Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Research on credit card anti-fraud based on data mining

Haolun Zhang *

Nanjing University of Science and Technology, China, Nanjing

*Corresponding author: 1069359277@qq.com

Abstract. The research topic of this paper is the field of credit card anti-fraud, using the European credit card data of the past two days in September 2018 for research, describing the data distribution through exploratory data analysis, data normalization, and finally using logistic regression for anti-fraud identification prediction.

Keywords: Credit card anti-fraud; Data mining; Data analysis; Logistic regression

1. Introduction

Financial technology is a hot issue in the financial field. With the development of data mining, big data, artificial intelligence and other related technologies, traditional offline risk control has been difficult to meet the anti-fraud identification of massive data. Big data risk control technology is playing an increasingly important role in the field of financial anti-fraud.

In order to promote the healthy and sustainable development of China's Internet consumer finance, excavate and meet customer "pain points", develop and create new consumers

Fee scenario, reliable structure of the credit database, improve the level of risk management should become the future of the Internet consumer finance field important development goals, and to achieve this series of goals is the key lies in the development and application of big data technology, big data technology should become the future development of the Internet consumer finance field security door and accelerator. [1]

The application of big data has changed the traditional financial service model. Financial enterprises and institutions can collect relevant information of users, and carry out quantitative processing and model construction. Based on the multi-dimensional analysis results of big data, it classifies users reasonably according to certain characteristics, so as to provide accurate and effective financial services to users according to their needs.[2]

This paper studies credit card anti-fraud. In the Internet era, credit card fraud not only "innovates" technical means, but also develops from "small team operations" to organized and premeditated industrial crimes. They collude through Internet channels, cross borders, have a clear division of labor, and involve a wide range of fields. Banks, third-party payment, network operators, e-commerce and other links. These criminals from the research of banking business and rules, to technical attacks, and then to customer information fraud, the process is clear, careful planning, seems to have formed a black industry chain. [3] In order to better realize anti-fraud identification, there are higher requirements for big data risk control technology.

The problem of information asymmetry in credit operation can be better solved through big data risk control technology [4]. Based on the rational use of data modeling technology, intelligence clues can be better explored and the operation principle measured [5] to effectively identify customers with high fraud risk.

2. Data background and description

2.1. Data source

This study chose open data set, data set address to http://www.kaggle.com/mlg-ulb.The data set studied the credit card transaction records of European cardholders within two days in September 2018. The data set was processed by PCA, and after data dimensionality reduction, there were no physical features. This study intends to study the anti-fraud identification of credit risk with these data.

2.2. Data content

After dimensionality reduction (PCA), the data contains only the numeric content, V1, V2,.... V28, are the most important content obtained through PCA. The only two pieces of data that are not processed by PCA are the amount of the transaction at the time of the transaction. Among them, the calculation of transaction time is calculated between each transaction and the time of the first transaction. Class represents a variable, if it is 1, then it means cheating occurred, otherwise it is 0.

2.3. Sample definition

The total sample size is 284,407, and each row records a transaction behavior. Class is the label of the data set. When the value is 1, it is a positive sample, indicating fraudulent transactions; when the value is 0, it is a negative sample, indicating normal transactions.

2.4. Data set partitioning

This paper uses simple random sampling to divide the data set. Simple random sampling, also known as simple random sampling, pure random sampling and SRS sampling, refers to a sampling method in which N units are randomly selected from the total n units as samples, so that the probability of each possible sample being selected is equal. In this paper, simple random sampling of data set is carried out according to 7:3 ratio, and train and test are divided. train is used as the data for training the model, test is used as the data for verifying the effect of the model, and the effect of the model on test reflects its generalization ability. The better the model's performance on test, the better its generalization ability.

2.5. Sample distribution

Table 2.1 Distribution of samples

Lable	Total sample	Train	Test
0	284315	199022	85293
1	492	342	150

After sampling, the distribution of positive and negative samples is shown in the figure above. In train, there are 342 positive samples and 199,022 negative samples; in test, there are 150 positive samples and 85,293 negative samples.

3. EDA

3.1. Exploratory data analysis for purposes

EDA is Exploratory Data Analysis, which refers to the exploration of existing data (especially the original data obtained from surveys or observations) with as few prior assumptions as possible. A data analysis method that explores data structures and rules by means of mapping, tabulation, equation fitting, and calculation of characteristic quantities, and analyzes the data to summarize its main characteristics. The difference with the traditional statistical analysis method is that: the

traditional statistical analysis method is based on probability theory, first assume that the data obey a certain distribution, and then estimate some parameters and statistics of the model according to the data sample, so as to understand the characteristics of the data. However, there are often many data in practice that do not conform to the assumed statistical model distribution, which leads to unsatisfactory data analysis results. EDA is a more realistic analysis method, which "sets aside" probability theory, starts from the data, emphasizes the visibility of the data, and lets the data itself "speak", helping us to understand the other value of the data in addition to formal modeling or hypothesis testing tasks.

This paper mainly explores and analyzes five aspects of saturation analysis, data type distribution, data situation, data distribution, and correlation coefficient matrix of numerical variables to describe the data as a whole.

3.2. Saturation analysis

A missing value is an unknown value, that is, we are not sure exactly what the value is, but a missing value is different from a null value, which means no value, and a missing value which means an unknown value. So the missing value has a large degree of uncertainty. If the missing value in the data is too high, the data is unstable and the model cannot be fitted from it. Therefore, the number of missing values can reflect the data quality of a data set to some extent.

$$N = \frac{A}{S} \tag{1}$$

$$COV = 1 - N \tag{2}$$

Where N represents the missing rate, A represents the number of missing values in a column, S represents the number of values in that column, and COV represents saturation.

The higher the saturation of a column in the data set, the lower the missing rate and the better the stability of the data in the column. We can sort by saturation in reverse order to filter out columns with low saturation.

As you can see from the figure below, the saturation is 100%, so no columns are filtered in the saturation analysis.

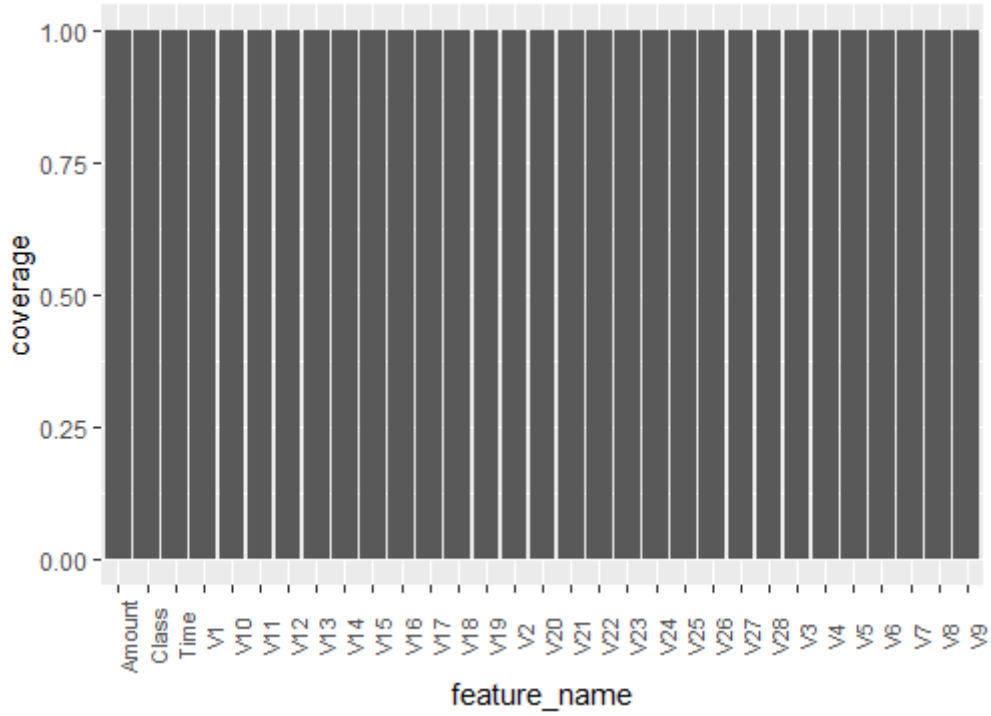

Figure 3.1. Data saturation

3.3. Data type distribution

There are 31 columns in the data set, among which class is the label column, that is, the feature column is 30 columns. V1... V28 is data that has been dimensionally reduced. The correlation between them is relatively small, and the correlation coefficient matrix will be observed for verification in the following steps.

Numerical variables: 30 column variables are numerical variables

Character variable: None

3.4. Data situation

The first 6 rows of data are as follows:

Table 3.1. Data situation

Industry	1	2	3	4	5	6
Time	0	0	1	1	2	2
V1	-1. 359	1.191	-1.358	-0.966	-1.158	-0.425
V2	-0.072	0.266	-1.340	-0.185	0.877	0.960
V3	2.536	0.166	1.773	1.792	1.548	1.141
V4	1.378	0.448	0.379	-0.863	0.403	-0.168
V5	-0.338	0.060	-0.503	-0.010	-0.407	0.420
V6	0.462	-0.082	1.800	1.247	0.095	-0.029
V7	0.239	-0.078	0.791	0.237	0.592	0.476
V8	0.098	0.085	0.247	0.377	-0.270	0.260
V9	0.363	-0.255	-1.514	-1.387	0.817	-0.568
V10	0.090	-0.166	0.207	-0.054	0.753	-0.371
V11	-0.551	1.612	0.624	-0.226	-0.822	1.341
V12	-0.417	1.065	0.066	0.178	0.538	0.359
V13	-0.991	0.489	0.717	0.507	1.345	-0.358
V14	-0.311	-0.143	-0.165	-0.287	-1.119	-0.137
V15	1.468	0.635	2.345	-0.631	0.175	0.517
V16	-0.470	0.463	-2.890	-1.059	-0.451	0.401
V17	0.207	-0.114	1.109	-0.684	-0.237	-0.058
V18	0.025	-0.183	-0.121	1.965	-0.038	0.068
V19	0.403	-0.145	-2.261	-1.232	0.803	-0.033
V20	0.251	-0.069	0.524	-0.208	0.408	0.084
V21	-0.018	-0.225	0.247	-0.108	-0.009	-0.208
V22	0.277	-0.638	0.771	0.005	0.798	-0.559
V23	-0.110	0.101	0.909	-0.190	-0.137	-0.026
V24	0.066	-0.339	-0.689	-1.175	0.141	-0.371
V25	0.128	0.167	-0.327	0.647	-0.206	-0.232
V26	-0.189	0.125	-0.139	-0.221	0.502	0.105
V27	0.133	-0.008	-0.055	0.062	0.219	0.253
V28	-0.021	0.014	-0.059	0.061	0.215	0.081
Amount	149.62	2.69	378.66	123.50	69.99	3.67
Class	0	0	0	0	0	0

3.5. Data distribution

Explore V1,... The data distribution of V28 is observed through the box diagram, and the data distribution results are shown in the figure below.

