RELATÓRIO DA ANÁLISE EXPLORATÓRIA DOS DADOS DA EMPRESA PORTO SEGURO

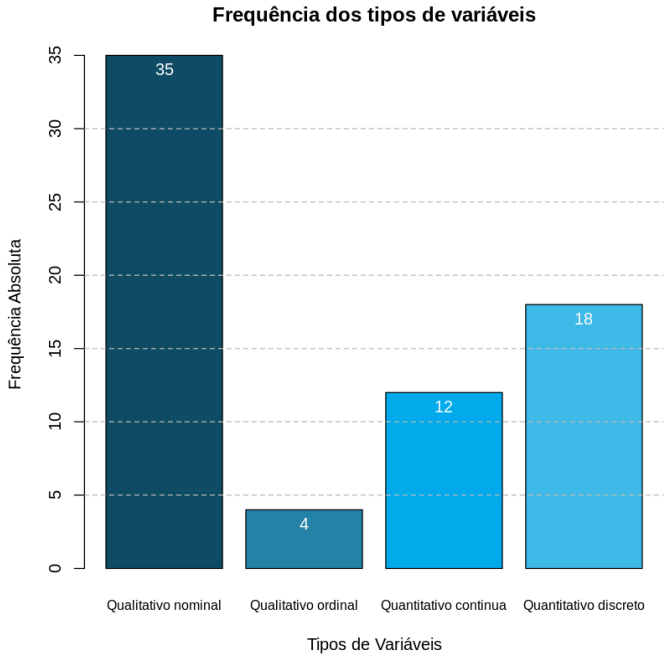
Síntese sobre os dados e propósito do estudo

A base de dados fornecida pela empresa Porto Seguro, contém 14.123 linhas e 70 colunas. Cada linha representa um indivíduo e nela estão contidos os valores das variáveis referentes à pessoa. Já as colunas, a primeira representa o "id", da segunda à sexagésima nona estão as variáveis explicativas e a última coluna, a variável resposta, sendo "1" caso a pessoa tenha efetuado a compra do produto e "0" caso contrário. Para fins da análise, cabe excluirmos a primeira coluna ("id"), para não influenciar numericamente em nosso estudo, assim a tabela fica com a dimensão de 14.123 linhas e 69 colunas [1].

O intuito da análise exploratória dos dados e suas relações, é predizer a probabilidade de um cliente concluir a compra do produto, baseado nas suas variáveis explicativas.

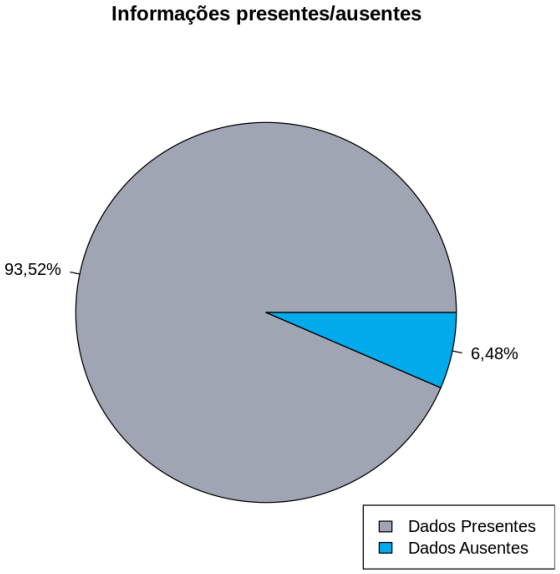
Estrutura dos dados e informações gerais dos arquivos "metadata.csv" e "train.csv"

Como já citado, analisaremos uma base de dimensão 14.123 × 69. Conforme o gráfico de barras abaixo, cada uma das 69 colunas dividem-se em quatro tipos de variáveis [2] :

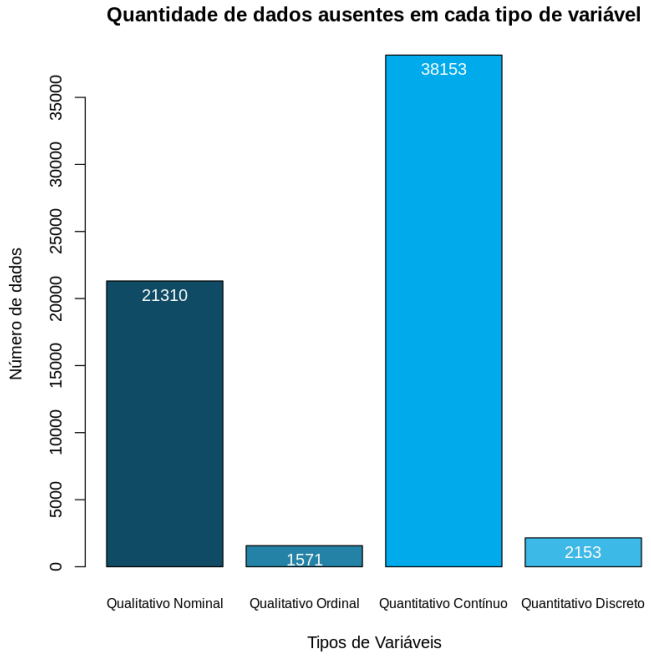


Nota-se, que das 69 variáveis, a superioridade se categoriza como qualitativa nominal, enquanto a minoria em qualitativa ordinal.

Outro ponto de importância, é verificarmos os dados faltantes, de modo que possamos averiguar se o volume desses pode prejudicar o ensaio [3].



Como temos a dimensão 14.123 × 69, isso significa um total de 974.487 células de dados. Nossa análise, demonstra que 63.187 é a quantidade de informações faltantes, implicando em valor relativo, cerca de 6,48% da totalidade dos dados, como mostra o gráfico acima [4]. Esta proporção não deve prejudicar o estudo de modo geral, contudo, a análise particular de uma variável pode correr risco, visto que, quanto menor a quantidade de dados, menor a qualidade do exame sobre as informações. O gráfico a seguir mostra a distribuição dos dados faltantes em cada classe de variável [5]:

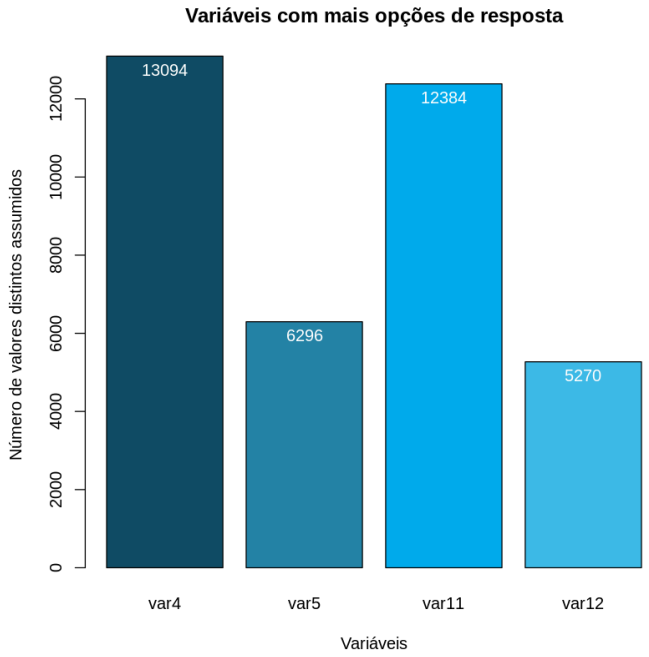


Por mais que a classe de dados qualitativos nominais seja maioria, as informações ausentes estão em superioridade nos dados do tipo quantitativo contínuo.

