Exemplo motivacional

Prof. Eduardo Vargas Ferreira

Curso de Especialização em
Data Science & Big Data
Universidade Federal do Paraná

16 de março de 2018

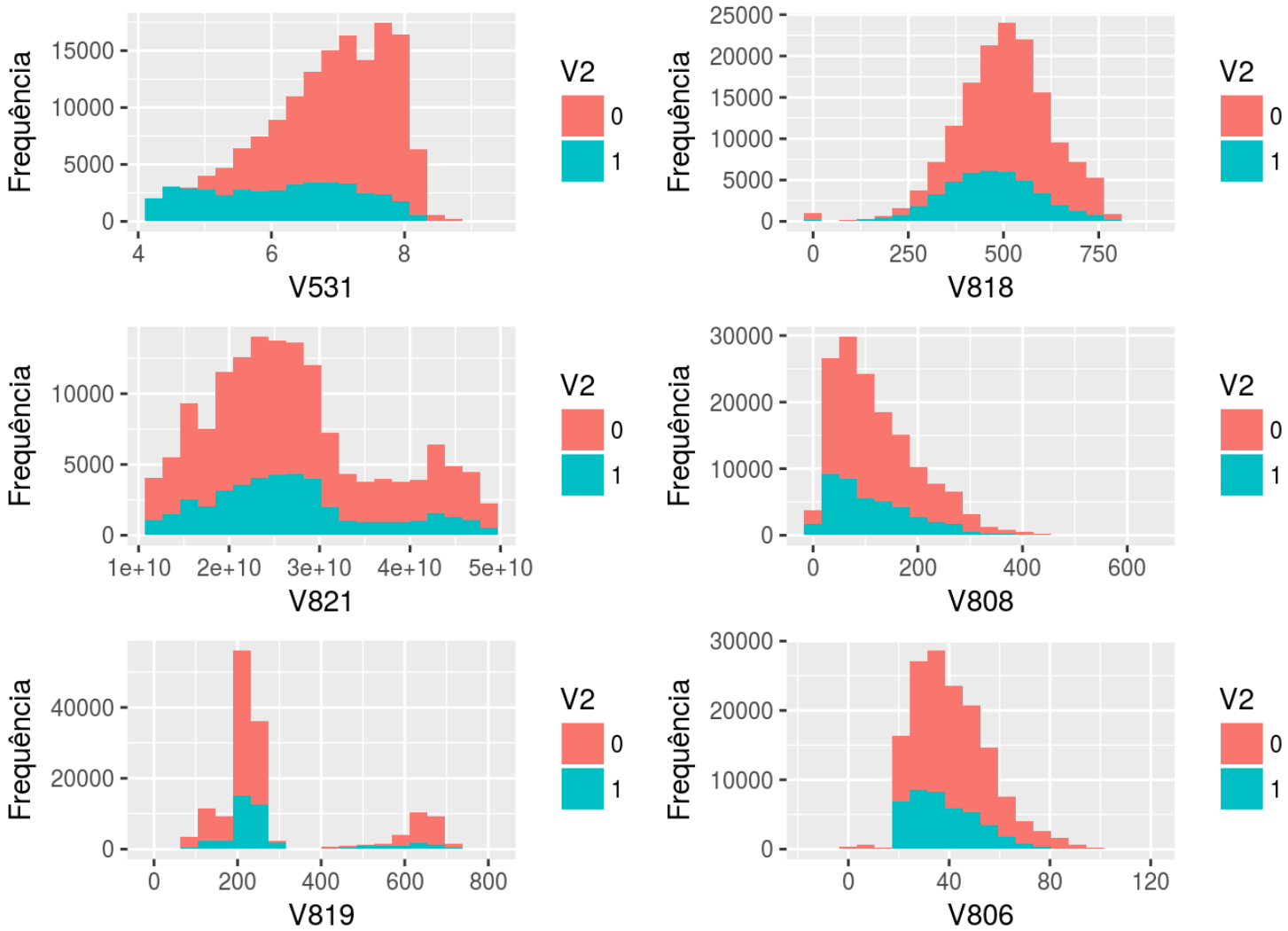
Machine Learning for medicine

- O problema apresentado se refere a um estudo sobre o tempo até a cura de uma doença.

- Acompanhou-se 145340 pacientes durante 24 meses. Todos recebendo o mesmo tipo de tratamento;

- O interesse é prever qual paciente responderá positivamente a essa intervenção e a velocidade de cura.

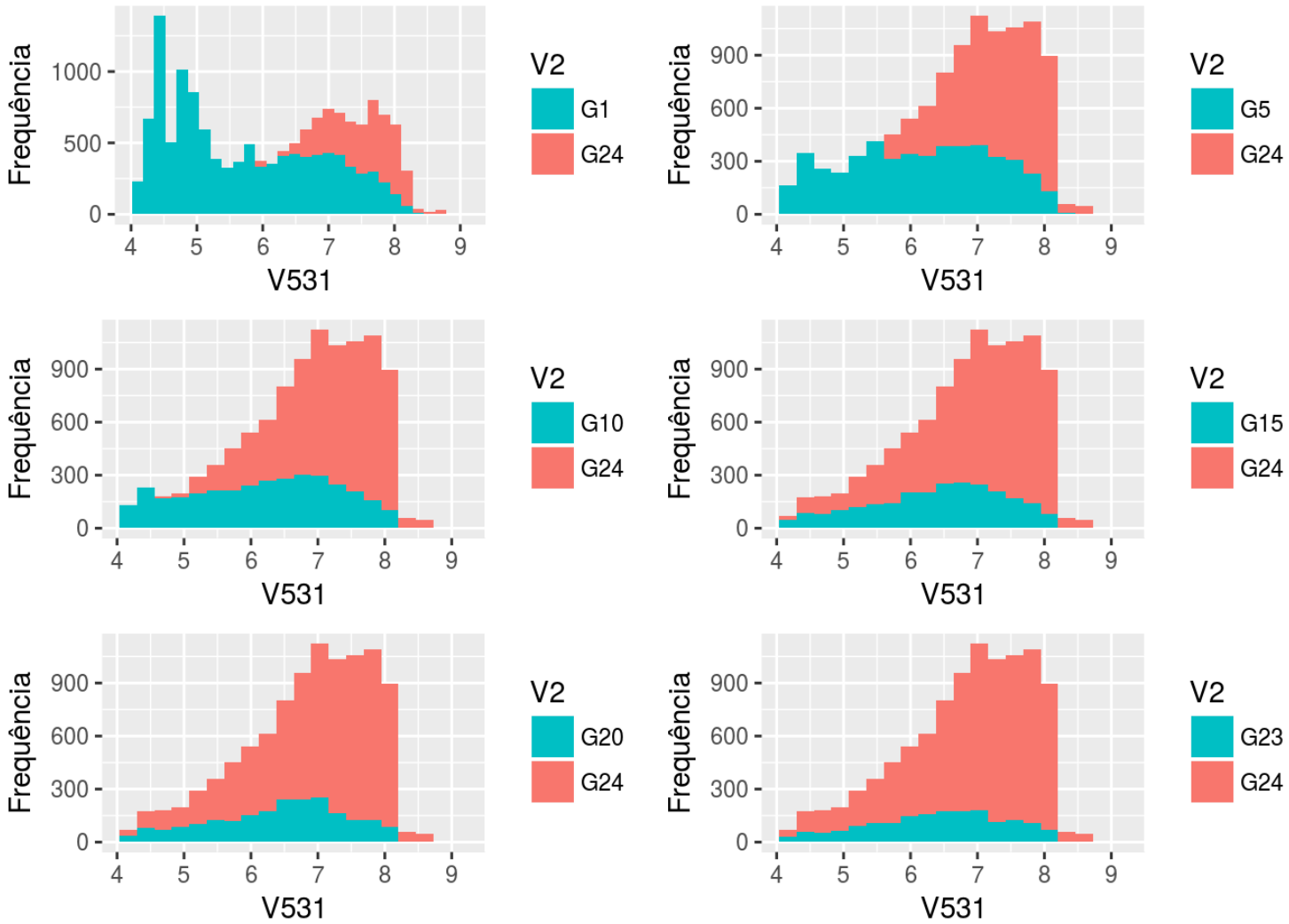
Análises


Frequência
15000
10000
5000
0
4
6
8
V531
V2
0
1
Frequência
25000
20000
15000
10000
5000
0
0
250
500
750
V818
V2
0
1
Frequência
10000
5000
0
1e+10
2e+10
3e+10
4e+10
5e+10
V821
V2
0
1
Frequência
30000
20000
10000
0
0
200
400
600
V808
V2
0
1
Frequência
40000
20000
0
0
200
400
600
800
V819
V2
0
1
Frequência
30000
20000
10000
0
0
40
80
120
V806
V2
0
1

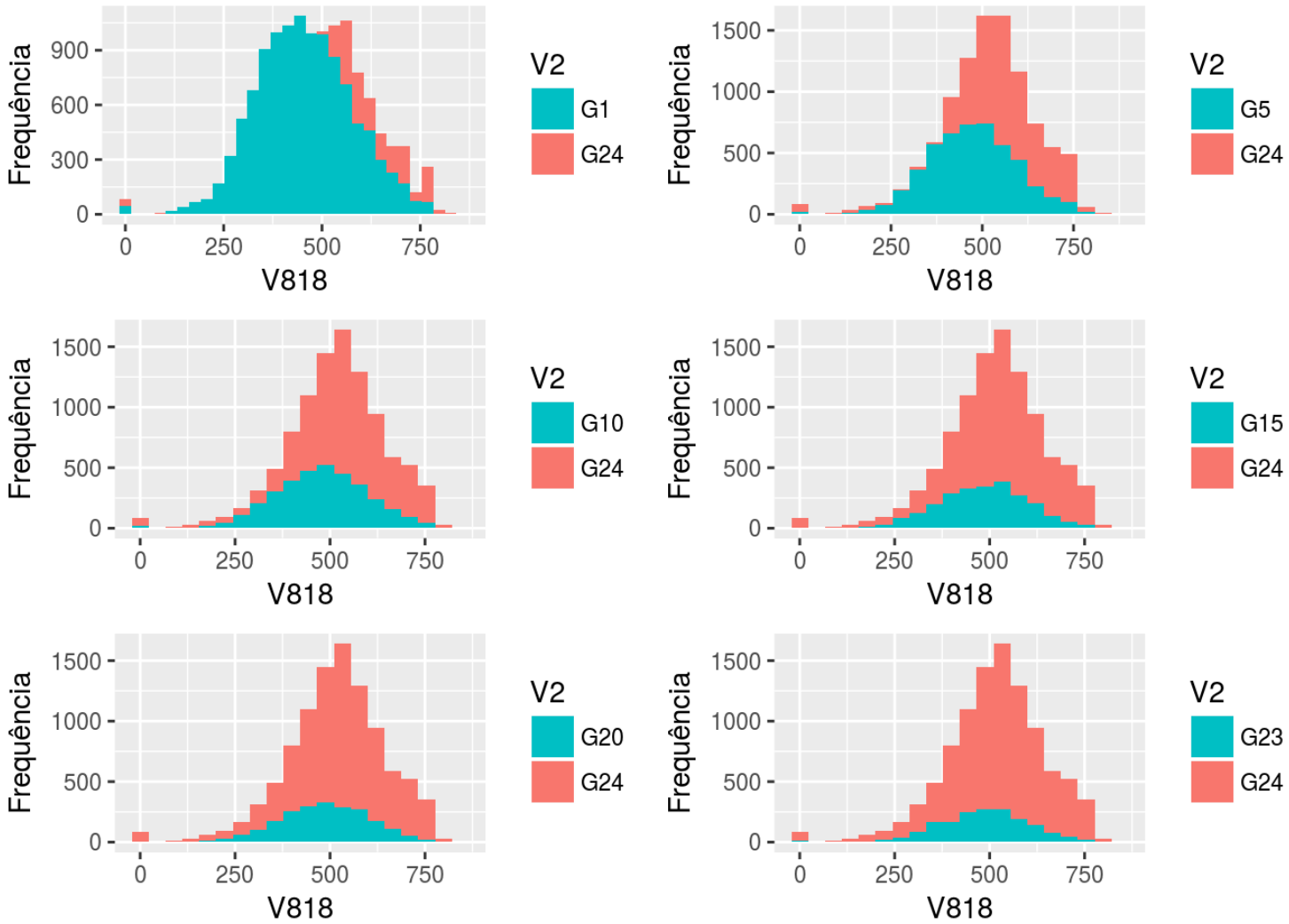
Pense um passo a frente

- A análise não pode parar no gráfico! Você deve extrair algo adicional que a máquina não é capaz de fazer.

