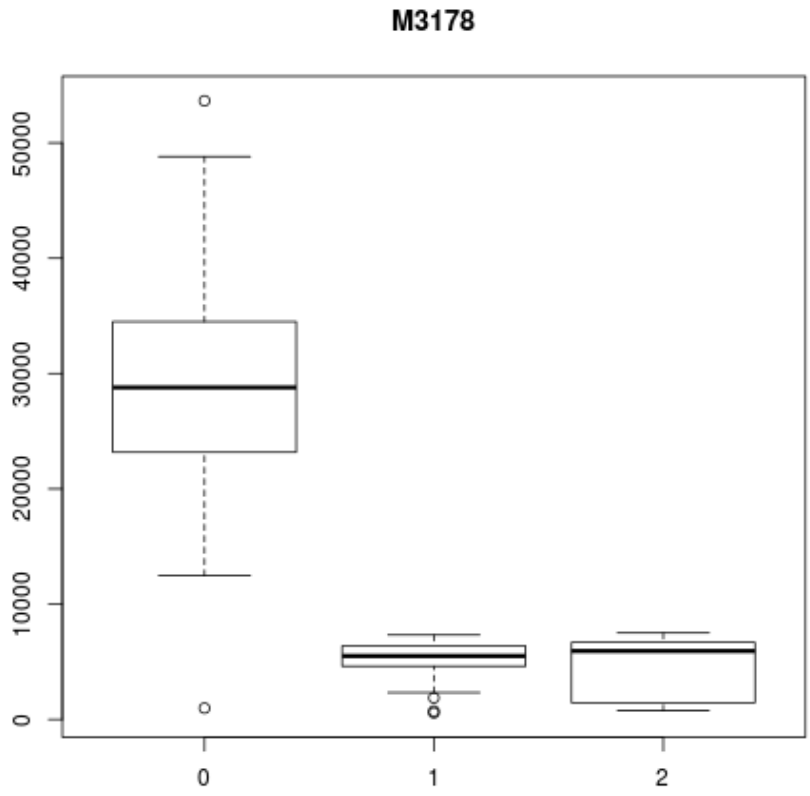

Figure 11. Distribution of the values of M3178 in the variable **Label**

Measurement 7. M2593

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	66128.74	1605.40	9940.62
Minimum	983.36	.00	706.63
Maximum	110474.83	36172.10	17274.17
Interquartile Range	28253.08	4454.46	9436.73

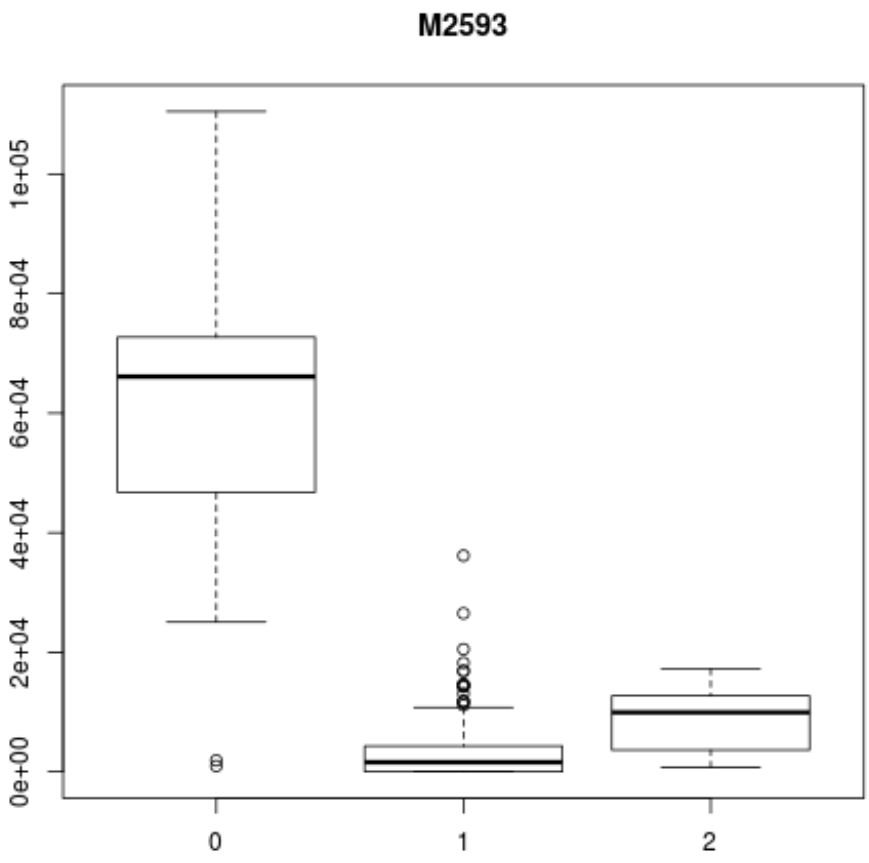

Figure 12. Distribution of the values of M2593 in the variable **Label**

Measurement 8. M2696

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	3404.35	78.69	23937.15
Minimum	446.96	0E-7	7448.12
Maximum	9674.41	30896.39	46727.95
Interquartile Range	3823.17	1884.25	19237.67

