SEMINARIO ANALISE MULTIVARIADA

IGOR BARBOSA NEGREIROS E JHEYMISSON THIAGO SOUSA SILVA

2023-05-03

ANÁLISE MULTIVARIADA NO BANCO DE DADOS SOBRE FRAUD NO CARTÃO DE CRÉDITO

BANCO DE DADOS COLETADO NO LINK:

HTTPS://WWW.KAGGLE.COM/CODE/SMNURUZZAMAN/FRAUD-DETECTION-WITH-SMOTE-AND-SHAP-XGB-99-99/INPUT

COM 6362620 OBSERVAÇÕES E 12 VARIÁVEIS, DAS QUAIS 5 SÃO NÚMERICAS FORAM SELECIONADAS DUAS DELAS

IMPORTANTANDO O BANCO DE DADOS

```
#dados ← read.csv("PS_20174392719_14912044439457_log.csv", header = TRUE, sep = ',', dec = '.')
names(dados)
[1] "step"           "type"           "amount"         "nameOrig"       "oldbalanceOrg"  "newbalanceOrig"
[7] "nameDest"       "oldbalanceDest" "newbalanceDest" "isFraud"        "isFlaggedFraud" "D2"
```

SUMARIO SOBRE AS VARIÁVEIS

```
> summary(dados)
      step            type              amount            nameOrig          oldbalanceOrg      newbalanceOrig
 Min.   :  1.0   Length:6362620     Min.   :       0   Length:6362620     Min.   :       0   Min.   :       0
 1st Qu.:156.0   Class :character   1st Qu.:   13390   Class :character   1st Qu.:       0   1st Qu.:       0
 Median :239.0   Mode  :character   Median :   74872   Mode  :character   Median :   14208   Median :       0
 Mean   :243.4                      Mean   :  179862                      Mean   :  833883   Mean   :  855114
 3rd Qu.:335.0                      3rd Qu.:  208721                      3rd Qu.:  107315   3rd Qu.:  144258
 Max.   :743.0                      Max.   :92445517                      Max.   :59585040   Max.   :49585040
   nameDest          oldbalanceDest     newbalanceDest         isFraud         isFlaggedFraud          D2
 Length:6362620     Min.   :       0   Min.   :       0   Min.   :0.000000   Min.   :0.0e+00   Min.   : 0.0000
 Class :character   1st Qu.:       0   1st Qu.:       0   1st Qu.:0.000000   1st Qu.:0.0e+00   1st Qu.: 0.7961
 Mode  :character   Median :  132706   Median :  214661   Median :0.000000   Median :0.0e+00   Median : 1.5090
                    Mean   : 1100702   Mean   : 1224996   Mean   :0.001291   Mean   :2.5e-06   Mean   : 2.0000
                    3rd Qu.:  943037   3rd Qu.: 1111909   3rd Qu.:0.000000   3rd Qu.:0.0e+00   3rd Qu.: 2.6076
                    Max.   :356015889   Max.   :356179279   Max.   :1.000000   Max.   :1.0e+00   Max.   :41.6851
```

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
1	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.00	0.00	0	0
2	1	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.00	0.00	0	0
3	1	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.00	0.00	1	0
4	1	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21182.00	0.00	1	0
5	1	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.00	0.00	0	0
6	1	PAYMENT	7817.71	C90045638	53860.00	46042.29	M573487274	0.00	0.00	0	0
7	1	PAYMENT	7107.77	C154988899	183195.00	176087.23	M408069119	0.00	0.00	0	0
8	1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M633326333	0.00	0.00	0	0
9	1	PAYMENT	4024.36	C1265012928	2671.00	0.00	M1176932104	0.00	0.00	0	0
10	1	DEBIT	5337.77	C712410124	41720.00	36382.23	C195600860	41898.00	40348.79	0	0

