

R

Estudo dirigido de linguagem R

Sergio Pedro R Oliveira

09 janeiro 2024

SUMÁRIO

1	OBJETIVO	1
2	CAP. 1 - INSTALAÇÃO DO R E RSTUDIO	2
3	CAP. 2 - PACOTE BASE E FUNÇÕES ESTATÍSTICAS BÁSICAS	3
3.1	Operações matematicas basicas	3
3.2	Vetor	3
3.3	Tabela de dados (data.frame) e matrizes	4
3.3.1	data.frame	4
3.3.2	Matrizes	4
3.4	Acessando valores em posições especificadas dos objetos - vetor , matriz e data.frame	5
3.4.1	Caso vetor e matriz	5
3.4.2	Caso data.frame	5
3.5	Visualizando dados	6
3.5.1	View() - visualização de dados	6
3.5.2	str() - estrutura de objetos	6
3.5.3	summary() - resumo de variáveis	6
3.5.4	class() - classe de objetos	7
3.6	Funções estatísticas básicas	8
4	CAP. 3 - PRINCIPAIS PACOTES	10
4.1	Instalação de pacotes	10
4.2	Pacotes	10
4.3	Carregamento de pacotes	11
4.4	Obter ajuda (informações) sobre pacotes	11
5	SITES PARA USO REMOTO DO R	12
6	CAP. 4 - R MARKDOWN	13
6.1	Preâmbulo	13
6.1.1	Titulo	13
6.1.2	Autor	13
6.1.3	Data	13
6.1.4	Tipo do Documento (<i>output</i>)	13
6.1.5	Sumário	14
6.1.6	Formatação desejada	14
6.1.7	Abstract	14
6.1.8	Bibliografia	14
6.2	<i>Chunks</i> (códigos embutidos)	16
6.2.1	Códigos embutidos no texto	16
6.2.2	Chunk	16
6.2.3	Configurando imagens e tabelas dentro do <i>chunk</i>	17
6.2.4	Global <i>Chunk</i>	18
6.3	Titulos e subtítulos	19
6.4	Pular linha	19
6.5	Listas	20
6.5.1	Listas numeradas	20
6.5.2	Listas não numeradas	20
6.6	Notas de rodapé (clicáveis)	21
6.7	Inserir tabelas	22
6.7.1	Formato de tabela padrão	22
6.7.2	Criador de tabelas online para R Markdown	22

6.7.3	Tabelas provenientes de banco de dados	23
6.7.3.1	Mostrar todos os dados	23
6.7.3.2	Mostrar apenas os primeiros dados	23
6.7.3.3	Bibliotecas para criação de Tabelas	24
6.7.3.4	Tabela para paginas web	31
6.8	Hiperlinks e imagens	32
6.8.1	Hiperlinks	32
6.8.2	Imagens	32
6.9	Fórmulas LaTeX	33
6.9.1	Equações	33
6.9.2	Superescrito e subscritos	33
6.9.3	Sublinhados, sobrelinhas e vetores	33
6.9.4	Frações, matrizes e chavetas	34
6.9.5	Expressões	36
6.9.6	Sinais e setas	37
6.10	Letras gregas	38
6.11	Formatação (Fontes)	39
6.12	Abas	39
7	CAP. 5 - PACOTES DO TIDYVERSE E IDENTIFICANDO/MUDANDO TIPOS DE VARIÁVEIS	40
7.1	Identificando/mudando tipos de variáveis	40
7.2	Pacotes do Tidyverse	41
7.3	Leitura de dados (readr)	42
7.3.1	Importação de dados via RStudio	42
7.3.2	Importação de dados via biblioteca readr	43
7.4	tibble	46
7.4.1	Visualização de tabelas tipo tibble	46
7.4.2	Criação de tabela tipo tibble	46
7.4.3	Funções tibble	47
7.5	Operador pipe	48
7.6	Manipulando dados com o dplyr	49
7.6.1	manipulação de dados:	49
7.6.2	combinando tabelas de dados:	55
7.7	Organizando dados com o tidyr	60
8	SINCRONIZAÇÃO COM BANCO DE DADOS	65
8.1	Pacotes de banco de dados	65
8.2	Sincronização com banco de dados	66
8.2.1	DBI	66
8.2.2	odbc	68
8.2.3	Drives	71
8.2.3.1	RSQLite	71
8.2.3.2	RMySQL ou RMariaDB	72
8.2.3.3	RPostgres	73
8.3	Importação de tabelas	74
8.3.1	Remoto	74
8.3.2	Local	74
8.3.2.1	Remoto para local - collect()	74
8.3.2.2	Local para remoto (exportar dados para o banco de dados) - copy_to()	74
8.4	Manipulação de tabelas	75

8.4.1	Manipular dados direto no banco de dados com R - <code>dbplyr</code>	75
8.4.2	Manipular dados localmente com SQL - <code>sqldf</code>	77
9 CAP. 6 - PACOTE DATA.TABLE		78
9.1	Teoria	78
9.2	Estrutura	78
9.3	Transformando <code>data.frame</code> em <code>data.table</code>	78
9.4	<code>data.table</code>	80
9.4.1	Manipulando linhas	80
9.4.2	Manipulando colunas	80
9.4.3	Sumarizando dados	81
9.4.4	Operando um subconjunto de dados	82
9.4.5	Modificando dados com <code>set</code>	83
10 CAP. 7 - GRÁFICOS PACOTE BÁSICO E PACOTE ggplot2		84
10.1	Gráficos com o pacote básico	84
10.1.1	Gráfico de barras (barplot)	87
10.1.1.1	Pré-requisitos	87
10.1.1.2	Preparação dos dados	87
10.1.1.3	Plotagem gráfico de barras (barplot)	88
10.1.2	Gráfico circular/pizza (pie)	90
10.1.2.1	Pré-requisitos	90
10.1.2.2	Preparação dos dados	90
10.1.2.3	Plotagem gráfico circular/pizza (pie)	91
10.1.3	Gráfico de linhas (plot lines)	93
10.1.3.1	Pré-requisitos	93
10.1.3.2	Preparação dos dados	93
10.1.3.3	Plotagem gráfico plot	94
10.1.3.4	Comparando séries de gráficos de linhas	96
10.1.4	Gráfico de dispersão (plot abline)	98
10.1.4.1	Pré-requisitos	98
10.1.4.2	Preparação dos dados	98
10.1.4.3	Plotagem gráfico plot abline	101
10.1.5	Diagrama de caixa (boxplot)	103
10.1.5.1	Separatriz	103
10.1.5.2	boxplot	105
10.1.5.3	Pré-requisitos	106
10.1.5.4	Preparação dos dados	106
10.1.5.5	Plotagem gráfico boxplot	106
10.1.6	Histograma (hist)	108
10.1.6.1	Pré-requisitos	109
10.1.6.2	Preparação dos dados	109
10.1.6.3	Plotagem histograma	109
10.2	Pacote <code>ggplot2</code>	111
10.2.1	O que precisa para fazer o gráfico?	111
10.2.2	Quais formatos podemos utilizar no <code>ggplot2 (geom_forma)</code> ?	111
10.2.3	Nome dos argumentos para adicionar efeito em gráficos do pacote <code>ggplot2</code>	112
10.2.4	Definindo um tema para o grafico <code>ggplot</code>	112
10.2.5	Pacote <code>ggthemes</code>	113
10.2.6	Inserindo títulos, subtítulos e rótulos aos eixos de um <code>ggplot</code>	114
10.2.6.1	Primeira forma	114
10.2.6.2	Segunda forma	114
10.2.7	Escalas no <code>ggplot2</code>	116
10.2.8	Cores nos gráficos <code>ggplot2</code>	120

10.2.8.1	Método para obter cores em R	121
10.2.8.2	Principais pacotes de paletas de cores do R	122
10.2.8.3	Tipos de paletas de cores	122
10.2.8.4	5 funções básicas do R que geram paletas de cores sequenciais	122
10.2.8.5	RColorBrewer paletas de cores disponíveis	123
10.2.8.6	Aplicando escala de cinza ao gráfico	125
10.2.9	Ajustando parâmetro de textos de um ggplot	126
10.2.10	Layout da janela gráfica e plotagem de vários gráficos em uma janela	128
10.2.10.1	Principais pacotes para configurar layout da janela gráfica	128
10.2.10.2	Pacote grid	128
10.2.10.3	Pacote patchwork	130
10.2.11	Gráficos usando pacote ggplot2	132
10.2.11.1	Gráfico de barras (geom_bar) com ggplot2	133
10.2.11.2	Histograma com ggplot2	137
10.2.11.3	boxplot (diagrama de caixa) com ggplot2	140
10.2.11.4	Gráfico circular (pizza) com ggplot2	147
10.2.11.5	Gráfico de pontos com ggplot2	149
10.2.11.6	Gráfico de linhas com ggplot2	154
10.2.11.7	Gráfico de pontos com ajuste por curva de tendência com ggplot2	158
10.2.11.8	Gráfico de dispersão com linha de tendência com ggplot2	160
10.2.11.9	Efeitos	166
10.2.12	Assistentes para ggplot2 - esquisse	174
10.2.12.1	Pacotes auxiliares ao ggplot2	174
10.2.12.2	Referências e modelos de assistente gráfico	174
10.2.12.3	Funções do assistente de gráficos ggplot2 - esquisse	175
10.2.12.4	Assistente gráfico esquisse	176
11 CAP. 8 - LIMPEZA RÁPIDA NOS DADOS		179
11.1	Pactoes	179
11.2	Dados “sujos”	179
11.3	Principais funções janitor	180
11.3.1	Limpando nomes do data.frame - clean_names()	180
11.3.2	Remova colunas ou linhas inúteis	181
11.3.3	Substitua valores perdidos/faltantes - mice()	183
11.3.4	Produzindo tabelas de frequência para uma variável - taby()	184
11.3.5	Tabulação cruzada - taby1()	186
11.3.5.1	Tabulação cruzada	186
11.3.5.2	Função taby1() para tabulação cruzada	186
11.3.6	Teste qui-quadrado para tabela cruzada - chisq.test()	191
11.3.7	Caça aos registros com valores duplicados - get_dupes()	194
11.3.8	Corrija número para data com a função excel_numeric_to_date()	195
11.3.9	Conte os níveis dos fatores - escala de <i>Likert</i>	196
11.3.9.1	Escala <i>Likert</i>	196
11.3.9.2	A função top_levels()	197
11.3.9.3	Plotagem de escala <i>Likert</i>	199
12 CAP. 9 - ANÁLISE DESCRIPTIVA DOS DADOS		205
12.1	Teoria	205
12.2	Tipos de variáveis	205
12.3	Tabulação dos dados	206
12.4	Estatística descritiva com o pacote DescTools	214
12.4.1	Teoria	214
12.4.2	Customizar os gráficos	216
12.4.3	Interpretação dos coeficientes	218