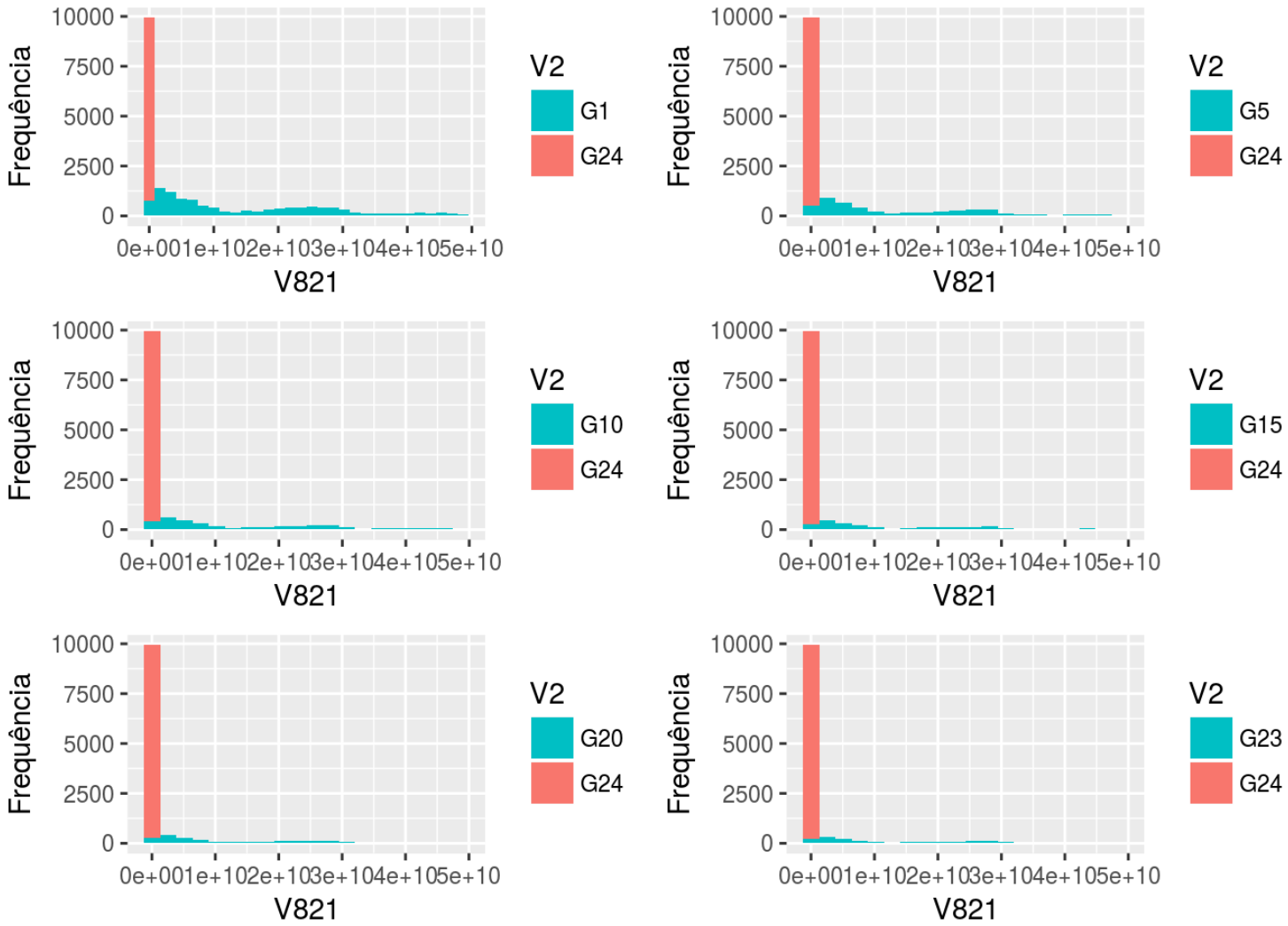
Análises



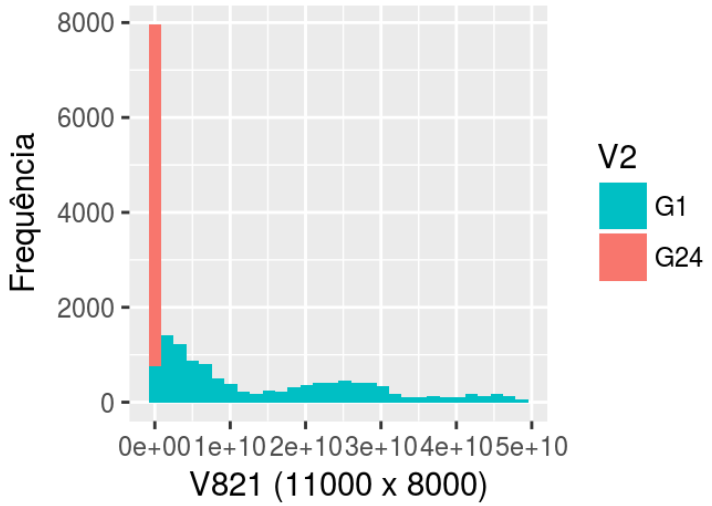
Análises

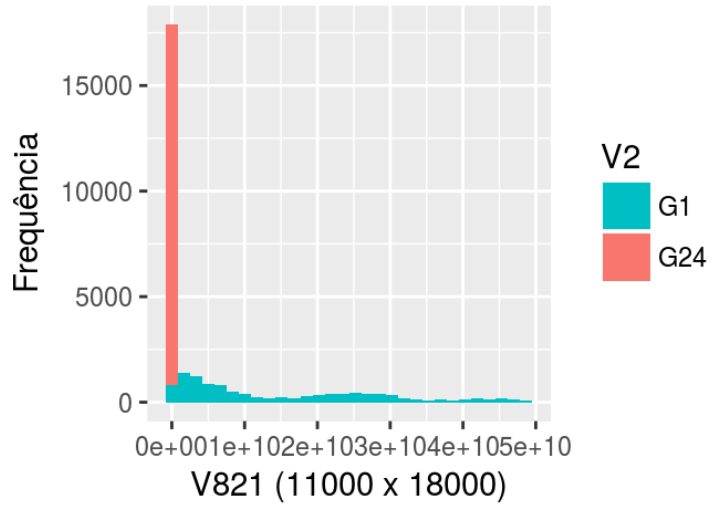


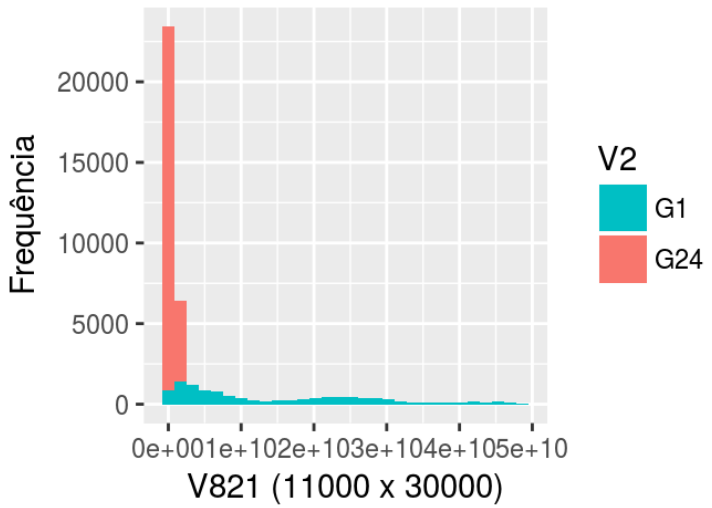
Análises

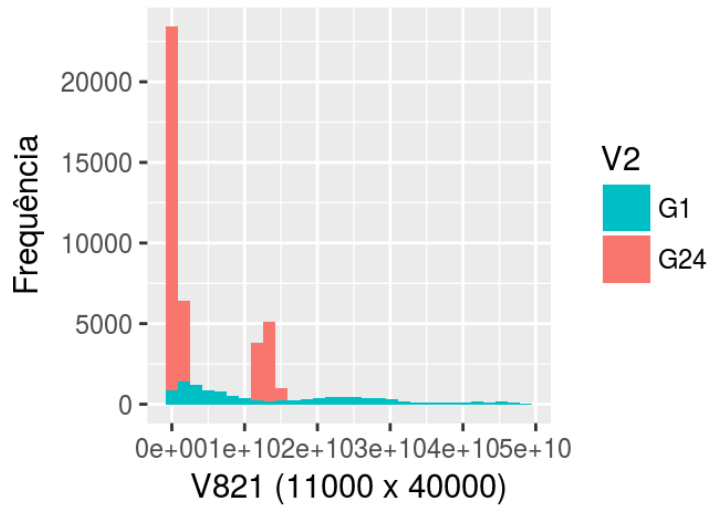

Frequência
10000
7500
5000
2500
0
0e+001e+102e+103e+104e+105e+10
V821
V2
G1
G24
Frequência
10000
7500
5000
2500
0
0e+001e+102e+103e+104e+105e+10
V821
V2
G5
G24
Frequência
10000
7500
5000
2500
0
0e+001e+102e+103e+104e+105e+10
V821
V2
G10
G24
Frequência
10000
7500
5000
2500
0
0e+001e+102e+103e+104e+105e+10
V821
V2
G15
G24
Frequência
10000
7500
5000
2500
0
0e+001e+102e+103e+104e+105e+10
V821
V2
G20
G24
Frequência
10000
7500
5000
2500
0
0e+001e+102e+103e+104e+105e+10
V821
V2
G23
G24

Análises

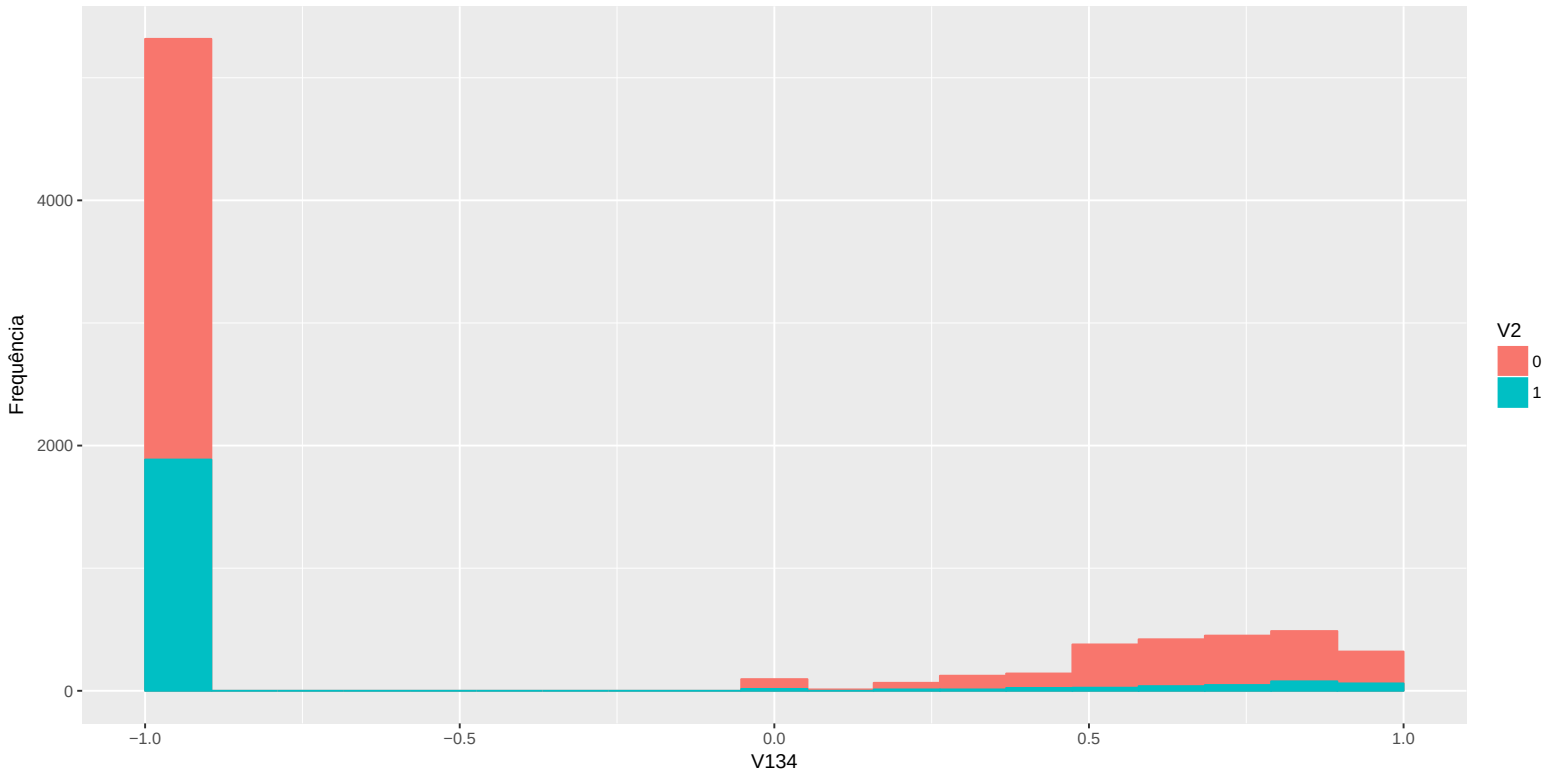

Frequência
8000
6000
4000
2000
0
0e+00 1e+10 2e+10 3e+10 4e+10 5e+10
V821 (11000 x 8000)
V2
G1
G24