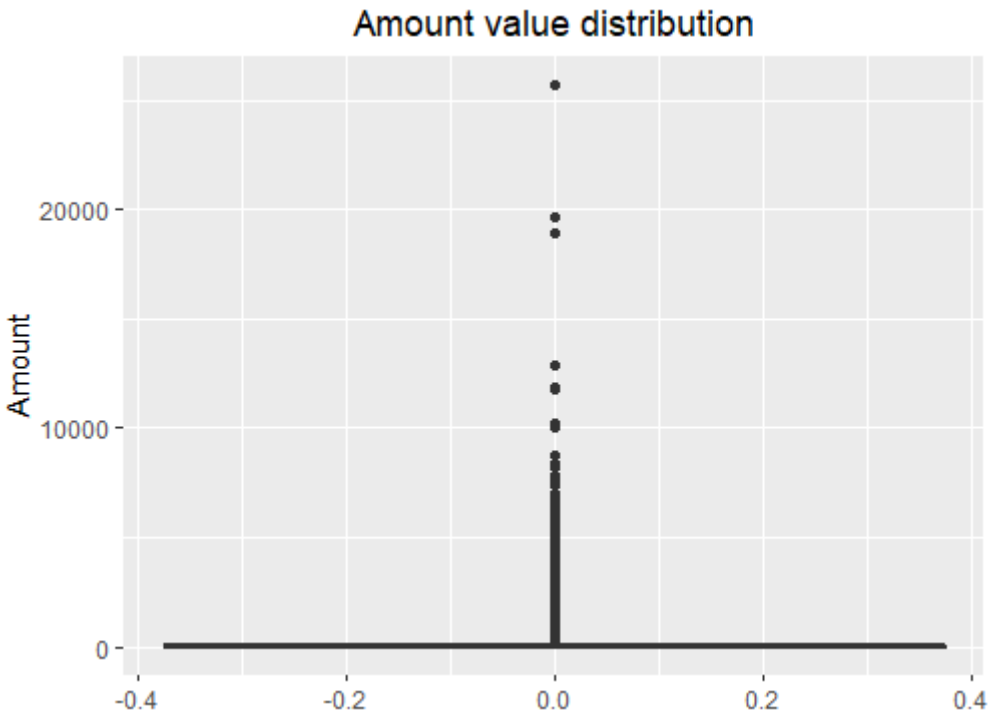

Figure 3.2. Amount distribution

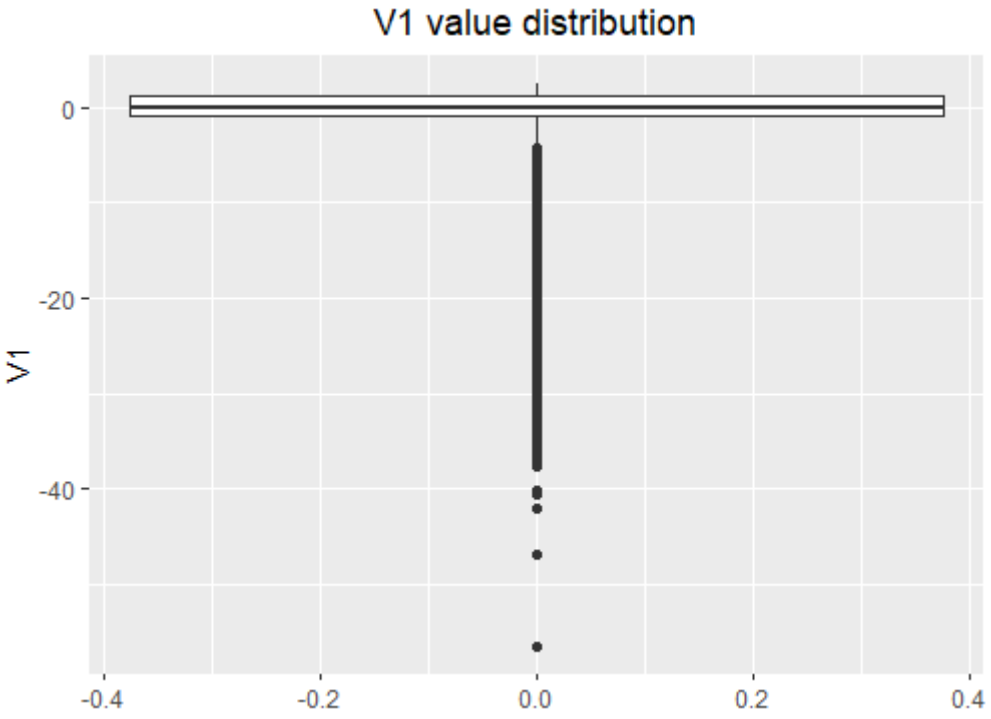

Figure 3.3. Distribution of V1

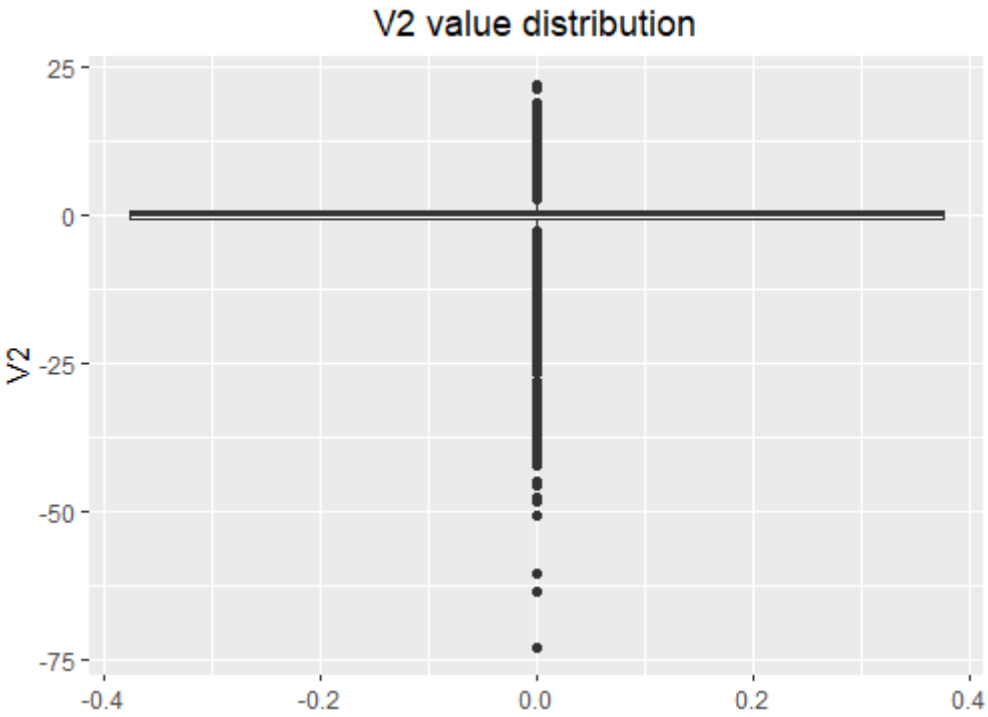

Figure 3.4. V2 distribution

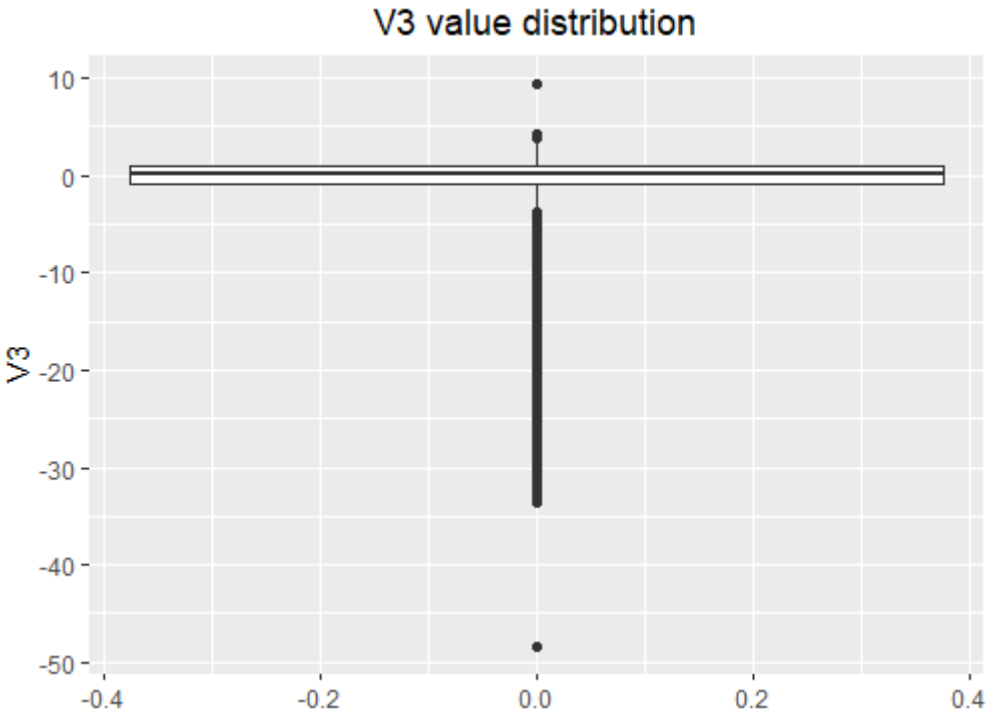

Figure 3.5. V3 distribution

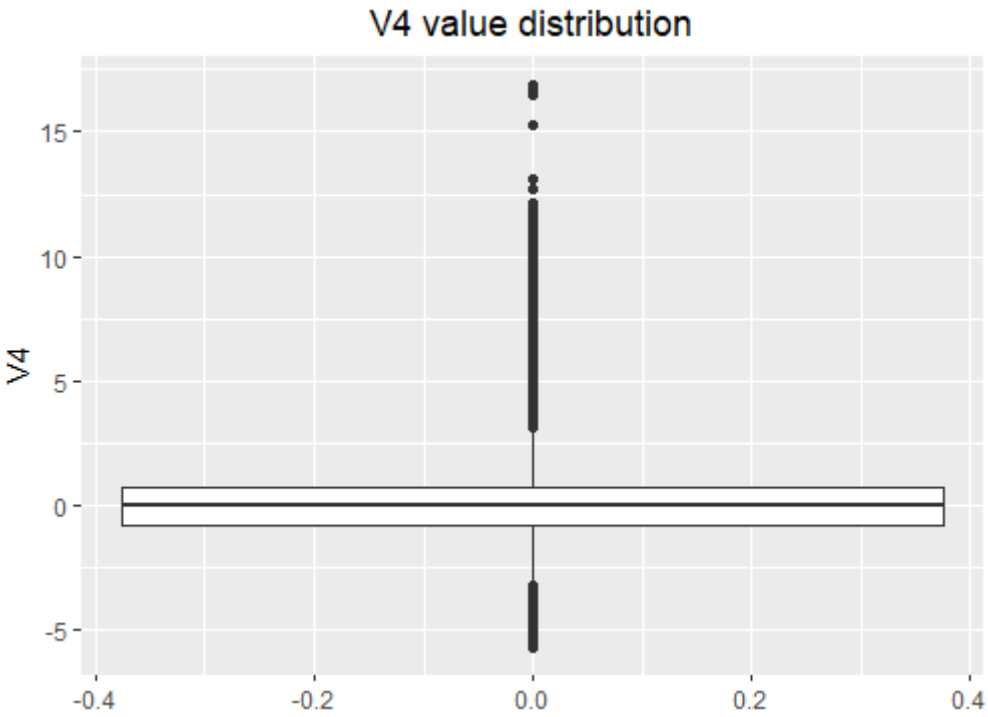

Figure 3.6. V4 distribution

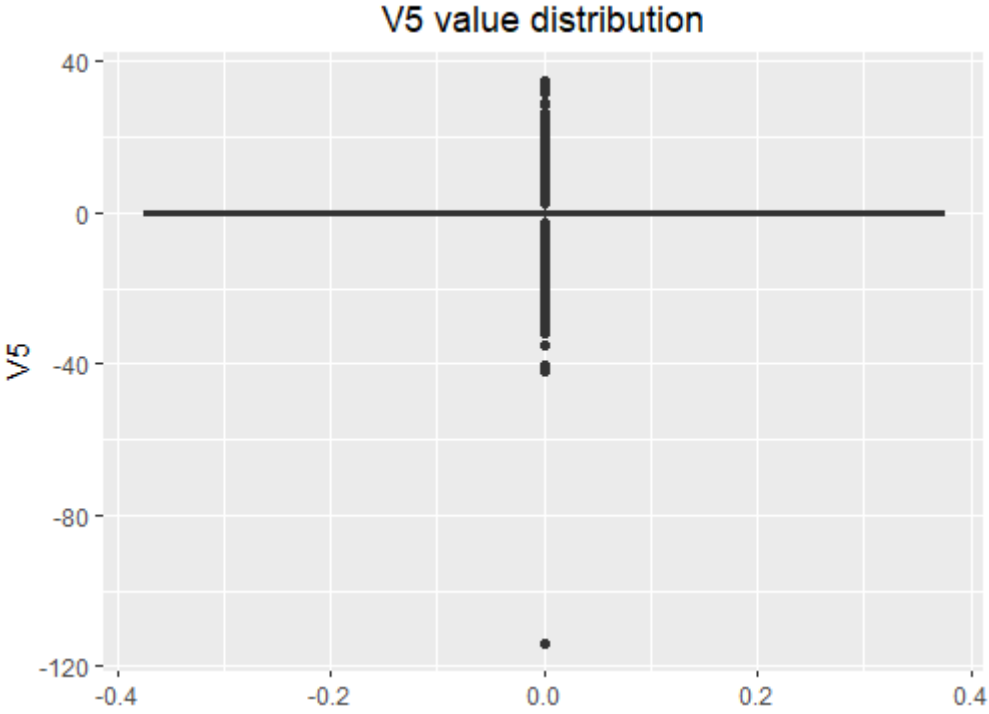

Figure 3.7. V5 distribution

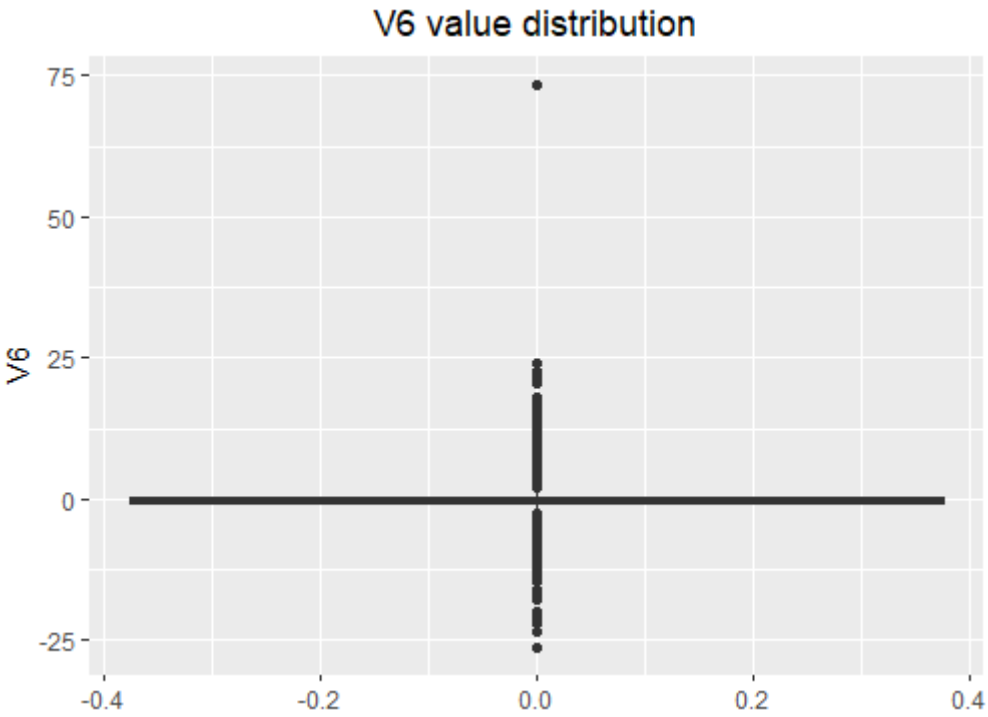

Figure 3.8. V6 distribution

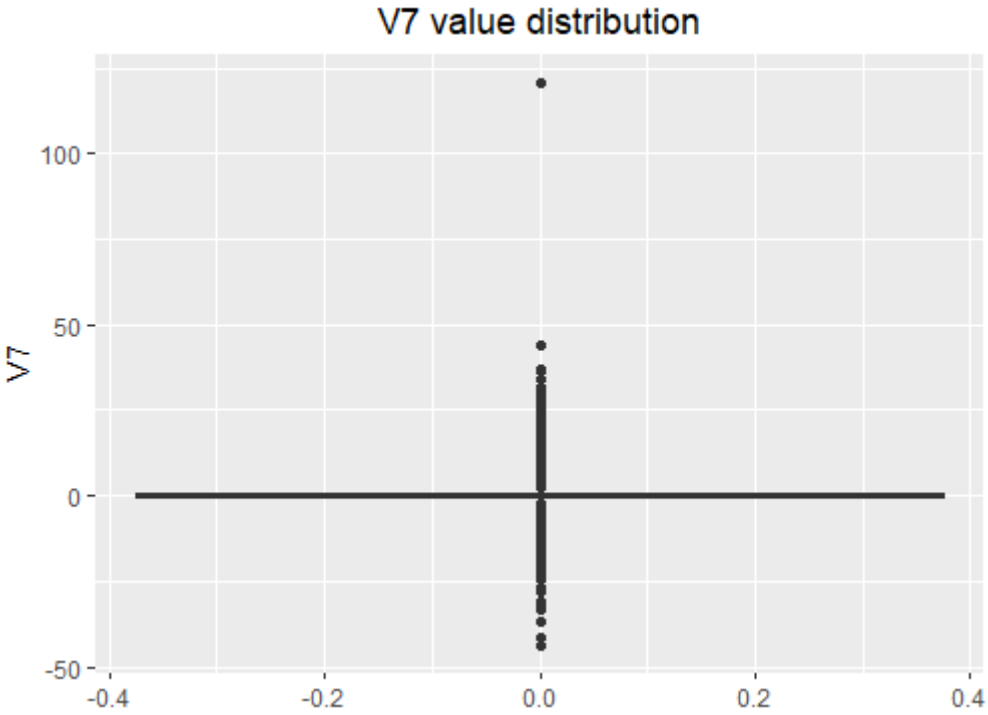

Figure 3.9. V7 distribution

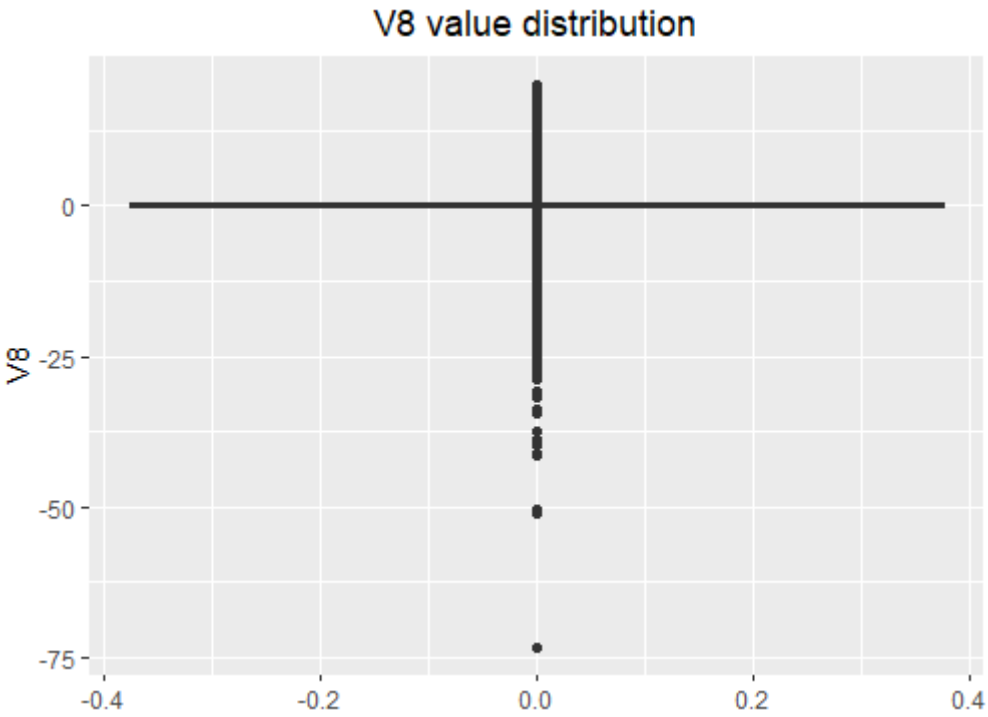

Figure 3.10. V8 distribution

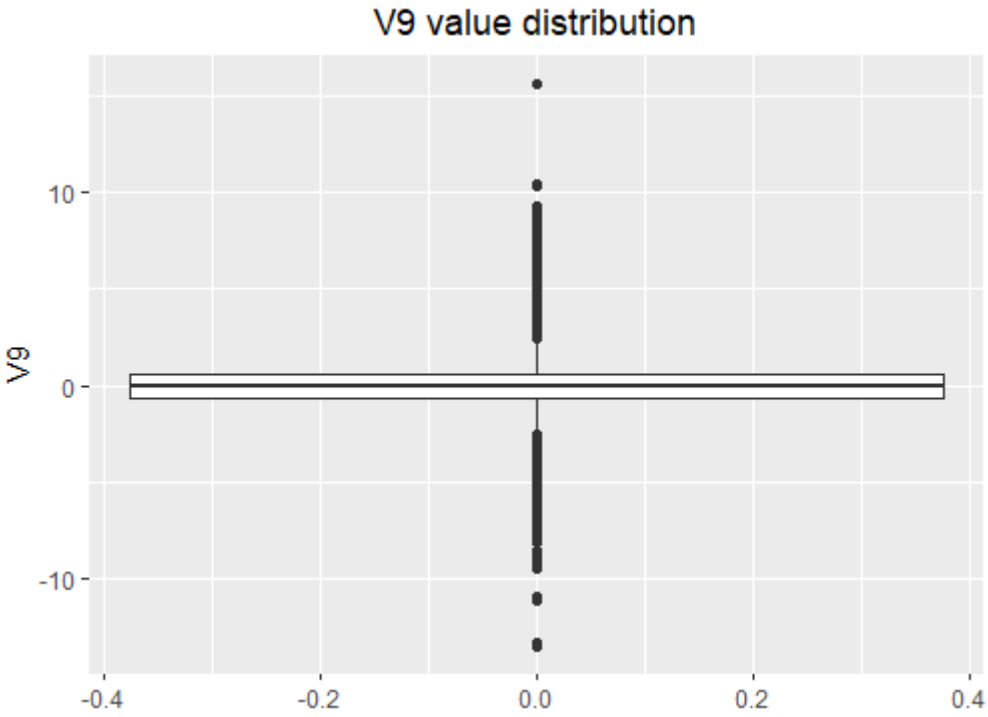

Figure 3.11. V9 distribution

Figure 3.12. V10 distribution

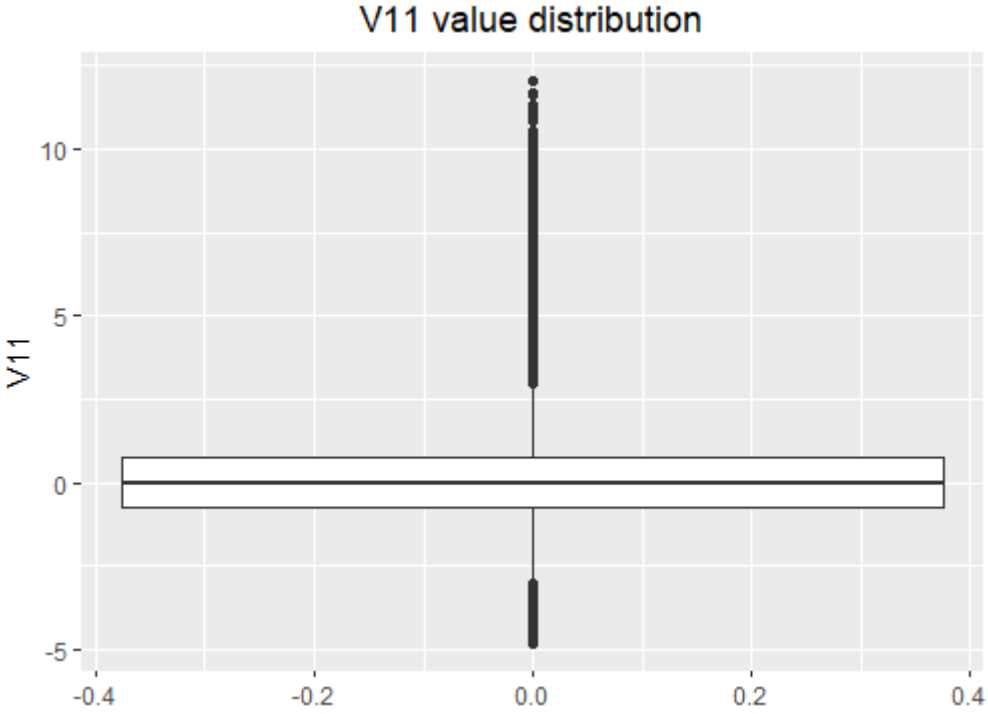

Figure 3.13. V11 distribution

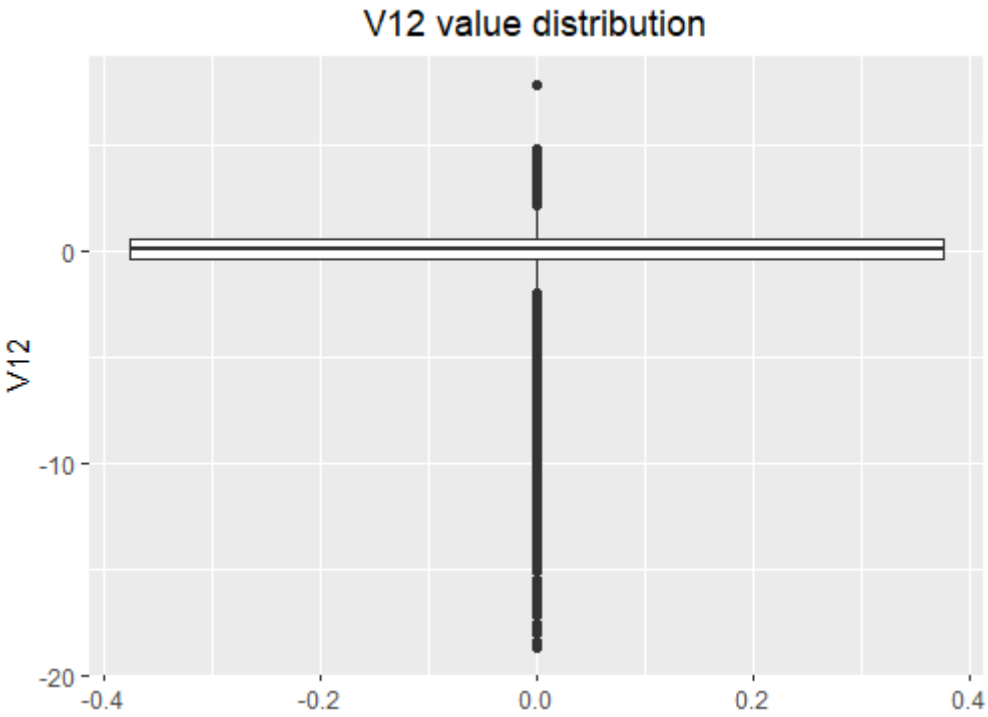

Figure 3.14. V12 distribution

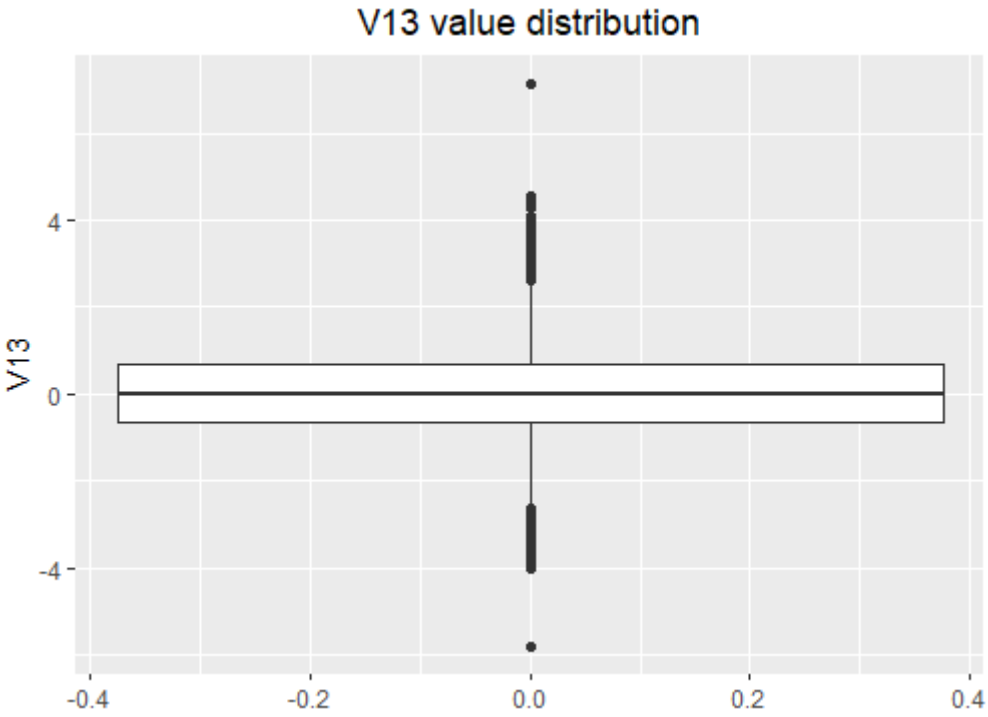

Figure 3.15. V13 distribution

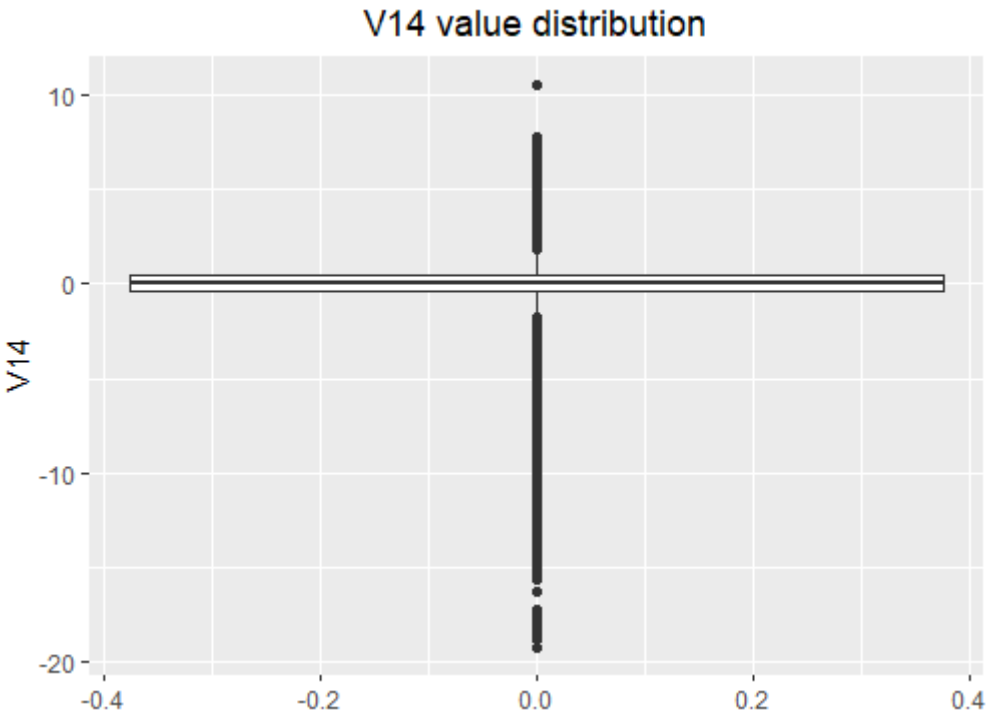

Figure 3.16. V14 distribution

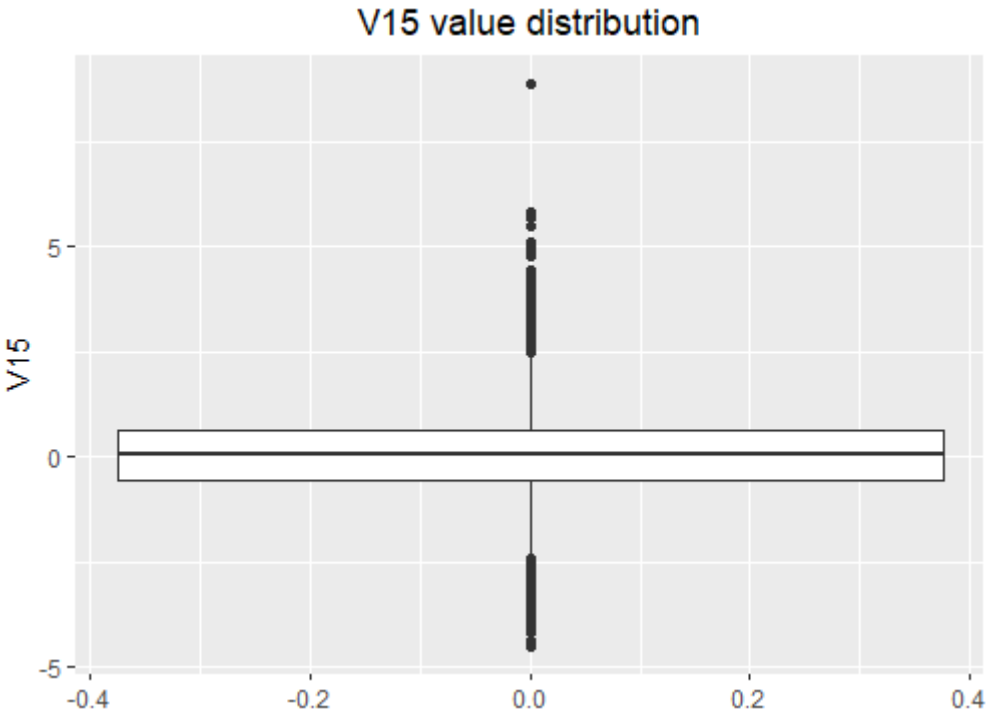

Figure 3.17. V15 distribution

Figure 3.18. V16 distribution

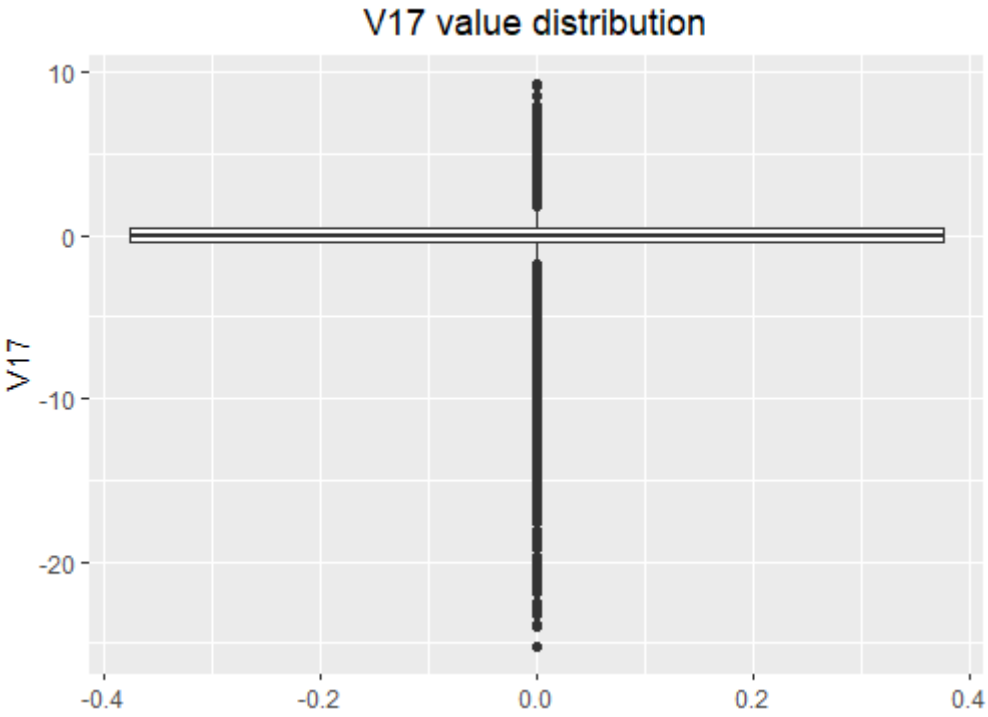

Figure 3.19. V17 distribution

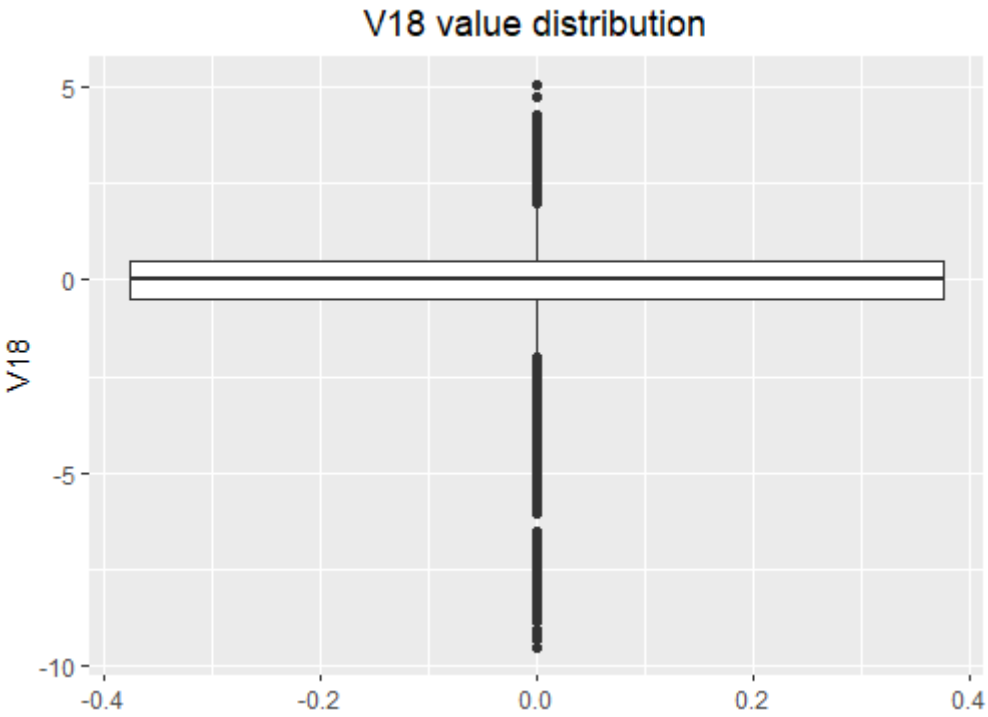

Figure 3.20. V18 distribution

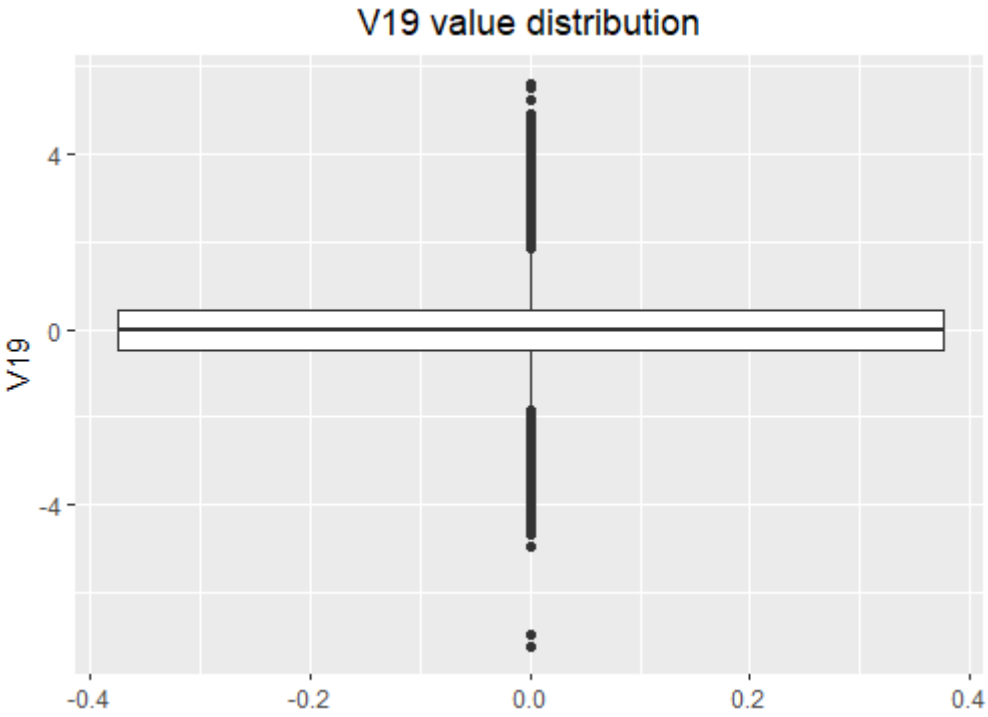

Figure 3.21. V19 distribution

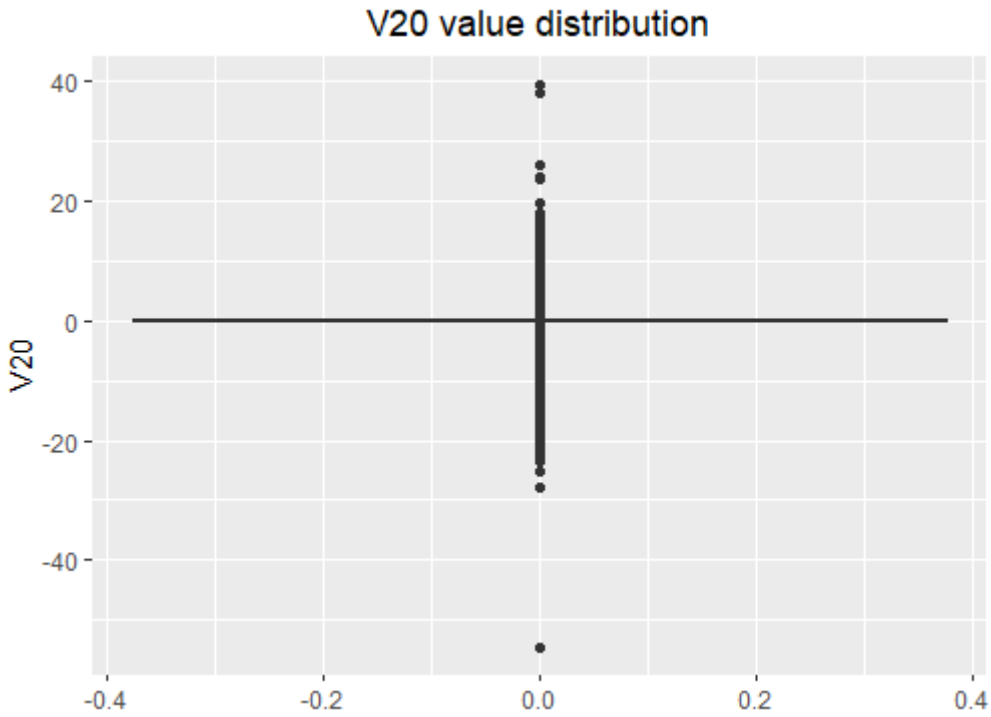

Figure 3.22. V20 distribution

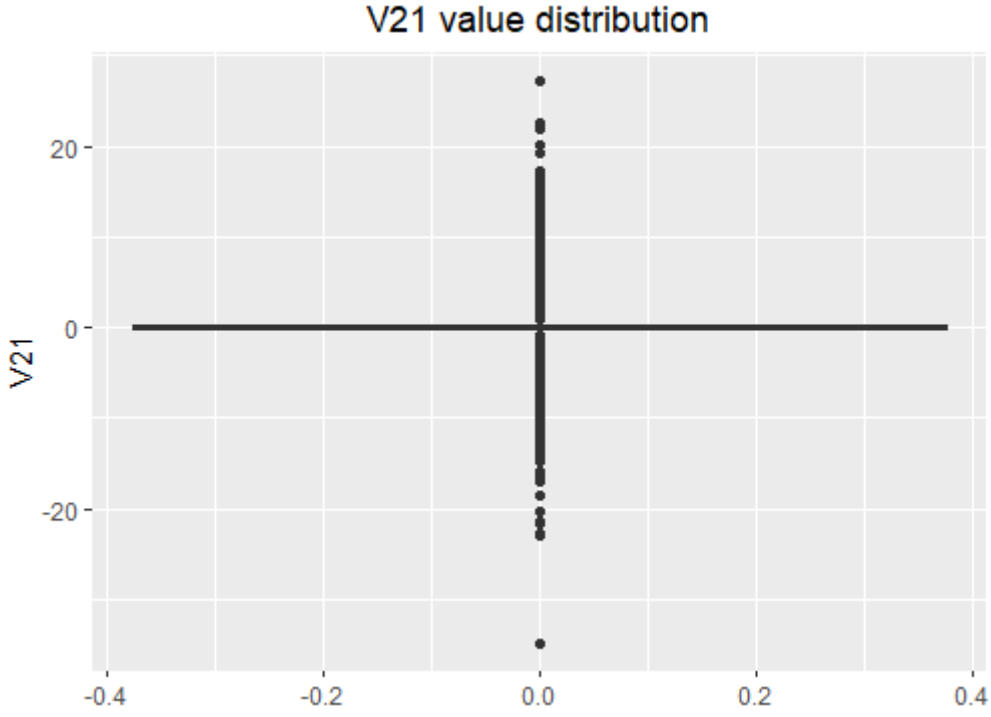

Figure 3.23. V21 distribution