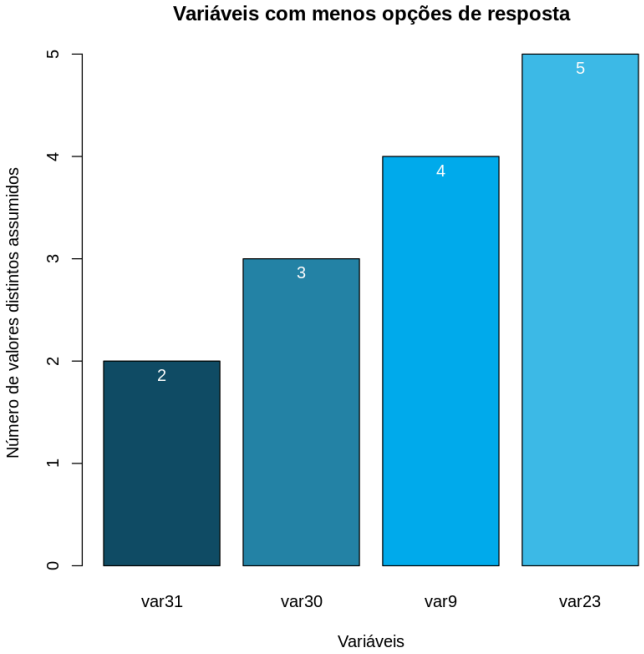
A seguir, mostra-se uma série de observações gerais de cada classe de variável e seus dados:

Qualitativas nominais

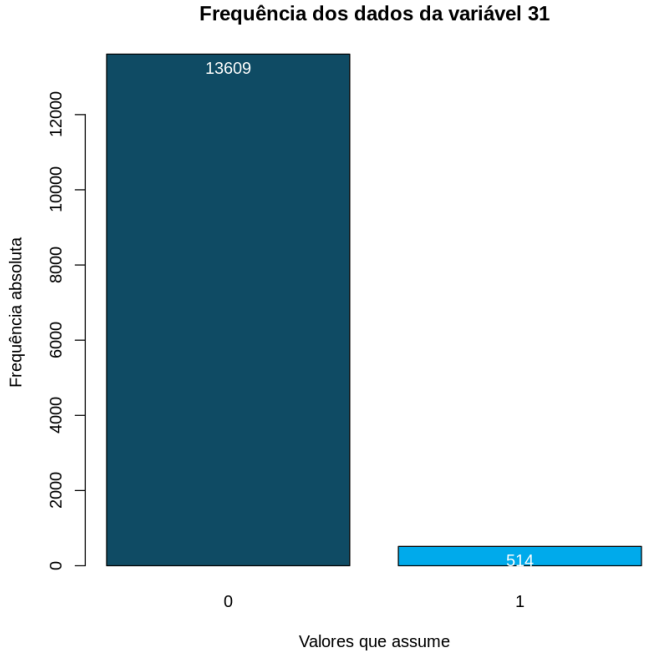
Volume de dados das variáveis [6]



Através do gráfico, podemos entender que a variável 4, dentre todas nominais, é a que mais assume valores distintos entre si, incluindo o valor 'NA'. Logo, dentro de 14123 dados da variável 4, cerca de 92% não são iguais entre si, enquanto o resto, é apenas algum valor repetido já contado nos 92%. Essa análise, mostra que é inviável plotar algum gráfico que mostre a frequência dos dados em sua totalidade, já que teremos muitos valores com frequência igual a 1. Além disso, nota-se a vasta extensão dos dados diferentes fornecidos pelos indivíduos. Em contrapartida, as variáveis 31, 30, 9 e 23 são o oposto da 4, tendo suas possibilidades de valores minimizadas em comparação com as variáveis 4, 5, 11 e 12, como mostra o gráfico de barras a seguir:



Diferentemente da variável 4, é viável criar o gráfico de frequência da variável 31 em sua totalidade de dados:

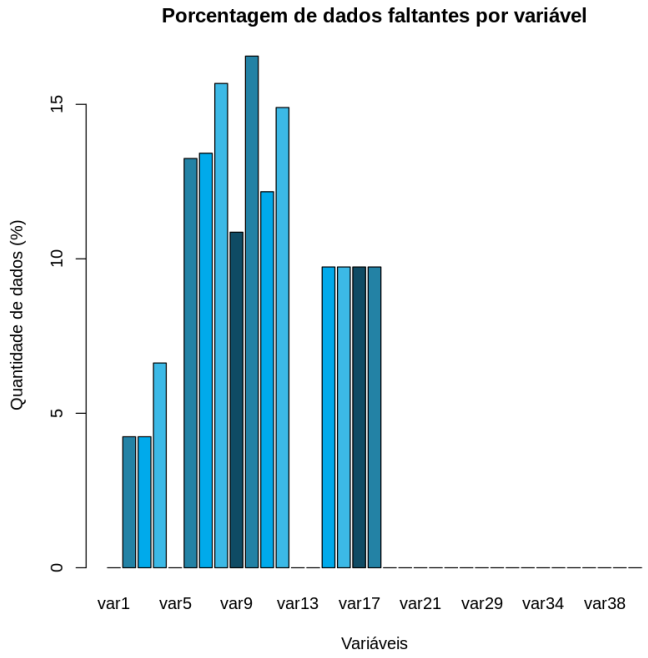


Moda [7]

	Variaveis	Moda	Nº de ocorrência
	<chr>	<chr>	<int>
33	var38	0	10748
34	var39	4	12084
35	var41	3	11738

Algumas modas chamam a atenção em seu número de frequência, sobretudo em tamanha magnitude, a variável 39 tem cerca de 85% de seus dados voltados para apenas um valor, a moda 4. Como visto no gráfico de frequência dos dados da variável 31, o valor de sua moda é zero e sua frequência é de 13609, equivalente a 96% dos dados da variável .

Dados ausentes por variável [8]



A variável nominal mais prejudicada pelos dados faltantes é a variável 10. Aproximadamente 16,5% de seus dados estão ausentes, o que em valor absoluto é 2339. Mas ainda assim, é uma quantidade considerada baixa em relação aos outros tipos de variáveis.

Associação entre a variável 30 e 31 [9]

Para analisarmos a associação entre essas variáveis, usaremos o teste de chi quadrado. Para isso, montamos a tabela de frequência observada (porcentagens aproximadas):

Var30/Var31	0	1	Total
0	0	514 (100%)	514 (100%)
1	7631 (100%)	0	7631 (100%)
2	5978 (100%)	0	5978 (100%)
Total	13609 (96%)	514 (4%)	14123 (100%)

Tabela de frequência esperada:

Var30/Var31	0	1	Total
0	494 (96%)	21 (4%)	514 (100%)
1	7326 (96%)	306 (4%)	7631 (100%)
2	5789 (96%)	187 (4%)	5978 (100%)
Total	13609 (96%)	514 (4%)	14123 (100%)

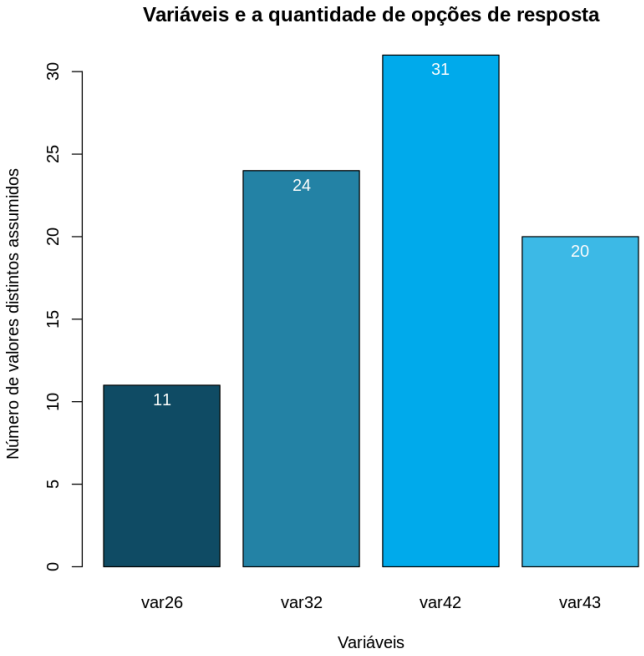
Desvio entre os valores observados e os esperados

Var30/Var31	0	1
0	-494	493
1	305	-306
2	189	-187

Por fim, temos que o valor de chi quadrado é 14123 e seu p-valor é menor que 2,2 x 10^(-16). Isso significa que não existe, ou quase não existe diferença entre as variáveis observadas e as variáveis esperadas, teóricas. Portanto, isso indica que existe associação entre as variáveis 30 e 31.