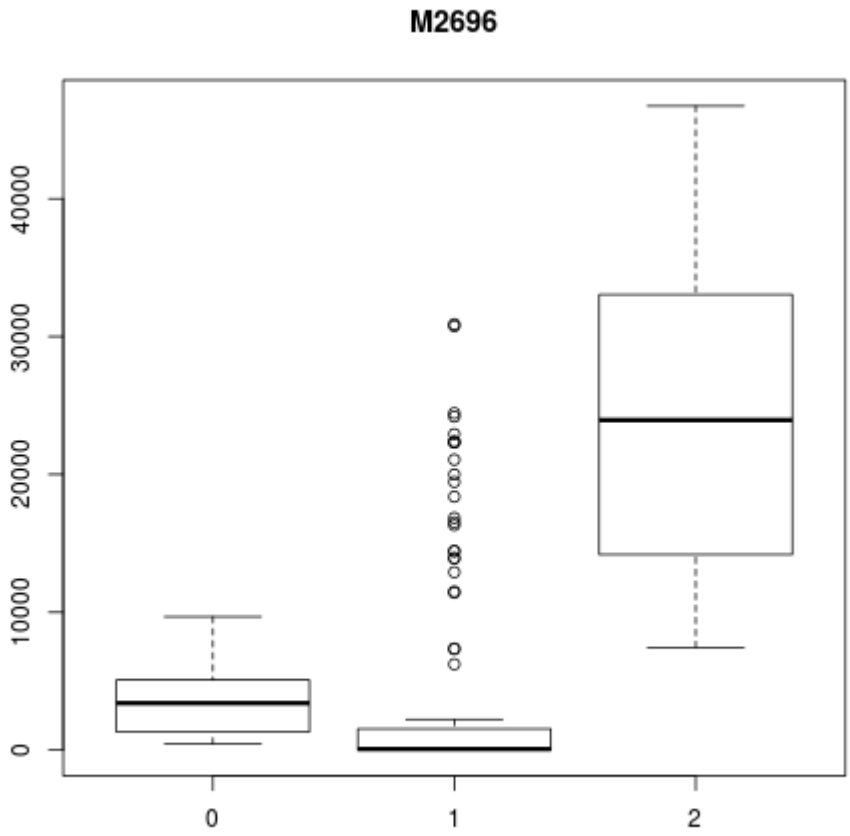

Figure 13. Distribution of the values of M2696 in the variable **Label**

Measurement 9. M3635

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	821.46	6896.90	26377.22
Minimum	oE-7	670.57	1813.79
Maximum	9313.31	26065.20	32320.47
Interquartile Range	1667.05	10922.81	3583.05

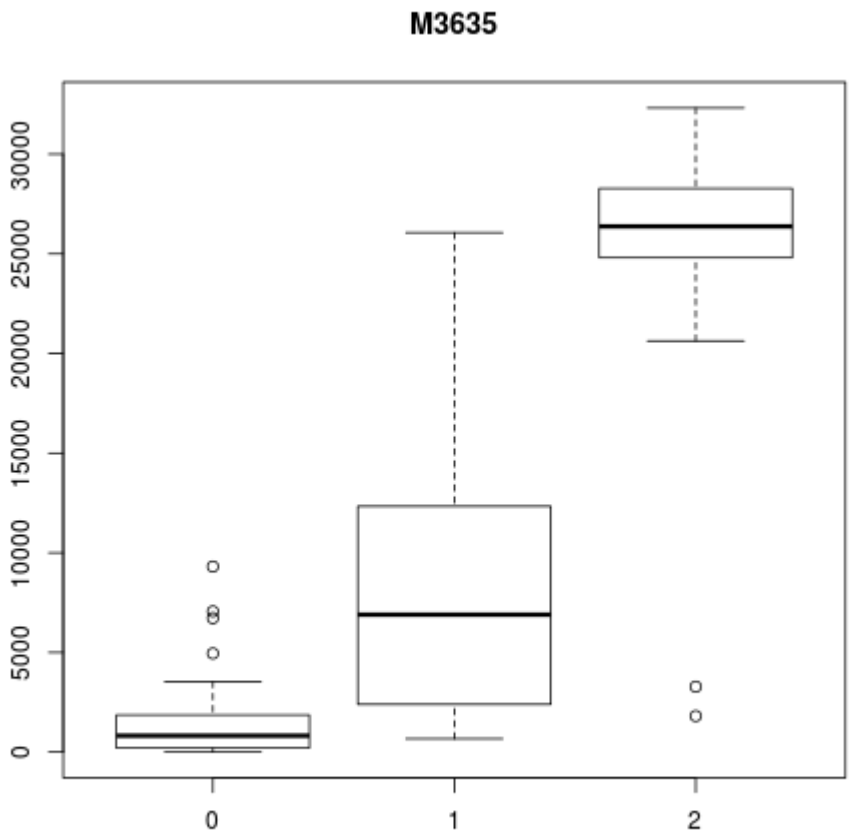

Figure 14. Distribution of the values of M3635 in the variable **Label**

Measurement 10. M4781

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	565.93	0E-9	5357.94
Minimum	0E-7	0E-7	4181.66
Maximum	2160.17	37268.76	27326.79
Interquartile Range	399.45	692.30	1090.63