12.5 Mapeamento de dados faltantes - <code>PlotMiss()</code>	221
12.6 Analisando datas com o pacote <code>DescTools</code>	222
12.6.1 Principais funções (<code>date</code>) do pacote <code>lubridate</code>	222
12.6.2 Operações com datas	225
12.6.3 Análise <code>DescTools</code> com datas	226
13 CAP. 10 - DISTRIBUIÇÕES DE PROBABILIDADES	229
13.1 Teoria e introdução	229
13.2 Distribuição normal	232
13.3 Aderência da distribuição normal	234
13.4 As hipóteses de um teste estatístico	237
13.5 Erros possíveis no teste de hipótese	239
13.6 Regra de decisão para todos os testes de hipóteses	240
13.7 Teste de normalidade	243
13.8 Análise gráfica	244
13.9 E quando os dados não são normais?	246
13.9.1 O que são dados com distribuição normal?	246
13.9.2 Normalização de dados	246
13.9.2.1 Distribuição normal padronizada	247
13.9.2.2 Transformação logarítmica	249
13.9.2.3 Transformação raiz quadrada	250
13.9.2.4 Transformação Box-Cox	251
13.10 Plot dois gráficos ou mais	252
14 CAP. 11 - MODELANDO A RELAÇÃO ENTRE DUAS VARIÁVEIS	253
14.1 Modelando a relação entre duas variáveis	253
14.2 Elaboração do modelo (casual)	253
14.3 Relações das variáveis	254
14.3.1 Numérica ~ categórica	254
14.3.1.1 Passos da criação do modelo numérica ~ categórica	254
14.3.1.2 Comparando gráficamente a distribuição de densidade de uma variável numérica condicionada a categorias de outra variável	258
14.3.2 Categórica ~ numérica (regressão logística)	261
14.3.2.1 O que é regressão logística?	261
14.3.2.2 Tipos de regressão logística	262
14.3.2.3 Modelo de regressão logística	263
14.3.2.4 A razão de chance ou <i>Odds Ratio (OR)</i>	271
14.3.2.5 Prevendo as probabilidades	274
14.3.2.6 Avaliando a significância dos parâmetros	275
14.3.2.7 Avaliando a acurácia do modelo	276
14.3.3 Categórica ~ categórica	278
14.3.4 Numérica ~ numérica	278
15 ANDAMENTO DOS ESTUDOS	279
REFERÊNCIAS	280

LISTA DE FIGURAS

1	Exemplo Tabela kable	25
2	Exemplo tabela paged_table	31
3	Logo do R Markdown	32
4	Gráfico de barras - Vertical (<code>barplot()</code>)	89
5	Gráfico de barras - Horizontal (<code>barplot(horizon = TRUE)</code>)	89
6	Gráfico circular “pizza” (<code>pie()</code>)	92
7	Gráfico de linhas (<code>plot() lines()</code>)	95
8	Gráfico de linha comparando séries	97
9	Correlações fortes e fracas	99
10	Tipos de Correlação	99
11	Tabela de correlação linear	100
12	Gráfico de dispersão “plot abline” (<code>plot() abline()</code>)	102
13	Exemplo de tabela de distribuição de frequências	103
14	Exemplo explicativo de boxplot	105
15	Gráfico de caixa (<code>boxplot()</code>)	107
16	Exemplo gráfico de barras	108
17	Exemplo histograma	108
18	Histograma (<code>hist()</code>)	110
19	Exemplo 1 - <code>scale_x_discrete</code>	118
20	Exemplo 2 - <code>scale_(x y)_continuous</code>	119
21	657 cores e seus repectivos nomes.	121
22	Gráfico com ajustes de texto.	127
23	<i>Layout</i> da janela gráfica com dois gráficos, usando biblioteca grid do R	129
24	<i>Layout</i> 1 da janela gráfica com dois gráficos, usando biblioteca patchwork do R	131
25	<i>Layout</i> 2 da janela gráfica com quatro gráficos, usando biblioteca patchwork do R	131
26	Gráfico de barras (<code>geom_bar()</code>)	134
27	Gráfico de Barras (<code>geom_bar()</code>) com duas ou mais categorias e <i>layout</i> com dois gráficos.	136
28	Exemplo gráfico de barras	137
29	Exemplo histograma	137
30	Histograma (<code>geom_histogram()</code>) com eixo x logarítmico.	139
31	Exemplo de tabela de distribuição de frequências	140
32	Exemplo explicativo de boxplot	142
33	boxplot (diagrama de caixa) com eixos invertidos.	143
34	Boxplot com efeito jitter.	144
35	boxplot dividido por facetas das regiões norte e nordeste, usando <code>facet_grid()</code>	145
36	Gráfico dividido por facetas das regiões, usando <code>facet_wrap()</code>	146
37	Gráfico circular por ggplot2, construído a partir das funções gráficas <code>geom_bar()</code> + <code>coord_polar()</code>	148
38	Gráfico de pontos, construído a partir da biblioteca ggplot2 usando a função <code>geom_point()</code>	150
39	Gráfico de pontos com efeito jitter, substitui <code>geom_point()</code> por <code>geom_jitter()</code>	151
40	Gráfico de pontos comparando sem efeito jitter (<code>geom_point()</code>) e com efeito jitter (<code>geom_jitter()</code>).	153
41	Gráfico de linhas, construído a partir da biblioteca ggplot2 usando as funções <code>geom_point()</code> + <code>geom_line()</code>	156
42	Gráfico de linha com ajuste por curva suavizada (<code>geom_point() + geom_smooth()</code>).	157
43	Gráfico de pontos com ajuste com curva de tendência suavizada smooth, com <code>span = 0.7</code>	159
44	Correlações fortes e fracas	161
45	Tipos de Correlação	161
46	Tabela de correlação linear	162
47	Gráfico de dispersão com linha de tendência (regressão linear).	165
48	Gráfico sem efeito jitter.	166
49	Gráfico com feito jitter.	166

50	Diferença entre <code>facet_grid()</code> e <code>facet_wrap()</code>	167
51	Gráfico por facetas com <code>facet_grid(rows = variável)</code> ou <code>facet_grid(variável ~.)</code> . .	169
52	Gráfico por facetas com <code>facet_grid(cols = variável)</code> ou <code>facet_grid(.~ variável)</code> . .	169
53	Gráfico por facetas com escala do eixo y livre. <code>facet_grid(scale = "free_y")</code>	170
54	Gráfico por facetas com rótulos com nome da variável. <code>facet_grid(labeler = label_both)</code>	170
55	Gráfico de tendência com curva suavizada smooth, área cinza indicador de confiabilidade. . .	172
56	Gráfico de tendência com curva suavizada smooth, orientada no eixo y <code>orientation = "y"</code> . .	172
57	Gráfico de tendência com curva suavizada smooth, com <code>span = 0.3</code> (mais sinuoso).	173
58	Gráfico de tendência com curva suavizada smooth, com <code>method = lm</code> , <code>se = FALSE</code> (regressão linear com área de confiança omitida)	173
59	Pacotes auxiliares (<code>ggThemeAssist</code> e <code>esquisse</code>) de construção de gráfico - <code>ggplot2</code> builder . .	174
60	Abrir o <code>esquisse</code> a partir do addins no RStudio	176
61	No <code>esquisse</code> , aba para editar textos do gráfico (título, subtítulo, legenda e rótulos)	176
62	No <code>esquisse</code> , aba para editar opções do gráfico.	177
63	No <code>esquisse</code> , aba para editar aparência do gráfico.	177
64	No <code>esquisse</code> , aba para exportar o gráfico (seja o código ou o .png)	178
65	No <code>esquisse</code> , aba para exportar o gráfico (seja o código ou o .png)	178
66	Exemplo tabela de frequência de uma variável.	184
67	Exemplo de tabulação cruzada	186
68	Exemplo de escala <i>Likert</i>	196
69	Gráfico das respostas em escala de <i>Likert</i>	200
70	Gráfico das respostas em escala de <i>Likert</i> - <code>type = "heat"</code>	202
71	Modelo 1 de visualização de escala <i>Likert</i>	203
72	Modelo 2 de visualização de escala <i>Likert</i> - <code>type = "heat"</code>	203
73	Modelo 3 de visualização de escala <i>Likert</i> - <code>grouping = bd\$categ</code>	204
74	Exemplo da função <code>PlotMiss()</code> para mapeamento de dados faltantes.	215
75	Gráficos de variável quantitativa conitínua (numérica) a partir de medidas estatística descritiva. <code>plot(Desc(dados\$variavel_numerica))</code>	217
76	Gráfico de variável quantitativa discreta (inteiro), dicotômica (dois níveis), a partir de medidas estatística descritiva. <code>plot(Desc(dados\$variavel_inteiro))</code>	217
77	Gráficos de variável qualitativa (categórica), politômica (mais de dois níveis), a partir de medidas estatística descritiva. <code>plot(Desc(dados\$variavel_categorica))</code>	217
78	Casos de assimetria.	219
79	Tipos de distribuição normal.	220
80	Mapeamento de dados faltantes - <code>PlotMiss()</code>	221
81	Analise de data, gráfico de dias.	227
82	Analise de data, gráfico de meses.	228
83	Analise de data, gráfico de dias da semana.	228
84	Probabilidade ou área sobre a curva.	233
85	Análise gráfica (histograma, boxplot, densidade empírica) através da função <code>PlotFdist()</code> , do pacote <code>DescTools</code>	235
86	Histograma para análise de normalidade dos dados - pacote básico.	244
87	Histograma para análise de normalidade dos dados - pacote <code>ggplot2</code>	244
88	<code>qqplot</code> para análise da normalidade dos dados - pacote <code>ggplot2</code>	245
89	<code>qqplot</code> para análise da normalidade dos dados - pacote <code>DescTools</code>	245
90	Distribuição normal padronizada, $X \sim N(\mu = 0, \sigma^2 = 1)$	248
91	Plotando dois ou mais gráficos de distribuição de probabilidade ao mesmo tempo, usando a função <code>par(mfrow = c(1,2))</code>	252
92	Gráfico referente ao teste de hipóteses variável numérica ~ variável categórica, do pacote <code>DescTools</code>	257
93	Gráfico residual, usando a função <code>plot(modelo, which = 5)</code> , para observar possíveis pontos <i>outliers</i> /pontos de alavancagem.	266
94	Painel para avaliar possíveis multicolinearidade, usando a função <code>pairs.panels()</code> , do pacote <code>psych</code>	266