Qualitativas ordinais

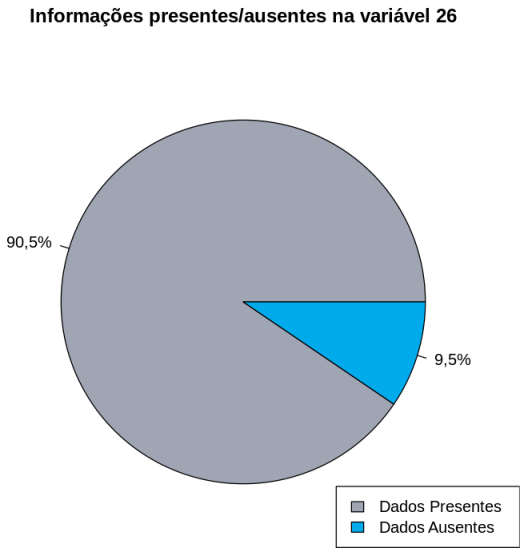
Volume de dados das variáveis [6]



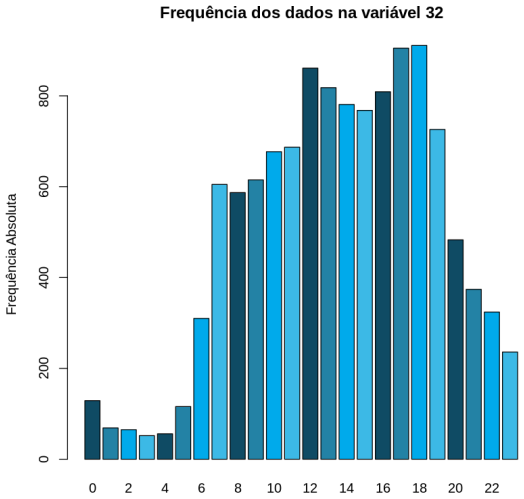
As variáveis de categoria qualitativa ordinal representam a menor porção dos dados. Conforme o gráfico acima, por exemplo, a variável 42 assumiu 31 valores distintos em 14123 dados, incluindo o valor de dado faltante (NA). Em comparação com o maior valor das qualitativas nominais, é muito baixo.

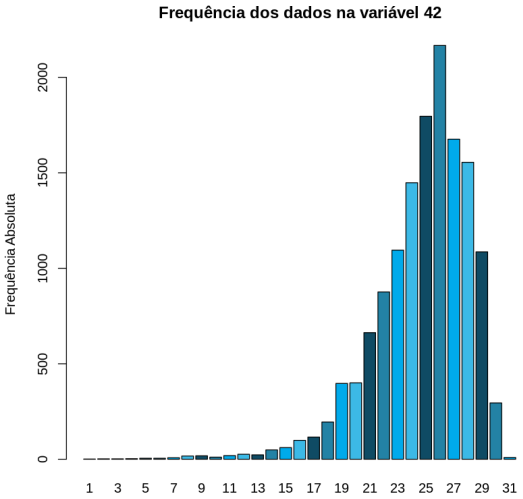
Dados ausentes por variável [8]

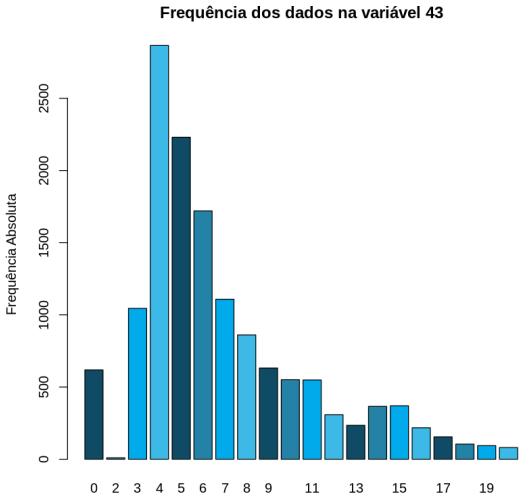
A única variável mais prejudicada devido à falta de dados, é a 26:

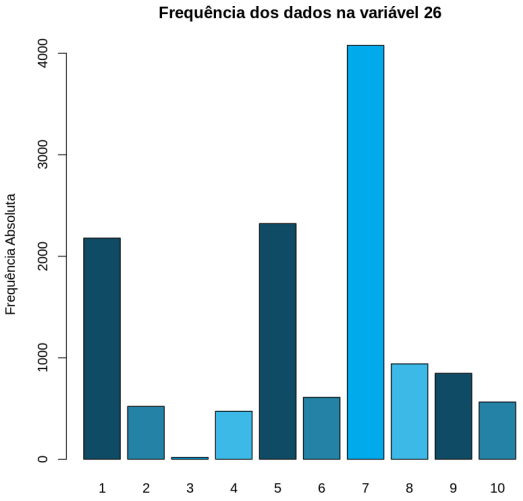


Frequência das variáveis [10]









A frequência da variável 32 parece estar mais concentrada em uma faixa a partir do valor 6 até o último valor, o 22. Aparenta ter uma leve simetria na frequência dos dados. Já o gráfico da variável 42, mostra que seus dados estão mais concentrados a partir do valor 17. A posição das barras se assemelham a curva de frequência de assimetria positiva. De modo contrário, a variável 43 tem seus dados concentrados nas primeiras opções de valores. De fato, é também observável, que a posição de suas barras se assemelham a curva de frequência com assimetria negativa. Por fim, a variável 26 tem uma distribuição de dados mais regular, com maior concentração nos valores 1, 5 e 7.

Moda [7]

É nítido a presença da moda de cada variável nos gráficos acima, mas para efeito de confirmação:

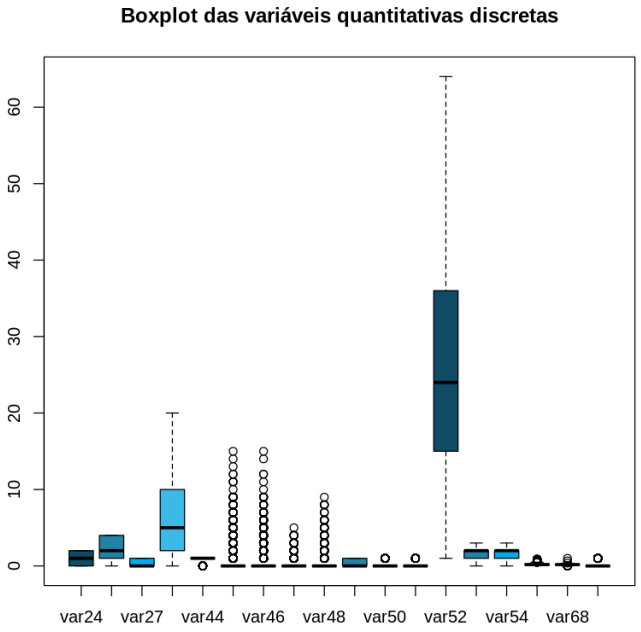
Variaveis	Moda	Nº de ocorrência
<chr>	<chr>	<int>
var26	7	3467
var32	18	911
var42	26	1833
var43	4	2423