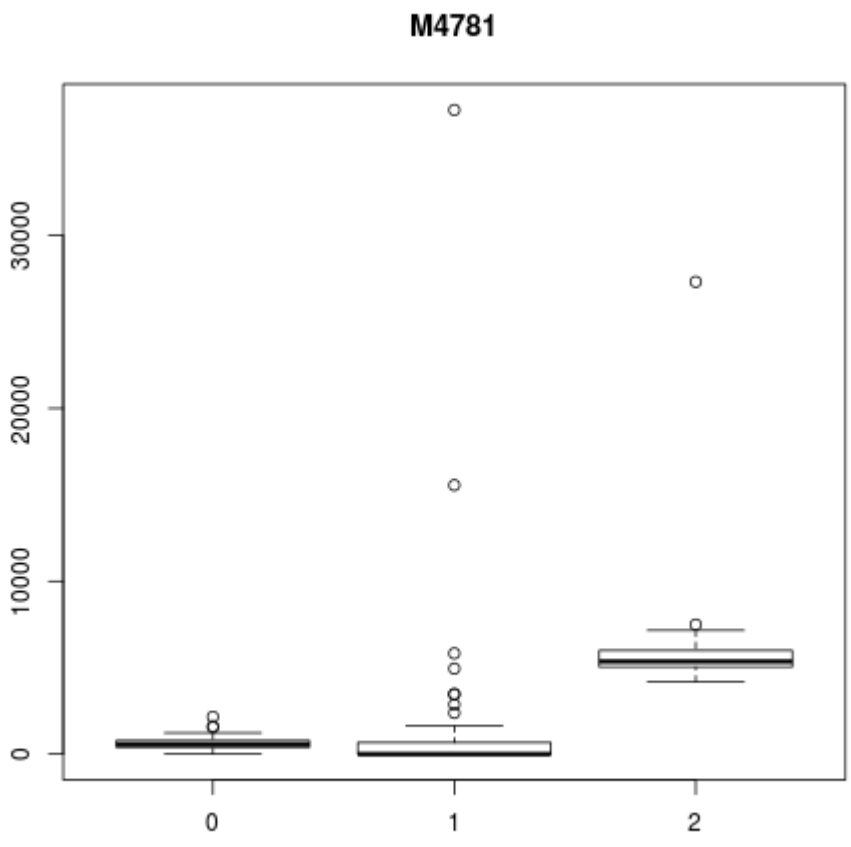

Figure 15. Distribution of the values of M4781 in the variable **Label**

Measurement 11. M3570

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	115331.40	0E-9	10862.88
Minimum	0E-7	0E-7	453.82
Maximum	191871.93	61370.99	68306.62
Interquartile Range	41618.73	2022.59	16469.77