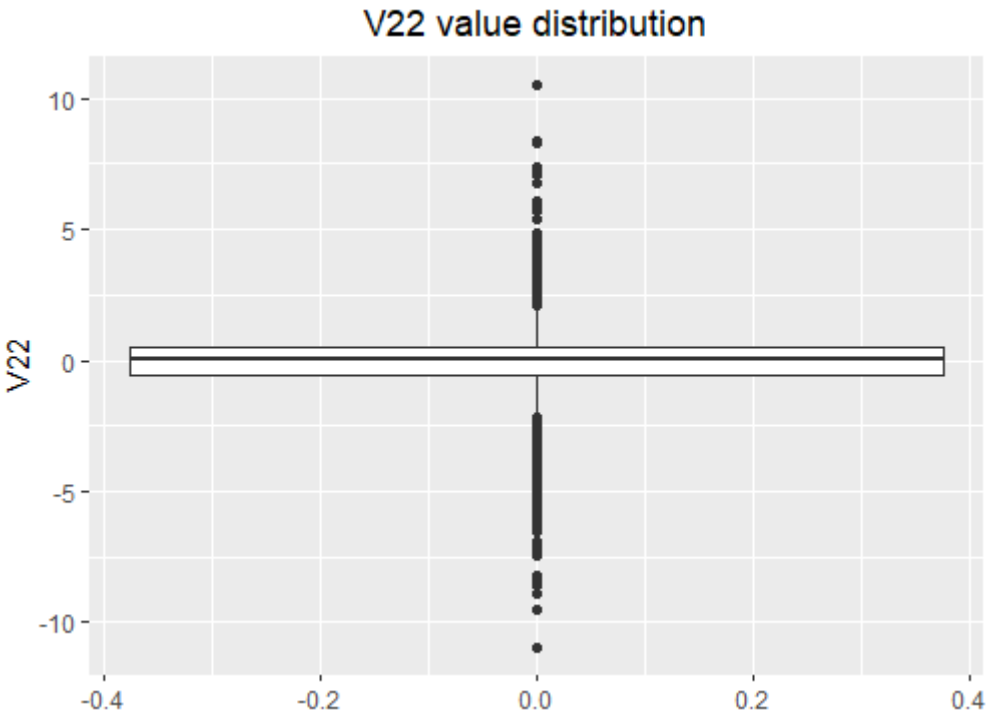

Figure 3.24. V22 distribution

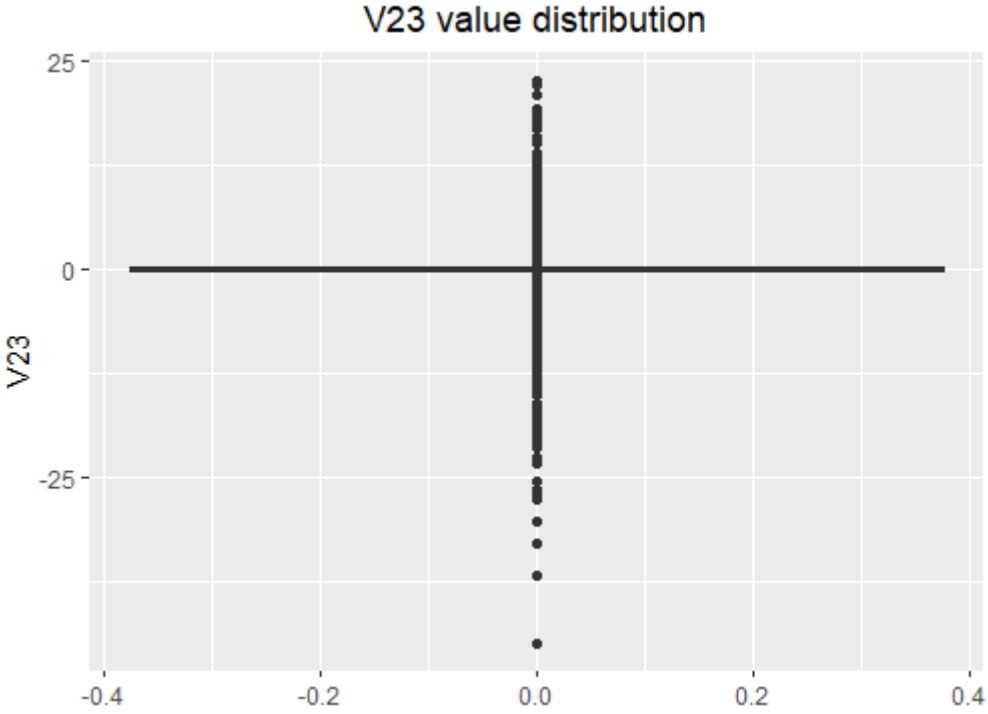

Figure 3.25. V22 distribution

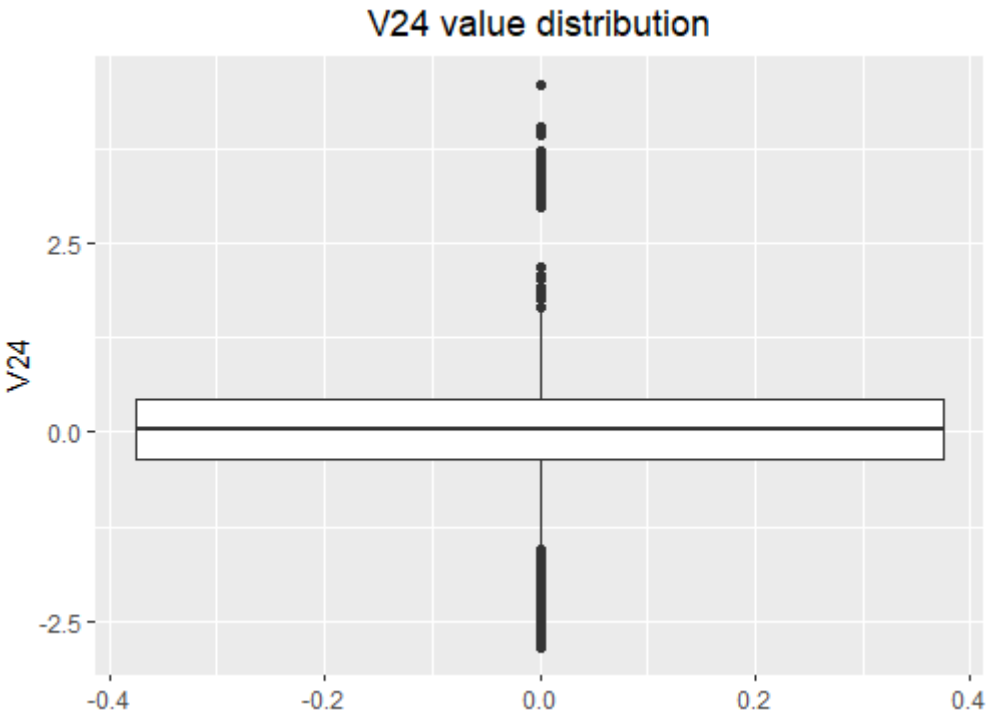

Figure 3.26. V24 distribution

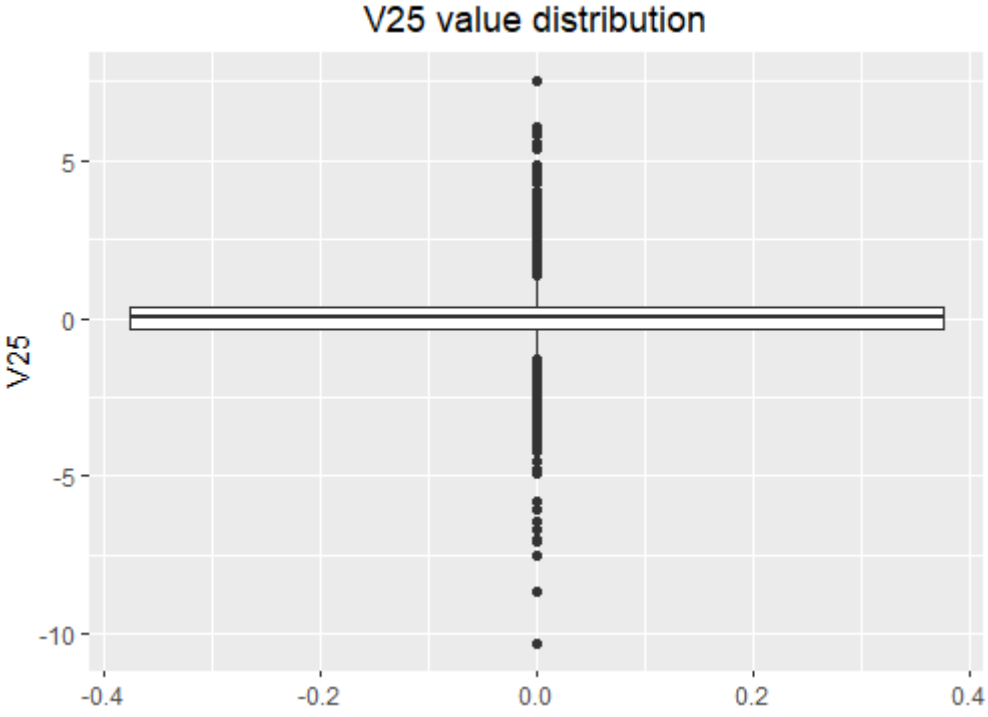

Figure 3.27. V25 distribution

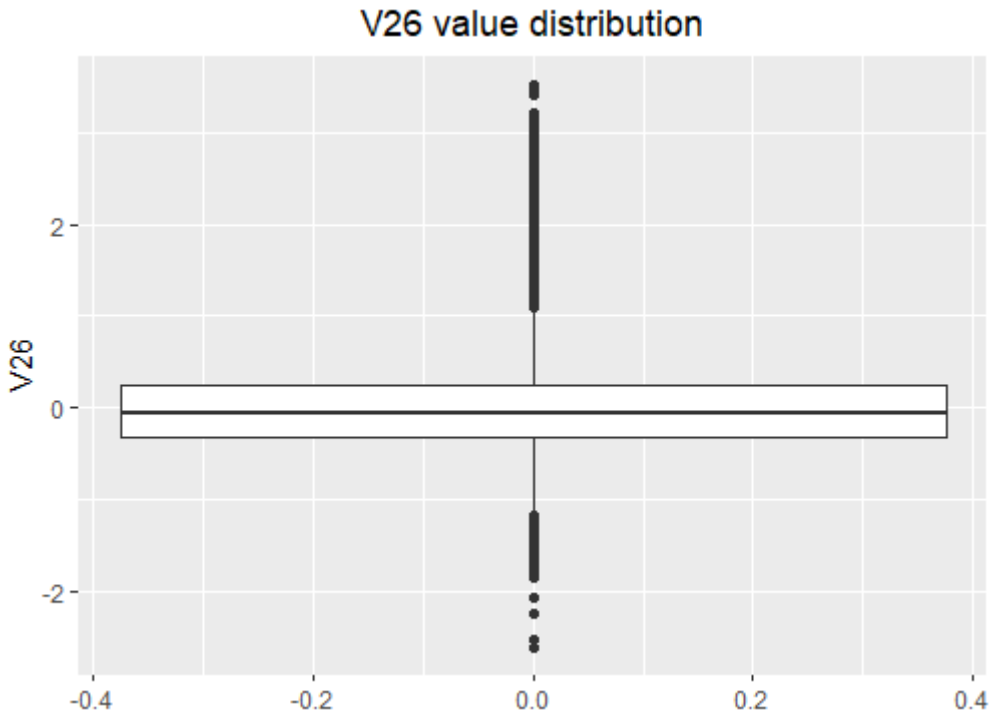

Figure 3.28. V26 distribution

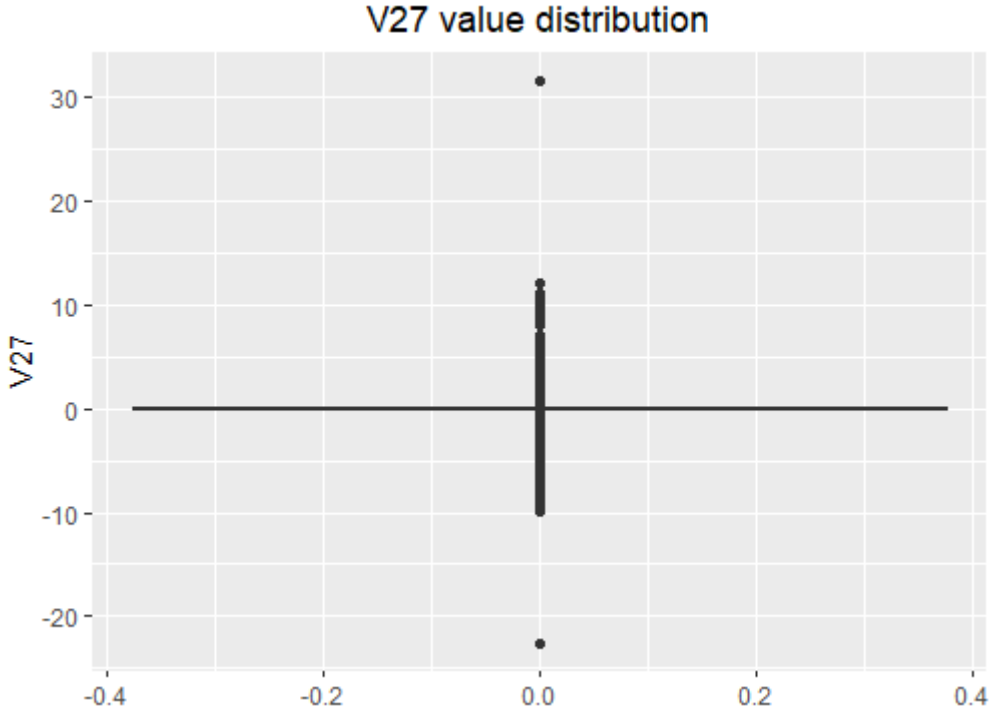

Figure 3.29. V27 distribution

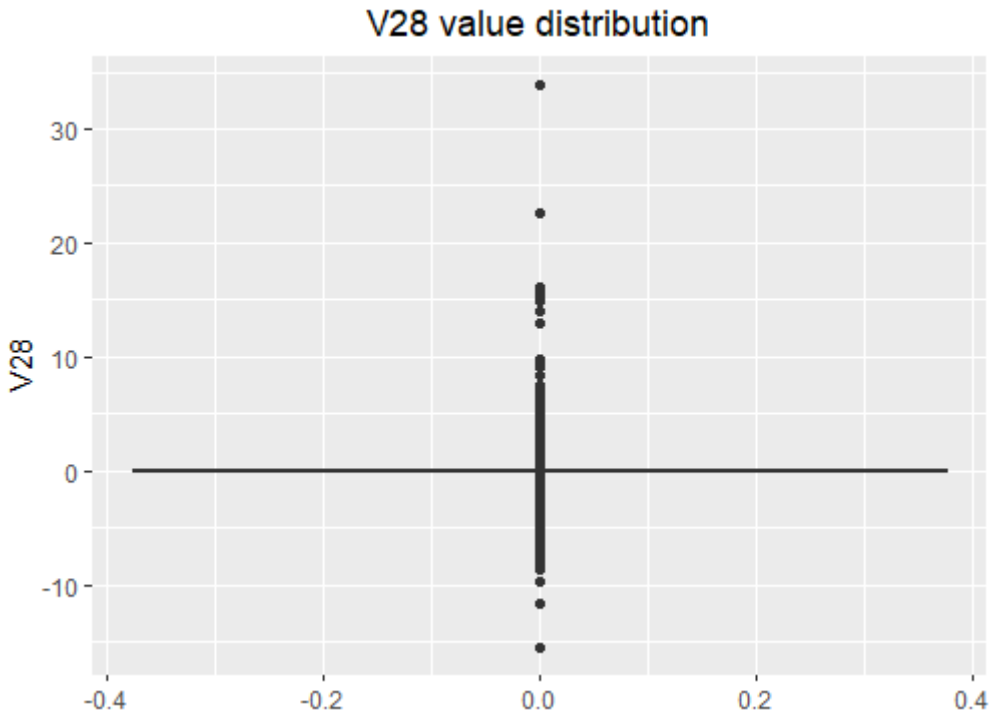

Figure 3.30. V28 distribution

3.6. Numerical variable correlation coefficient matrix

Correlation coefficient is an index to measure the degree of linearity between variables. Simply put, if variable X gradually increases, variable Y also gradually increases, it means that X and Y are positively correlated; otherwise, X gradually increases and Y gradually decreases, it means that X and Y are negatively correlated. If the change in X does not affect the change in Y, that is, X and Y are independent of each other, then X and Y are independent. The correlation coefficient matrix is a matrix composed of the correlation coefficients of pairwise variables in the data set. For example, if there are two columns X and Y, and the correlation coefficient between X and Y is 0.7, the correlation coefficient matrix is shown in the following table.

For the correlation coefficient, let's start with its formula. In general, the formula for the correlation coefficient is:

$$cor = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

In fact, the correlation coefficient is the covariance of X and Y divided by the standard deviation of X and the standard deviation of Y, so the correlation coefficient can be regarded as covariance, a special covariance after eliminating the impact of two variables' dimensions and standardization.

Then the corresponding correlation coefficient can reflect whether the changes of the two variables are the same or opposite. If the changes are the same, the result is positive; otherwise, it is negative. And a more important property is that it simply reflects the degree of similarity between the two variables. Take Table 3.2 as an example.

Table 3.2. Correlation coefficient matrix

	X	Y
X	1	0.7
Y	0.7	1

As can be seen from the figure above, the correlation coefficient between X and X itself is 1, and the correlation coefficient between X and Y is equal to the correlation coefficient between Y and X.

In addition, another characteristic of the correlation coefficient is that its value can only change between +1 and -1, because the correlation coefficient is covariance divided by standard deviation. Even if the fluctuation range of X or Y is large, the corresponding change of standard deviation and covariance will be larger. In this way, the numerator and denominator will be larger, then the change trend will be cancelled out, and the corresponding fluctuation will be similar to that in the small case. Therefore, the value of correlation coefficient can only be within this interval.