Quantitativas discretas [11]

Variaveis	Média	Mediana	Desvio Padrão	Desvio Absoluto Mediano	Max.	Min.
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
var24	0.95864901	1.0000000	0.91556321	1.00000000	2.0000000	0.000000000
var25	2.09318134	2.0000000	1.38868139	1.00000000	4.0000000	0.000000000
var27	0.37003470	0.0000000	0.48283074	0.00000000	1.0000000	0.000000000
var40	6.01033775	5.0000000	4.78317943	4.00000000	20.0000000	0.000000000
var44	0.86575090	1.0000000	0.34093182	0.00000000	1.0000000	0.000000000
var45	0.27522481	0.0000000	0.91468689	0.00000000	15.0000000	0.000000000
var46	0.22162430	0.0000000	0.80336690	0.00000000	15.0000000	0.000000000
var47	0.04984777	0.0000000	0.26913489	0.00000000	5.0000000	0.000000000
var48	0.10606812	0.0000000	0.48756742	0.00000000	9.0000000	0.000000000
var49	0.34369468	0.0000000	0.47495749	0.00000000	1.0000000	0.000000000
var50	0.14833959	0.0000000	0.35544887	0.00000000	1.0000000	0.000000000
var51	0.05529987	0.0000000	0.22857272	0.00000000	1.0000000	0.000000000
var52	26.10485344	24.0000000	13.77744133	10.00000000	64.0000000	1.000000000
var53	1.63527579	2.0000000	0.74965439	0.00000000	3.0000000	0.000000000
var54	1.51922396	2.0000000	0.71400950	0.00000000	3.0000000	0.000000000
var67	0.20688940	0.1764706	0.12264738	0.05882353	0.9117647	0.029411765
var68	0.17987256	0.1764706	0.05596251	0.04044118	1.0000000	0.003676471

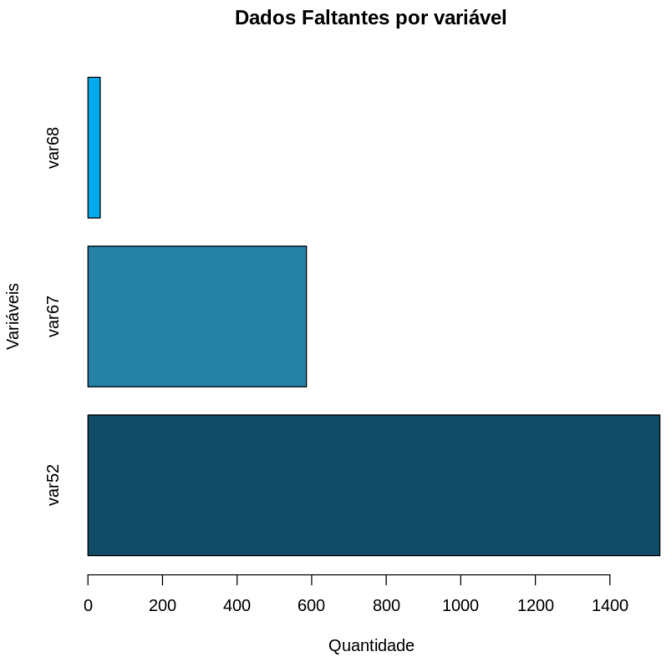
Nesta tabela podemos retirar algumas informações importantes sobre as variáveis. A variável 52 se destaca pela amplitude de seu intervalo de valores, logo atrás estão as variáveis 40, 45 e 46, com amplitudes 20, 15 e 15, respectivamente. Contudo, uma grande amplitude,

pode indicar que existem valores discrepantes. Analisando o boxplot [12] a seguir, das variáveis citadas, notamos que apenas a variável 40 e a 52 não possuem valores discrepantes.

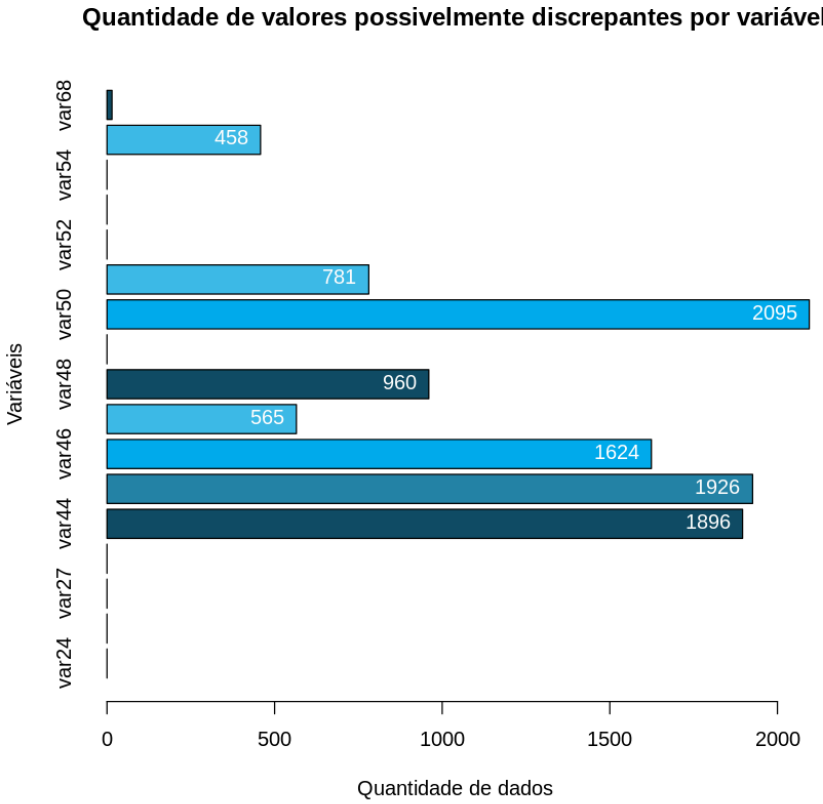


Observa-se cerca de 6 variáveis com valores compactados em uma faixa menor, entre 0 e 1. Além disso, é nítido que algumas variáveis apresentam valores possivelmente discrepantes, como mostra o boxplot.

Dados faltantes e Dados possivelmente discrepantes [8]



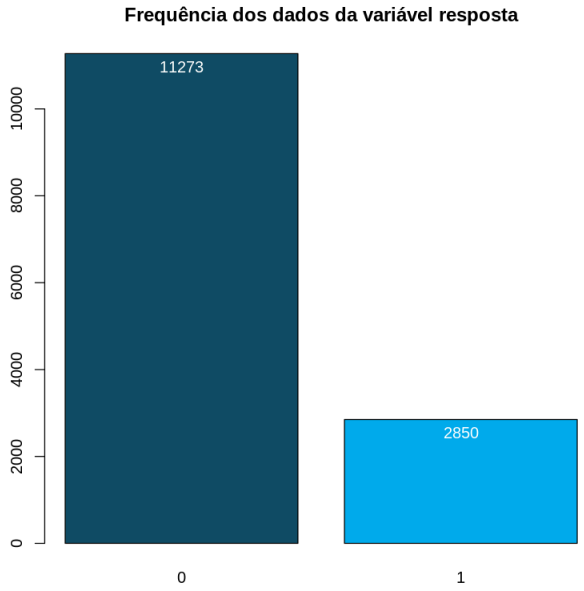
Apenas as variáveis 52, 67 e 68 apresentam dados faltantes na categoria quantitativa discreta. Ainda assim, a quantidade não prejudica em alto nível as variáveis [13].