Frequência
15000
10000
5000
0
0e+00 1e+10 2e+10 3e+10 4e+10 5e+10
V821 (11000 x 18000)
V2
G1
G24

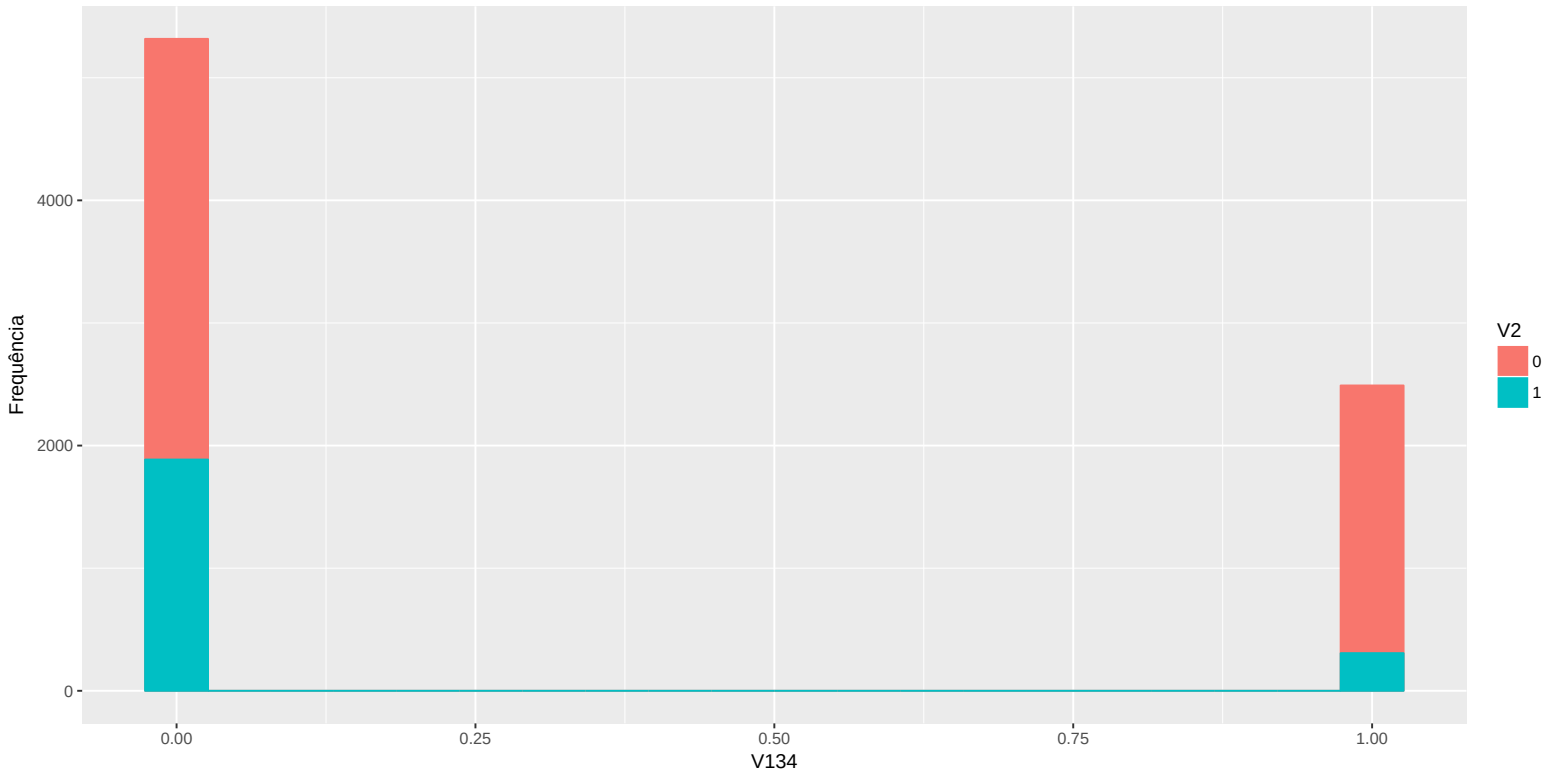

Frequência
20000
15000
10000
5000
0
0e+00 1e+10 2e+10 3e+10 4e+10 5e+10
V821 (11000 x 30000)
V2
G1
G24


Frequência
20000
15000
10000
5000
0
0e+00 1e+10 2e+10 3e+10 4e+10 5e+10
V821 (11000 x 40000)
V2
G1
G24

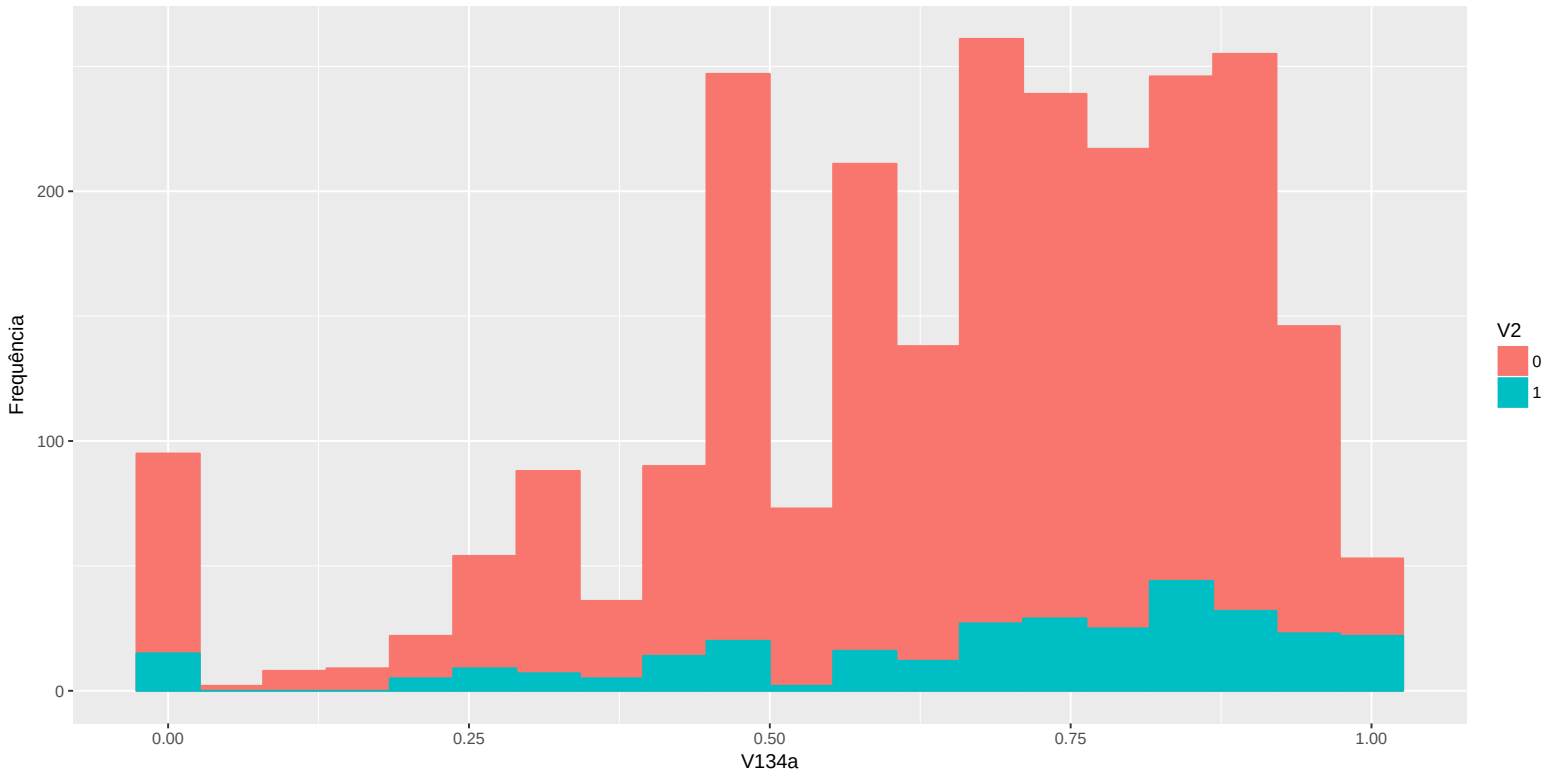
Análises



Análises

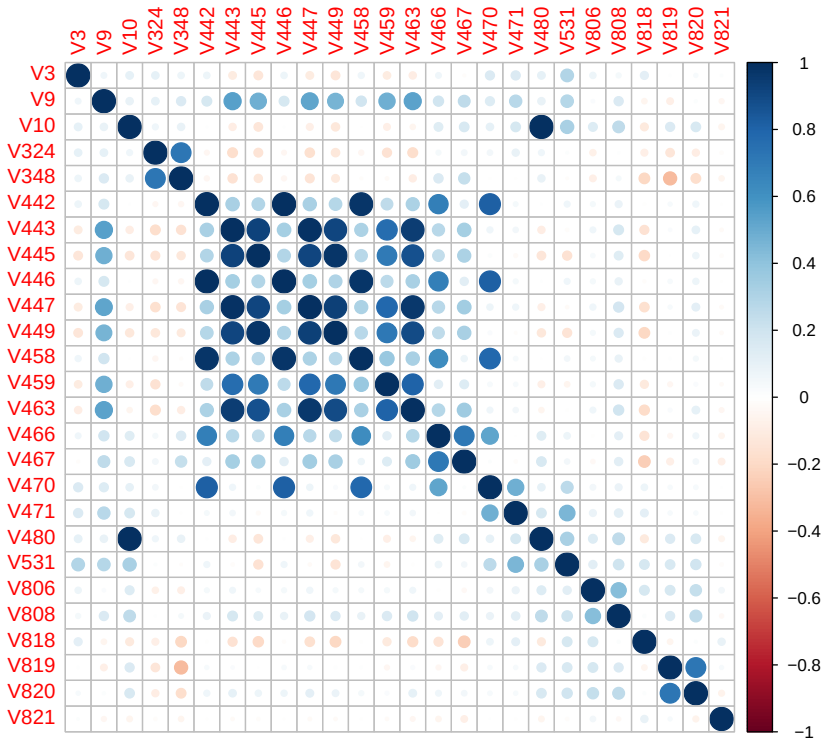


Análises

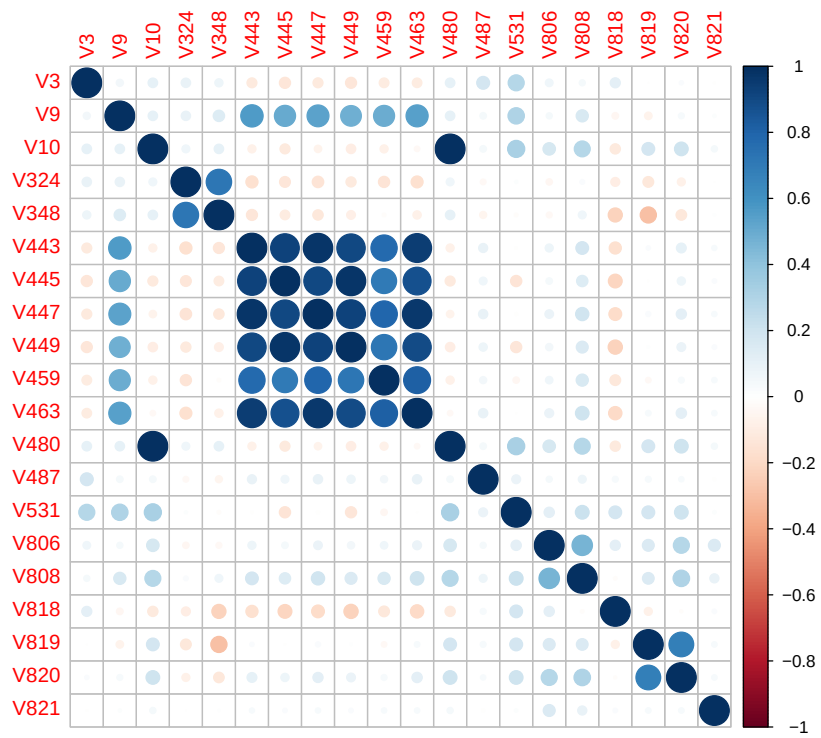


Análises

Dados completos 1



Dados completos 2



Predição versus Inferência

- **Data Modeling Culture**
  - Domina a comunidade estatística;
  - O principal objetivo está na interpretação dos parâmetros;
  - Testar suposições é fundamental.