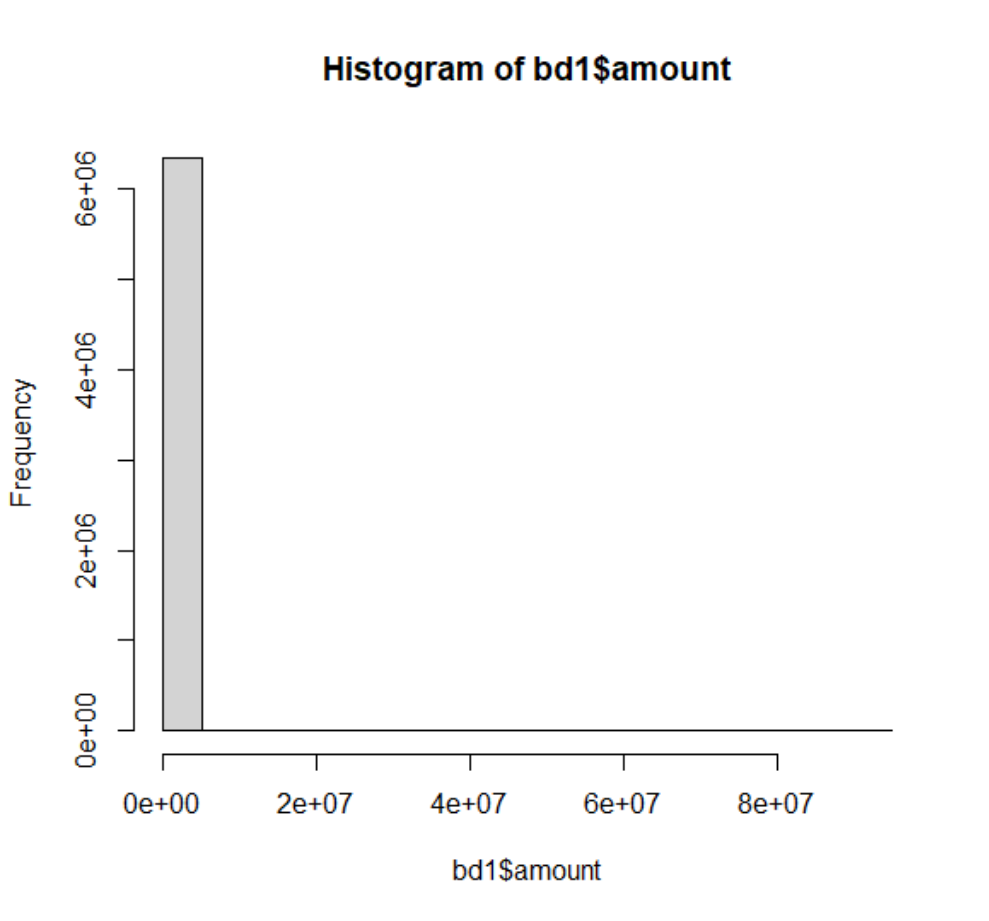
SELECIONADO AS VARIÁVEIS PARA OS TESTES

FOI SELECIONADO AS VARIÁVEIS “AMOUNT” E “OLDBALANCEORG”

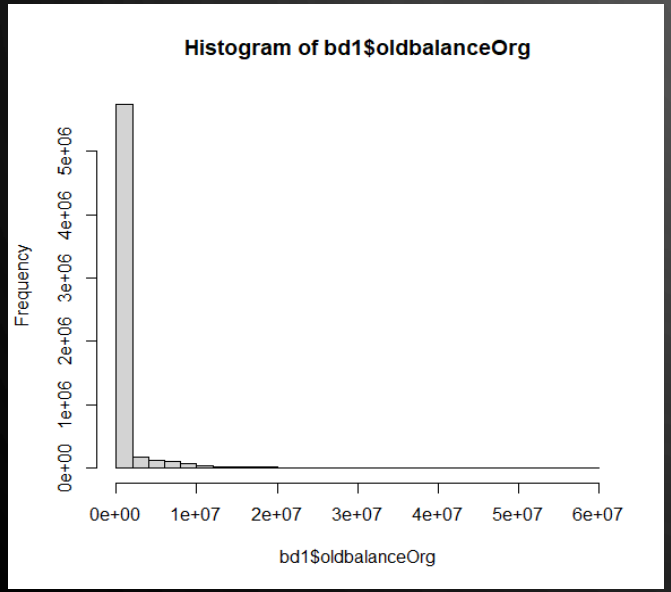
```
> bd = dados[c("amount", "oldbalanceOrg")]
> bd = na.omit(bd) #removendo os NA
```

TESTES DE NORMALIDADE

PRIMEIRAMENTE FOI FEITO O HISTOGRAMAS DE AMBAS AS VARIÁVEIS



HISTOGRAMA OLDBALANCEORG



Como se pode observar nenhuma das duas tem comportamento de seguir uma distribuição normal