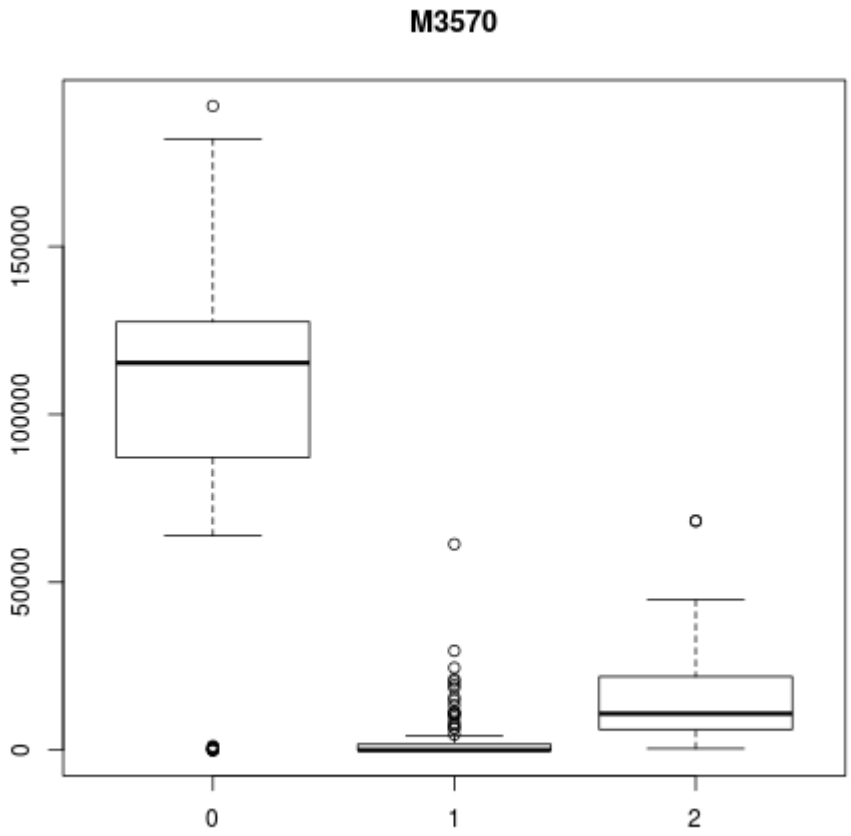

Figure 16. Distribution of the values of M3570 in the variable **Label**

Measurement 12. M4992

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	2236.45	78.80	41271.54
Minimum	0E-8	0E-8	27722.21
Maximum	151731.91	88947.62	90283.07
InterquartileRange	3235.51	985.94	12436.69

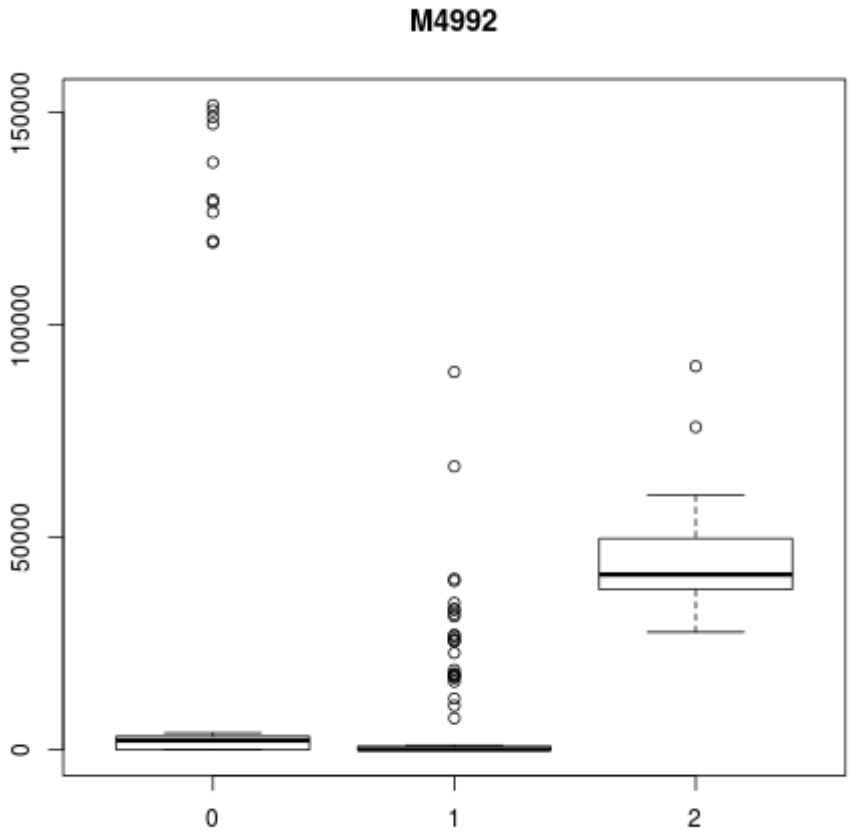

Figure 17. Distribution of the values of M4992 in the variable **Label**

Measurement 13. M1638

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	597.48	2140.24	13214.40
Minimum	0E-7	oE-7	1731.48
Maximum	4322.62	12399.53	33656.04
InterquartileRange	987.76	1722.63	6425.22