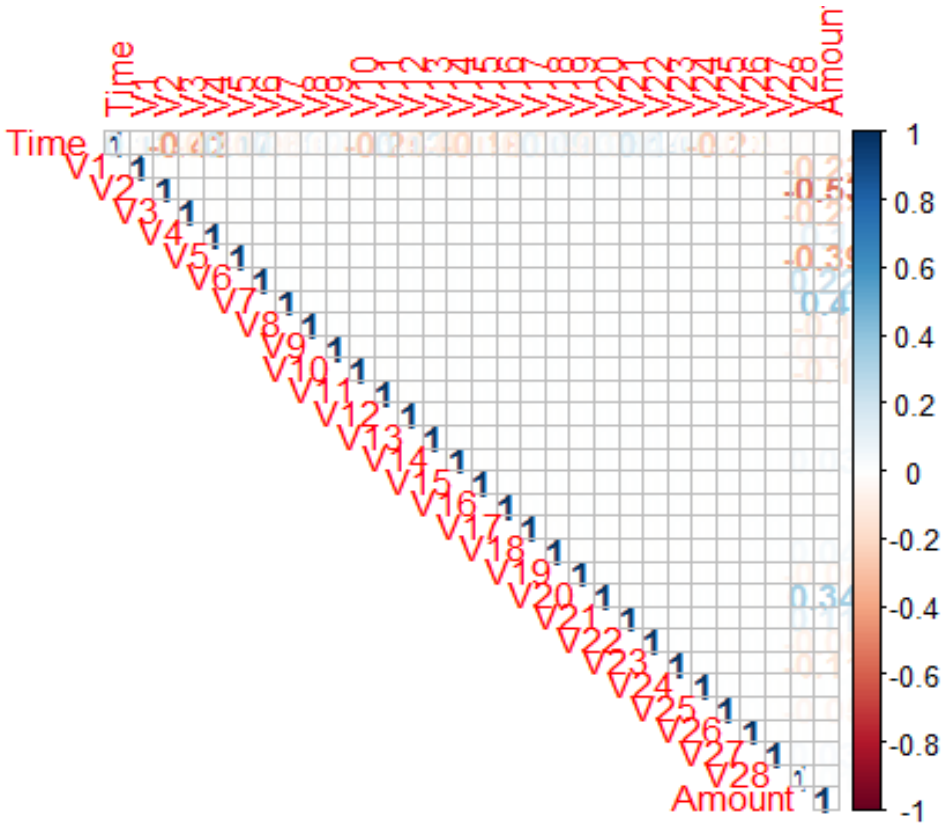

Figure 3.31. Correlation coefficient

In this paper, the correlation coefficient matrix is shown in Figure 3.31, where blue represents positive correlation and red represents negative correlation, where the absolute value of correlation coefficient is higher than 0.6 is correlation, and no column with correlation coefficient higher than 0.6 is shown in the figure below. That is, there is no multicollinearity problem in the 30 variables. Since the data is processed by dimensionality reduction, there is no multicollinearity in theory, which is consistent with the description of the phase relation matrix.

4. Data processing

4.1. Normalized processing

Normalization is to limit the data that needs to be processed to a certain range after processing. First of all, normalization is for the convenience of later data processing, and second, to ensure that the program runs faster convergence. The specific function of normalization is to generalize the statistical distribution of uniform samples. In this paper, the maximum-minimum value normalization is adopted to control all the features uniformly between 0-1. The data distribution after normalization is shown in the figure below.