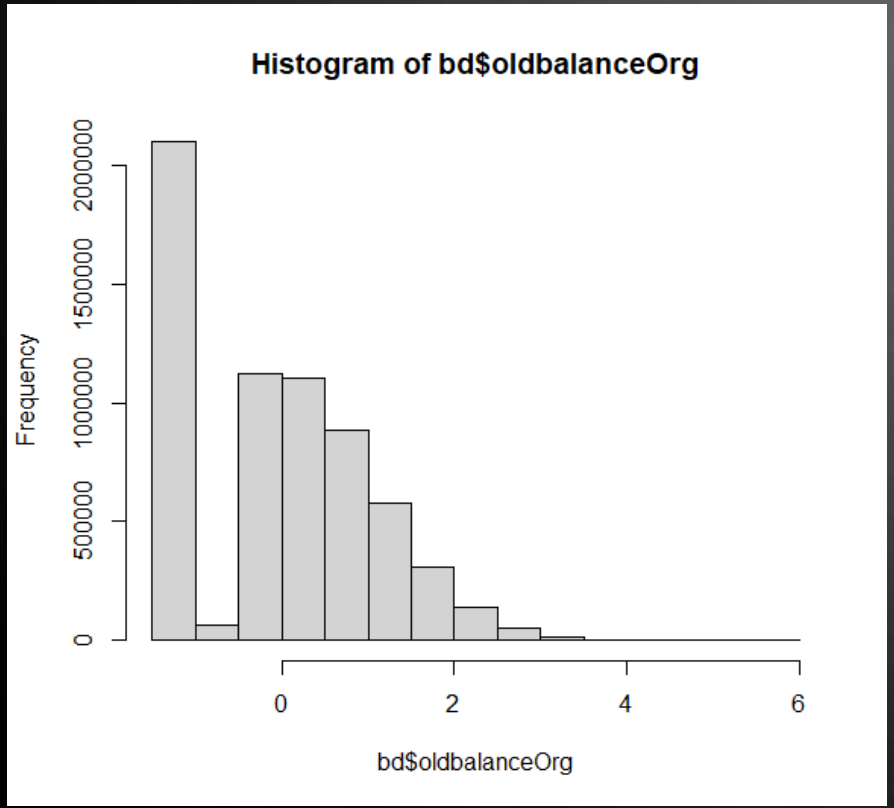
TRANSFORMAÇÃO PARA VARIÁVEL NORMAL

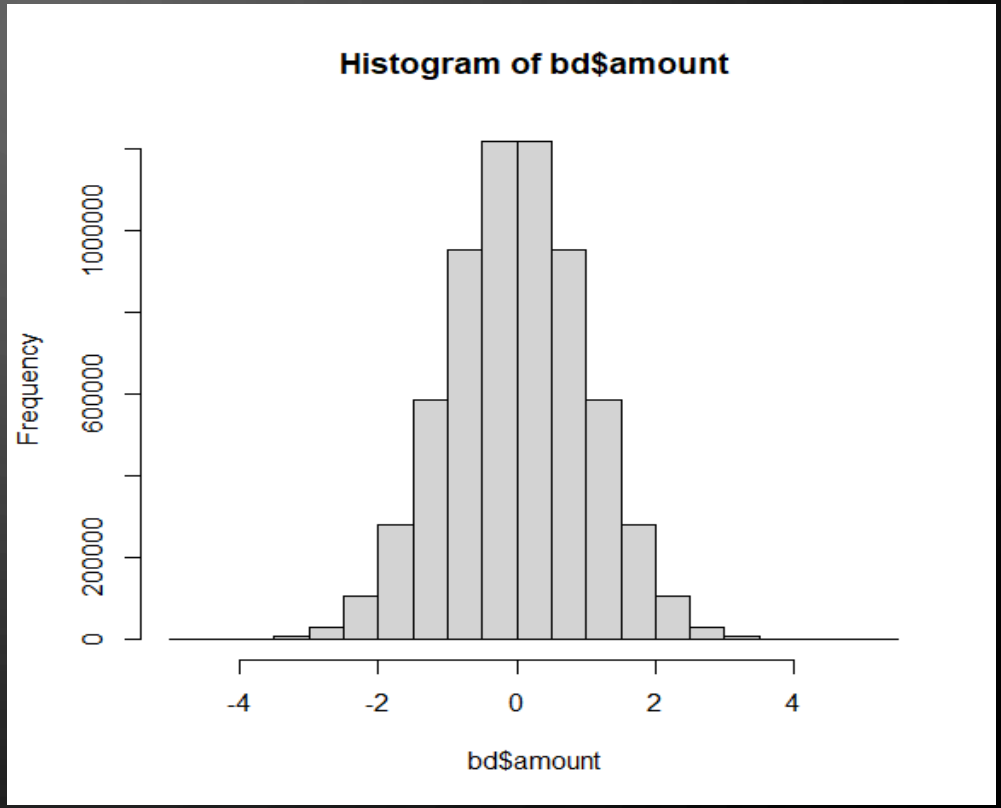
UTILIZANDO O PACOTE BESTNORMALIZE PARA FAZER AS TRANSFORMAÇÃO DAS VARIAVEIS AMOUNT E OLDBALANCEORG

```
BD$AMOUNT = BESTNORMALIZE::BESTNORMALIZE(BD$AMOUNT)$X.T
```

```
BD$OLDBALANCEORG = BESTNORMALIZE::BESTNORMALIZE(BD$OLDBALANCEORG)$X.T
```

APÓS AS TRANSFORMAÇÕES FORAM FEITOS NOVOS HISTOGRAMAS.