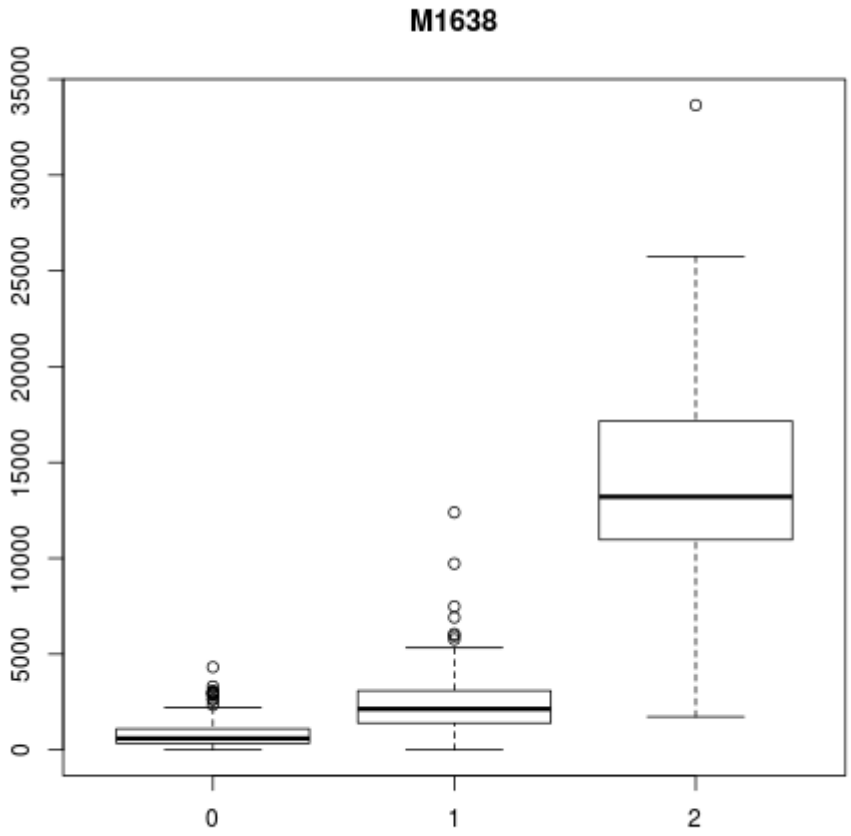

Figure 18. Distribution of the values of M1638 in the variable **Label**

Measurement 14. M4392

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	60205.56	0E-10	0E-10
Minimum	0E-8	0E-8	0E-8
Maximum	99186.40	10477.08	1830.51
InterquartileRange	17313.32	0E-8	86.76

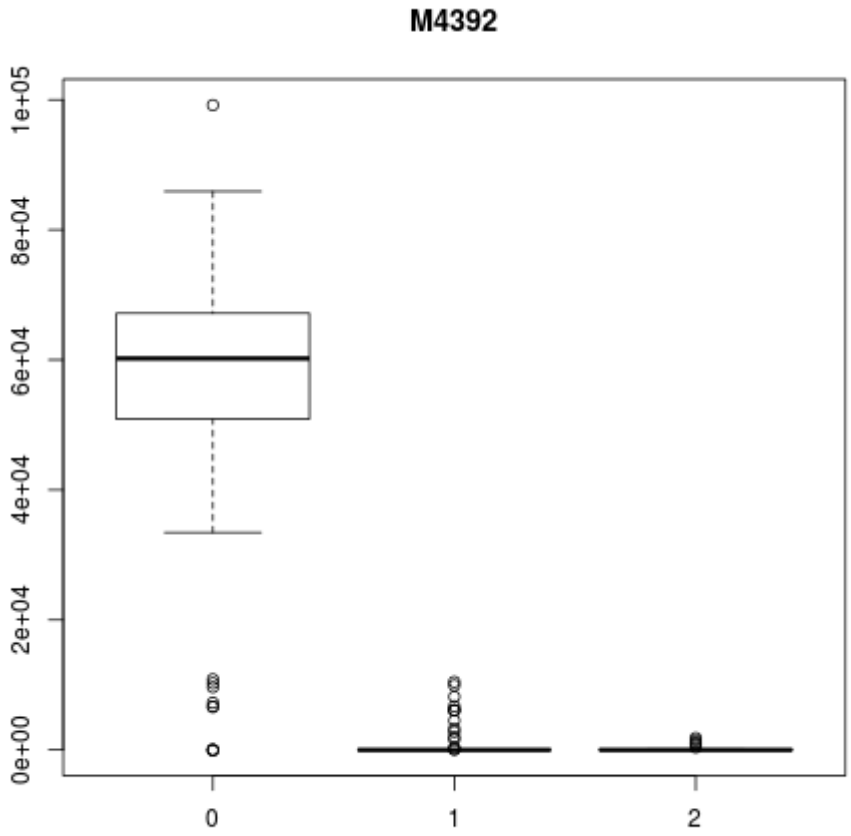

Figure 19. Distribution of the values of M4392 in the variable **Label**

Measurement 15. M554

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	86435.95	51564.89	39306.14
Minimum	48894.70	18376.11	32843.23
Maximum	169604.90	74901.64	48264.25
InterquartileRange	22489.56	14991.22	7086.38