- **Algorithmic Modeling Culture**
  - Domina a comunidade de Machine Learning;
  - O modelo é utilizado para criar bons algoritmos preditivos;
  - Interpretamos os resultados, mas esse - em geral - não é o foco.

L. Breiman. Statistical modeling: The two cultures. Statistical Science, 16(3):199-231, 2001

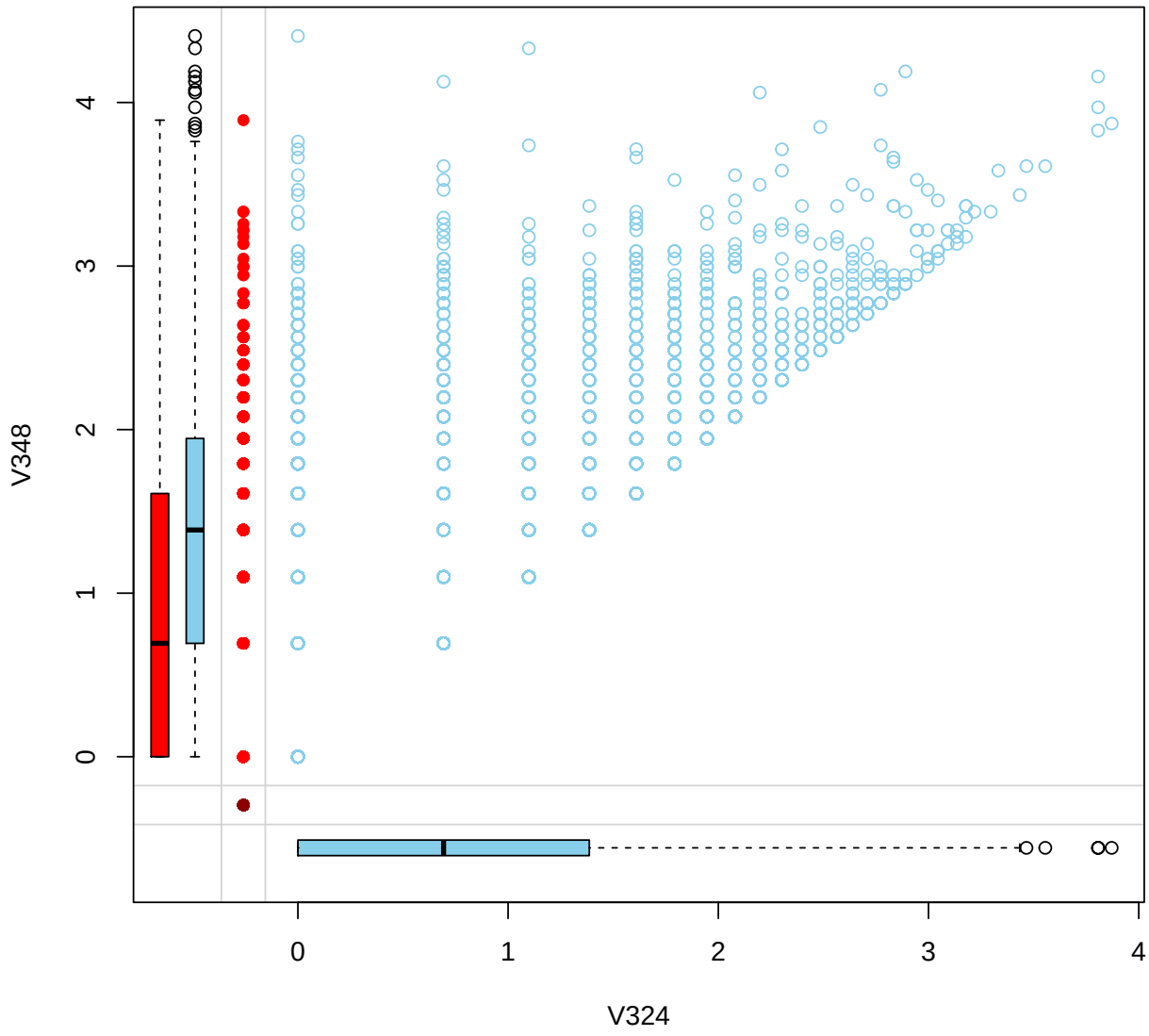
Dados faltantes

- **Missing completely at random:** quando a probabilidade dos dados faltantes é a mesma entre as observações, p ex.:
  - Dados perdidos por um *backup* incorreto;
- **Missing at random:** quando os dados faltantes variam de acordo com outras variáveis, p. ex.:
  - *Missing* sobre idade pode ser diferente entre mulheres e homens;
- **Missing not at random:** quando a probabilidade de *missing* está relacionada com o *missing*, p. ex.,:
  - Dependendo da renda do cliente, é mais provável que ele não responda sobre a renda;
  - Indivíduo não comparece ao teste de droga, porque a utilizou na noite anterior.

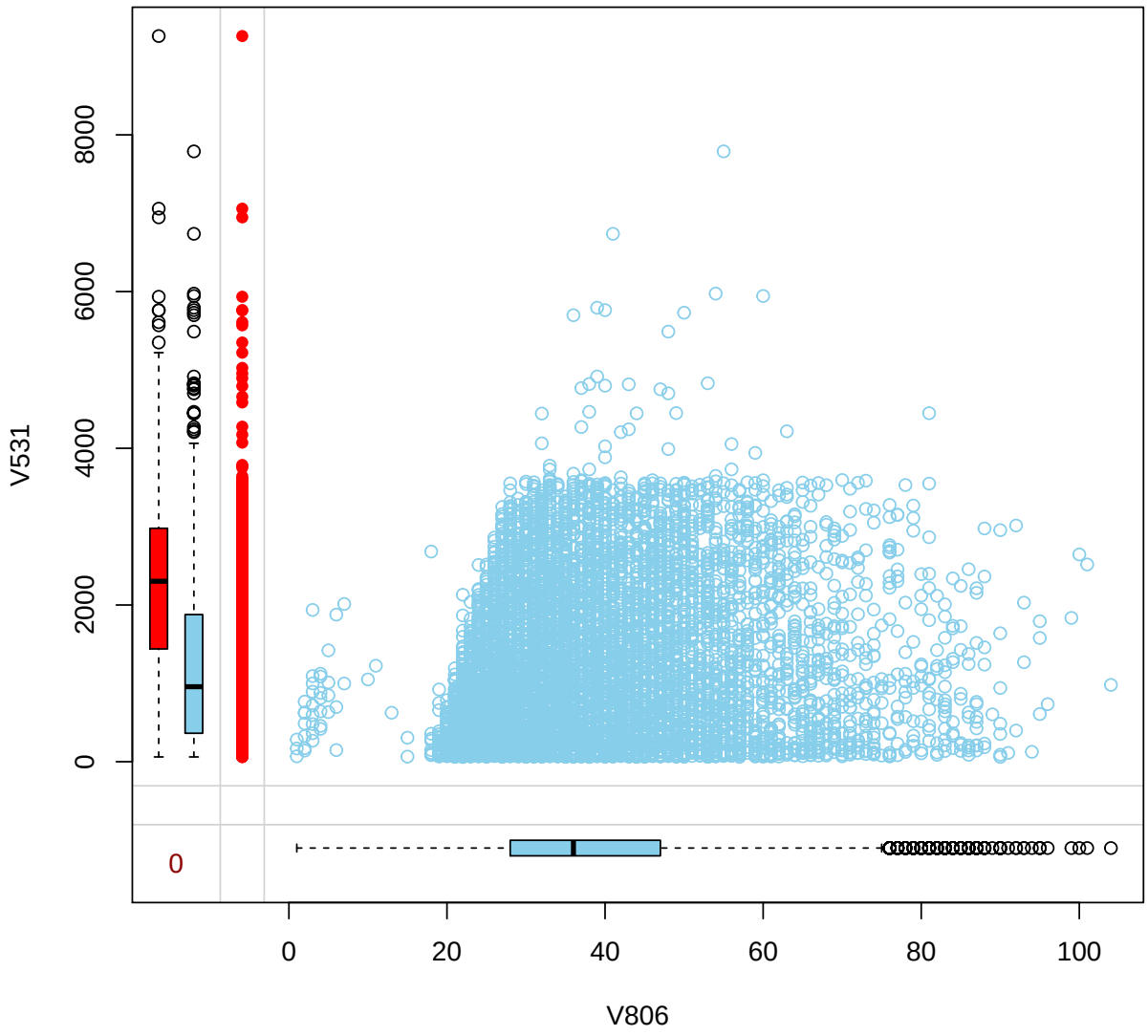
Tratamento dos dados faltantes

- **Deletar:** utilizado quando a natureza do "*missing*" é **completely at random**.
  - Podemos eliminar a linha inteira. É uma abordagem simples, mas retira poder dos dados, devido à redução do tamanho da amostra;
  - Ou utilizar os dados completos, de acordo - somente - com as variáveis de interesse.
- **Imputação:** utilizado quando trata-se de **missing at random** ou **missing not at random**.
  - Média, mediana, moda.
  - Modelo preditivo;
  - KNN.