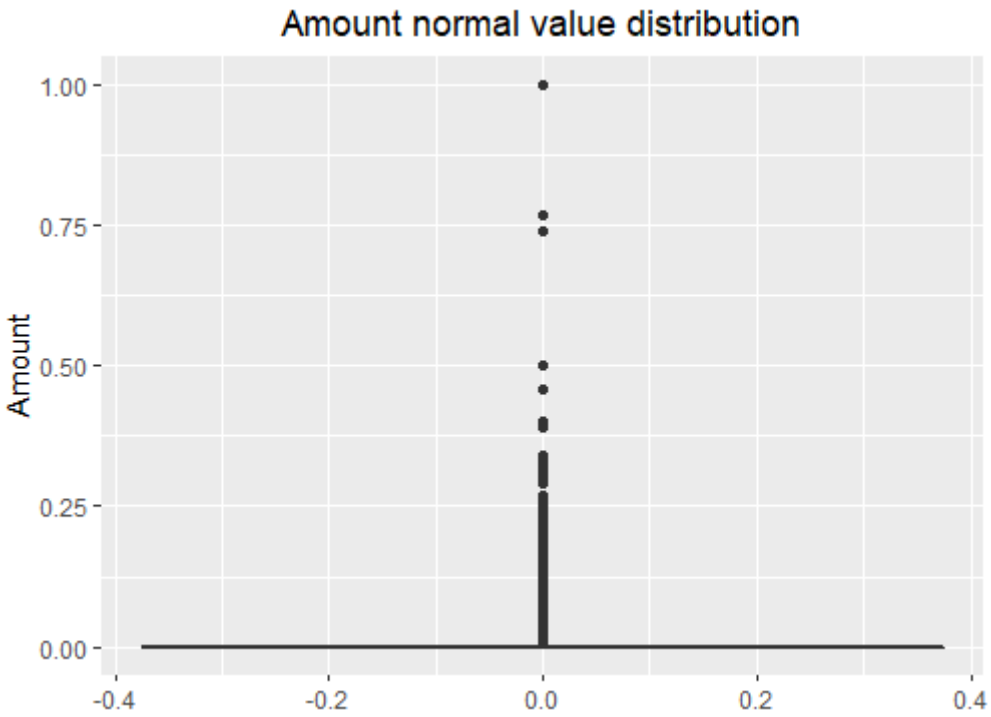

Figure 4.1. Amount distribution

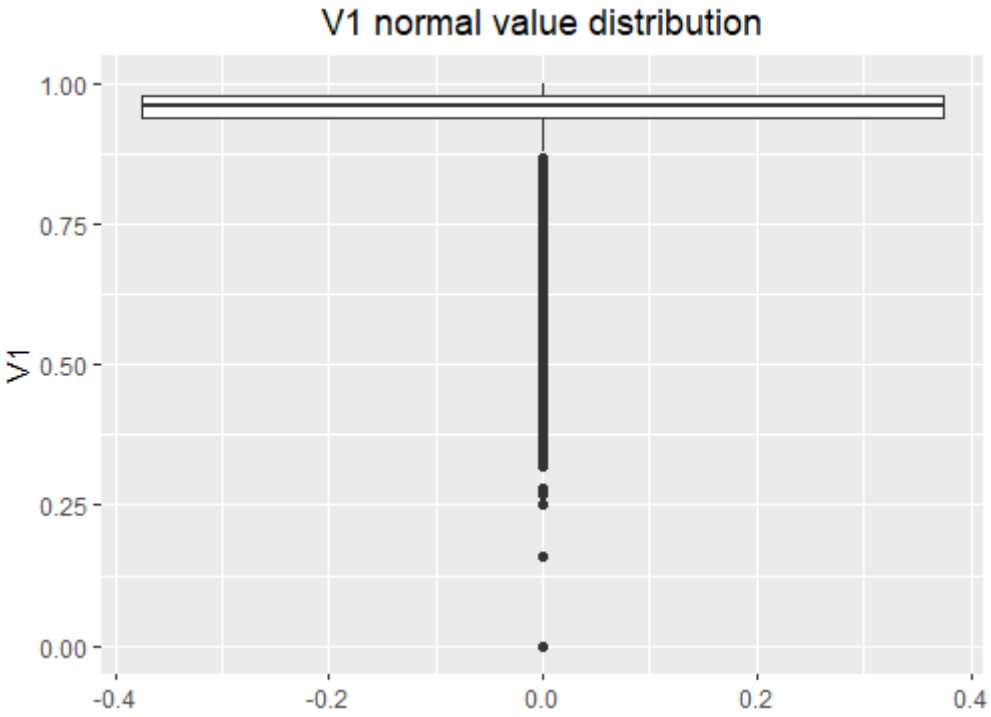

Figure 4.2. V1 distribution

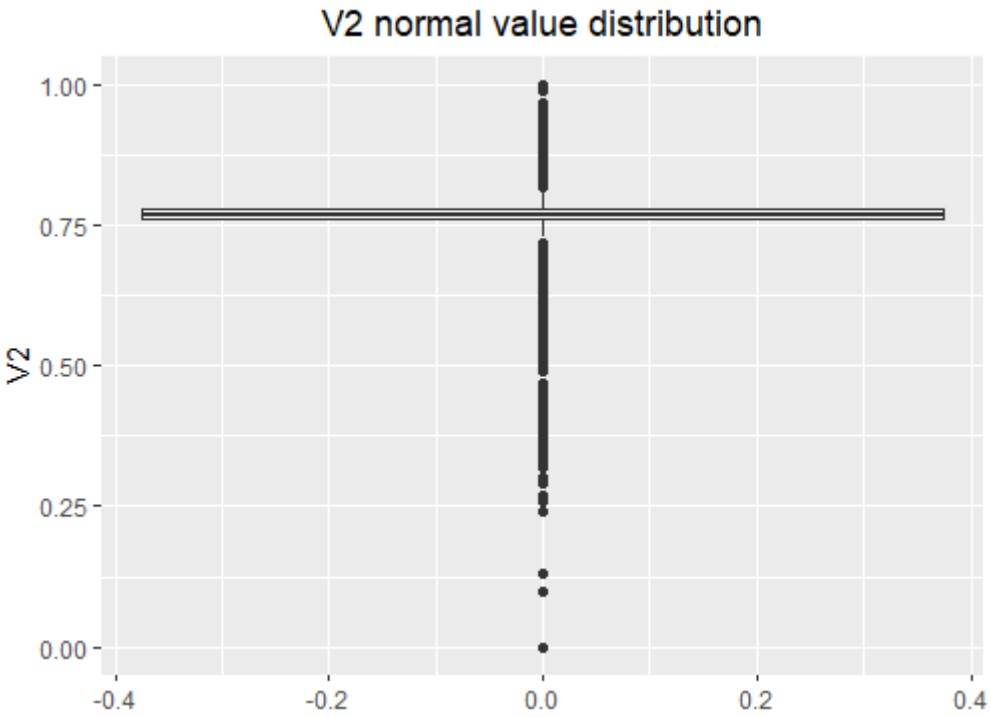

Figure 4.3. V3 distribution

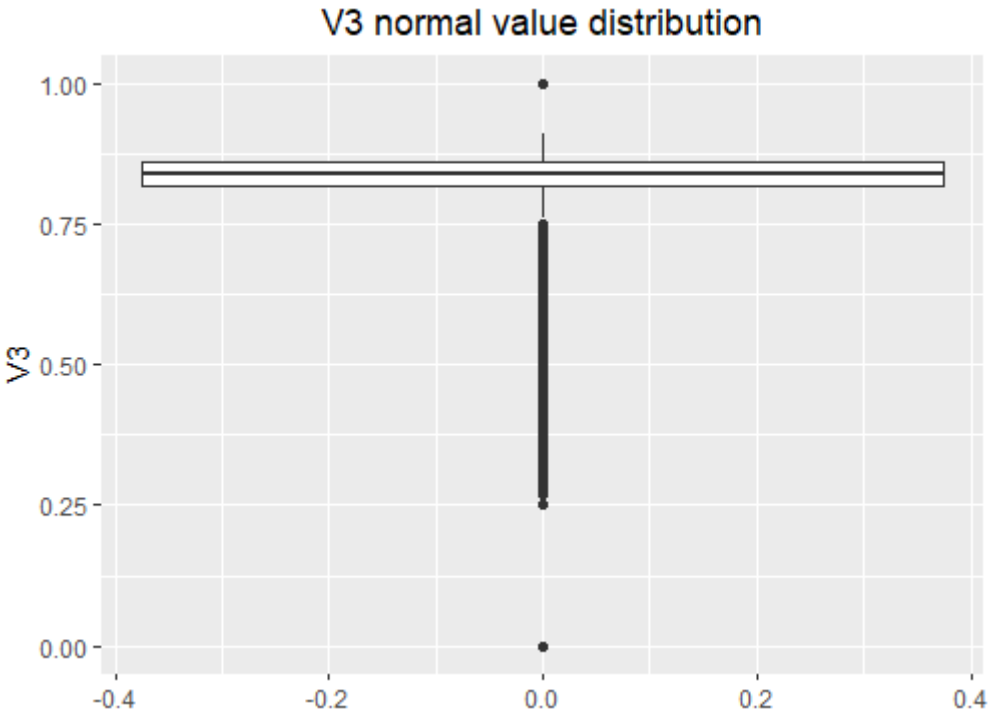

Figure 4.4. V4 distribution

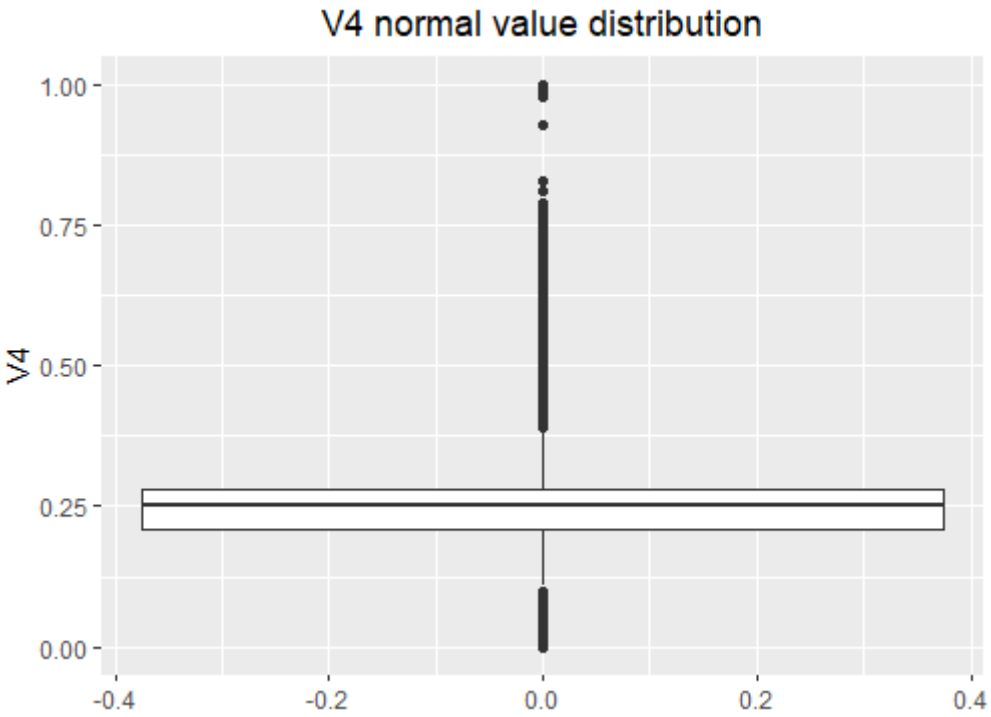

Figure 4.5. V4 distribution

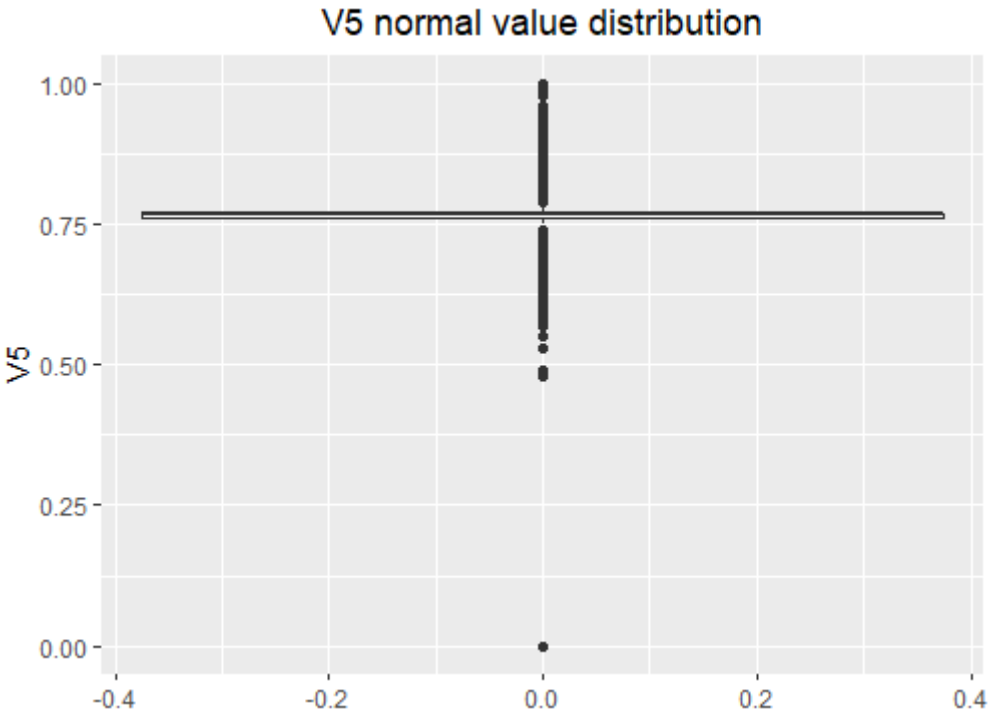

Figure 4.6. V5 distribution

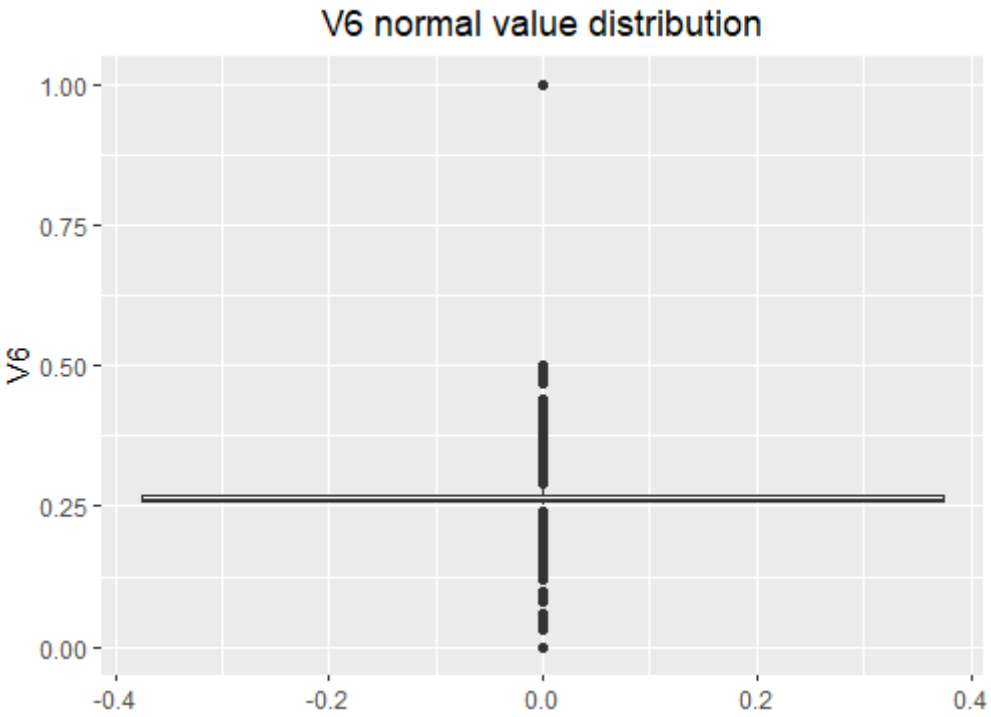

Figure 4.7. V6 distribution

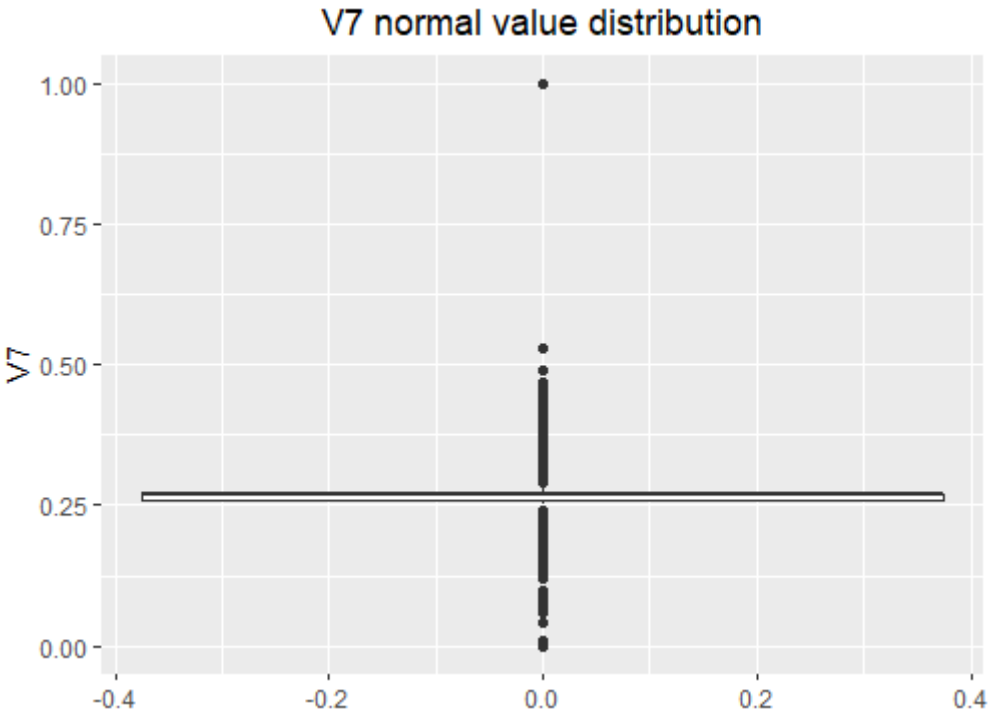

Figure 4.8. V7 distribution

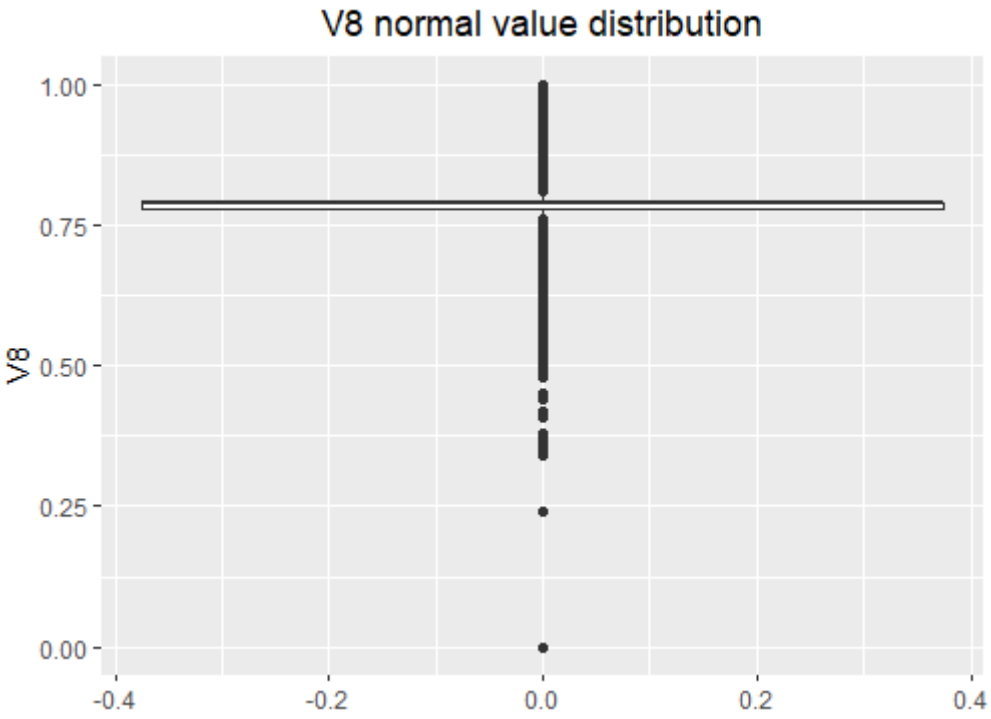

Figure 4.9. V8 distribution

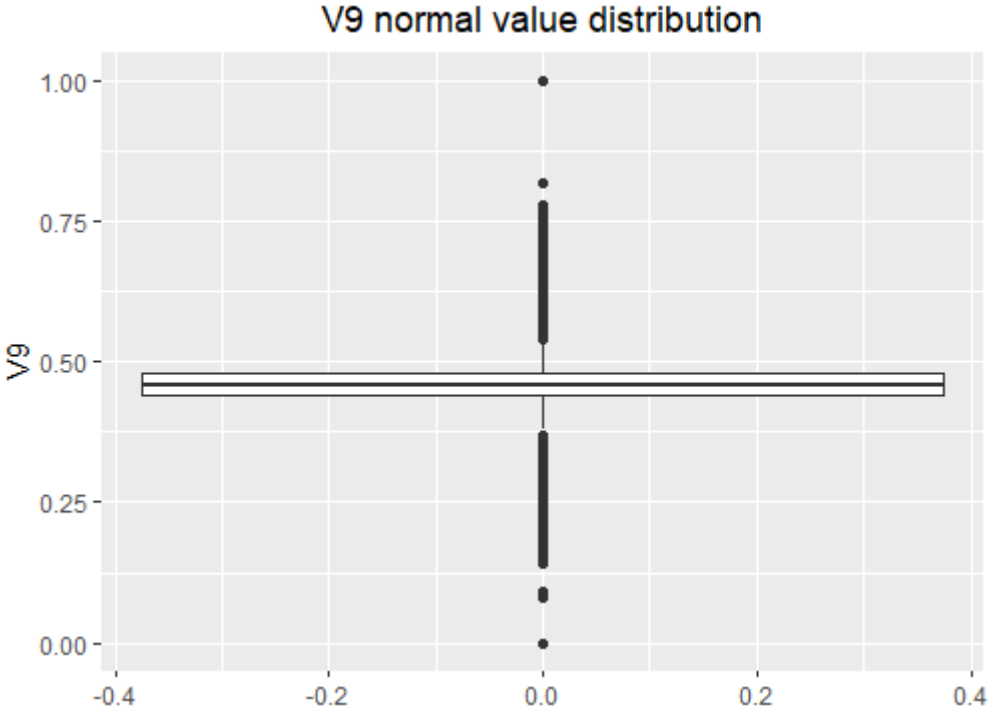

Figure 4.10. V9 distribution

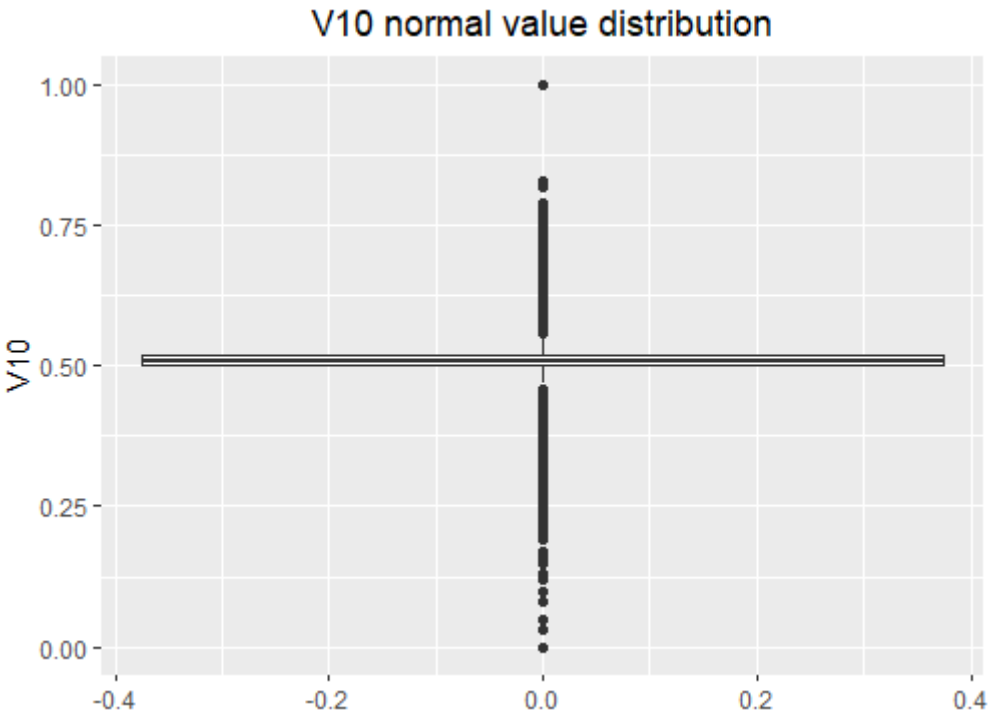

Figure 4.11. V10 distribution

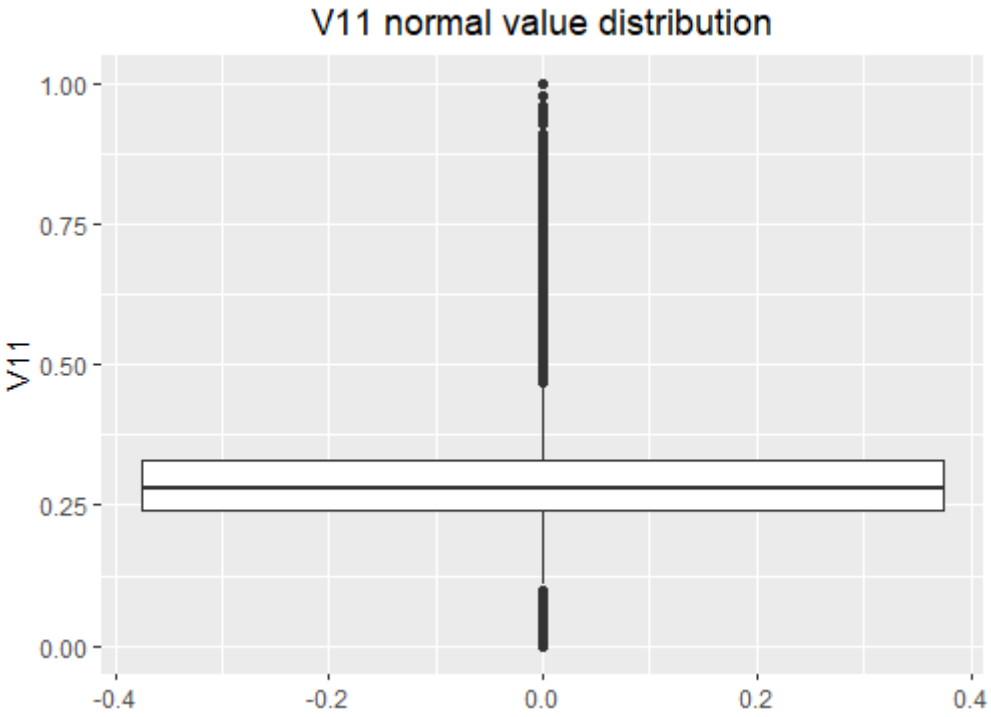

Figure 4.12. V11 distribution

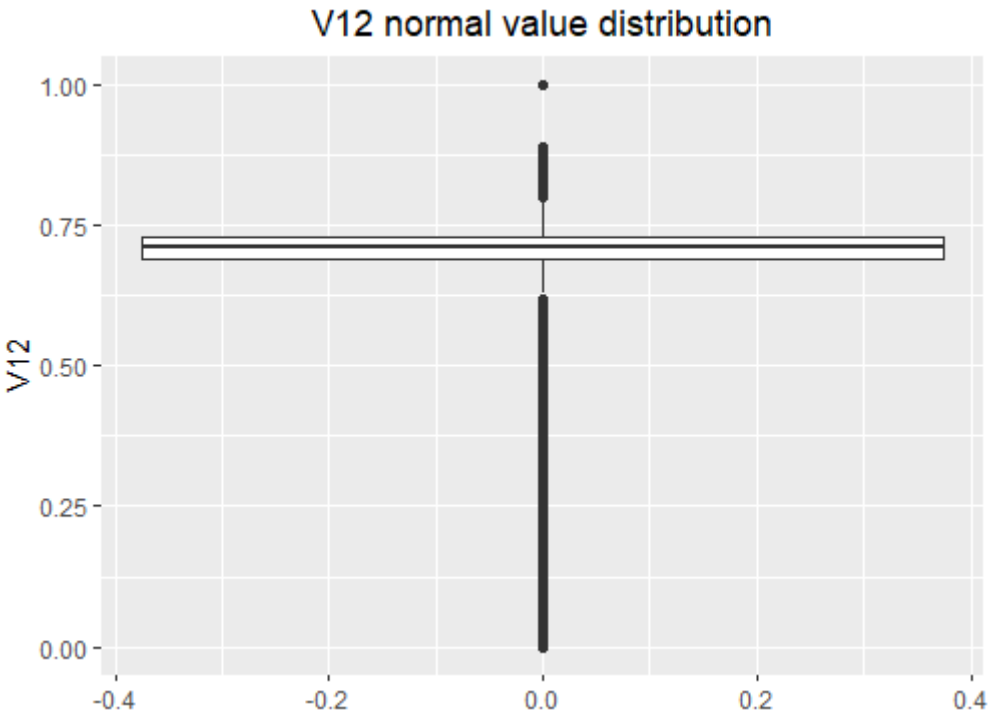

Figure 4.13. V12 distribution

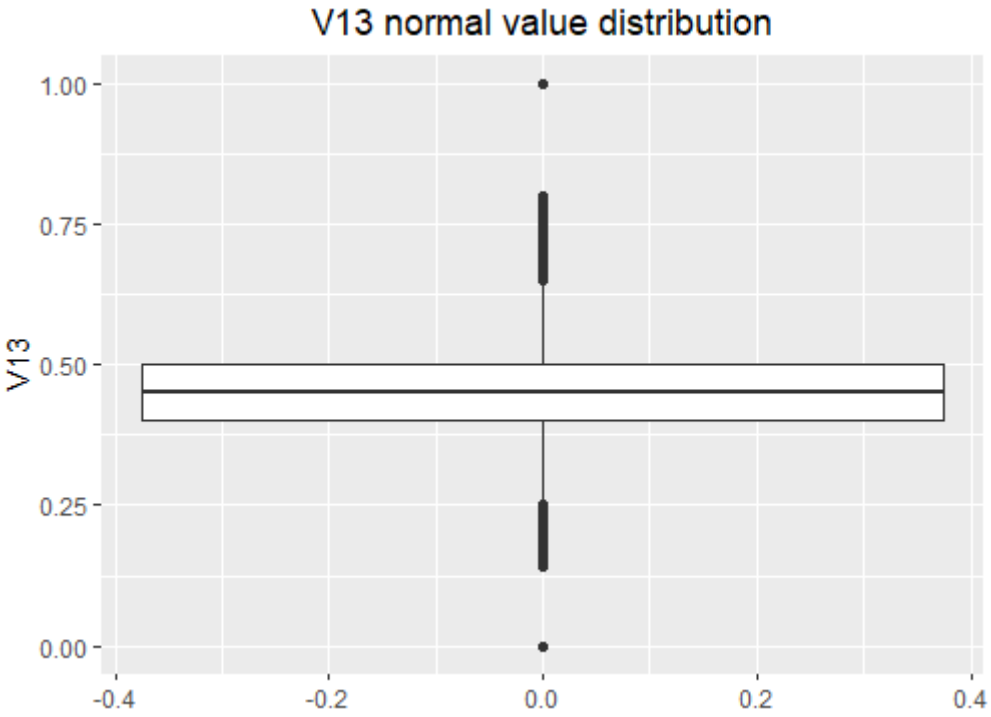

Figure 4.14. V13 distribution

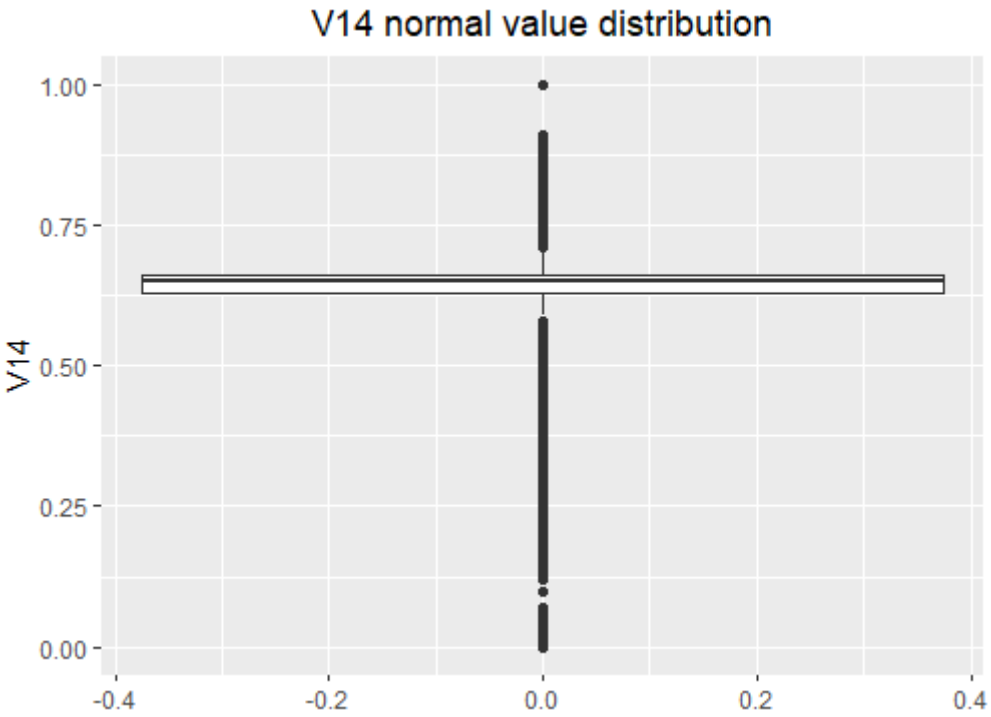

Figure 4.15. V14 distribution

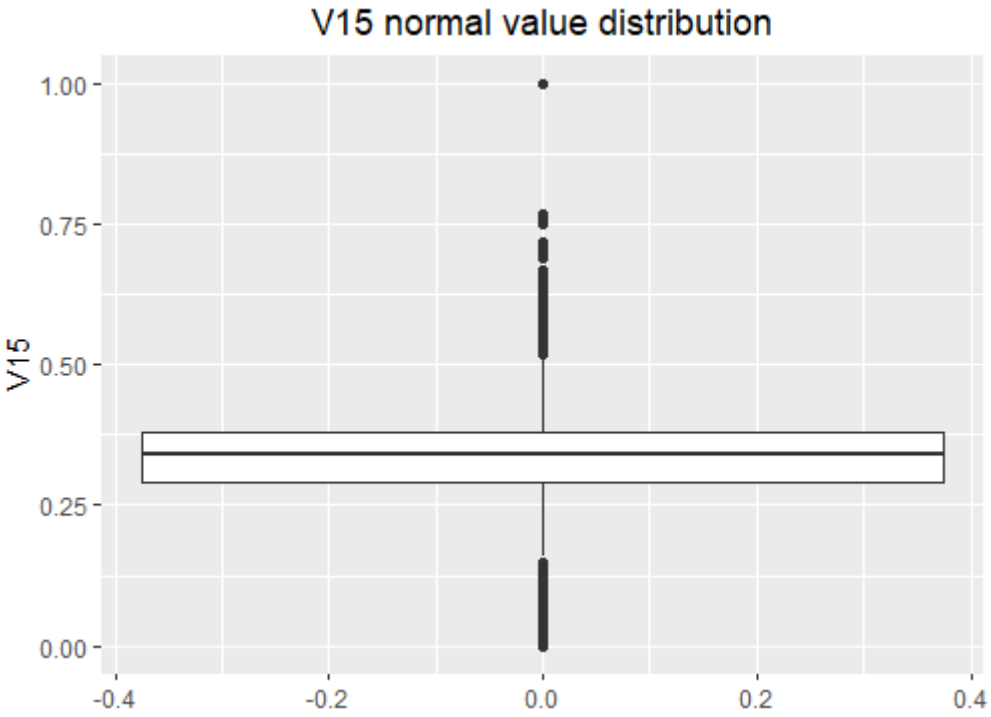

Figure 4.16. V15 distribution

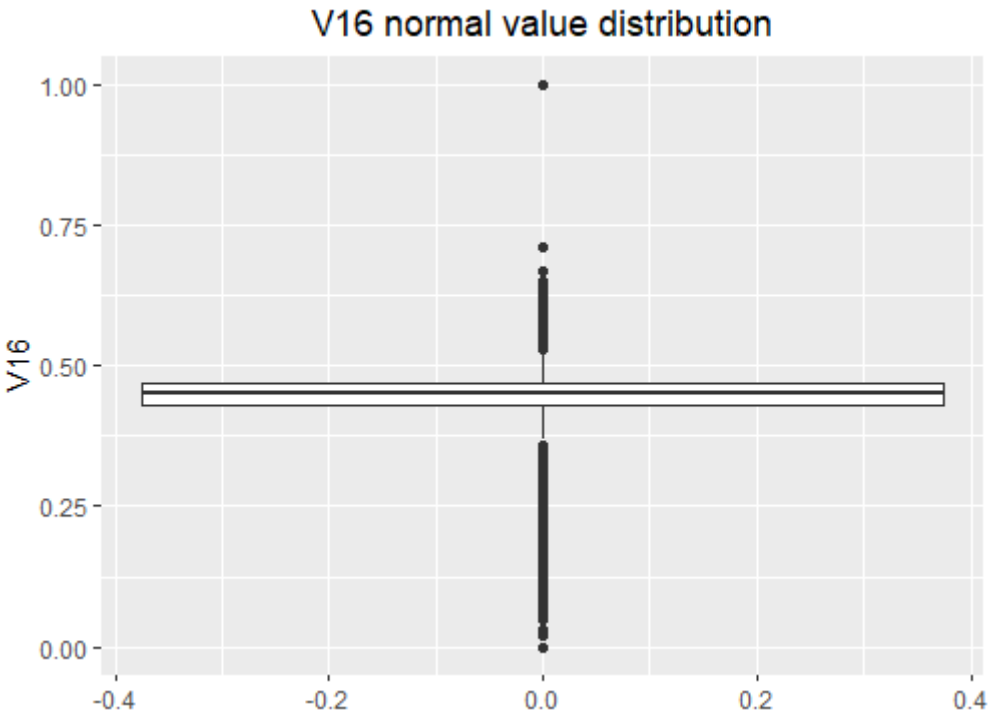

Figure 4.17. V16 distribution

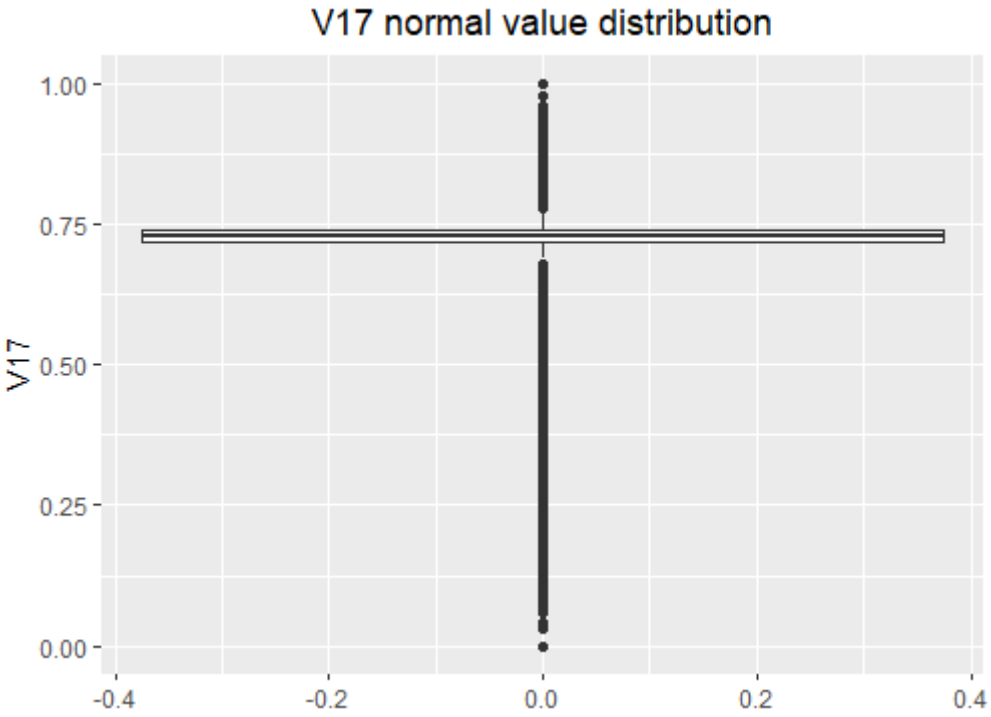

Figure 4.18. V17 distribution

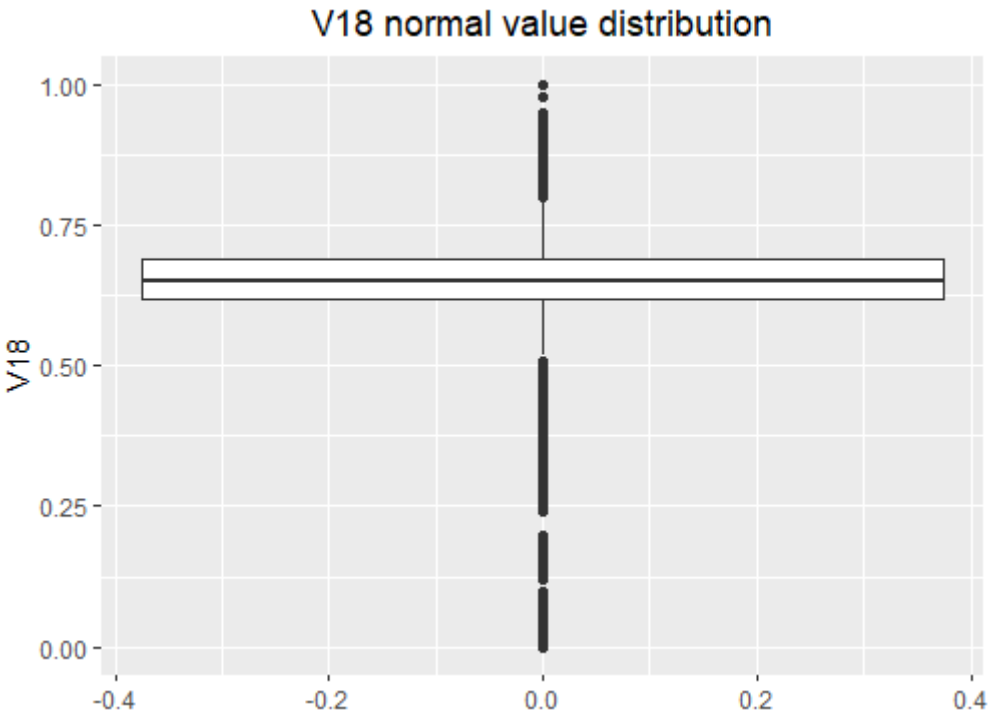

Figure 4.19. V18 distribution

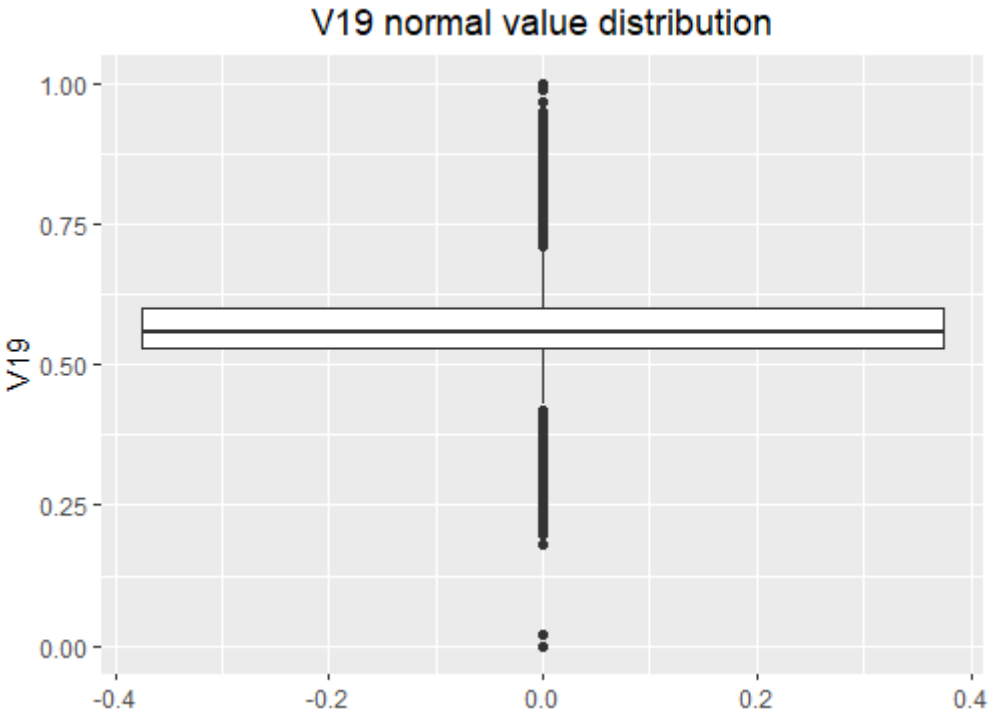

Figure 4.20. V19 distribution

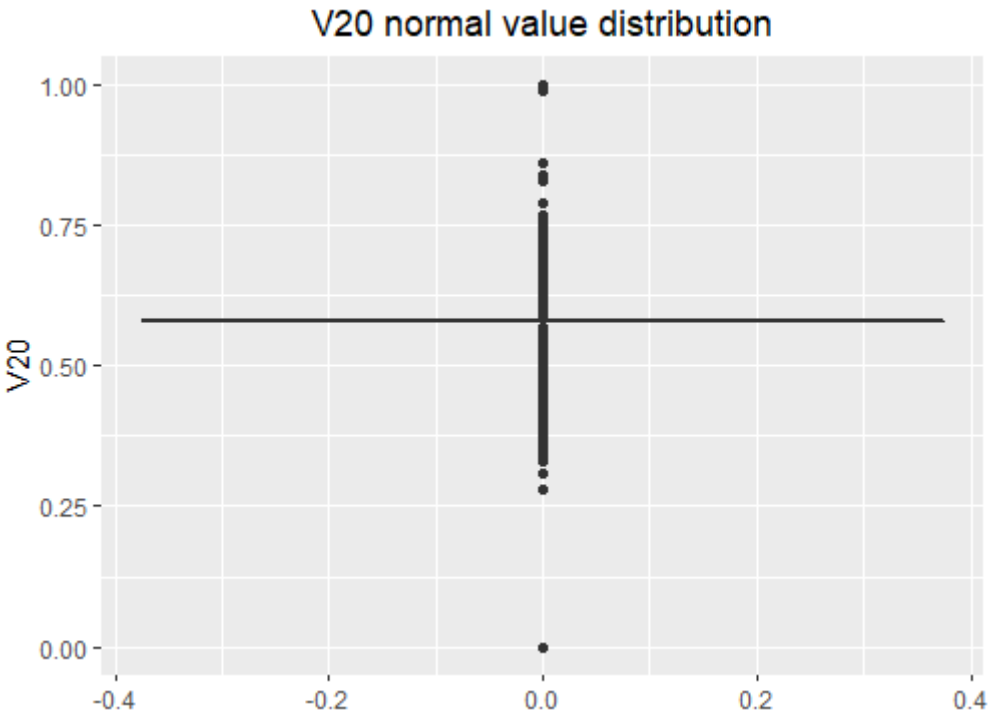

Figure 4.21. V20 distribution

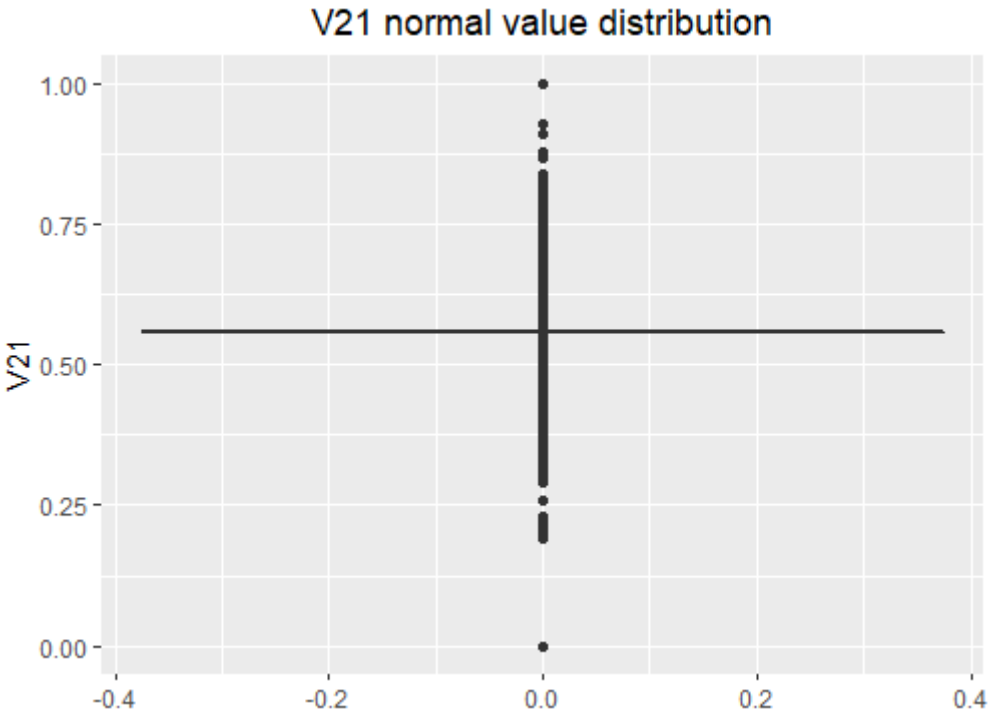

Figure 4.22. V21 distribution

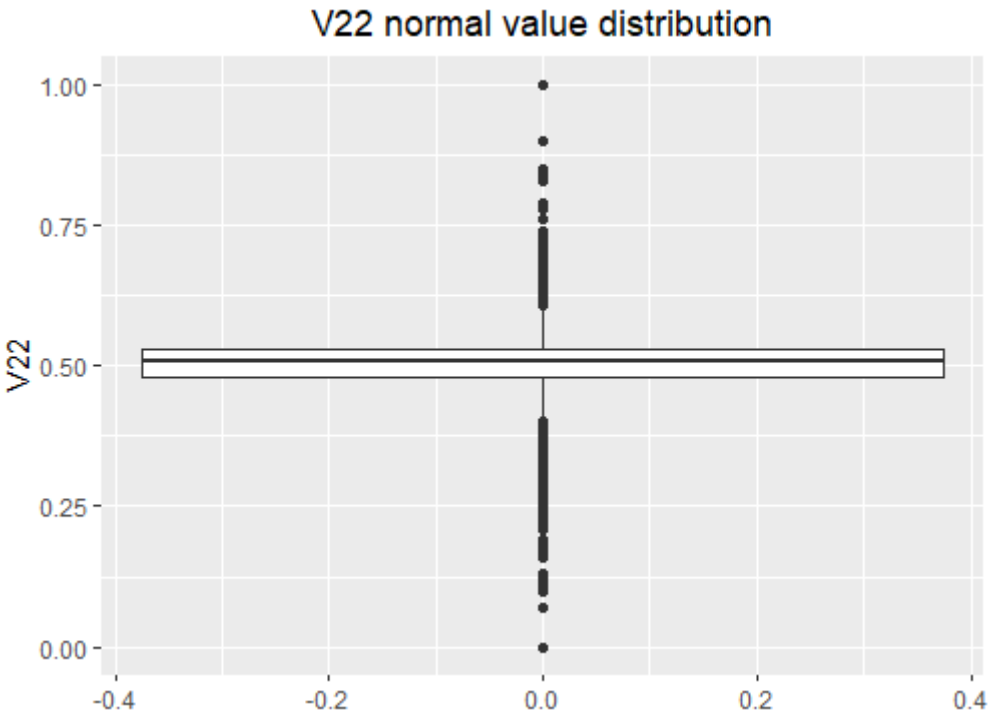

Figure 4.23. V22distribution

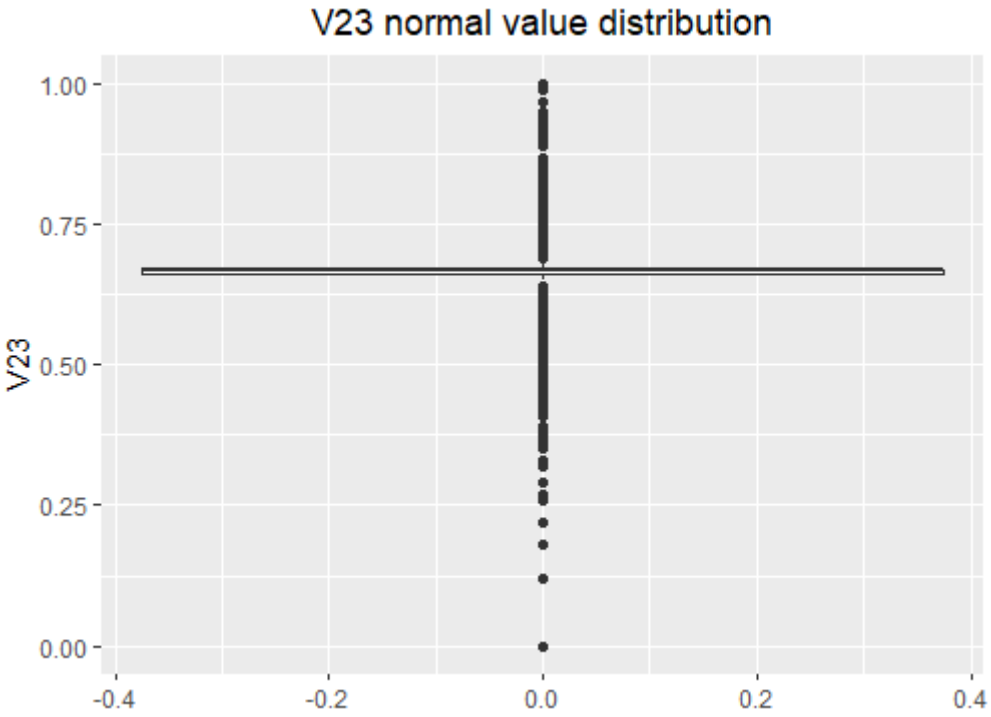

Figure 4.24. V23distribution

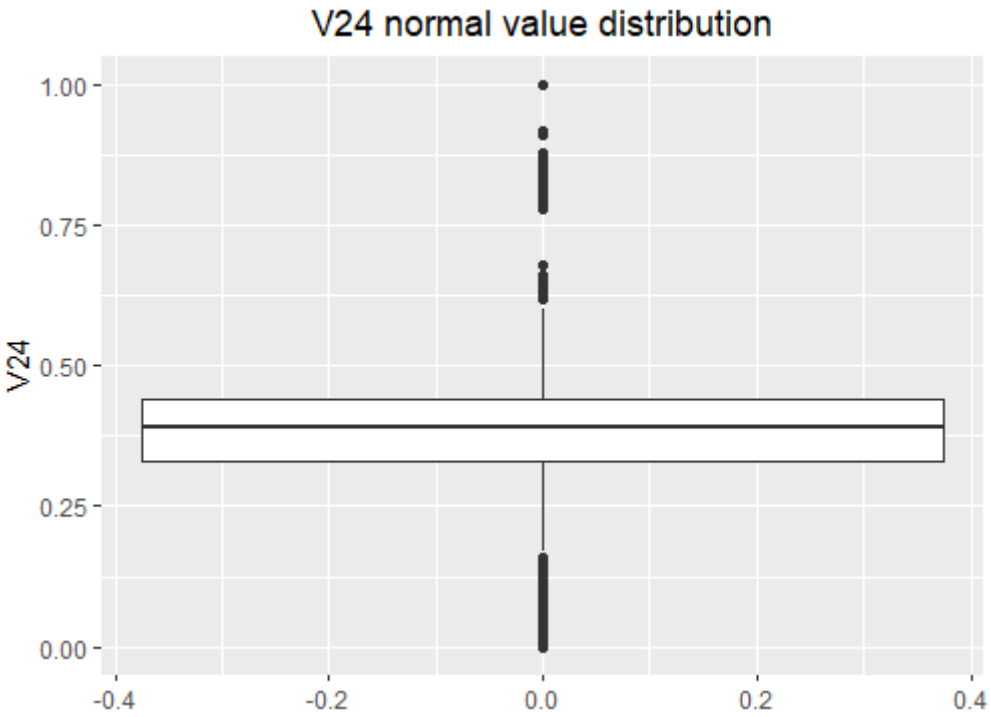

Figure 4.25. V24distribution

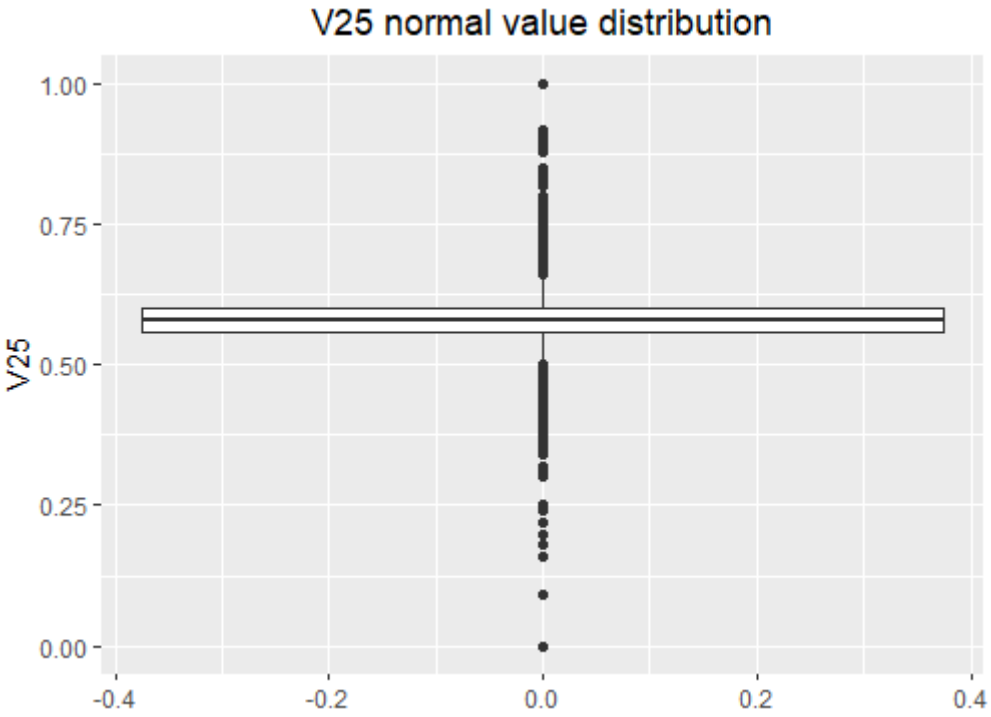

Figure 4.26. V25distribution

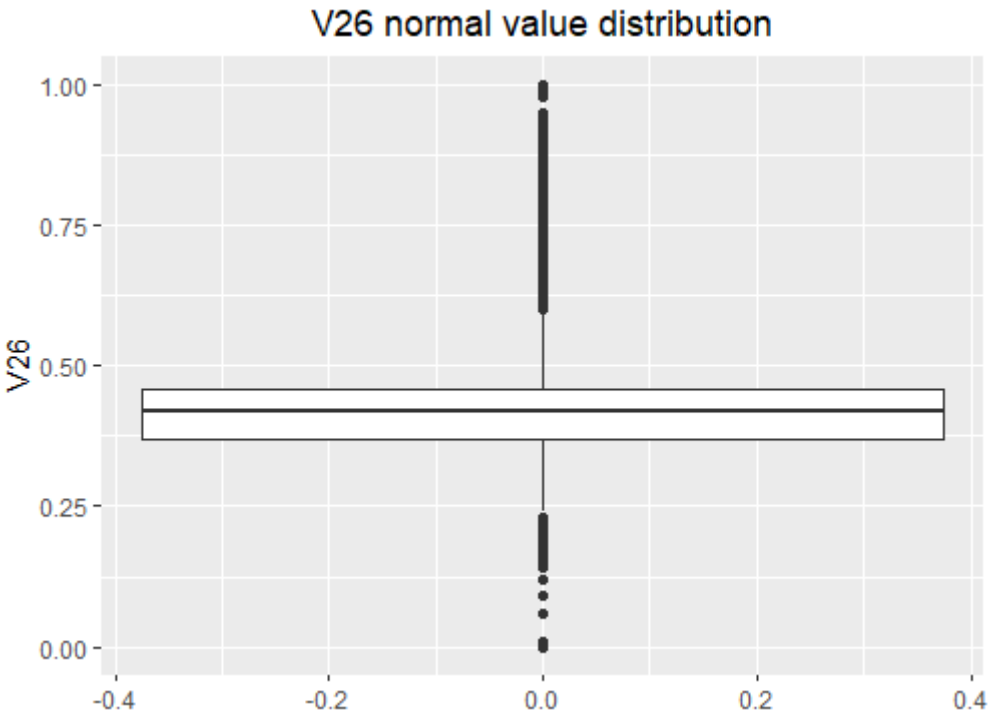

Figure 4.27. V26distribution

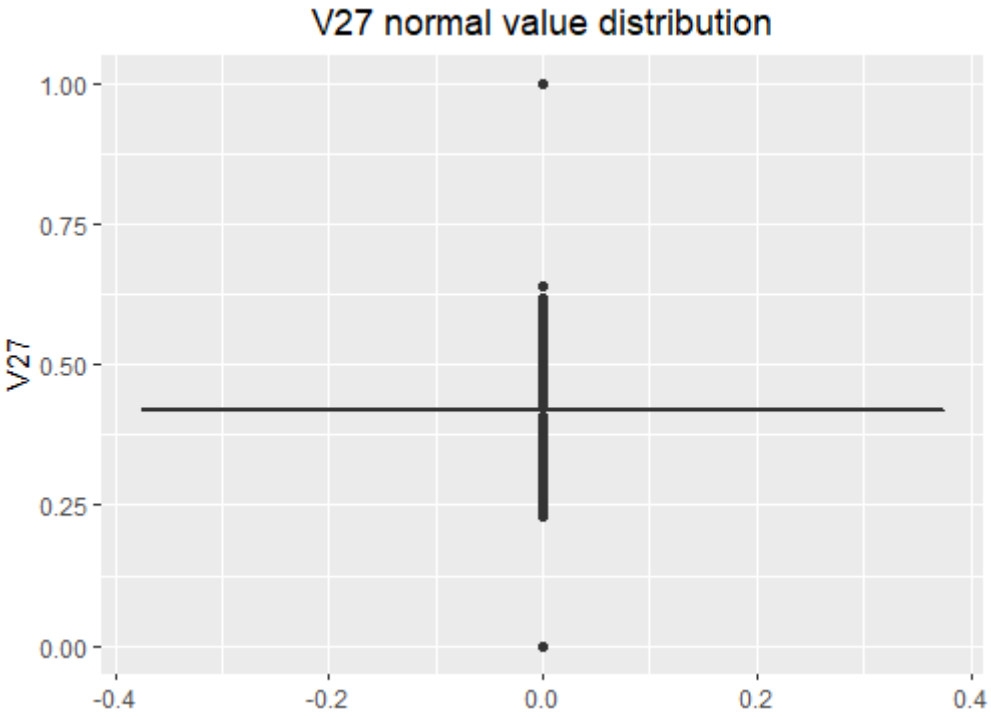

Figure 4.28. V27distribution

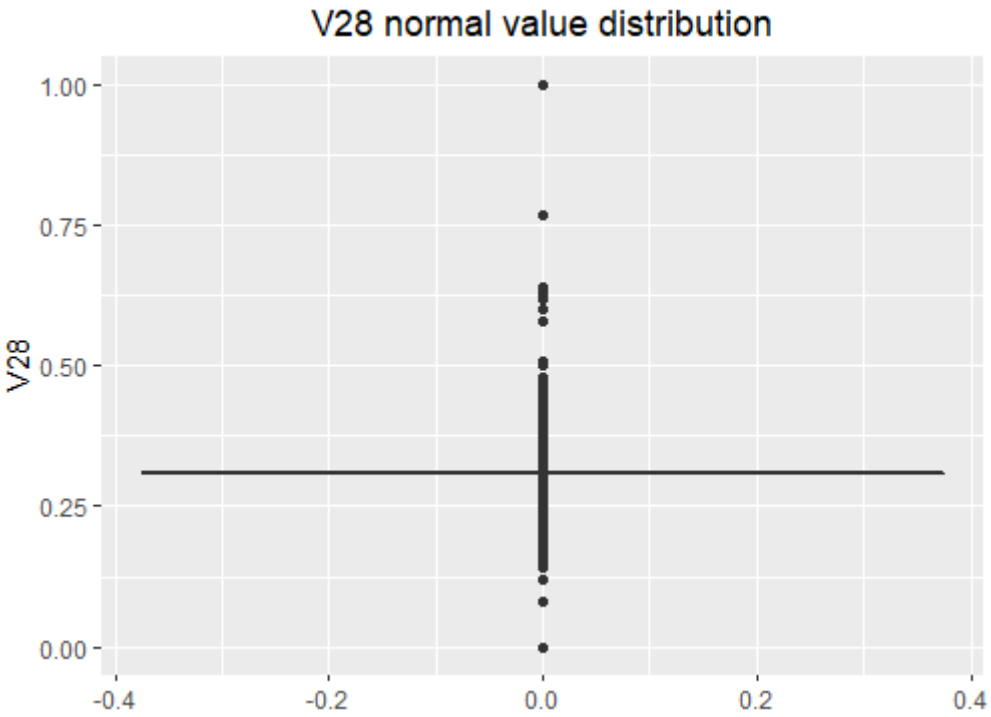

Figure 4.29. V28distribution

5. Sample processing

5.1. Whether there is sample imbalance

The data used in this paper has the problem of sample imbalance. In the data set, the sample population has positive sample 284315 and negative sample 492. The ratio of positive and negative samples was 1:578. In train, there were 342 positive samples and 199,022 negative samples, and the ratio of positive and negative samples was 1:581; in test, there were 150 positive samples and 85,293 negative samples, and the ratio of positive and negative samples was 1:569. When the positive and negative samples remain 1:1 or the ratio is close to 1:1, the sample is considered to be balanced, and the sample is unbalanced. Do not process the unbalanced data temporarily and observe the effect of the model.