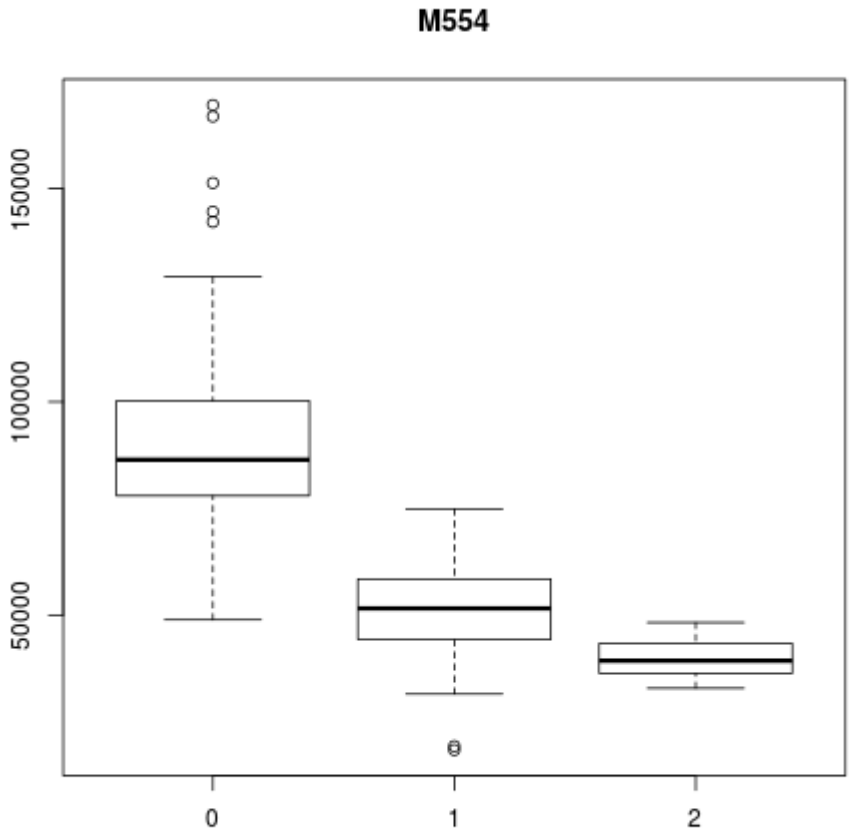

Figure 20. Distribution of the values of M554 in the variable **Label**

Measurement 16. M99

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	2457.11	4382.09	12454.42
Minimum	505.62	0E-7	126.77
Maximum	4593.23	13740.38	22164.47
InterquartileRange	1041.93	6888.89	6994.58

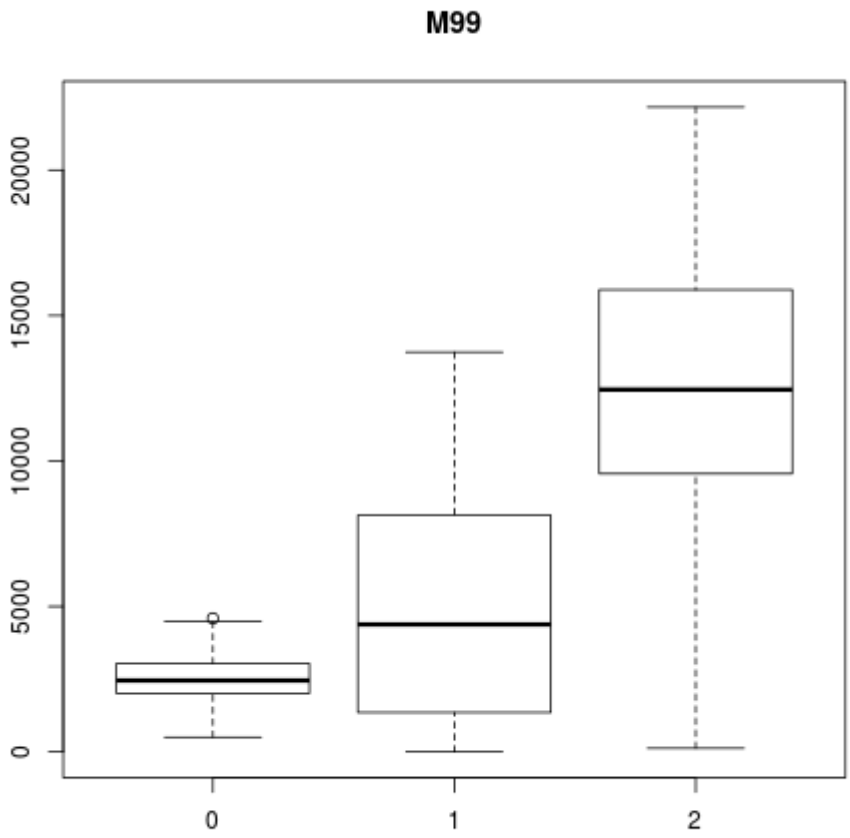

Figure 21. Distribution of the values of M99 in the variable **Label**

Measurement 17. M309

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	830.91	2063.27	6520.59
Minimum	0E-7	0E-7	480.97
Maximum	3337.18	6584.37	7860.68
InterquartileRange	850.40	1455.15	1232.36