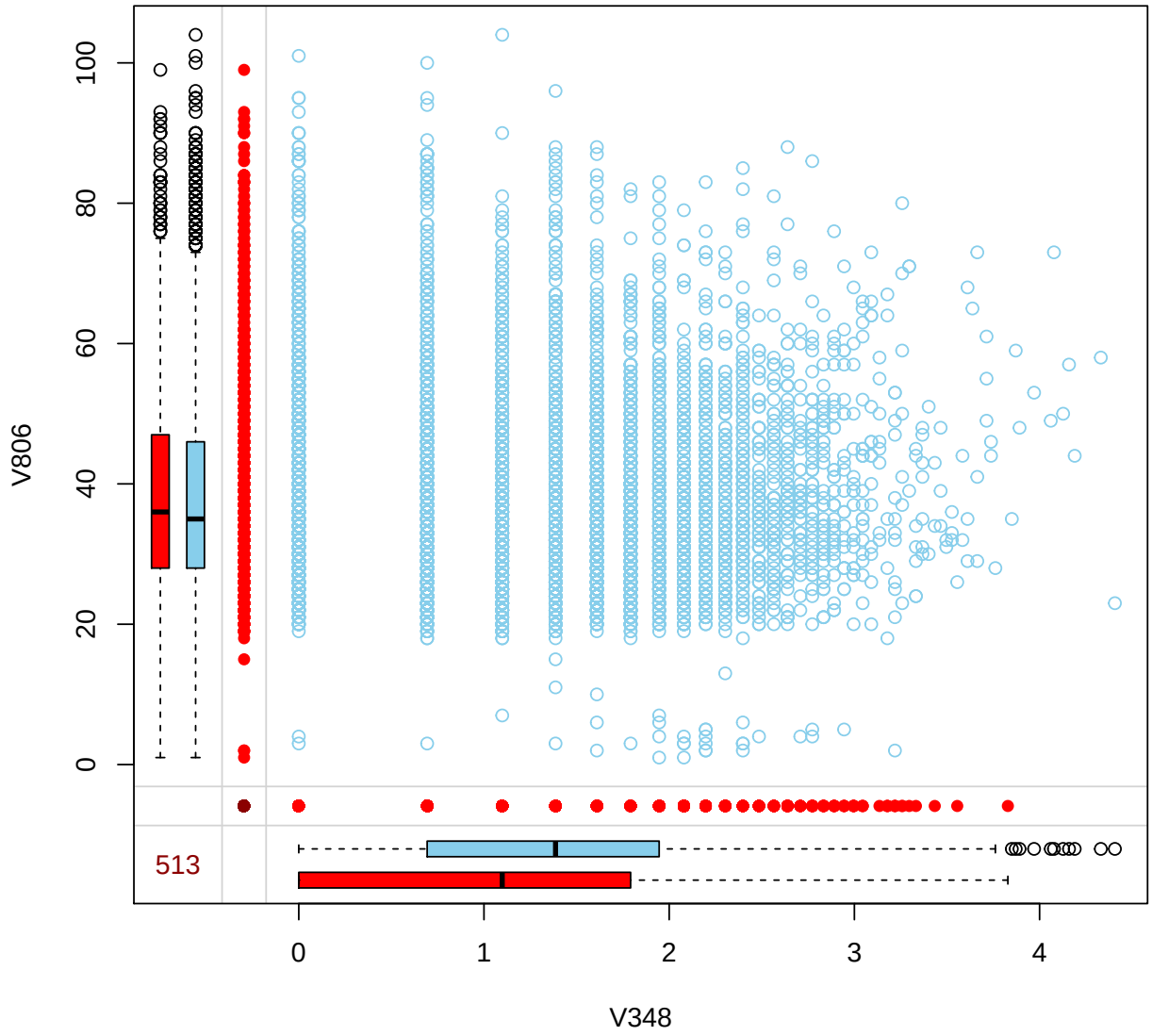
Análises



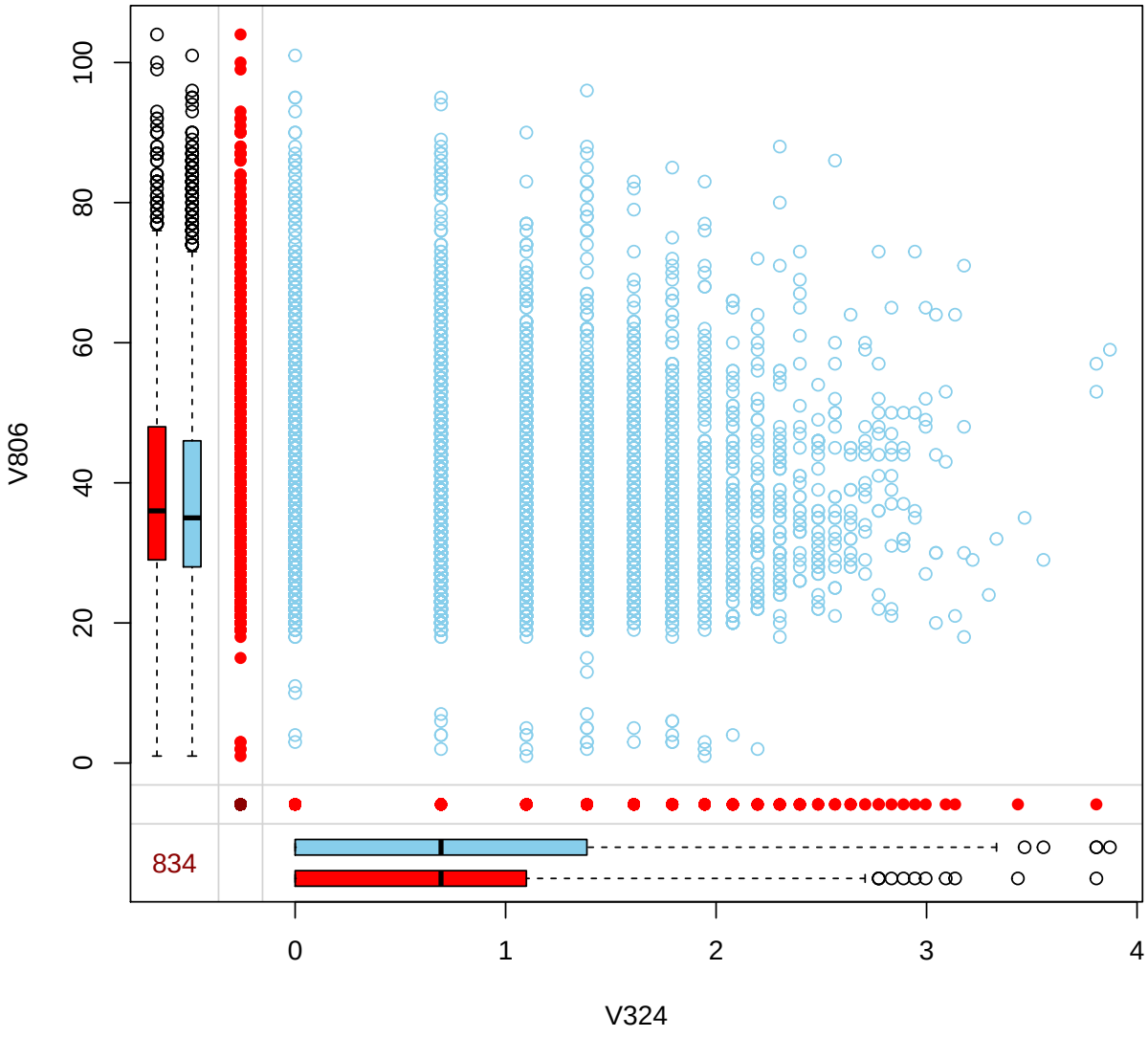
Análises



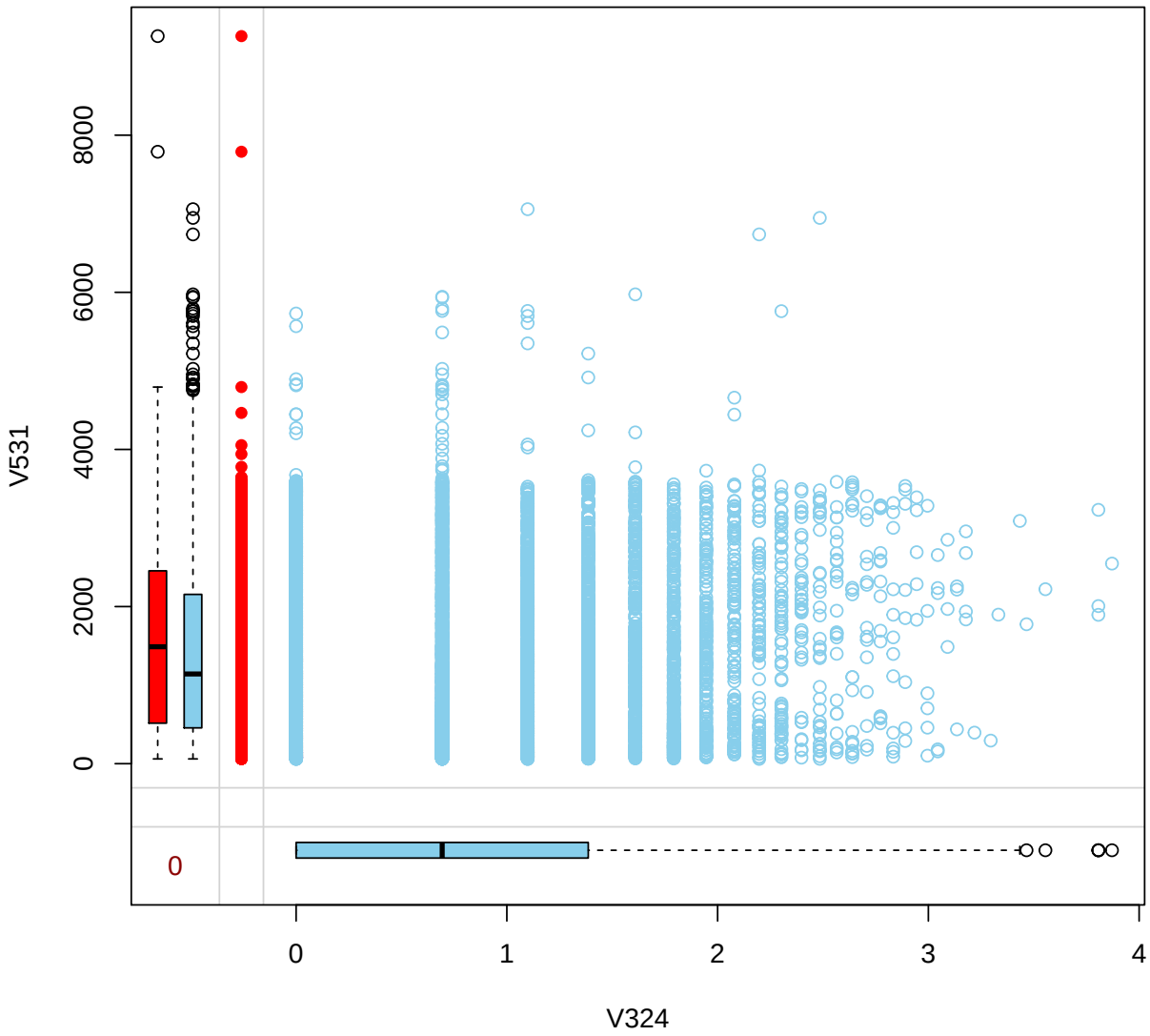
Análises



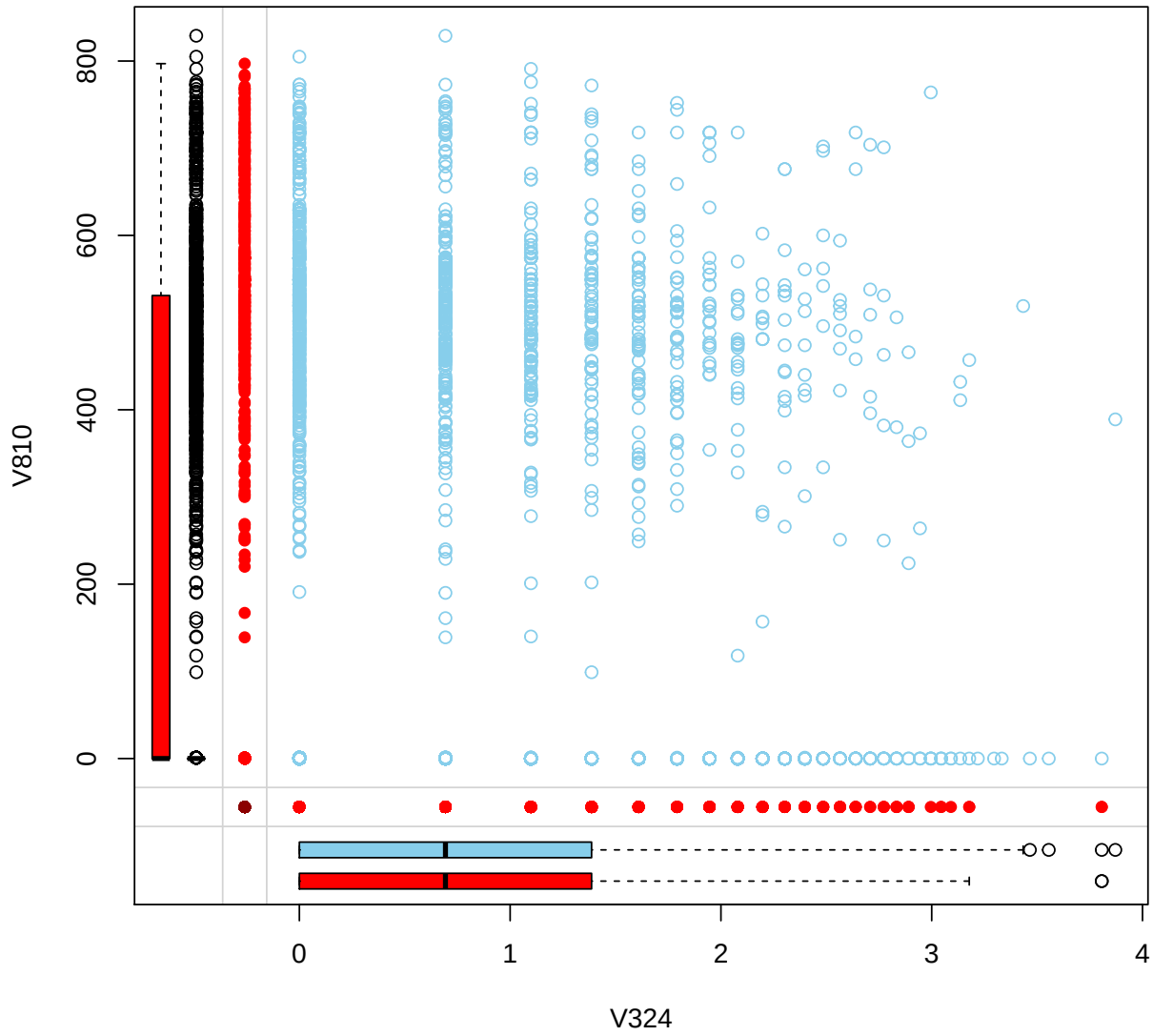
Análises



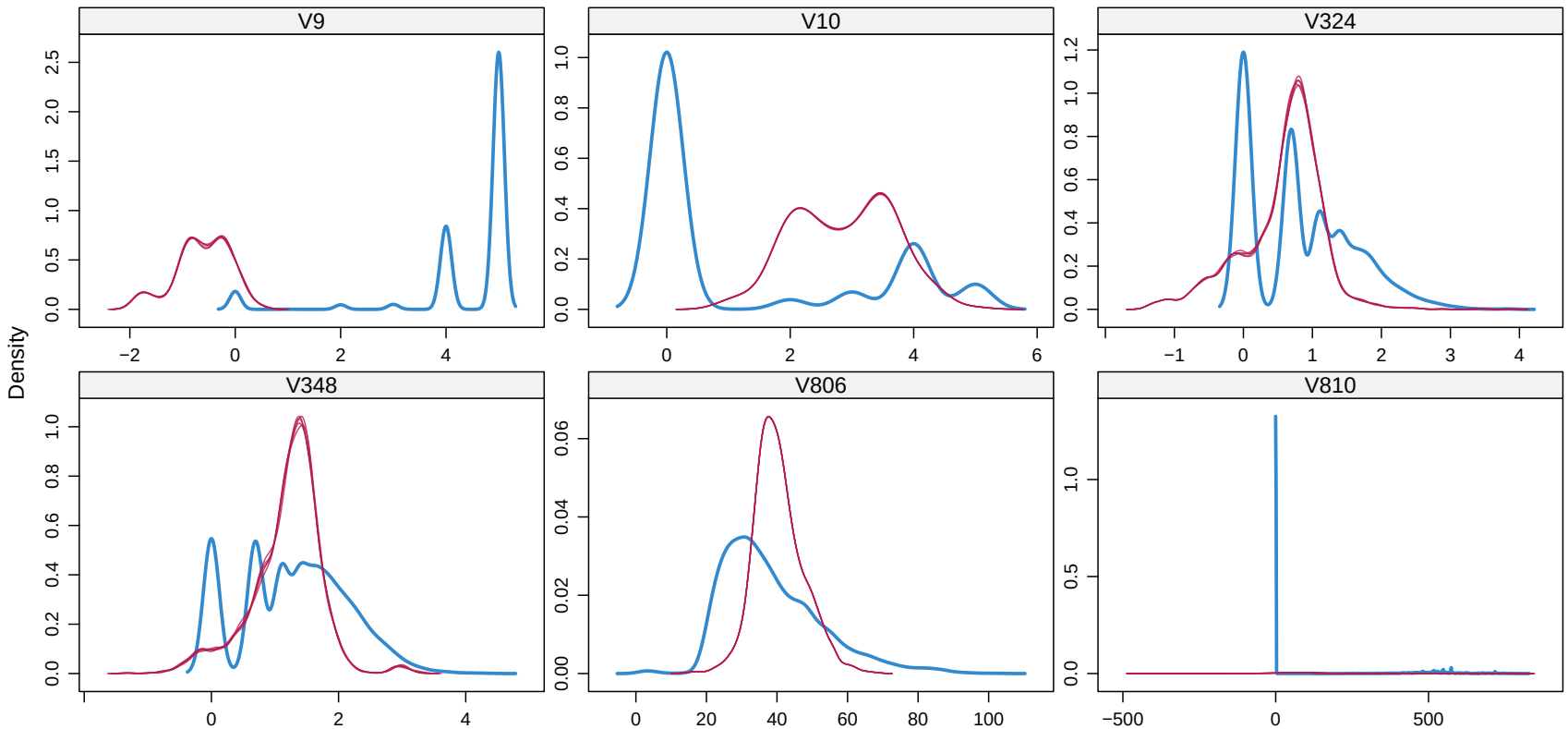
Análises



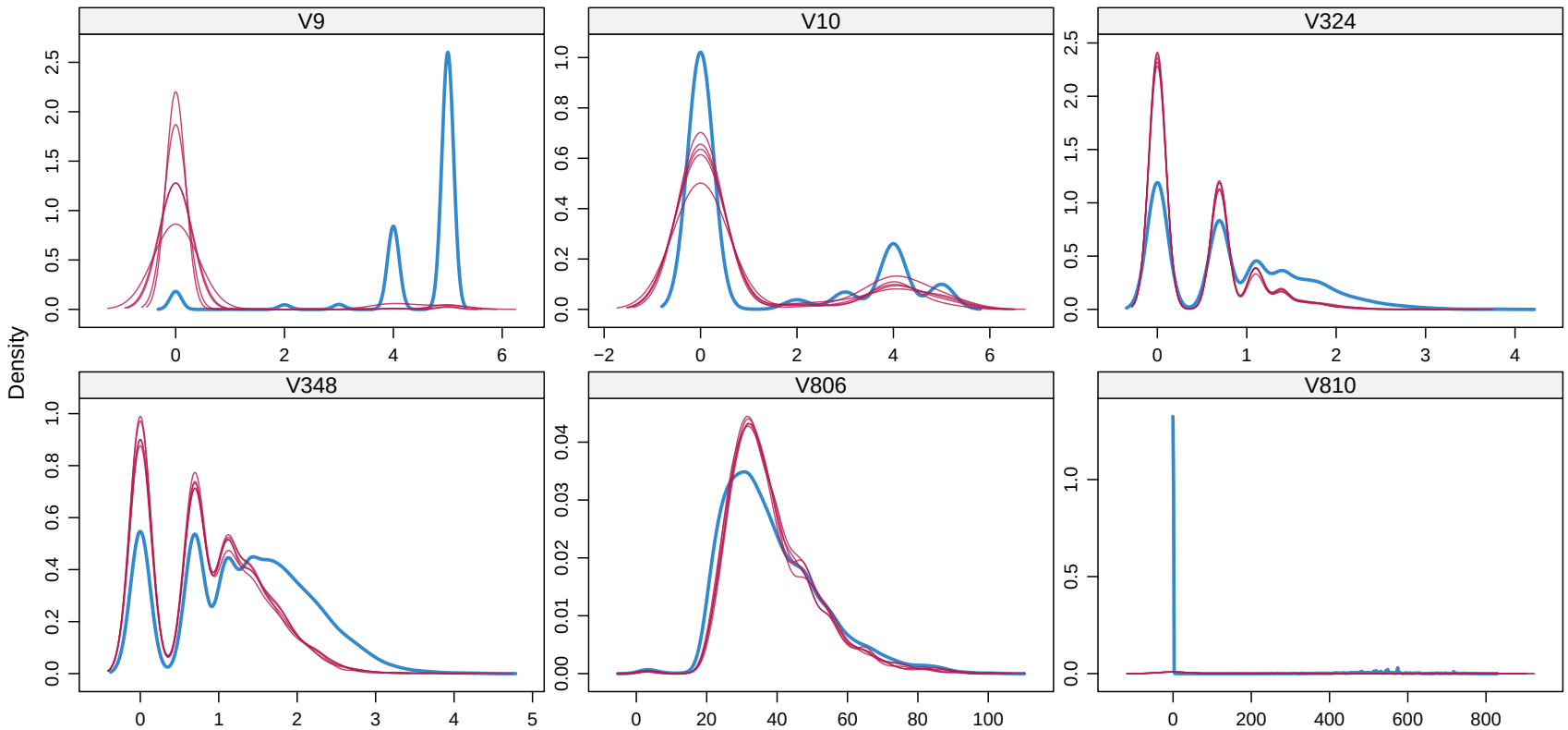
Análises



Análises



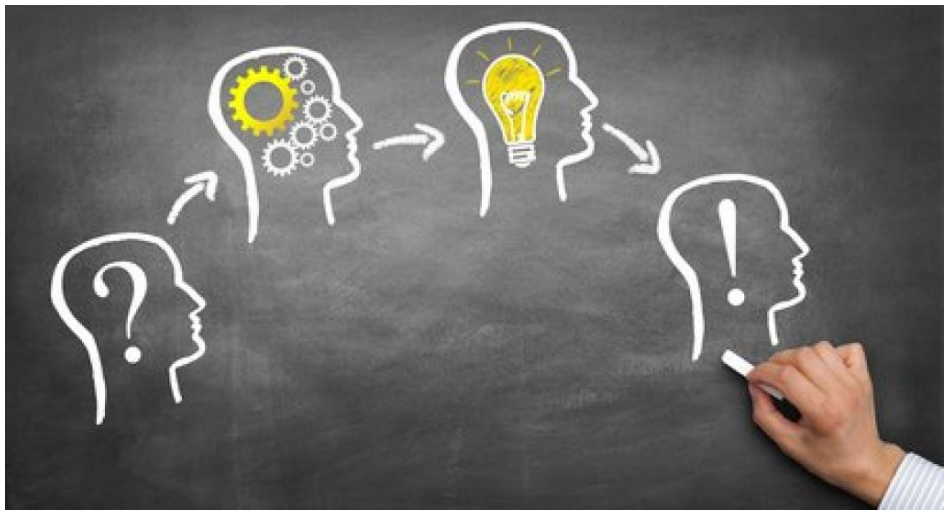
Análises