6. Build model

6.1. Model used

logistic regression:

$$h_\theta(x) = \frac{1}{1+e^{\theta^T x}} \tag{4}$$

Logistic regression is a simple and common binary classification model, which obtains the class of an object by input the sequence of attributes of the object of unknown class. For common binary classification, logistic regression is divided by an interval distribution, that is, if the Y value is greater than or equal to 0.5, it belongs to the positive sample, if the Y value is less than 0.5, it belongs to the negative sample, so that the logistic regression model can be obtained.

6.2. Model effect (based on variable adjustment after feature selection)

AUC (Area Under Curve) is defined as the area under the ROC curve. The value of AUC is often used as the evaluation criterion of the model because the ROC curve cannot clearly explain which

classifier has the better effect in many cases. As a value, the classifier with larger AUC has the better effect.

```
Call:
glm(formula = label ~ ., family = binomial(link = "logit"), data = traindata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0916  -0.0294  -0.0199  -0.0135   4.2182

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  49.4173    19.0326   2.596  0.00942 **
Time         -0.1576     0.4560  -0.346  0.72965
V1            8.0793     3.0688   2.633  0.00847 **
V2            5.2334     8.0654   0.649  0.51642
V3            1.7877     3.6729   0.487  0.62645
V4           15.3247     1.8525   8.273  < 2e-16 ***
V5           19.2251    11.4588   1.678  0.09340 .
V6          -10.7968     8.4274  -1.281  0.20014
V7          -27.3869    13.1999  -2.075  0.03801 *
V8          -13.9979     3.6584  -3.826  0.00013 ***
V9           -8.1101     3.3995  -2.386  0.01705 *
V10         -32.7404     5.2963  -6.182 6.34e-10 ***
V11           1.4437     1.6112   0.896  0.37022
V12          -1.1598     2.5259  -0.459  0.64612
V13          -4.8460     1.2128  -3.996 6.45e-05 ***
V14         -13.8150     2.0540  -6.726 1.75e-11 ***
V15          -1.9814     1.3549  -1.462  0.14363
V16          -6.9993     4.2141  -1.661  0.09672 .
V17          -0.3898     2.7324  -0.143  0.88657
V18           0.6723     2.0510   0.328  0.74305
V19           1.7040     1.4081   1.210  0.22624
V20         -48.0173     9.0802  -5.288 1.24e-07 ***
V21          19.4069     4.4303   4.381 1.18e-05 ***
V22          10.6580     3.2629   3.266  0.00109 **
V23          -3.7935     5.5015  -0.690  0.49048
V24           0.7675     1.2913   0.594  0.55225