95	Painel para avaliar possíveis multicolinearidade, usando a função <code>pairs.panels()</code> , do pacote <code>psych</code> .	269
----	--	-----

LISTA DE TABELAS

1	Operações básicas do R	3
6	Principais tipos de dados	40
7	Tabela dos principais operadores lógicos usados na função filter.	50
8	Principais funções de summarise	53
9	Tabela de chegada de turistas no formato larga	60
10	Tabela de chegada de turistas no formato longo	60
11	Tabela em formato longo dieta de pacientes	62
12	Tabela em formato larga dieta de pacientes	62
13	5 formas de manipulação de linhas no data.table	80
14	8 formas de manipulação de colunas no data.table	80
15	Argumentos para operações em um DT aplicados a uma ou mais colunas	81
16	Funções set para modificação de dados no formato data.table	83
17	Nome das principais funções para construção de gráficos do pacote base.	84
18	Nome dos argumentos para adicionar efeito em gráficos.	85
19	Nome das principais formas geométricas para construção de gráficos do pacote ggplot2	111
20	Nome dos argumentos para adicionar efeito em gráficos do pacote ggplot2.	112
21	Temas do pacote ggthemes	113
22	Pacote RColorBrewer: Nome das paletas sequencial.	123
23	Pacote RColorBrewer: Nome das paletas divergente.	123
24	Pacote RColorBrewer: Nome das paletas qualitativas e número de cores possíveis em cada paleta.	124
25	Tabela de frequência para variável categórica	208
26	Tabela de frequência para variável numérica contínua, usando o método de separação de classes nclass.Sturges()	211
27	Tabela de frequência para variável numérica continua, com separação de classes inserido manualmente e limites com aberturas invertidas usando right = FALSE	213
28	Principais saídas da função Desc para variáveis numéricas.	214
29	Nome das funções correspondentes às principais distribuições de probabilidade	230
30	Lista de funções para realização de teste de hipóteses do pacote stats.	241
31	Exemplo de coeficiente estimados do modelo de regressão logística	275
32	Matriz de confusão do modelo de regressão logística.	276

1 OBJETIVO

Estudo dirigido de linguagem R.

2 CAP. 1 - INSTALAÇÃO DO R E RSTUDIO

- Download da linguagem R:
<https://www.r-project.org/>
- Download Rstudio IDE:
<https://posit.co/downloads/>

3 CAP. 2 - PACOTE BASE E FUNÇÕES ESTATÍSTICAS BÁSICAS

3.1 Operações matematicas basicas

Table 1: Operações básicas do R

Nome da operação	Operação	Resultado
Adição	5+4	9
Subtração	6-2	4
Multipliação	7*3	21
Divisão	45/9	5
Potência	2^2	4
Raiz	sqrt(121)	11
Exponencial	exp(0)	1
Log na base e	log(1)	0
Log na base 10	log10(1)	0
Log na base 2	log2(4)	2
Log na base 3 ou qualquer outra	log(9,3)	2

3.2 Vetor

- Para criar um vetor usamos a função `c()`.
- Os argumentos são separados por vírgula dentro do parênteses.
- strings devem estar entre aspas duplas.
Ex.: `c("um","sete","nove")`
- Vetores são compostos de elementos todos do mesmo tipo.
- Armazenando vetores em um objeto:
Ex.: `obj_qualquer <- c(1,2,3)`

3.3 Tabela de dados (`data.frame`) e matrizes

3.3.1 `data.frame`

- Uma tabela onde cada coluna é um vetor.
- Como cada coluna é um vetor, cada coluna pode ser de um tipo diferente.
Ex.: `nome_data.frame <- data.frame(vetor_1, vetor_2)`
- Acrescentando uma nova coluna ao `data.frame`.
Ex.: `nome_data.frame <- data.frame(nome_data.frame, vetor_3)`
- Para visualizar um `data.frame` podemos usar a função `View()`.
Ex.: `View(nome_data.frame)`

3.3.2 Matrizes

- A diferença entre **matrizes** e **data.frames**, é que no caso das matrizes todas as colunas e linhas devem ser do mesmo tipo. Enquanto nos **data.frames** as colunas podem ser de tipos diferentes.
- Para adicionar uma coluna numa matriz, usamos a função `cbind()`.
Ex.: `nome_matriz <- cbind(vetor_1, vetor_2, ...)`
- Para adicionar uma linha numa matriz, usamos a função `rbind()`.
Ex.: `nome_matriz <- rbind(vetor_3, vetor_4, ...)`
- Quando inserimos dados (vetor) de naturezas diferentes (tipos) numa matriz, ela converte todos os dados para um único tipo. A principio *string* (*chr*).

3.4 Acessando valores em posições especificadas dos objetos - vetor, matriz e data.frame

3.4.1 Caso vetor e matriz

- Podemos acessar os valores do objeto tipo **vetor** e **matriz**, informando a posição entre colchetes [].
- Para os **vetores** precisamos apenas informar a posição. A contagem da posição começa a partir do 1.
Ex.:
`vetor <- c(5,18,89)`
`vetor[1]`
- Para as **matrizes**, é necessário informar a posição [*linha, coluna*]. A contagem da posição começa a partir do 1.
Ex.:
`Mc[1,2]`
- Para acessar todos os valores de uma *linha* da **matriz**, podemos determinar a *linha* e deixar a *coluna* em branco.
Ex.:
`Mc[1,]`
- Para acessar todos os valores de uma *coluna* da **matriz**, podemos determinar a *coluna* e deixar a *linha* em branco.
Ex.:
`Mc[,2]`

3.4.2 Caso data.frame

- No caso do **data.frame** podemos acessar os valores das colunas informando, “nome do **data.frame**” “\$” “nome da coluna”.
Sintaxe:
`nome_dataframe$nome_coluna`
- O **data.frame** também aceita as mesmas formas de acessar posições que as **matrizes**.

3.5 Visualizando dados

3.5.1 View() - visualização de dados

- Podemos visualizar dados de duas formas:
 - Escrevendo o nome da variável
O valor dela será impresso na tela.
 - Atraves da função **View()**
Ao chamar a função **View()** e colocar dentro a variavel que queremos ver, será exibido uma nova janela com o valor da variável numa tabela.