Pergunta

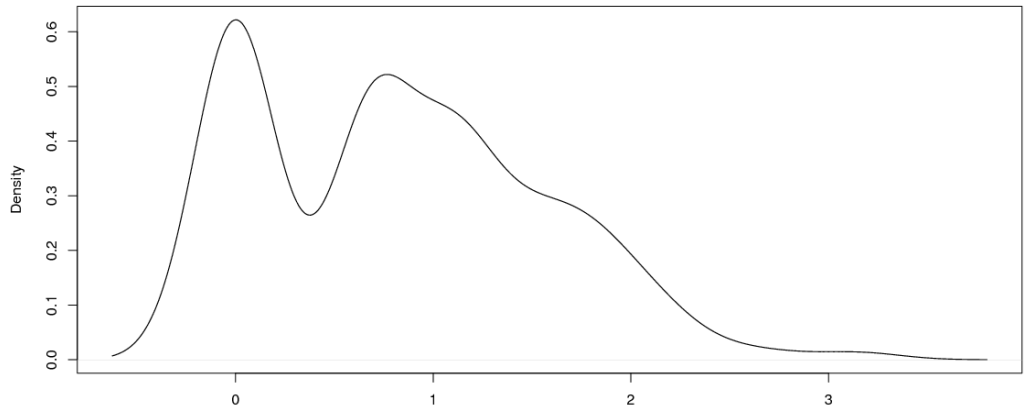
- Você confiaria em modelos estatísticos produzidos a partir da eliminação dos dados faltantes? E a partir da imputação dos dados faltantes?

- Você confiaria em modelo estatístico cujos dados possuem outliers?

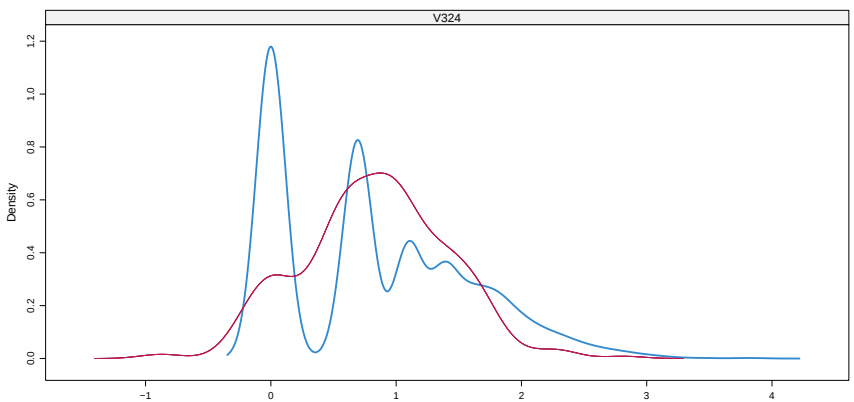


- Em probabilidade é que desenvolvemos as percepções sobre os dados!