A gama de valores que aparentam ser discrepantes, é notável em mais variáveis que as informações sobre valores ausentes, como visto no boxplot acima. Nota-se que, a variável 50, é a que mais apresenta valores possivelmente discrepantes [14].

Entretanto, ao verificar a frequência dos dados, conclui-se que é uma variável que assume apenas dois valores, o que explica tantos dados outliers em seu boxplot. Vale ressaltar, que as estatísticas foram geradas com base nas informações do boxplot. Por isso, elas indicam valores possivelmente discrepantes, o que deve ser verificado posteriormente.

Variável resposta (y) e sua associação com a variável 31



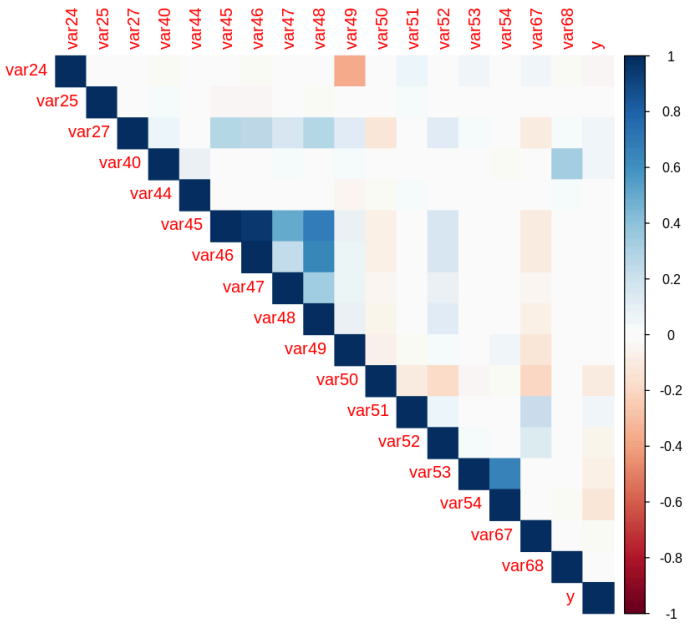
Todos os dados sobre a variável resposta estão presentes. Nota-se que a maioria das pessoas não efetuou a compra do produto, indicando uma dificuldade no objetivo principal [15]. Agora, analisaremos sua associação com a variável 31. A tabela de valores observados [16]:

y/var31	0	1	Total
0	10858 (96%)	415 (4%)	11273 (100%)
1	2751 (96%)	99 (4%)	2850 (100%)
Total	13609 (96%)	514 (4%)	14123 (100%)

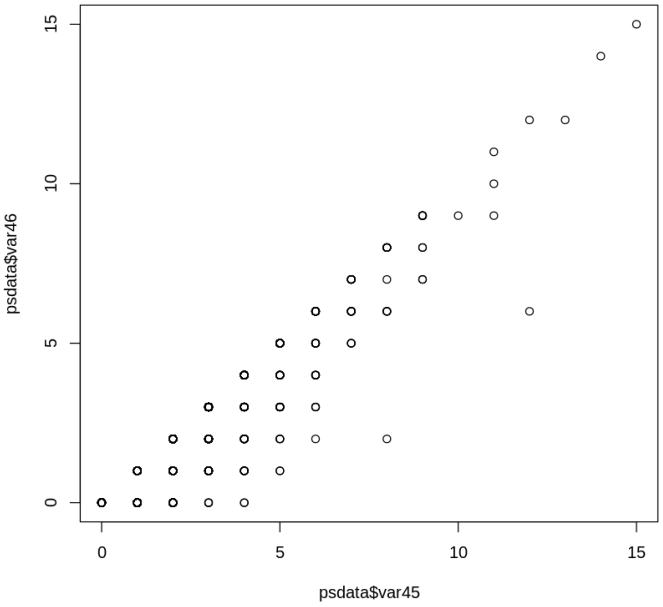
Nem precisamos montar a tabela de valores esperados. Pois, nota-se que a própria tabela de valores observados coincide com a de esperados. Portanto, os dados observados não diferem, ou diferem pouco, dos dados esperados. Obtemos um valor de chi quadrado igual a 0,22369 e um p-valor igual a 0,6362. Se usarmos a convenção da porcentagem de confiança igual a 5%, podemos sugerir então que não há associação entre as variáveis, pois p-valor > 5%.

Correlação [17]

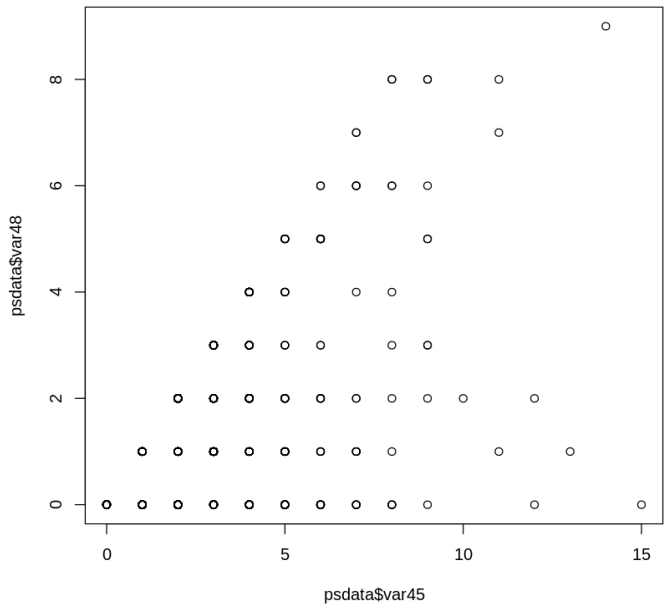
Podemos calcular a correlação entre as variáveis quantitativas e obter o gráfico abaixo:



Segundo o gráfico que exibe a correlação entre as variáveis, os pares que mais possuem correlação são (var45, var46), (var45, var48). Plotados os gráficos de dispersão [18]:



Visto que, o par (var45, var46) é o que apresenta o maior nível de correlação, é possível enxergar uma relação linear, que segue em um crescimento proporcional entre as variáveis .



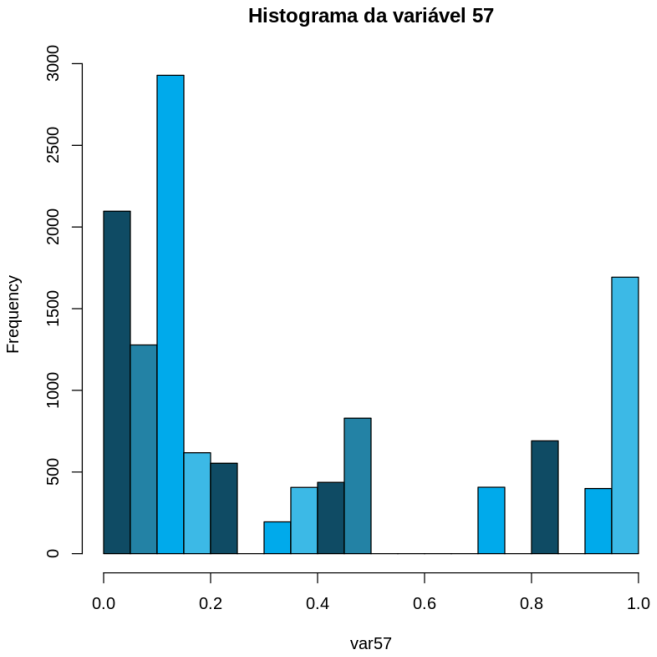
Por mais que seja o segundo maior nível de correlação, a relação entre as variáveis parece um pouco mais abstrata, mas ainda assim segue uma relação de crescimento proporcional.