3.5.2 str() - estrutura de objetos

- A função “**str()**” retorna a estrutura do objeto do argumento.
- Retorna diversos dados, entre eles:
 - A classe do objeto.
 - Tamanho do objeto.
 - A lista, ou vetor, dos campos com o tipo e tamanho.
- Sintaxe:
str(argumento)

3.5.3 summary() - resumo de variáveis

- A função **summary()** retorna o resumo de variaveis.
- O retorno depende do argumento (se for um vetor, uma lista, um **data.frame**).
- O retorno para uma matriz ou **data.frame**, vai ser os metodos aplicados a cada campo/coluna.
- O retorno da função, no geral, retorna diversos metodos aplicados aos dados, tais como:
 - valor mínimo
 - 1º quantil
 - valor da mediana
 - valor da media
 - 3º quantil
 - valor máximo
- Sintaxe:
summary(nome_variavel)

3.5.4 class() - classe de objetos

- A função “**class()**” retorna a que classe do objeto do argemunto pertence.
- Basicamente diz se o objeto é numerico, string, vetor, lista, data.frame, matriz, ...
- Sintaxe:
`class(argumento)`

3.6 Funções estatísticas básicas

Função	Descrição
apply(D,i,f)	Retorna os valores resultantes da aplicação da função f ao objeto D, linhas i=1, ou colunas i=2.
c(valor1, valor2, valor3)	Concatena uma sequência de valores seja númerico ou de caracteres. Neste último caso os valores devem estar entre aspas.
cbind(x1, x2, ..., xn)	Cria uma matriz com n colunas formada pelos vetores x1, x2, ..., xn.
ceiling(x)	Retorna o menor inteiro maior ou igual ao valor x.
cor(x,y)	Calcula o coeficiente de correlação.
cumsum(x)	Retorna um vetor com valores acumulados em soma sobre os elementos de x.
cumprod(x)	Retorna um vetor com valores acumulados em produto sobre os elementos de x.
cummin(x)	Retorna um vetor com valores acumulados em mínimo sobre os elementos de x.
cummax(x)	Retorna um vetor com valores acumulados em máximo sobre os elementos de x.
data.frame(x1, x2, ..., xn)	Cria um dataframe com os vetores x1, x2, ..., xn.
det(M)	Calcula o determinante da matriz quadrada M.
dim(M)	Retorna as dimensões do objeto M.
diff(x)	Retorna um vetor com a diferença entre os valores de x.
eigen(M)	Retorna os autovalores e os autovetores da matriz quadrada M.
floor(x)	Retorna o maior inteiro menor ou igual a x.
identical(x,y)	Verifica se os vetores são idênticos.
intersect(x,y)	Realiza a interseção de dois conjuntos.
head(D)	Mostra o cabeçalho do objeto D.
length(x)	Calcula o comprimento do vetor x.
mean(x)	Calcula a média do vetor x.
median(x)	Calcula a mediana do vetor x.
min(x)	Calcula o mínimo de x.
max(x)	Calcula o máximo de x.
ncol(M)	Retorna o número de colunas da matriz M.
nrow(M)	Retorna o número de linhas da matriz M.
polyroot(x)	Encontra as raízes do polinômio de ordem n cujos coeficientes são representados no vetor x em ordem decrescente.
prod(x)	Multiplica os valores de x.
quantile(x,k)	Calcula o percentil de ordem $0 \leq k \leq 1$ dos valores de x.
Re(x)	Retorna a parte real de um vetor x.
rep(x,k)	Cria um vetor repetindo a sequência x k vezes.
round(x,k)	Arredonda o valor x com k casas decimais.
sd(x)	Calcula o desvio-padrão do vetor x.
seq(i,j,k)	Cria uma sequência de i ate j com tamanho de passo k.
setdiff(x,y)	Retorna um vetor contendo os elementos do conjunto diferença entre x e y.
setequal(x,y)	Verifica se os elementos dos vetores x e y são iguais, independentemente da frequência em que aparecem no vetor.
solve(A,b)	Resolve Ax=b, retornando x.
sort(x)	Ordena os valores de vetor x em ordem crescente.
sort(x, decreasing = T)	Ordena os valores de x em ordem decrescente.

Função	Descrição
str(D)	Retorna a estrutura do objeto D.
sum(x)	Soma os valores de x.
union(x,y)	Retorna os elementos da união entre x e y.
var(x)	Calcula a variância do vetor x.
var(x,y)	Calcula a covariância entre x e y.
View(D)	Mostra o dataframe em janela separada.

4 CAP. 3 - PRINCIPAIS PACOTES

4.1 Instalação de pacotes

- sintaxe de instalação:
`install.packages("nome do pacote")`
- sintaxe de varias instalações simultaneas:
`install.packages(c("nome do pacote", "nome do pacote", ...), dependencies = TRUE)`

4.2 Pacotes

1. Principais pacotes:

- **stringr**
Pacote para trabalhar com strings (texto).
- **Rmarkdown**
Produção de relatorios (html, pdf, doc, md).
- **knitr**
Interpretação e compilação do documento rmd.
- **data.table**
Exploração de data.frames.
- **janitor**
Limpeza de dados.
- **DescTools**
Analise descritiva de dados.
- **tidyverse**
conjunto de pacotes.
 - **readr**
Importação e leitura de arquivos de dados.
 - **tibble**
estruturação de data.frame.
 - **dplyr**
Manipulação de data.frame.
 - **tidyr**
Organização de data.frame.

- **ggplot2**
Visualização de dados, produção de gráficos.
- **purr**
Manipulação de vetores e listas.
- **foreign**
Leitura e gravação de dados armazenados por algumas versões de “Epi Info”, “Octave”, “Minitab”, “S”, “SAS”, “SPSS”, “Stata”, “Systat”, “Weka” e para leitura e gravação de alguns “dBase” arquivos.
- **devtools**
Para instalar pacotes que não estejam no **CRAN**.

2. Pacotes auxiliares ao pacote **ggplot2**:

- **ggthemes**
- **grid**

4.3 Carregamento de pacotes

- Para poder utilizar o conjunto de funções de um determinado pacote, não basta apenas instalar o pacote, é preciso carrega-lo no script.
- As principais formas de carregar um pacote no script é através dos comandos *library()* e *require()*.
`library(nome_pacote)`
`require(nome_pacote)`
- Outra possibilidade, é ao usar um função especificar a qual pacote ela pertence.
`nome_pacote::função`.

4.4 Obter ajuda (informações) sobre pacotes

Duas formas de se conseguir informações sobre determinado pacote é através dos comandos:

1. `package?nome_pacote`
2. `help(package = "nome_pacote")`

5 SITES PARA USO REMOTO DO R

- Alguns sites que possibilitam utilizar o R básico, sem que seja necessário instalá-lo no computador.
 - Uma ótima saída quando necessário utilizar em algum computador público (lan houses, hotéis, laboratórios, ...)
1. <http://rstudio.cloud/>
 2. <http://jupyter.org/try>
 3. http://www.tutorialspoint.com/execute_r_online.php
 4. http://github.com/datacamp/datacamp_light
 5. <http://rdrr.io/snippets>
 6. <http://www.jdoodle.com/execute-r-online>
 7. http://rextester.com/l/r_online_compiler
 8. <http://rnotebook.io>

6 CAP. 4 - R MARKDOWN

6.1 Preâmbulo

6.1.1 Título

title: “Título desejado”

6.1.2 Autor

- Para inserir um autor:
author: “Nome do autor”
- Para inserir varios autores:
author:
 - autor_1^ [instituto]
 - autor_2^ [instituto]

6.1.3 Data

- O comando “*date*”, adiciona uma data ao documento.
- Podemos adicionar uma data qualquer para o documento no formato “dd/mm/aaaa”.
date: “dd/mm/aaaa”
- Outra possibilidade é usar uma função dentro de um *chunk* “r Sys.Date()”, para adicionar a data atual do sistema (modelo inglês).
date: “r Sys.Date()”
- Outra opção é usar o a função dentro de um *chunk* “r format(Sys.time(), ‘%d %B %Y’)”. A data será gerada no modelo: 02 agosto 2004.
date: “r format(Sys.time(), ‘%d %B %Y’)”
Obs.: *chunk* deve ser colocado entre acentos graves.

6.1.4 Tipo do Documento (*output*)

- *output*: o tipo de saída, podem ser:
 - Documentos:
 - * *pdf_document*
 - * *md_document*
 - * *html_document*
 - * *word_document*
 - * *odt_document*

- * *rtf_document*
 - Apresentação:
 - * *powerpoint_presentation*
 - * *ioslides_presentation*
 - * *beamer_presentation*
 - mais:
 - * *flexdashboard::flex_dashboard*
 - * *github_document*

6.1.5 Sumário

- Para inserir o sumário no documento, basta colocar o comando “*doc: yes*” indentado dentro do tipo de saída.
- O comando **number_sections: true** adiciona numeração aos capítulos do sumário.

6.1.6 Formatação desejada

Para determinar a formatação desejada, basta salvar um arquivo com o nome *estilo.docx*, que contenha a formatação e referenciar o arquivo, indentado dentro do tipo de arquivo, através do comando “*reference_docx: caminho/.../estilo.docx*”.

6.1.7 Abstract

Abstract: “Texto de abstract”.

6.1.8 Bibliografia

- Ter um arquivo *.bib com as referencias.
- Adicionar o arquivo *.bib no preâmbulo do **R Markdown**, através do comando:
bibliography: caminho/arquivo.bib
 - Um arquivo *.csl com o estilo da citação.
 Este arquivo pode ser obtido no site:
<https://www.zotero.org/styles>
 Pesquisar por: “abnt”
 Opção: “Instituto de Pesquisa Econômica Aplicada - ABNT (Português - Brasil)”
 - Adicionar o arquivo *.csl no preâmbulo do R Markdown, através do comando:
csl: caminho/arquivo.csl
 - É necessário criar um capítulo no final para as referências. A bibliografia vai ser alocada no final do documento, logo neste último capítulo. A bibliografia é sempre inserida ao final do documento.

- Por fim, para aparecer as referencias elas precisam ser citadas no texto.
As principais formas de citar uma referência num texto de **R Markdown** é:

- Uma citação:

- Exemplo do comando: [@ chave_da_referencia]

- Exemplo de como fica no arquivo final: (Alcoforado, 2021).

- Mais de uma citação ao mesmo tempo:

- Exemplo do comando: [@ chave_da_referencia_1, @ chave_da_referencia_2]

6.2 *Chunks* (códigos embutidos)

6.2.1 Códigos embutidos no texto

- Podemos embutir códigos ao longo do texto.
- Para inserir um código que será rodado no meio do texto, usamos um sinal de crase para abrir, definimos a linguagem (normalmente r), o comando que desejamos e um sinal de crase para fechar o código.
`Este é um código embutido`
- Para rodar pequenos comandos no meio do texto códigos embutidos é uma ótima opção.
- Exemplo:
O resultado do comando 1:3 é criar uma sequencia com os valores 1:3. A soma destes valores é `sum(1:3)`.
O resultado do comando 1:3 é criar uma sequencia com os valores 1, 2, 3. A soma destes valores é 6.