V25           0.3372     2.8155   0.120  0.90466
V26          -1.0711     1.3914  -0.770  0.44144
V27         -39.7640     7.8344  -5.076 3.86e-07 ***
V28         -11.1308     5.4552  -2.040  0.04131 *
Amount       37.7724    16.3271   2.313  0.02070 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5039.6  on 199486  degrees of freedom
Residual deviance: 1584.3  on 199456  degrees of freedom
AIC: 1646.3

Number of Fisher Scoring iterations: 11
```

Figure 6.1. Model overview

When the AUC is 0.5, it means that the classifier has no effect and is approximately random. The AUC in this report is shown below.

Training set AUC 0.97, test set AUC0.98

Train auc

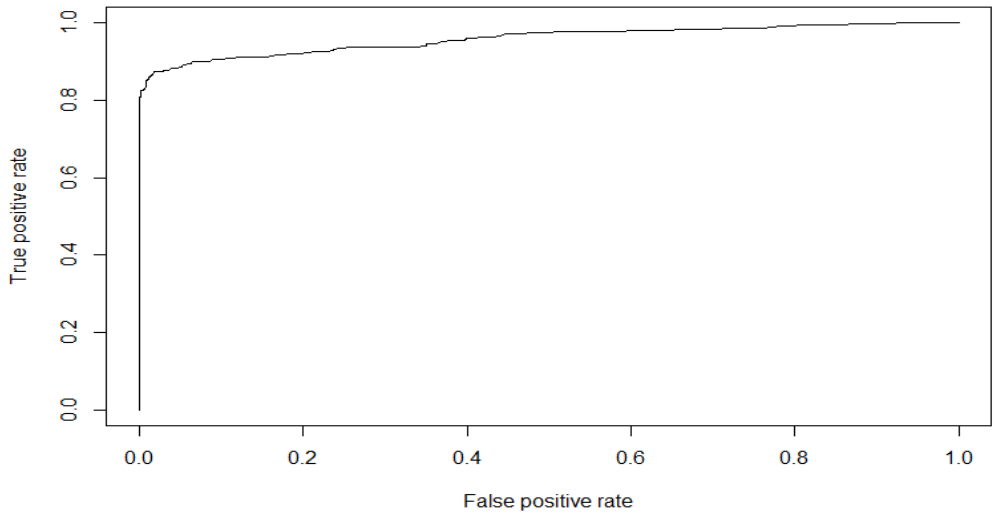

Figure 6.2. Training set AUC curve

In the logistic regression model, the correlation coefficient between features and y is shown in the figure below.

	Feature	Imporance
1	(Intercept)	49.42
2	Amount	37.77
3	V21	19.41
4	V5	19.23
5	V4	15.32
6	V22	10.66
7	V1	8.08
8	V2	5.23
9	V3	1.79
10	V19	1.70
11	V11	1.44
12	V24	0.77
13	V18	0.67
14	V25	0.34
15	Time	-0.16
16	V17	-0.39
17	V26	-1.07
18	V12	-1.16
19	V15	-1.98
20	V23	-3.79
21	V13	-4.85
22	V16	-7.00
23	V9	-8.11
24	V6	-10.80
25	V28	-11.13
26	V14	-13.81
27	V8	-14.00
28	V7	-27.39
29	V10	-32.74
30	V27	-39.76
31	V20	-48.02

Figure 6.3. correlation coefficient

7. Summary

Consider using undersampling + voting to deal with data imbalance. In this report, we can try to conduct multiple samples, each time undersampling according to positive and negative samples 1:1, and the data after each sample is used as train to train the model. Finally, N models are trained. When predicting test, N models are loaded to make predictions in turn, and the average value of N results is finally taken as the overall prediction effect. From the correlation coefficient between features and y, it can be seen that some of the correlation coefficients are negative, and the variables with negative correlation coefficients can be considered to be eliminated during optimization.

Acknowledgements

The authors gratefully acknowledge the financial support .

References

[1] Zhao Dawei. Research on Internet Consumer Finance driven by Big Data technology [J]. Finance and Economics, 2019:41-45.

[2] Zhuang Xudong. Library Catalog [J]. Information Forum, 2019:29-32.

[3] Yang Ying, ZHAO Shouxiang. Library Catalog [J]. Research on Credit Card Anti-fraud System under Internet Environment, 2019:50-52.

[4] Liu Jiawei. Reconstruction of credit risk control system by financial technology [J]. Financial Technology, 2018:74-75.

[5] Lu Sixing. Library Catalog [M]. Intelligence Leads Anti-fraud dual sword building miracle, 2019:38-39.