Quantitativas contínuas [11]

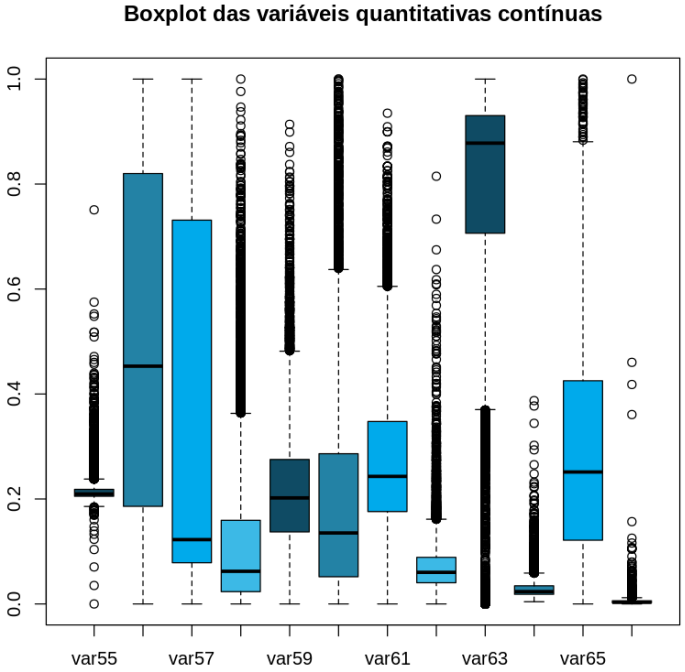
Variaveis	Média	Mediana	Desvio Padrão	Desvio Absoluto Mediano	Max.	Min.
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
var55	0.216031191	0.20946498	0.02264158	0.005370749	0.7508578	0.00000000
var56	0.490075463	0.45300000	0.32526993	0.302000000	1.0000000	0.00000000
var57	0.345775190	0.12258065	0.35699975	0.122580645	1.0000000	0.00000000
var58	0.116597928	0.06224898	0.13878019	0.046114232	1.0000000	0.00000000
var59	0.221587802	0.20204574	0.12000767	0.068259171	0.9137730	0.00000000
var60	0.203007885	0.13526722	0.20486673	0.099806595	1.0000000	0.00000000
var61	0.270888062	0.24305264	0.13831680	0.079908978	0.9349940	0.00000000
var62	0.073086007	0.06026501	0.05396327	0.022765030	0.8148459	0.00000000
var63	0.779220518	0.87790111	0.23406807	0.067608476	1.0000000	0.00000000
var64	0.028777806	0.02347952	0.01868114	0.006681059	0.3873865	0.00426668
var65	0.301399586	0.25142725	0.22725378	0.144576454	1.0000000	0.00000000
var66	0.007116482	0.00319933	0.02998180	0.001825796	1.0000000	0.00000000

Algo singular ocorre com a variável 57. O seu valor da mediana e desvio absoluto mediano coincidem. Isso significa que a mediana é um valor característico entre os dados, é algo bem peculiar, pois temos uma medida de centralidade e uma de dispersão assumindo o

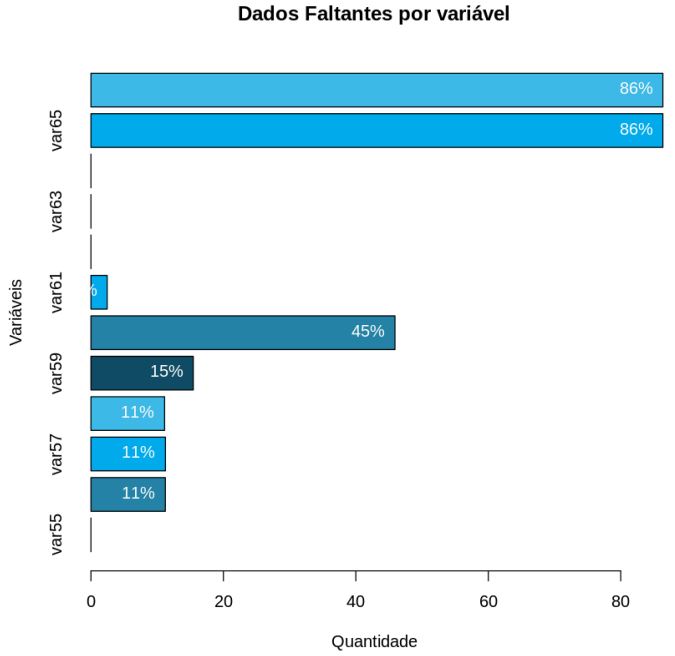
mesmo valor descritivo. Diferentemente, sua média e desvio padrão não são iguais, mas possuem uma alta proximidade. Há uma assimetria em seus dados, como mostra o histograma, o que também será confirmado no boxplot [19].



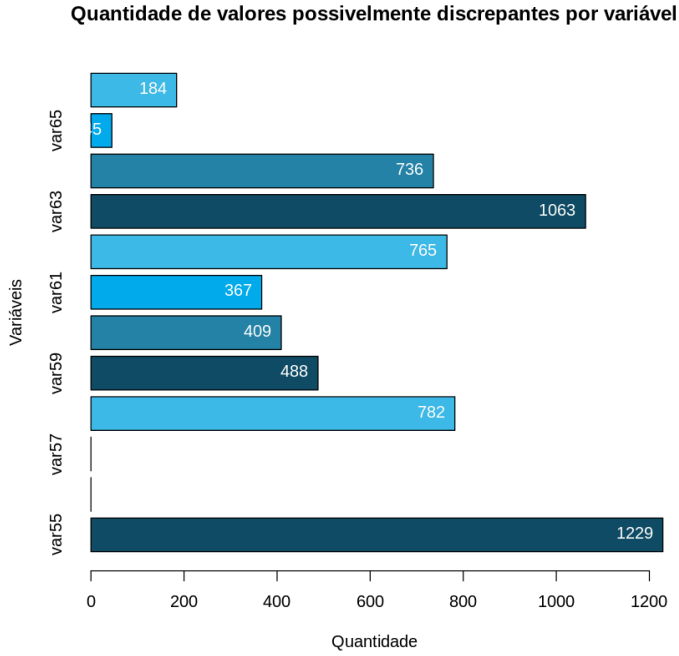
Os dados das variáveis quantitativas contínuas estão limitados no intervalo de 0 a 1. Além disso, um excesso de valores possivelmente discrepantes é visível no boxplot, aparentemente, apenas as variáveis 56 e 57 não possuem valores outliers [20].



Dados faltantes e Dados possivelmente discrepantes [21]

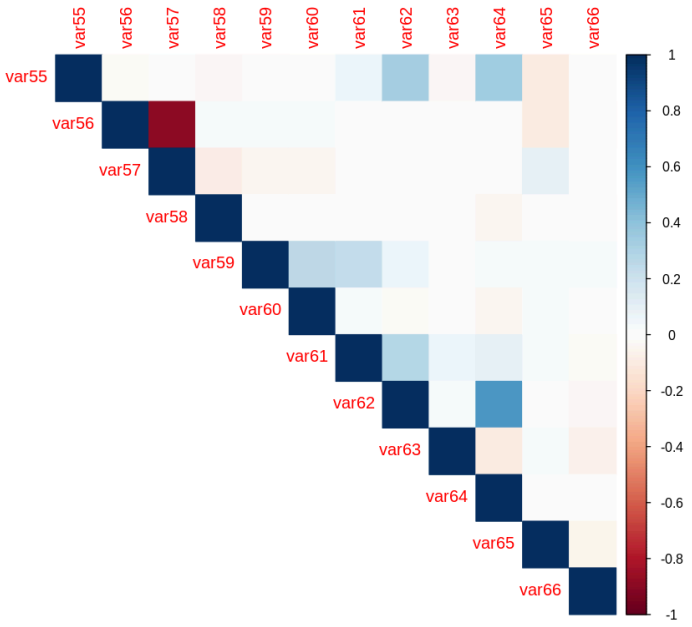


As variáveis 65 e 66 são as mais prejudicadas devido ao dado faltante. Apenas 1978 dados estão presentes em ambas variáveis, dificultando a análise individual da variável. Ademais, dos 1978, 184 são outliers, o que demonstra que a variável 66 aponta ser a pior em questão de quantidade de dados [22].