6.2.2 Chunk

- Códigos em R, ou em outras linguagens, podem ser inseridos nos documentos através de *chunks*.
- *Chunks* são blocos de programação.
- A principal forma de inserir *chunks* é:
 - Três sinais de acento grave (crases) para abrir o *chunk*.
 - Na primeira linha, definir a linguagem do bloco de programação:
 - **R**
 - **Python**
 - **Julia**
 - **C++**
 - **SQL**
 - ...
 - Para dar um nome ao *chunk*, após definir a linguagem de programação basta colocar o nome do *chunk*. Nomear o *chunk* facilita determinar sua função dentro do relatório/documento.
 - Ainda na primeira linha, considerações sobre o bloco de programação (*chunk options*):
 - *include*
Mostra (*true*), ou não (*false*), o código e os resultados no arquivo finalizado. O R Markdown ainda executa o código e o resultado dele ainda pode ser usado em outro bloco de programação.
include = false | true
 - *echo*
Impede (*false*), ou não (*true*), que o código apareça, não afeta o resultado.

- ***echo*** = *false* | *true*
 - ***results***
“*hide*” mostra o código e omite o resultado.
results = “*hide*”
 - ***message***
Imprede (*false*), ou não (*true*), que mensagens geradas por código apareçam no arquivo finalizado.
message = *false* | *true*
 - ***warning***
Imprede (*false*), ou não (*true*), que avisos gerados pelo código apareçam no final.
warning = *false* | *true*
 - ***fig.cap***
Adiciona uma legenda aos resultados gráficos.
fig.cap = “...”
- Bloco de programação, escrito na linguagem definida.
- Três sinais de acento grave (crases) para fechar o *chunk*.
- Outras formas de inserir *chunks* é através do botão *Insert*, na área superior da tela do script, do **RStudio**.
- Observação:
message e *warning* igual a *false* é muito utilizado quando se carrega bibliotecas (**library**) no *chunk*, evita que as mensagens do carregamento apareçam.

6.2.3 Configurando imagens e tabelas dentro do *chunk*

- Os comandos de configuração de imagem no *chunk* são inseridos no cabeçalho do *chunk*.

- Principais comando de configuração de imagens com *chunk*:

- ***fig.width*** =
Largura da figura em cm na janela gráfica.
- ***fig.height*** =
Altura da figura em cm na janela gráfica.
- ***fig.align*** =
Alinha a figura no arquivo final (“left”, “right” ou “center”).
- ***fig.cap*** = “ ”
Texto para legenda.
- ***dpi*** =
Valor referente a qualidade da imagem, padrão é 72.
- ***out.width ou out.height*** =
Porcentagem do tamanho original da imagem.

6.2.4 Global *Chunk*

- Para definir as opções globais que se aplicam a cada parte do seu arquivo, chame `knitr::opts_chunk$set` em uma parte do código.
- O `knitr` tratará cada opção que você passar para `knitr::opts_chunk$set` como um padrão global que pode ser substituído em cabeçalhos de blocos individuais.

6.3 Títulos e subtítulos

- Ao utilizar o comando `#` e em seguencia um texto, geramos um título.
`# Titulo`
- A cada `#` que adicionamos, diminuimos uma camada de subtítulos.
`## Subtitulo`

6.4 Pular linha

- Para que duas frases fiquem em linhas separadas, dê dois espaços entre elas.
- Os dois espaços funcionam também para deixar uma linha em branco.
- Outra forma é adicionar “`\`”, tem o mesmo efeito.

6.5 Listas

6.5.1 Listas numeradas

- Basta inserir o número seguido de ponto e espaço.
 1. Tópico da lista numerada
- A ordem das principais camadas de lista numeradas são:
 - Número
 1. Primeira camada
 - Algarismos romanos
 - i) Segunda camada
 - Letra
 - A. Terceira camada
- Para inserir uma lista dentro de uma outra lista, é necessário indentar os tópicos.

6.5.2 Listas não numeradas

- Os principais símbolos (na ordem de utilização) da lista não numerada:
 - Asterisco(*)
 - Mais(+)
 - Traço(-)
- Para inserir uma lista dentro de uma outra lista, é necessário indentar os tópicos.

6.6 Notas de rodapé (clicáveis)

- Há duas opções para criar uma nota de rodapé:
 1. Escrever ao final do texto `[^1]` e então (pode ser logo abaixo, ou depois) escrever a nota de rodapé:
“Essa informação não é um consenso `[^1]`”
`[^1]: Esta é uma nota de rodapé.`
 2. Colocar a informação da nota de rodapé no meio do texto, e o R numerará automaticamente:
“Essa informação não é um consenso `^`[`Esta é uma nota de rodapé`]
- Observação:
A informação da nota de rodapé deve estar separado do texto por uma linha, no primeiro caso, ou contida na nota no link clicável, como no segundo caso.
- Exemplo:
O RMarkdown é uma ferramenta excelente para documentar seus códigos e apresentar os resultados. As muitas funcionalidades dele são descritas detalhadamente no livro R Markdown: The Definitive Guide ¹.

¹R Markdown: The Definitive Guide. Yihui Xie, J. J. Allaire, Garrett Grolemund. Disponível em:
<https://bookdown.org/yihui/rmarkdown/>

6.7 Inserir tabelas

6.7.1 Formato de tabela padrão

- A tabela mais simples é através do padrão:
 - Primeira linha:
Cabeçalho das colunas, separado por barra vertical(|).
 - Segunda linha:
 - * Tracejados (pelo menos 3), para representar cada coluna, com dois pontos onde se espera que o texto esteja alinhado:
 - Dois pontos no inicio do tracejado para representar alinhamento do texto a esquerda.
 - Dois pontos no inicio e no fim do tracejado para representar alinhamento centralizado do texto.
 - Dois pontos no final do tracejado para representar alinhamento do texto a direita.
 - * Cada coluna separada por barra vertical.
 - Terceira linha em diante:
Cada dado de linha em uma linha, com os dados de cada coluna separado por barras verticais.

6.7.2 Criador de tabelas online para R Markdown

Site que ajuda a construir tabelas para **R Markdown**:

https://tablesgenerator.com/markdown_tables

6.7.3 Tabelas provenientes de banco de dados

6.7.3.1 Mostrar todos os dados Dentro do *chunk* chamar a variável que contém um **dataframe**, para imprimir ela na tela.

6.7.3.2 Mostrar apenas os primeiros dados

- Dentro do *chunk* chamar a variável que contém um **dataframe**, e usar a função **head()** que mostra as 5 primeiras linhas. Podemos adicionar o parâmetro de quantidade de linhas desejamos apresentar.
- Exemplo:
`head(var_dataframe, n_linha)`

6.7.3.3 Bibliotecas para criação de Tabelas

6.7.3.3.1 kable

- Dentro do *chunk*, podemos chamar a biblioteca **knitr**, e usar a função **kable()** onde podemos chamar como argumento a variável **dataframe** (e outras funções).

- A função **kable()**, apresenta uma tabela em formato mais profissional.

- Argumentos do **kable**:

- **format**

Tipos de formatos que a tabela pode ser representada.

```
knitr::kable(head(mtcars[, 1:4]), "pipe")
```

- * pipe

- * simple

- * latex

- * html

- * rst

- **col.names**

O nome das colunas.

Podemos usar o argumento **col.names** para substituir os nomes das colunas por um vetor de novos nomes.

```
knitr::kable(iris, col.names = c('We', 'Need', 'Five', 'Names', 'Here'))
```

- **row.names**

Adiciona nome as linhas.

- **align**

Para alterar o alinhamento das colunas da tabela.

Podemos usar um vetor contendo os valores consistindo dos caracteres **l** (esquerda), **c** (centro) e **r** (direita).

```
kable(..., align = c("l","c",...))
```

ou

```
knitr::kable(iris2, align = "lccrr")
```

- **caption**

Adiciona uma legenda a tabela.

```
knitr::kable(iris2, caption = "An example table caption.")
```

- **digits**

Define o número máximo de casas decimais.

```
knitr::kable(d, digits = 4)
```

```
knitr::kable(d, digits = c(5, 0, 2))
```

- **format.args**

Define o formato me que os números serão apresentados.

- * scientific
Se é no formato científico (**true** ou **false**).