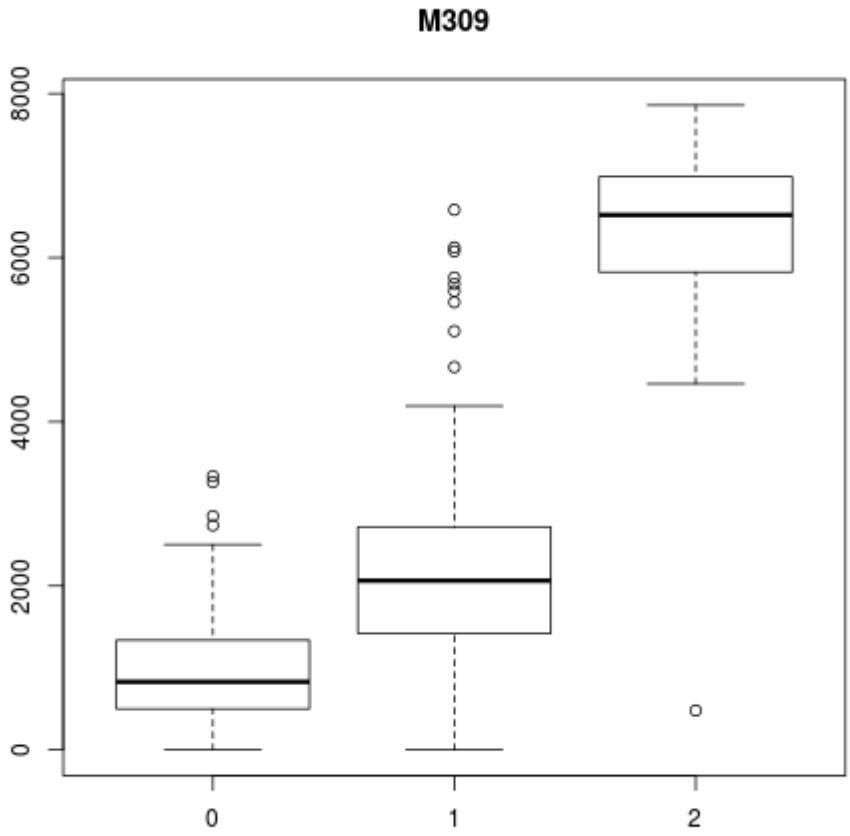

Figure 22. Distribution of the values of M309 in the variable **Label**

Measurement 18. M1191

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	16451.13	45956.76	60016.32
Minimum	9297.16	18859.26	25012.43
Maximum	38840.29	67056.92	66888.62
InterquartileRange	9736.86	24580.96	6604.88

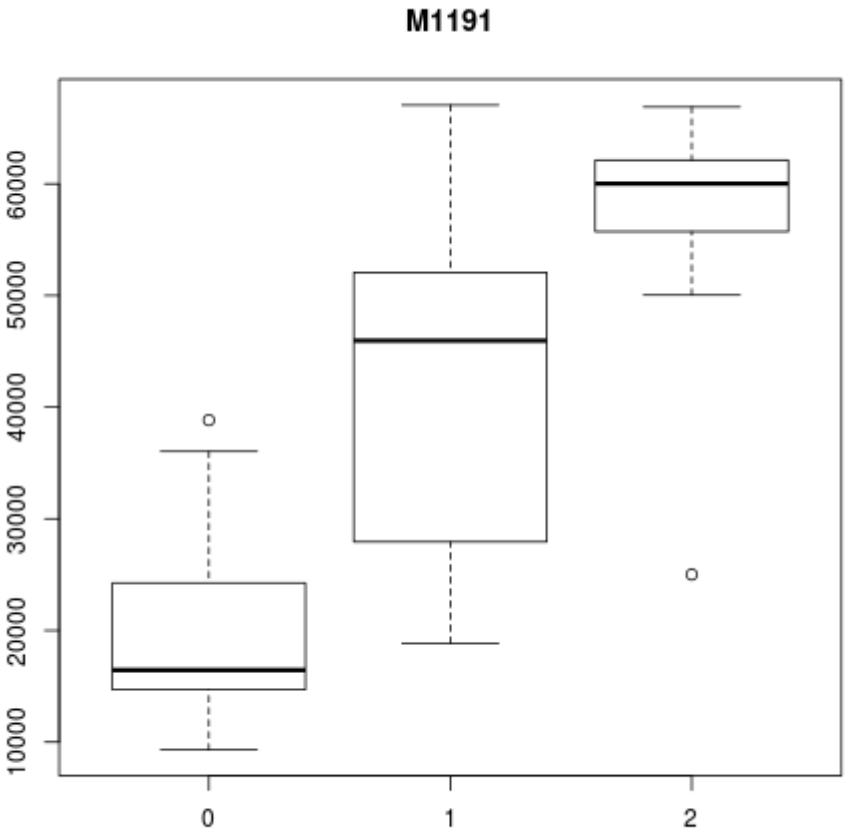

Figure 23. Distribution of the values of M1191 in the variable **Label**

Measurement 19. M5017

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	6218.78	0E-10	1682.23
Minimum	0E-8	0E-8	967.87
Maximum	12001.87	9120.80	4674.30
InterquartileRange	4503.41	621.17	705.69