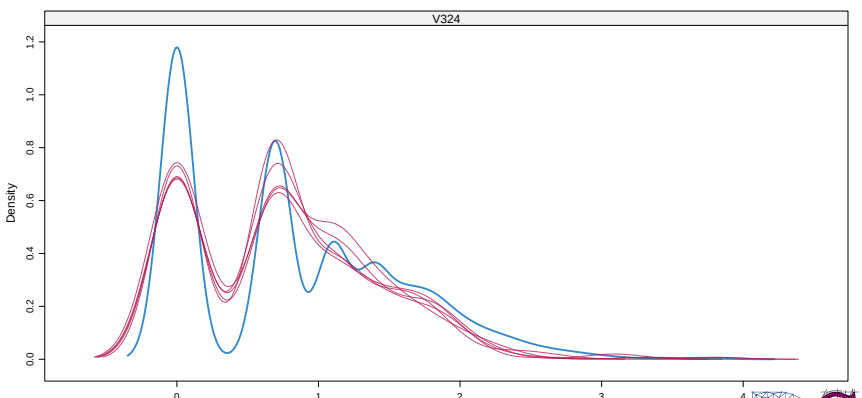
Análises



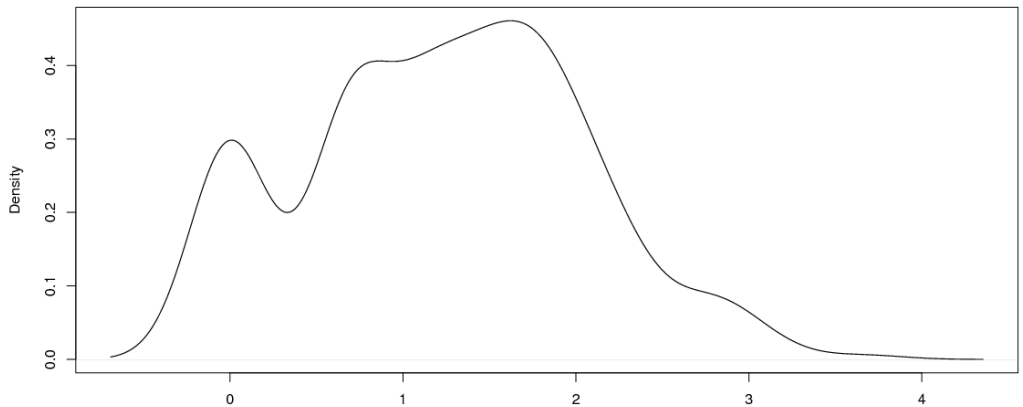
Regressão



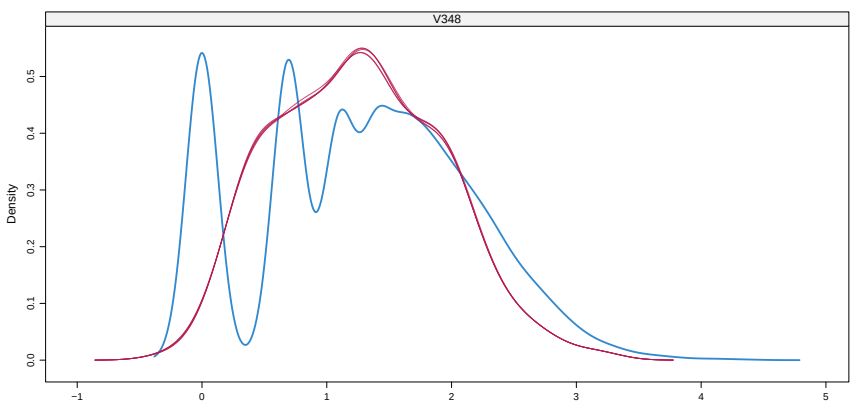
Random Forest



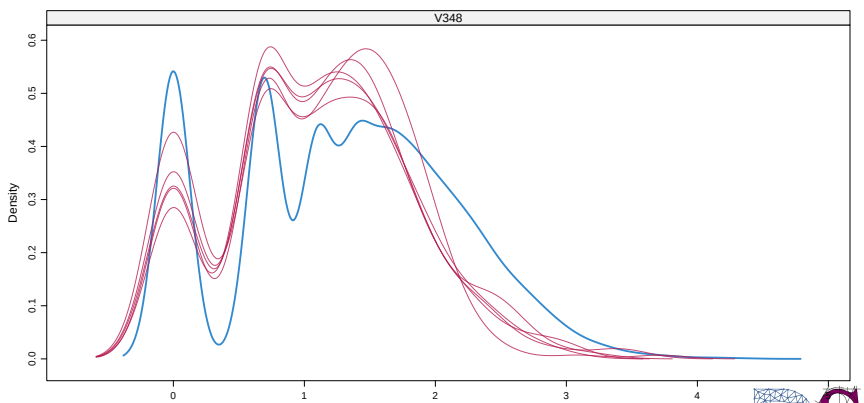
Análises



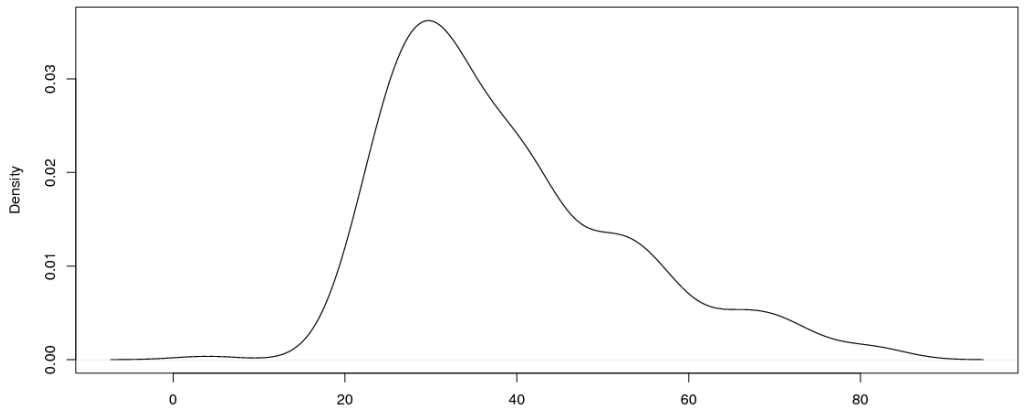
Regressão



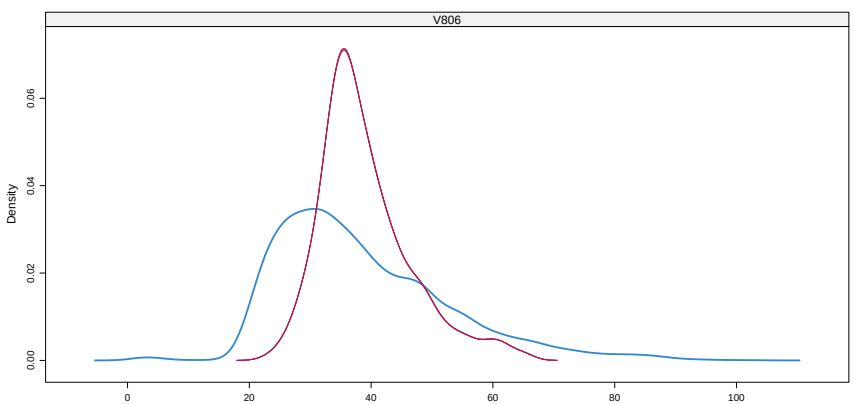
Random Forest



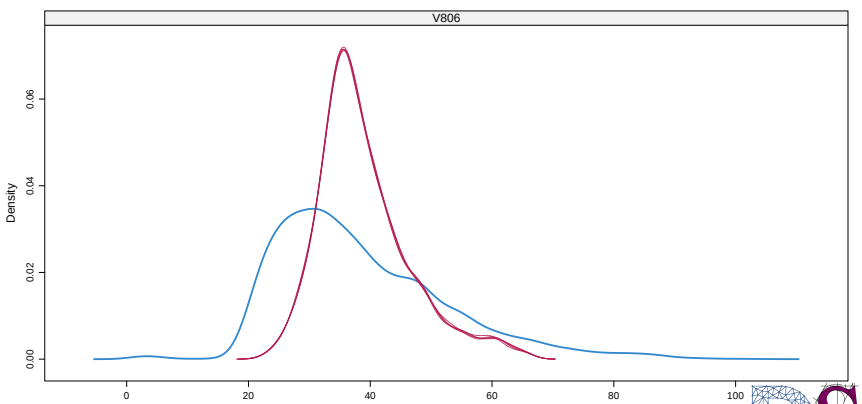
Análises



Regressão

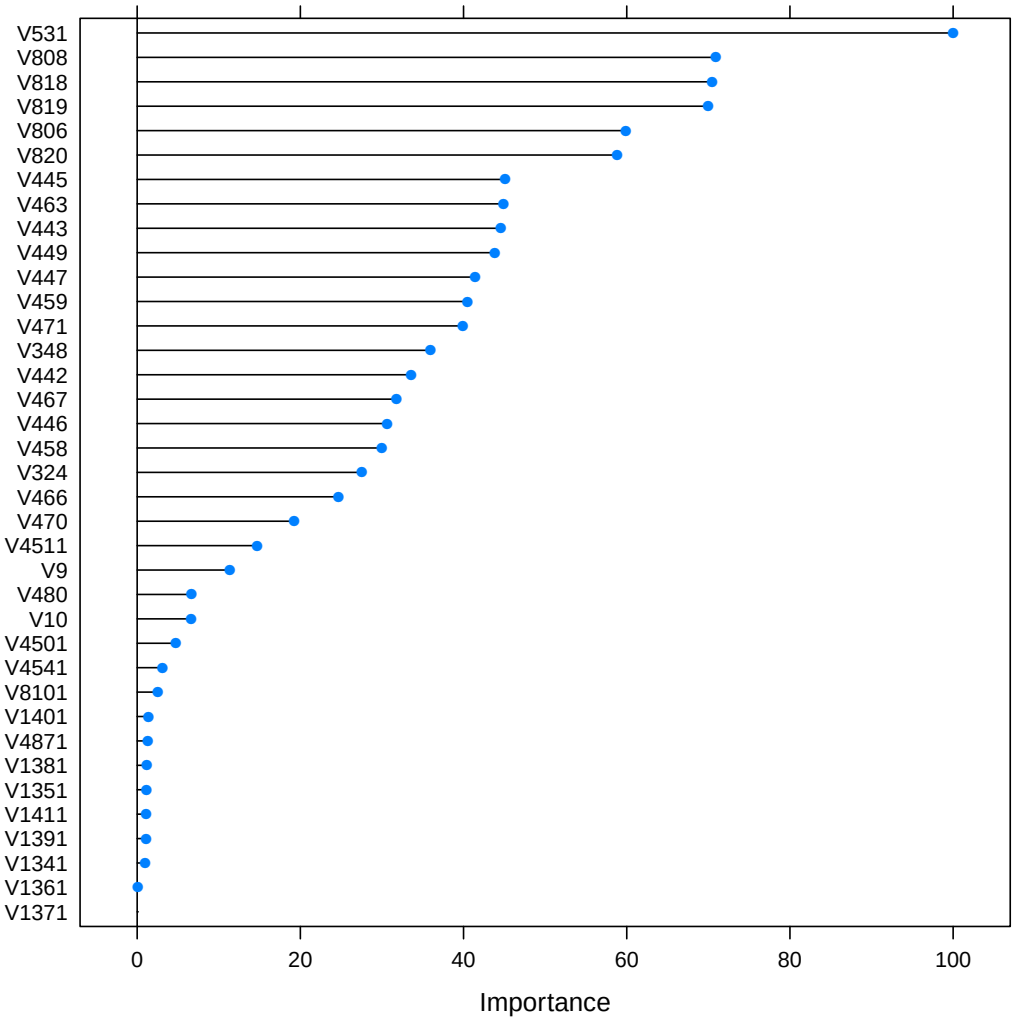


Random Forest



Análises

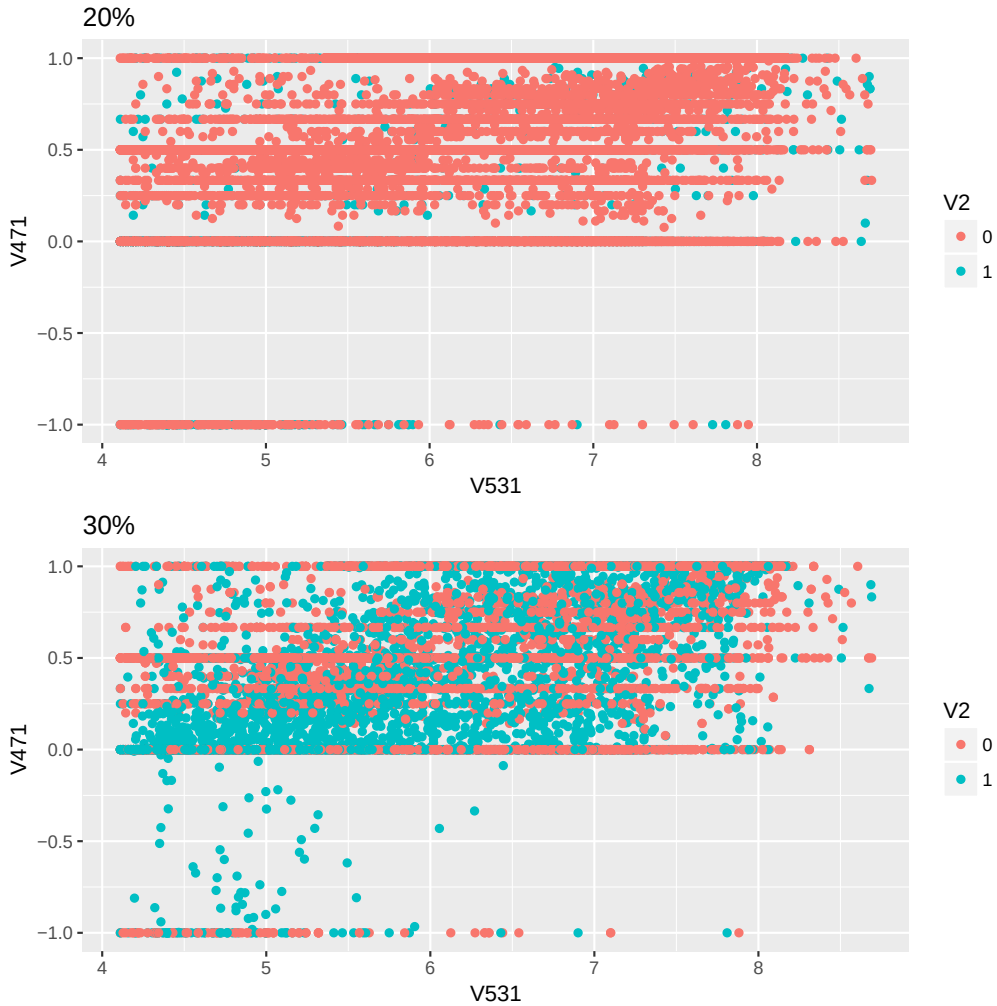
SMOTE: 20% sim e 80% não (dados originais)




```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4168  890
         1  142  238

      Accuracy : 0.8102
        95% CI : (0.79, 0.82)

   Sensitivity : 0.9671
   Specificity : 0.2110
Pos Pred Value : 0.8240
Neg Pred Value : 0.6263
```