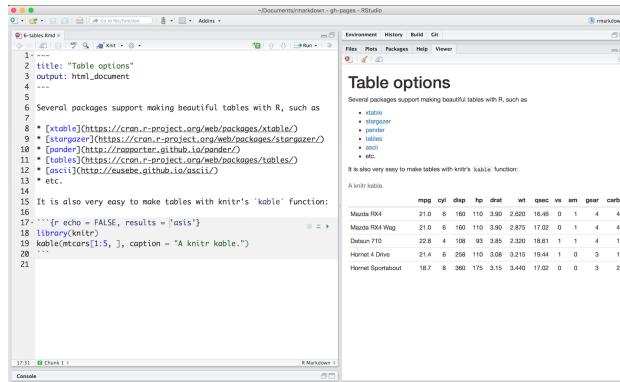
```
knitr::kable(d, digits = 3, format.args = list(scientific = FALSE))
```
- * big.mark
Como será a separação para números grandes.

```
knitr::kable(d, digits = 3, format.args = list(big.mark = ",", scientific = FALSE))
```
- escape
Ativa (**TRUE**) e desativa (**FALSE**) os caracteres especiais.

```
knitr::kable(d, format = "latex", escape = TRUE)
```

- Exemplo:

```
library(knitr)
kable(head(var_dataframe, 10))
```



The screenshot shows the RStudio interface. In the top-left, there's a browser window titled 'RStudio.html' displaying the R code. In the bottom-right, the 'Output' pane shows the generated HTML table. The table has 10 rows and columns for 'mpg', 'cyl', 'disp', 'hp', 'drat', 'wt', 'qsec', 'vs', 'am', 'gear', and 'carb'. The first few rows are as follows:

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
21.0	6	160	110	3.90	2.675	17.02	0	1	4	4
22.8	4	108	93	3.85	2.320	18.01	1	1	4	1
21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

Figure 1: Exemplo Tabela kable

6.7.3.3.2 kableExtra

- Para mais opções de formatação do `knitr::kable`, temos o pacote `kableExtra`.
- `kableExtra` é um pacote complementar ao `knitr::kable`, por conta disto, é necessário chamar a função `kable` (primeiramente), e concatenar as funções do pacote `kableExtra` pelo operador pipe `%>%`.

```
library(knitr)
library(kableExtra)
kable(iris) %>%
  kable_styling(latex_options = "striped")
```

- Definir o tamanho da fonte:

```
kable(head(iris, 5), booktabs = TRUE) %>%
  kable_styling(font_size = 8)
```

- Estilizar linhas e colunas específicas:

- Funções:

- * **row_spec**
Especifica a linha que vai ser estilizada.
 - * **column_spec**
Especifica a coluna que vai ser estilizada.

- Estilizações:

- * negrito (**bold**)
 - * itálico (**italic**)
 - * fundo preto (**background**)
 - * fonte branca (**color**)
 - * sublinhado (**underline**)
 - * espaçamento (**monospace**)
 - * ângulo (**angle**)

```
kable(head(iris, 5), align = 'c', booktabs = TRUE) %>%
  row_spec(1, bold = TRUE, italic = TRUE) %>%
  row_spec(2:3, color = 'white', background = 'black') %>%
  row_spec(4, underline = TRUE, monospace = TRUE) %>%
  row_spec(5, angle = 45) %>%
  column_spec(5, strikeout = TRUE)
```

- Alterar o tamanho da tabela, preenche todo espaço disponível (**full_width**).

```
kable(head(dados, 10), col.names = c("Gênero", "Álcool", "Memória", "Latência")) %>%
  kable_styling(full_width = FALSE)
```

- **bootstap_options**

– Cores alternadas entre linhas (**bootstap_options** = c("striped")).

```
kable(head(dados, 10), col.names = c("Gênero", "Álcool", "Memória", "Latência")) %>%
  kable_styling(full_width = F, bootstrap_options = c("striped"))
  – Deixando a tabela mais condensada/junta (bootstap_options = c("striped", "condensed")).
```

```
kable(head(dados, 10), col.names = c("Gênero", "Álcool", "Memória", "Latência")) %>%
  kable_styling(full_width = F, bootstrap_options = c("striped", "condensed"))
```

- Agrupar linhas e colunas.

Podemos agrupar conjunto de linhas, ou colunas, e dar um cabeçalho para elas.

– Agrupar colunas:

Através da função **add_header_above** podemos dar nome aos agrupamentos e definir o número de colunas agrupadas.

```
iris2 <- iris[1:5, c(1, 3, 2, 4, 5)]
names(iris2) <- gsub('[.]', '', names(iris2))
kable(iris2, booktabs = TRUE) %>%
  add_header_above(c("Length" = 2, "Width" = 2, " " = 1)) %>%
  add_header_above(c("Measurements" = 4, "More attributes" = 1))
```

– Agrupar linhas:

Através da função **pack_rows** e do argumento **index** podemos dar nome aos agrupamentos e definir o número de linhas agrupadas.

```
iris3 <- iris[c(1:2, 51:54, 101:103), ]
kable(iris3[, 1:4], booktabs = TRUE) %>%
  pack_rows(index = c("setosa" = 2, "versicolor" = 4, "virginica" = 3))
```

6.7.3.3.3 xtable

- A biblioteca **xtable** converte um objeto R em um objeto **xtable**, que pode ser expresso como uma tabela **LaTeX** ou **HTML**.
- Dentro do *chunk*, podemos chamar a biblioteca **xtable**, e usar a função **xtable()**, que recebe como argumentos a variável **dataframe** (e outras funções) e o *tipo* da saída para a tabela (**LaTeX** ou **HTML**).

```
library(xtable)
xtable(dataframe, type = "latex")
```

```
library(xtable)

coluna1 <- c(1,2,3,4,5,6)
coluna2<- c("a","b","c","d","e","f")
tab <- data.frame(coluna1,coluna2)

xtable(tab,type = "latex")
xtable(tab,type = "html")
```

6.7.3.3.4 pander

- O principal objetivo do pacote **pander** R é oferecer uma ferramenta de fácil renderização de objetos R no markdown do Pandoc.
- Um dos recursos mais populares do **pander** é **pandoc.table**, renderizando a maioria dos objetos R tabulares em tabelas de remarcão com várias opções de configuração:

- *Style (Estilo)*

- * “*simple*”
style = "simple"
 - * “*grid*”
style = "grid"
 - * “*markdown*”
style = "markdown"

- *Caption (Legenda)*

- `caption = "Legenda"`

- *Highlighting cells (Celulas destacadas)*

- Comandos para destacar linhas, colunas e celulas.

- As celulas pode estar em negrito e italicico ao mesmo tempo.

- * Italics (*italicico*):

- `emphasize.italics.rows(1)`
`emphasize.italics.cols(2)`
`emphasize.italics.cells(which(t > 20, arr.ind = TRUE))`

- * strong (**negrito**):

- `emphasize.strong.rows(1)`
`emphasize.strong.cols(1)`
`emphasize.strong.cells(which(t > 20, arr.ind = TRUE))`

- * verbatim (estilo literal):

- `emphasize.verbatim.rows(1)`
`emphasize.verbatim.cols(2)`
`emphasize.verbatim.cells(which(t > 20, arr.ind = TRUE))`

- Exemplo:

- `emphasize.italics.cols(1)`
`emphasize.italics.rows(1)`
`emphasize.strong.cells(which(t > 20, arr.ind = TRUE))`
`pandoc.table(t)`

- *Justify (Alinhamento da celula)*

- * Opções de alinhamento de celula:

- “*right*”

- “*left*”

- “center”
- * Formas de alinhamento de celula:
- Alinhando tudo de uma vez:
`justify = "right"`
 - Alinhando cada coluna separadamente:
`justify = c("right","center","left")`
- *Table and Cell width (Largura)*
- * `split.table` (**Largura tabela**) A largura máxima da tabela são 80 caracteres, caso ultrapasse esse tamanho, a tabela será quebrada e a parte excedente será inserida abaixo, como uma continuação. Para desligar essa opção e aumentar o tamanho da tabela, basta adicionar a opção *Inf*.
`split.table = Inf`
- * `split.cell` (**Largura celula**) O tamanho máximo da celula são 30 caracteres, caso ultrapasse esse tamanho, o texto será quebrado e adicionado a baixo, ainda na celula.
Para ajustar o tamanho da celula (definir o número de caracteres) existem três opções:
- Todas de uma vez.
`split.cell = 40`
 - Coluna por coluna.
`split.cell = c(40,20,5)`
 - Em termos de porcentagem.
`split.cell = "40%"`
`split.cell = c("80%","20%","40%)`
- Exemplo:
`library(pander)`
`pandoc.table(dataframe, justify = "center", caption = "Exemplo de tabela")`

6.7.3.4 Tabela para paginas web

- Dentro do *chunk*, podemos chamar a biblioteca **rmarkdown**, e usar a função **paged_table()**, onde podemos chamar como argumento a variável **dataframe**.
- Esse tipo de tabela é ideal para aplicações *web*.
- Separa os dados por páginas, de maneira dinâmica e com interação do usuário.
- Mostra dez linhas por página.

• Exemplo:

```
library(rmarkdown)  
paged_table(var_dataframe)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	►
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	>
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	

1-5 of 32 rows | 1-10 of 12 columns

Previous [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) Next

Figure 2: Exemplo tabela paged_table

6.8 Hiperlinks e imagens

6.8.1 Hiperlinks

- Sintaxe:
[Nome do Link] (Endereço do Link)
- Exemplo:
Canal do YouTube

6.8.2 Imagens

- Existem duas formas de pegar uma imagem são elas:
 - Pegar a imagem de um endereço da web (igual a hiperlink).
! [Legenda] (https://miro.medium.com/max/600/1*sCJzUnDlAuvGrlllJeXKw.jpeg)
 - Pegar a imagem de uma pasta no computador (adicionar caminho ate a imagem).
! [Legenda] (Cap4-R_markdown/RMarkdown.png)
- Sintaxe:
! [Legenda] (Endereço da Imagem)
- Exemplo:



Figure 3: Logo do R Markdown

6.9 Fórmulas LaTeX

6.9.1 Equações

- As equações no **R Markdown** são escritas com a linguagem **LaTeX**.
- Para que a equação apareça no meio do texto, devemos escrevê-la entre dois cifrões: `$equação$`
- Para que a equação apareça no formato destacado (display), deve ser colocada entre quatro cifrões: `$$equação$$`

6.9.2 Superescrito e subscritos

- Superescrito $a^2 = a^2$
- Subscrito $a_2 = a_2$
- Agrupado $a^{2+2} = a^{2+2}$
- Subscrito dois índices $a_{i,j} = a_{i,j}$
- Combinando super e subscrito $a_2^3 = a_2^3$
- Derivadas $x' = x'$

6.9.3 Sublinhados, sobrelinhas e vetores

Fórmula	Símbolo
<code>\hat{a}</code>	\hat{a}
<code>\bar{b}</code>	\bar{b}
<code>\overrightarrow{ab}</code>	\overrightarrow{ab}
<code>\overleftarrow{cd}</code>	\overleftarrow{cd}
<code>\widehat{def}</code>	\widehat{def}
<code>\overline{ghi}</code>	\overline{ghi}
<code>\underline{jkl}</code>	\underline{jkl}

6.9.4 Frações, matrizes e chavetas

- Fração:

$$\$\\frac{1}{2}$$

- pmatrix:

```
$\\begin{pmatrix}
```

x & y \\

z & v \\

\\end{pmatrix}\$

$$\begin{pmatrix} x & y \\ z & v \end{pmatrix}$$

- bmatrix:

```
$\\begin{bmatrix}
```

0 & \cdots & 0 \\

\vdots & \ddots & \vdots \\

0 & \cdots & 0 \\

\\end{bmatrix}\$

$$\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$

- Bmatrix:

```
$\\begin{Bmatrix}
```

x & y \\

z & v \\

\\end{Bmatrix}\$

$$\begin{Bmatrix} x & y \\ z & v \end{Bmatrix}$$

- vmatrix:

```
$\\begin{vmatrix}
```

x & y \\

z & v \\

\\end{vmatrix}\$

$$\begin{vmatrix} x & y \\ z & v \end{vmatrix}$$

- Vmatrix:

```
$\\begin{Vmatrix}
```

x & y \\

z & v \\

\\end{Vmatrix}\$

$$\begin{vmatrix} x & y \\ z & v \end{vmatrix}$$

- matrix:

```
$\begin{matrix} x & y \\ z & v \end{matrix}$
```

$$\begin{matrix} x & y \\ z & v \end{matrix}$$

6.9.5 Expressões

- Combinação

$$\$ \{n \choose k} \$ = {n \choose k}$$

- Função piso

$$\$ \lfloor x \rfloor \$ = \lfloor x \rfloor$$

- Função teto

$$\$ \lceil x \rceil \$ = \lceil x \rceil$$

- Sobrechaves

$$\$ \begin{matrix} 5050 \\ \underbrace{1+2+\cdots+100} \\ 5050 \end{matrix} \$ = \overbrace{1+2+\cdots+100}^{5050}$$

- Sobchaves

$$\$ \begin{matrix} \underbrace{1+2+\cdots+100} \\ 5050 \end{matrix} \$ = \underbrace{1+2+\cdots+100}_{5050}$$

- Função por partes

$$\$ f(n) = \begin{cases} n/2, & \text{se } n \text{ é par} \\ 3n + 1, & \text{se } n \text{ é ímpar} \end{cases} \$$$

- Limites

$$\$ \lim_{n \rightarrow \infty} x_n \$ = \lim_{n \rightarrow \infty} x_n$$

- Integral

$$\$ \int_{-N}^N e^x dx \$$$

- Integral Linear

$$\$ \oint_C x^3 dx + 4y^2 dy \$ = \oint_C x^3 dx + 4y^2 dy$$

- Integral Múltipla

$$\$ \iiint_V \mu(u, v, w) du dv dw \$ = \iiint_V \mu(u, v, w) du dv dw$$

- Somatório

$$\$ \sum_{k=1}^N k^2 \$ = \sum_{k=1}^N k^2$$

- Somatório com dois índices

$$\$ \sum_{\substack{0 < i < m \\ 0 < j < n}} k_{i,j} \$ = \sum_{\substack{0 < i < m \\ 0 < j < n}} k_{i,j}$$

- Produtório

$$\$ \prod_{i=1}^N x_i \$ = \prod_{i=1}^N x_i$$

- Raiz n-ésima

$$\$ f(x) \approx \sqrt[n]{x} \$ = f(x) \approx \sqrt[n]{x}$$

6.9.6 Sinais e setas

- Principais sinais e setas:

Fórmula	Símbolo
\sim	\sim
\simeq	\simeq
\cong	\cong
\leq	\leq
\geq	\geq
\equiv	\equiv
\approx	\approx
\neq	\neq
\leftarrow	\leftarrow
\rightarrow	\rightarrow
\leftrightarrow	\leftrightarrow
\longleftarrow	\longleftarrow
\longrightarrow	\longrightarrow
\mapsto	\mapsto
\longmapsto	\longmapsto
\nearrow	\nearrow
\searrow	\searrow
\swarrow	\swarrow
\nwarrow	\nwarrow
\uparrow	\uparrow
\downarrow	\downarrow
\updownarrow	\updownarrow

- Guia de fórmulas:
http://pt.wikipedia.org/wiki/Ajuda:Guia_de_ediç~ao/Fórmulas_TeX

6.10 Letras gregas

- Expressões matemáticas, ou letras gregas, devem vir entre símbolos de \$.

Fórmula	Símbolo
$\$\\alpha$$	α
$\$\\beta$$	β
$\$\\gamma$$	γ
$\$\\delta$$	δ
$\$\\epsilon$$	ϵ
$\$\\varepsilon$$	ε
$\$\\zeta$$	ζ
$\$\\eta$$	η
$\$\\theta$$	θ
$\$\\vartheta$$	ϑ
$\$\\iota$$	ι
$\$\\kappa$$	κ
$\$\\lambda$$	λ
$\$\\mu$$	μ
$\$\\nu$$	ν
$\$\\xi$$	ξ
$\$\\pi$$	π
$\$\\varpi$$	ϖ
$\$\\rho$$	ρ
$\$\\varrho$$	ϱ
$\$\\sigma$$	σ
$\$\\varsigma$$	ς
$\$\\tau$$	τ
$\$\\upsilon$$	υ
$\$\\phi$$	ϕ
$\$\\varphi$$	φ
$\$\\chi$$	χ
$\$\\psi$$	ψ
$\$\\omega$$	ω

- Para letra maiúscula, inicie a letra na fórmula com letra maiúscula.

$\delta = \$\\delta$

$\Delta = \$\\Delta$

6.11 Formatação (Fontes)

- Para deixar uma palavra em **negrito**, coloque-a entre quatro asteriscos: ****negrito****.
- Para deixar uma palavra em *itálico*, coloque-a entre dois asteriscos: ***itálico***.
- Para deixar uma palavra em ~~tachado~~, coloque-a entre dois til: **~~tachado~~**.
- Para deixar caracteres ^{sobrescritos}, coloque-os entre acentos circunflexos: **^1^**.
- Para deixar caracteres _{subscritos}, coloque-os entre til: **~1~**.
- Outra forma de escrever subscritos₂ (forma *LaTeX*), colocar no formato subscrito equação do *LaTeX*: **subscrito\$_{2}\$_**.
- Para destacar um termo como **código**, coloque-o entre crases (backticks): **`código`**.
- Para criar uma citação (quote), escreva o texto após um sinal de maior: **> Citação**.

6.12 Abas

- Aplica a um **#titulo** um comando (**{.tabset}**) que transforma em abas os **##subtitulo** com os graficos e tabelas contidos neles.
- Muito útil para relatórios dinâmicos (**html**).
- Exemplo:
titulo {.tabset}

7 CAP. 5 - PACOTES DO TIDYVERSE E IDENTIFICANDO/MUDANDO TIPOS DE VARIÁVEIS

7.1 Identificando/mudando tipos de variáveis

- i. Principais tipos de variáveis:

Table 6: Principais tipos de dados

Tipo	Descrição
numeric	Pode ser tanto inteiro (int, ou integer) quanto float (dbl).
character	São caracteres (chr).
factor	São variáveis do tipo fator.
logical	Variáveis do tipo lógico: TRUE ou FALSE.
complex	No formato de números complexo: 4 + 5i.

- ii. Identificando o tipo da variável:

Uso do `is`.

```
is.numeric(variavel)
[TRUE]
```

- iii. Mudando o tipo da variável:

Uso do `as`.

```
as.character(variavel)
```

- iv. Observações:

- Não é possível transformar uma variável do tipo caracter(character) direto para tipo número (numeric), é preciso transformar de caracter (character) para fator (factor) e de fator (factor) para número (numeric).

```
is.character(variavel)
[TRUE]
as.factor(variavel)
as.numeric(variavel)
```

- O contrário, transformar de número (numeric) para caracter (character) é possível.