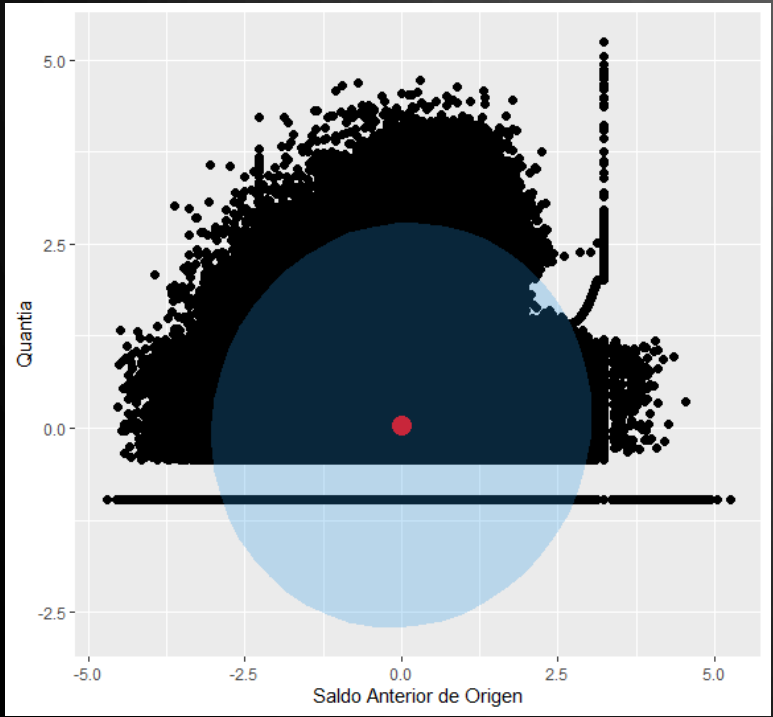




VETORES DE MÉDIA E MATRIZ DE CORRELAÇÃO E COVARIÂNCIA

```
> # Vetor de medias
> Vmean  = colMeans(bd)
> Vmean
       amount oldbalanceOrg
-3.519306e-06  3.390802e-17
> # matriz de covariancia
> Mcov = cov(bd)
> Mcov
                 amount oldbalanceOrg
amount        0.99997264    0.04671722
oldbalanceOrg 0.04671722    1.00000000
> ## Matriz de correlação
> cor(bd)
                 amount oldbalanceOrg
amount        1.00000000    0.04671786
oldbalanceOrg 0.04671786    1.00000000
```

SCATTER PLOT



Como observado na Covariância é praticamente 0 assim não foi formado uma elipse e ser praticamente um círculo com um nível de confiança de 0.99

CALCULANDO A DISTANCIA DE MAHALANOBIS

CALCULAMOS A DISTANCIA DE MAHALANOBIS UTILIZANDO A FUNÇÃO MAHALANOBIS(BANCO DADOS, MÉDIA, COVARIÂNCIA) COMO SENDO AS VARIÁVEIS DA FUNÇÃO E FOI FEITO TAMBÉM O PONTO DE CORTE COMO A FUNÇÃO QCHISQ.

```
D2 <- MAHALANOBIS(BD, VMEAN, MCOV);D2
```

```
DADOS$D2<- D2 PONTO_CORTE = QCHISQ(P = 0.01, DF = 2, LOWER.TAIL = F)
```

COMPARANDO AS MAIORES DISTANCIAS

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud	D2
1	425	TRANSFER	1e+07	C40489106	59585040	49585040	C650095152	0	0	1	0	41.68509
2	730	TRANSFER	1e+07	C726730575	57316255	47316255	C1364745638	0	0	1	0	39.18680
3	646	TRANSFER	1e+07	C590657619	50399045	40399045	C1971187430	0	0	1	0	38.02874
4	425	TRANSFER	1e+07	C1551381510	49585040	39585040	C1042012237	0	0	1	0	37.26728
5	730	TRANSFER	1e+07	C507645439	47316255	37316255	C270374999	0	0	1	0	36.69927
6	741	TRANSFER	1e+07	C780743034	45674548	35674548	C491519946	0	0	1	0	36.24619

CONCLUISÕES

COMO OBSERVADO NA TABELA ANTERIOR OS VALORES COM AS DISTANCIAS GRANDES TEM VALORES GRANDE ASSOCIADOS TAMBÉM O QUE APARENTEMENTE TEM UMA CORRELAÇÃO COM AS FRAUDES.