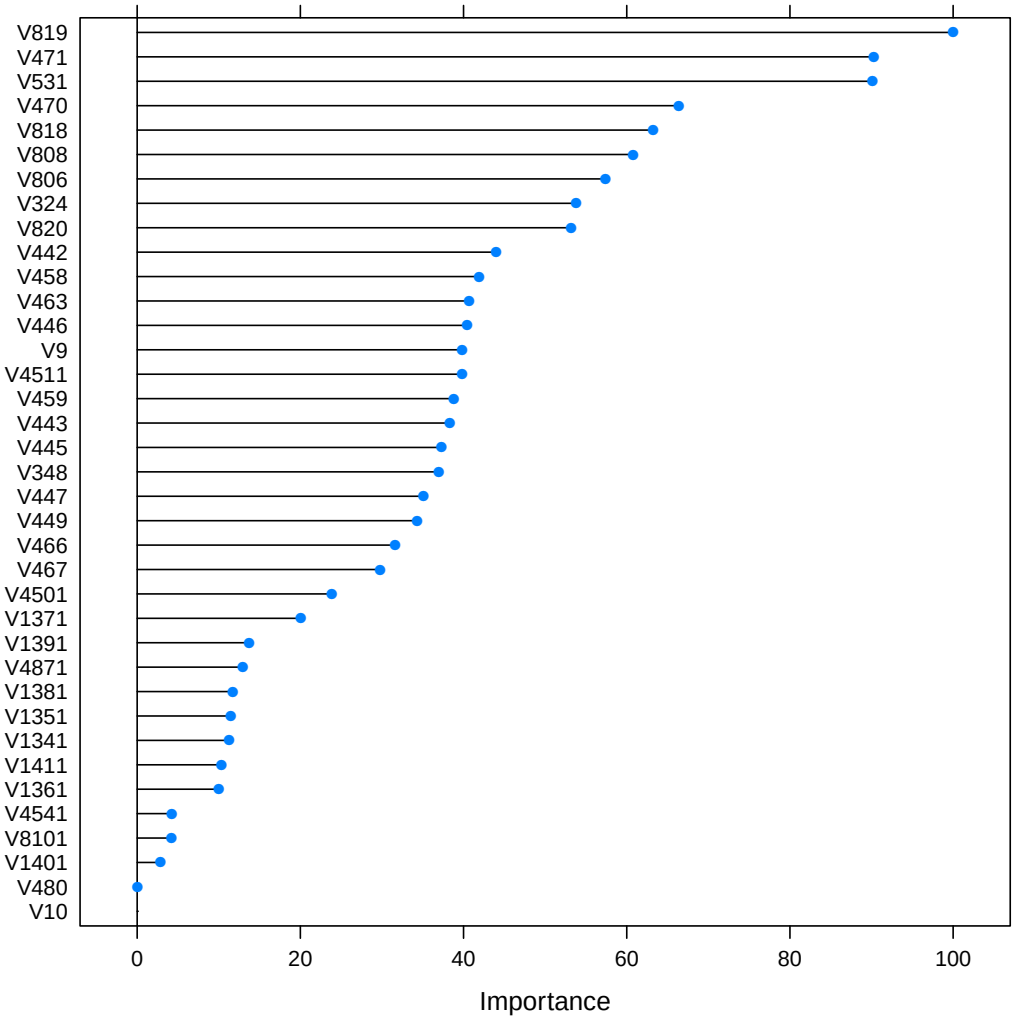
SMOTE

- SMOTE: Synthetic Minority Over-sampling Technique



Análises

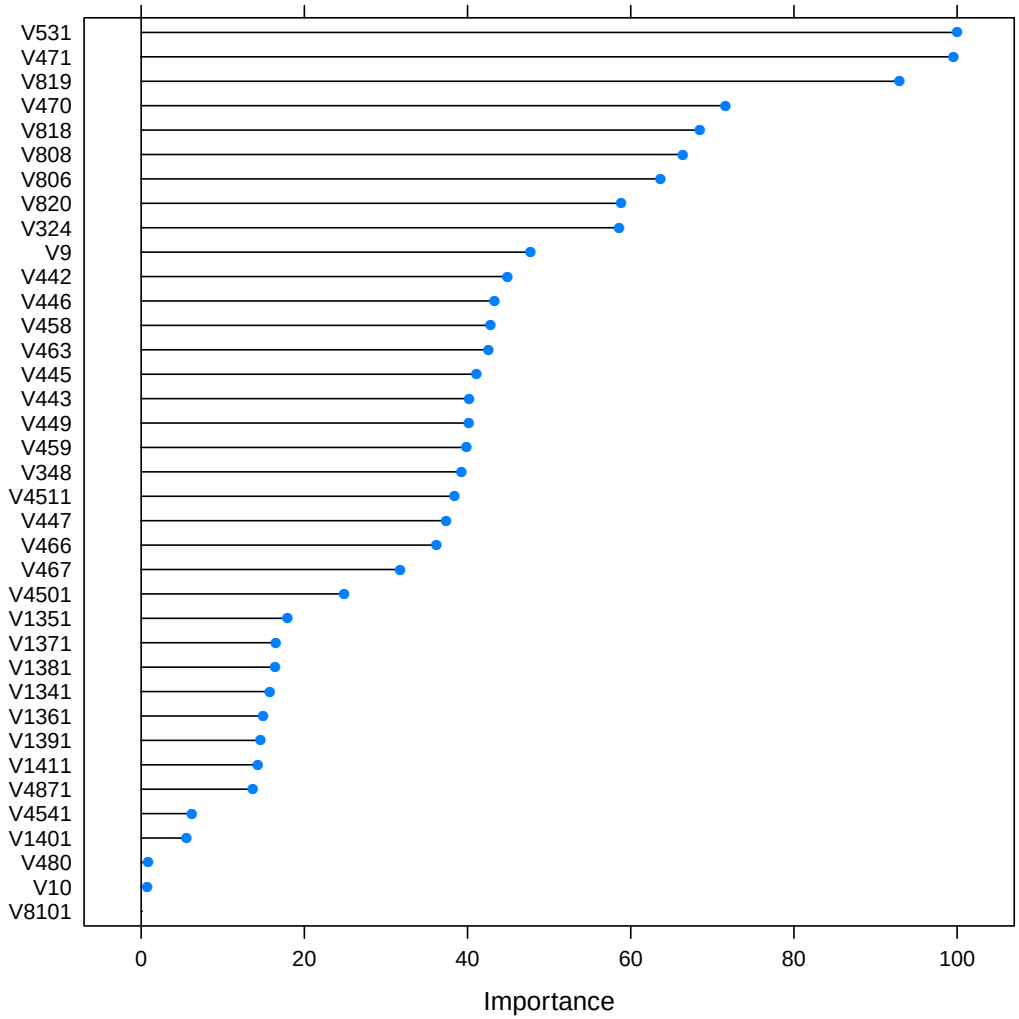
SMOTE: 25% sim e 75% não




```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4258  729
         1   52  399

      Accuracy : 0.8564
        95% CI : (0.84, 0.86)

   Sensitivity : 0.9879
   Specificity : 0.3537
Pos Pred Value : 0.8538
Neg Pred Value : 0.8847
```

Análises

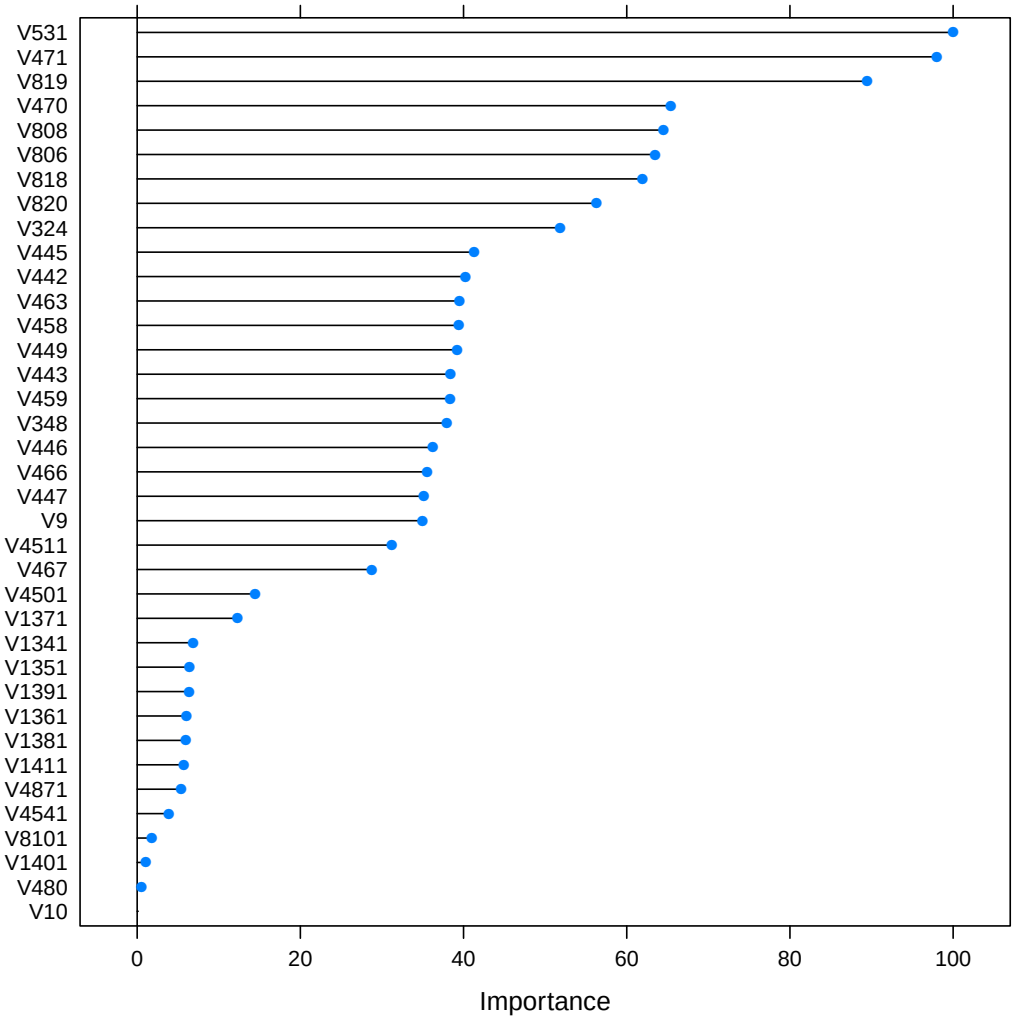
SMOTE: 30% sim e 70% não




```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4189  612
         1  121  516

      Accuracy : 0.8652
        95% CI : (0.85, 0.87)

   Sensitivity : 0.9719
   Specificity : 0.4574
Pos Pred Value : 0.8725
Neg Pred Value : 0.8100
```

Análises

SMOTE: 50% sim e 50% não




```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3861  368
         1  449  760

               Accuracy : 0.8498
                 95% CI : (0.84, 0.85)

   Sensitivity : 0.8958
   Specificity : 0.6738
Pos Pred Value : 0.9130
Neg Pred Value : 0.6286
```