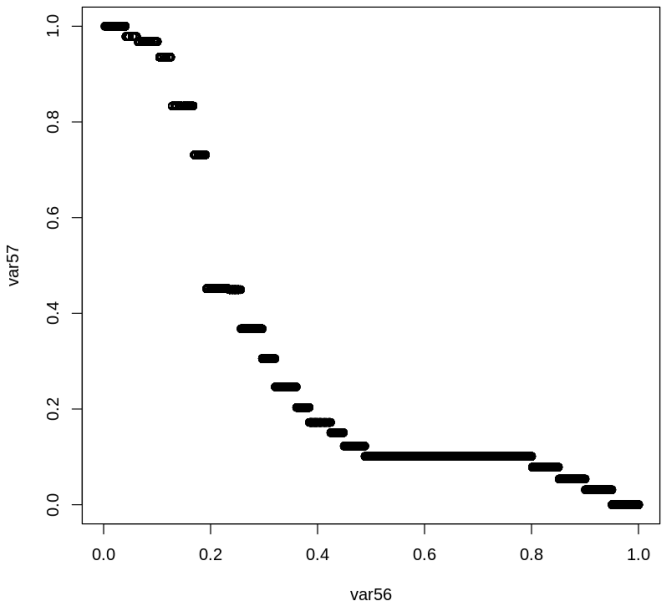


Correlação 23

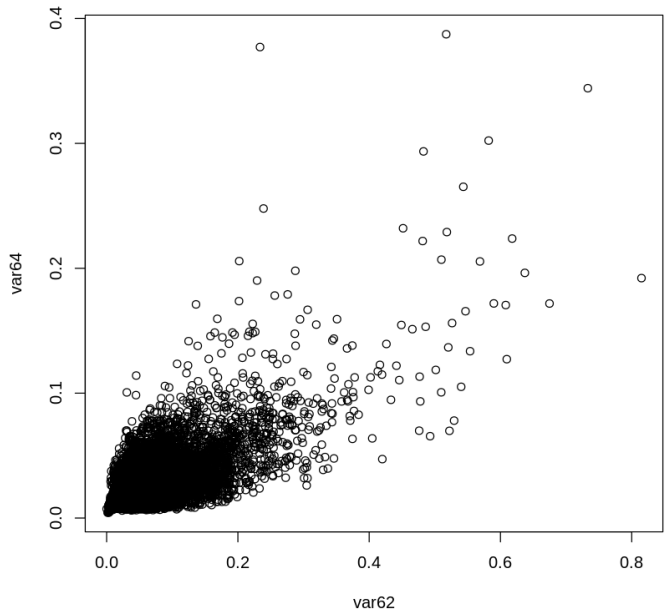
Podemos calcular a correlação entre as variáveis quantitativas contínuas e obter o gráfico abaixo:



Observa-se que, as variáveis que possuem um nível considerável de dependência entre si, são os pares (var56, var57) e (var62,var64). Podemos confirmar a correlação através dos seus gráficos de dispersão:



Percebe-se que a medida que a variável 56 cresce, a 57 diminui, configurando uma relação inversamente proporcional. Já para o par (var62,var64):



Ao contrário da relação anterior, o gráfico mostra uma relação de proporcionalidade, ou seja, quando uma variável cresce a outra também tende a crescer.

APÊNDICE DOS CÓDIGOS EM R

1. Exclusão da primeira coluna e dimensão dos dados, “train.csv” :

```
psdata <- read.csv('train.csv', na.strings = '-999')
psmeta <- read.csv('metadata.csv')
psdata <- psdata[,-1]
psmeta <- psmeta[-1,]
dim(psdata)
```

2. Contagem dos tipos de dados, “meta.csv”:

```
text(
barplot(
  table(psmeta[,2]),
  cex.names= .8,
  xlab = "Tipos de Variáveis",
  ylab = "Frequência Absoluta",
  xpd = F,
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6"),
  )
  , y = c(35, 4, 12, 18), labels = c(35, 4, 12, 18), pos = 1, col = "white")
  abline(h = c(5, 10, 15, 20, 25, 30), col = "gray", lty = "dashed")
```

3. Gráfico de pizza, dados presentes/ausentes:

```
colors = c("#a1a6b4","#01AEEF")
pie(
  table(is.na(psdata)),
  main = "Informações presentes ou ausentes",
  label = c("93,52%", "6,48%"),
  col = colors
)
legend("bottomright", c("Dados Presentes", "Dados Ausentes"), fill = colors)
```

4. Cálculo para os dados faltantes:

```
#Dados Faltantes
dadosFaltantes = sum(1*is.na(psdata)) # 63187
dFrelativo = (dadosFaltantes/(14123*69))*100
dFrelativo
```

5. Número de dados faltantes em cada tipo de variável:

```
QN =sum(1*is.na(psdata[QualNom]))
QO =sum(1*is.na(psdata[QualOrd]))
QC = sum(1*is.na(psdata[QuantCont]))
QD = sum(1*is.na(psdata[QuantDisc]))
valores = c(QN, QO, QC, QD)
nomes = c("Qualitativo Nominal", "Qualitativo Ordinal", "Quantitativo Contínuo",
"Quantitativo Discreto")

text(
barplot(
  valores,
  names.arg = nomes,
  main = "Quantidade de dados ausentes em cada tipo de variável",
  cex.names= .8,
  xlab = "Tipos de Variáveis",
  ylab = "Número de dados",
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6")
  )
  , y = valores, labels = valores, pos = 1, col = "white"
  )
```

6. Função que calcula o volume de dados, quantidade de dados únicos:

```
OpcoesDeValoresVarQual <- function(tipoDeVariavel){
  #Monta uma tabela mostrando os valores únicos assumidos por variável
  tabelaTipoVariavel = psdata[tipoDeVariavel]
  df = data.frame(
    variavel = c(names(tabelaTipoVariavel))
  )
  for(v in 1:ncol(tabelaTipoVariavel)){
                df[v,"Qtd_de_valores_diferentes_que_a_variável_pode_assumir"]    =
length(unique(tabelaTipoVariavel[,v]))
  }
  return (df)
}
```

7. Função que calcula a Moda:

```
Moda <- function(tipoDeVariavel){
  #Função que calcula a moda para todas as variáveis do tipo especificado
  tabelaTipoVariavel = psdata[tipoDeVariavel]
  medidasResumo = data.frame(
    Variaveis = c(names(tabelaTipoVariavel))
  )

  for(v in 1:ncol(tabelaTipoVariavel)){
    colunaAtual = tabelaTipoVariavel[,v]
    tabelaFrequencia = table(tabelaTipoVariavel[,v])
     medidasResumo[v, "Moda"] = toString(names(tabelaFrequencia[tabelaFrequencia ==
max(tabelaFrequencia, na.rm = T)]))
    medidasResumo[v, "N° de ocorrência"] = max(tabelaFrequencia, na.rm = T)
  }

  return(medidasResumo)
}
```