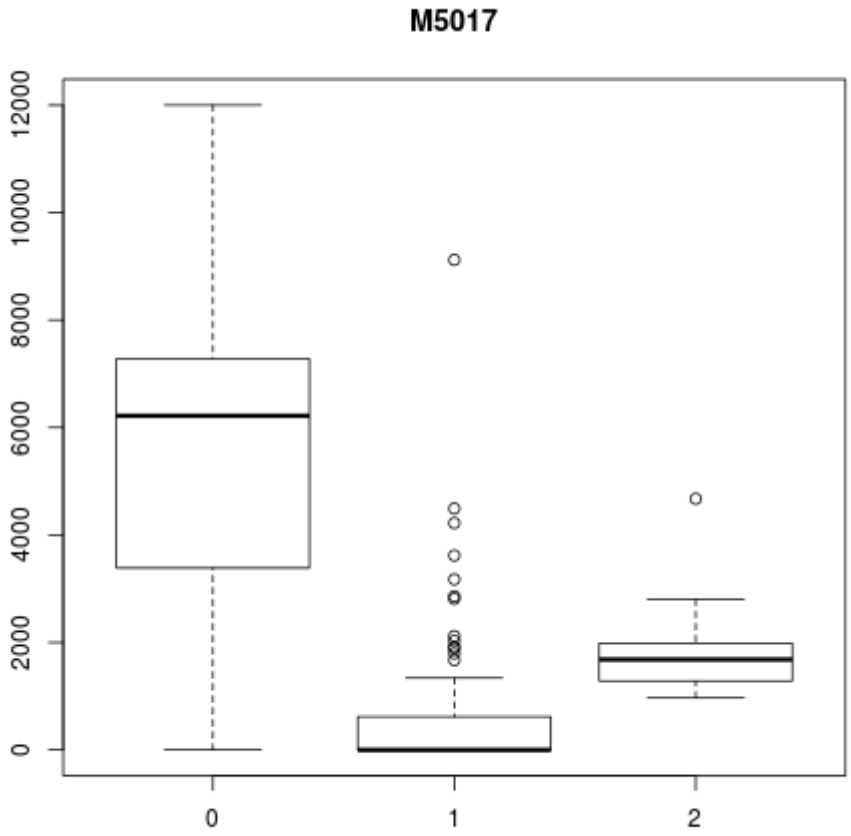

Figure 24. Distribution of the values of M5017 in the variable **Label**

Measurement 20. M2695

	Class 0	Class 1	Class 2
Number of Cases	58	105	30
Median	2106.24	0E-8	15661.66
Minimum	.00	.00	5360.96
Maximum	7404.30	25104.24	31844.02
InterquartileRange	2714.84	1743.45	11928.39

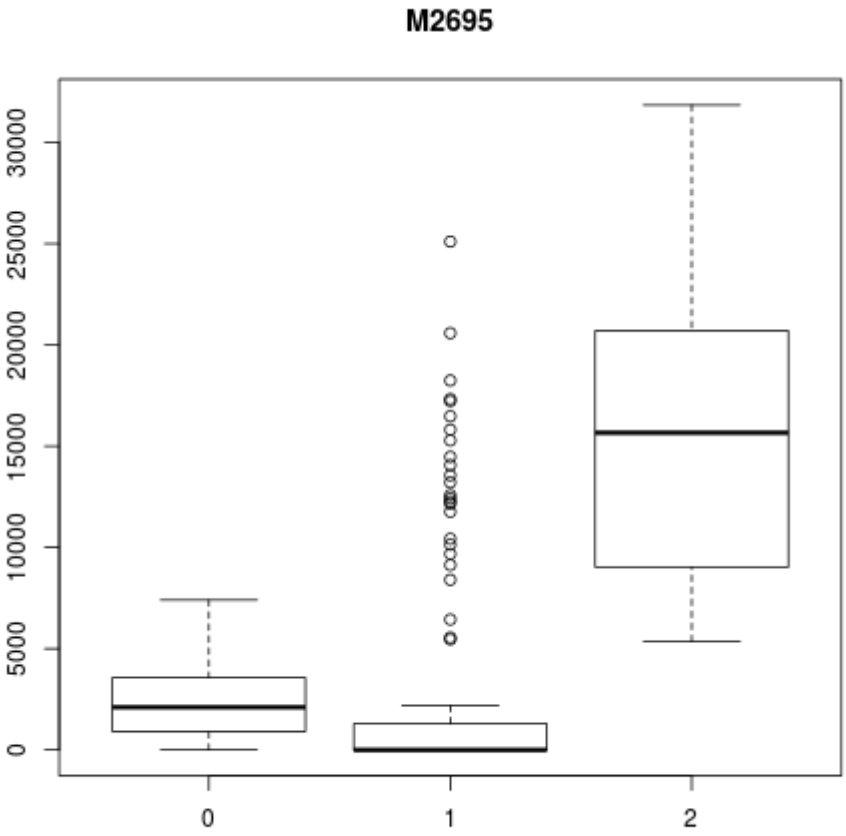

Figure 25. Distribution of the values of M2695 in the variable **Label**