8. Função que calcula os dados ausentes por variável:

```
DadosFaltantesPorVariavel <- function(tipoDeVariavel){
  #Seleciona a tabela do tipo de variável especificada
  #Retorna uma tabela com a quantidade de dados faltantes por variável
  tabelaTipoVariavel = psdata[tipoDeVariavel]

  tabela_Dados_Faltantes = data.frame(
    variaveis = c(names(tabelaTipoVariavel))
  )

  for( v in 1:ncol(tabelaTipoVariavel)){
    colunaAtual = tabelaTipoVariavel[,v]
    tabela_Dados_Faltantes[v,"Qtd. Dados Ausentes"] = sum(1*is.na(colunaAtual))
  }

     tabela_Dados_Faltantes[,"Porcentagem | (Dados Ausentes ÷ 14123)%"] =
(tabela_Dados_Faltantes[,"Qtd. Dados Ausentes"]/14123)*100

  return(tabela_Dados_Faltantes)
}
```

9. Teste chi quadrado “var30” e “var31”:

```
#Qualitativas Nominais
table(psdata[c("var30", "var31")])
chisq.test(table(psdata[c("var30", "var31")]))
```

10. Plot das frequência das variáveis qualitativas ordinais:

```
barplot(
  table(psdata$var26),
  main = "Frequência dos dados na variável 26",
  ylab = "Frequência Absoluta",
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6")
)

barplot(
  table(psdata$var32),
  main = "Frequência dos dados na variável 32",
  ylab = "Frequência Absoluta",
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6")
)

barplot(
  table(psdata$var42),
  main = "Frequência dos dados na variável 42",
  ylab = "Frequência Absoluta",
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6")
)

barplot(
  table(psdata$var43),
  main = "Frequência dos dados na variável 43",
  ylab = "Frequência Absoluta",
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6")
)
```

11. Função tabela de medidas resumo:

```
TabelaDeMedidasResumo <- function(tipoDeVariavel){
  # Retorna uma tabela com insformações das medidas-resumos de posição e dispersão

  tabelaTipoVariavel = psdata[tipoDeVariavel]
  medidasResumo = data.frame(
    Variaveis = c(names(tabelaTipoVariavel))
  )

  for(v in 1:ncol(tabelaTipoVariavel)){
      colunaAtual = tabelaTipoVariavel[,v]
      medidasResumo[v, "Média"] = mean(colunaAtual, na.rm = T)
      medidasResumo[v, "Mediana"] = median(colunaAtual, na.rm = T)
      medidasResumo[v,"Desvio Padrão"] = sd(colunaAtual, na.rm = T)
      medidasResumo[v, "Coeficiente de variacao (%)"] =
          medidasResumo[v, "Desvio Absoluto Mediano"] = median(abs(colunaAtual -
median(colunaAtual, na.rm = T)), na.rm = T)
      medidasResumo[v, "Max."] = max(colunaAtual, na.rm = T)
      medidasResumo[v,"Min."] = min(colunaAtual, na.rm = T)
    }
```

```
  return(medidasResumo)
  }
```

12. Boxplot das variáveis quantitativas discretas:

```
boxplot(
  psdata[QuantDisc],
  main = "Boxplot das variáveis quantitativas discretas",
  col =  c("#114F66", "#2585A8", "#01AEEF", "#40BAE6")
  )
```

13. Função que calcula a quantidade de dados possivelmente discrepantes:

```
QuantidadeDeValoresPossivelmenteDiscrepantes <- function (tipoDeVariavel){
  # Retorna uma tabela em que a primeira coluna se refere ao nome da variável e as
outras os seus valores discrepantes

  tabelaTipoVariavel = psdata[tipoDeVariavel]

   Quantidade_De_Possiveis_Valores_Discrepantes = data.frame(
      variáveis = c(names(tabelaTipoVariavel))
   )

  for(v in 1:ncol(tabelaTipoVariavel)){
    coluna = tabelaTipoVariavel[ ,v] # Acessando coluna
    resumoColuna = boxplot.stats(coluna) # Acessando os valores de LI e LS
    LI = resumoColuna$stats[1] #Limite Inferior
    LS = resumoColuna$stats[5] #Limite Superior
       Quantidade_De_Possiveis_Valores_Discrepantes[v,"Quantidade Acima do LS"] =
length(coluna[(coluna > LS) & (!is.na(coluna) )])
      Quantidade_De_Possiveis_Valores_Discrepantes[v,"Quantidade Abaixo do LI"] =
length(coluna[(coluna < LI) & (!is.na(coluna) )])
  }

          Quantidade_De_Possiveis_Valores_Discrepantes[,"total_discrepantes"]     =
Quantidade_De_Possiveis_Valores_Discrepantes[,"Quantidade Acima do LS"] +
    Quantidade_De_Possiveis_Valores_Discrepantes[,"Quantidade Abaixo do LI"]

  return(Quantidade_De_Possiveis_Valores_Discrepantes)
}
```

14. Boxplot da Variável 50:

```
boxplot(
  psdata$var50,
  main = "Boxplot da Variável 50",
  col = "#2585A8"
  )
```

15. Frequência dos dados da variável y

```
taby = table(psdata$y)
taby
text(
barplot(
  taby,
  main = "Frequência dos dados da variável resposta",
  col = c("#114F66","#01AEEF"),
  ), y = c(11273, 2850), labels = c(11273, 2850), pos = 1, col = "white"
)
```

16. Teste de chi quadrado entre “y” e “var31”:

```
table(psdata[c("y", "var31")])
chisq.test(table(psdata[c("y", "var31")]))
```

17. Gráfico de correlação, variáveis quantitativas discretas

```
cor(psdata[,QuantDisc], use = "complete.obs")
corrplot(cor(psdata[,QuantDisc], use = "complete.obs"), type = "upper", method = "color")
```

18. Plot dos gráficos de dispersão:

```
plot(psdata$var45, psdata$var46)
plot(psdata$var45, psdata$var48)
```

19. Histograma da variável 57:

```
var57 = psdata$var57
hist(
  var57,
  main = "Histograma da variável 57",
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6")
)
```

20. Boxplot das variáveis quantitativas contínuas:

```
boxplot(
  psdata[QuantCont],
  main = "Boxplot das variáveis quantitativas contínuas",
  col =  c("#114F66", "#2585A8", "#01AEEF", "#40BAE6")
  )
```

21. Dados possivelmente discrepantes, QuantCont:

```
#DADOS Possivelmete DISCREPANTES
qvpdQC = QuantidadeDeValoresPossivelmenteDiscrepantes(QuantCont)

text(
barplot(
  qvpdQC[,4],
  names.arg = qvpdQC[,1],
```

```
  main = "Quantidade de valores possivelmente discrepantes por variável",
  ylab = "Variáveis",
  xlab = "Quantidade",
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6"),
  horiz = T
), x = qvpdQC[,4], labels = qvpdQC[,4], pos = 2, col = "white"
)
```

22. Dados faltantes, QuantCont:

```
dfpvQC = DadosFaltantesPorVariavel(QuantCont)
# dfpvQC
text(
barplot(
  dfpvQC[,3],
  names.arg = dfpvQC[,1],
  main = "Dados Faltantes por variável",
  ylab = "Variáveis",
  xlab = "Quantidade",
  col = c("#114F66", "#2585A8", "#01AEEF", "#40BAE6"),
  horiz = T
), x = dfpvQC[,3], labels = c(0, '11%', '11%', '11%', '15%', '45%', '2%', 0, 0, 0,
'86%', '86%'), pos = 2, col = "white"
)
```

23. Gráfico de correlação e plots das variáveis.

```
# VARIÁVEIS QUANTITATIVAS CONTÍNUAS
cor(psdata[,QuantCont], use = "complete.obs")
corrplot(cor(psdata[,QuantCont], use = "complete.obs"), type = "upper", method =
"color")
# (62, 64) e (56, 57)
var62 = psdata$var62
var64 = psdata$var64
plot(var62, var64)

var56 = psdata$var56
var57 = psdata$var57
plot(var56,var57 )
```