```
is.numeric(variavel)
[TRUE]
as.character(variavel)
```

7.2 Pacotes do Tidyverse

- **readr**
Leitura de dados.
- **tibble**
Tipo de data.frame.
- **magrittr**
Operador pipe %>%, concatena linhas de comando.
- **dplyr**
Manipulação de dados.
- **tidyverse**
Organização de dados.
- **ggplot2**
Elaboração de gráficos.

7.3 Leitura de dados (readr)

- Os principais formatos de importação de dados são:

- *csv*
- *xls*
- *xlsx*
- *sav*
- *dta*
- *por*
- *sas*
- *stata*

- Entre os principais formatos de importação de dados o mais usado é o *csv*.

7.3.1 Importação de dados via RStudio

- No “Environment” tem a opção “Import Dataset”, que pode ser usada para importação de dados “.csv”.

“Environment” > “Import Dataset” > “From Text (Readr)”

- Dentro de “Import Text Data”:

- **File/URL**

O caminho ate o arquivo “.csv”.

- **Data Preview**

Mostra uma prévia de como os dados serão lidos (ficarão organizados no **R**). Se não estiver vizualizando, aperte o botão **update**.

- **Import Options**

São as configurações que podem ser modificadas para garantir a integridade da importação dos dados.

Definindo, por exemplo, se o que separa casas decimais nos dados é virgula ou ponto.

- **Code Preview**

Apresenta o código que está sendo construído pela automatização da janela. Este código pode ser copiado e executado fora da janela.

- **Import**

Botão para concluir a operação da importação dos dados.

7.3.2 Importação de dados via biblioteca `readr`

- As principais funções de importação de arquivo `.csv` são:
 - `read.csv`
É uma função basica do **R**, não precisa chamar nenhuma biblioteca. Usa o separador de campos virgula.
 - `read.csv2`
É uma função basica do **R**, não precisa chamar nenhuma biblioteca. Usa o separador de campos ponto e virgula.
 - `readr::read_csv`
É uma função do pacote **readr**, por isso o uso de “`readr::`” para chamar a função. Usa o separador de campos virgula.
 - `readr::read_csv2`
É uma função do pacote **readr**, por isso o uso de “`readr::`” para chamar a função. Usa o separador de campos ponto e virgula.
 - `readr::read_tsv`
É uma função do pacote **readr**, por isso o uso de “`readr::`” para chamar a função. Usa o separador de campos tabulação.
 - `readr::read_delim`
É uma função do pacote **readr**, por isso o uso de “`readr::`” para chamar a função. Usa o separador de campos generico, deve ser especificado pelo parâmetro `delim` =.
- Principais parâmetros, das funções de importação, do pacote **readr**:
 - `file` =
Define o caminho (ou **URL**), que deve ser pecorrido, e o arquivo, no formato `.csv`, a ser importado.
Deve estar entre aspas.
Exemplo: `file = "Caminho/arquivo.csv"`
 - `col_names` =
Indica se a primeira linha contém, ou não, o nome das colunas. Também pode ser usado para renomear colunas.
Se a primeira linha contém o nome das colunas = **TRUE**.
Para nomear, ou renomear, colunas podemos usar um vetor contendo os nomes.
Exemplo:
`col_names = TRUE`
`col_names = c("coluna_1", "coluna_2", ...)`
 - `col_types` =
Caso alguma coluna tenha sido importada com a classe errada, podemos usar esse parâmetro para mudar e especificar o tipo de cada coluna.
Podemos especificar através de uma lista contendo as classes de cada coluna, ou uma cadeia de caracteres com caracteres simbolicos para cada classe de cada coluna.
Outra possibilidade é mudar as classes através de funções de mudança de classe, usando a função `cols()`, onde `.default` = indica a classe default de importação, para casos não especificados, e caso precisarmos identificar uma coluna em específico `nome_da_coluna` =.

* Caracteres simbólicos:

- c = character
- i = integer
- n = number
- d = double
- l = logical
- f = factor
- D = date
- T = date time
- t = time
- ? = guess
- _or_ = skip

* Funções de mudança de classe:

- col_character()
- col_date()
- col_time()
- col_datetime()
- col_double()
- col_factor()
- col_integer()
- col_logical()
- col_number()
- col_skip()

* Exemplos:

```
col_types = "iccd"
col_types = cols(.default = "i", xxx = "c")
col_types = cols(.default = col_integer(), xxx = col_character())
```

- skip =

Pula linhas do começo do arquivo antes de iniciar a importação. Util quando tem algum texto explicativo na primeira linha do arquivo.

Exemplos:

```
skip = 0  
skip = 1
```

- **na** =

Indica quais *strings* deverão ser tratadas como **NA** na hora da importação.

Exemplo: `na = c("", "NA")`

- **delim** =

No caso da função `read_delim`, podemos definir através deste parâmetro o tipo de delimitador de campos usado no arquivo. O caractere usado como delimitador de campo deve estar entre aspas.

Exemplo: `delim = ","`

- Sintaxe:

```
library(readr)  
dt <- read_csv2(file = "~/caminho/arquivo.csv",  
  col_names = TRUE,  
  col_types = "iccD",  
  na = c("", "NA"),  
  skip = 0)
```

- Observação:

Por padrão *csv* usa separação por vírgula, porém no Brasil como a vírgula é usada para separação de casas decimais, o padrão *csv* brasileiro o separador de campo é o ponto e vírgula, sendo assim, para importar dados em formato *csv* no Brasil a melhor escolha é o pacote `readr::read_csv2`.

7.4 tibble

7.4.1 Visualização de tabelas tipo tibble

- *tibble* é um tipo especial de tabela equivalente ao *data.frame*, porém mais compacta e com mais informações.
- O *tibble* exibe informações sobre os tipos de cada variável:
 - *factor(fct)*
 - *character(chr)*
 - *integer(int)*
 - *double(dbl)*
- Visualização da tabela:
 - O *tibble* também omite linhas quando a tabela é muito numerosa, para melhor visualização.
 - O *tibble* por **default** exibira ate 10 linhas.

```
library(tibble)
dt <- tibble(dados)
dt
```

- Caso necessite ver mais linhas basta especificar.
`print(dt, n=15)`

7.4.2 Criação de tabela tipo tibble

- Primeiramente é necessário chamar a biblioteca **tibble**
`library(tibble)`
- De forma semelhante ao **data.frame**, podemos criar tabelas do tipo **tibble**.
`x = tibble(coluna1 = c(...), coluna2 = c(...), ...)`

7.4.3 Funções tibble

- `as_tibble()`
 - Transforma um `data.frame` em tipo `tibble`, através da função `as_tibble()`.
`x <- as_tibble(x)`
- `is_tibble()`
 - Verifica se uma tabela é tipo `tibble`, através da função `is_tibble()`. Retorna `TRUE` (se verdadeiro), ou `FALSE` (se falso).
`is_tibble(x)`
- `add_column()`
 - Adiciona novas colunas.

`dados1 %>%
add_column(nome = valor)`
 - Também é possível definir a posição onde a nova coluna vai se encaixar, indicando a posição (`.before = 1` ou `.after = 1`).

`dados1 %>%
add_column(nome = valor, .before = 1)`
- `add_row()`
 - Adiciona novas linhas.

 - Também é possível definir a posição da nova linha através dos comandos `.before` ou `.after`.

 - É necessário adicionar as informações e referenciar as colunas.

`dados1 %>%
add_row(cupom = 100, filial = "A",
valor_compra = 10, n_itens = 1,
desconto_perc = 0, quinzena = 1,
.before = 1)`

7.5 Operador pipe

- Esta contido do pacote `magrittr`.
- Funciona como uma função composta, tornando a leitura das linhas de comando mais lógica e natural.
- Trata-se de um operador cuja notação é `%>%`. Com ele podemos encadear (concatenar) linhas de comandos na ordem de execução.
- Atalho no teclado `ctrl+shift+M`.
- Exemplo:

```
library(magrittr)
library(dplyr)

dados1 %>%
  select(filial, quinzena) %>%
  filter(quinzena == 1)
```

7.6 Manipulando dados com o dplyr

7.6.1 manipulação de dados:

- `select()`

- Seleciona e retorna as colunas selecionadas da tabela.
- Retorna as colunas selecionadas no formato tabela.
- Pode retornar mais de uma coluna.
- Exemplo:

```
library(dplyr)
library(magrittr)
dados1 %>%
  select(filial, quinzena, valor_compra)
```

- `pull()`

- Extrai uma coluna de uma tabela de dados e retorna ela como vetor.
- A coluna identificada para extração pode ser tanto pelo nome, quanto pela posição.
- Retorna apenas uma coluna, no formato vetor.
- Exemplo:

```
library(dplyr)
library(magrittr)
vetor <- dados1 %>%
  pull(filial)
ou,
pull(2)
ou,
pull(-5)
```

- **filter()**

- Filtra linhas.

- Exemplo:

```
library(dplyr)
library(magrittr)
dados1 %>%
filter(filial == "A")
```

- Principais operadores lógicos:

Table 7: Tabela dos principais operadores lógicos usados na função filter.

Operador lógico	Descrição
==	Igualdade
!=	Diferença
>	Maior que
<	Menor que
>=	Maior ou igual que
<=	Menor ou igual que
&	E
	OU
!	Negação

- **distinct()**

- Remove linhas com valores repetidos de determinadas colunas.

- Podemos extrair todas as linhas distintas , do banco de dados, pelo comando **distinct()**, apenas não especificando as colunas.

- Exemplo:

```
library(dplyr)
library(magrittr)
dados1 %>%
distinct(filial)
ou,
distinct(filial, quinzena, desconto_perc)
ou,
distinct()
```

- **arrange()**

- Reordena em determinadas colunas as linhas.

- Pode reordenar mais de uma coluna por vez.

- `arrange(coluna_1,coluna_2,...)`

- ou

```
arrange(coluna_1) %>%  
arrange(coluna_2)
```

- A ordem das colunas na função determina a prioridade na ordenação.

- Por **default** reordena as linhas em ordem crescente.

- Podemos também reordenar as linhas em ordem decrescente:

- * `arrange(-nome_coluna)`

- Colocando um sinal de negativo na frente da coluna é informar ordenar em decrescente.

- * `arrange(desc(nome_coluna))`

- Usando a função `desc()`.

- Exemplo:

```
library(magrittr)  
library(dplyr)  
  
dados1 %>%  
arrange(n_itens,valor_compra) %>%  
filter(valor_compra > 100) %>%  
select(filial,n_itens,valor_compra)
```

- **mutate()**

- Cria novas colunas, na base de dados.

- Exemplo:

```
library(magrittr)
library(dplyr)

dados1 %>%
  mutate(vmci = round(valor_compra/n_itens))
```

- **transmute()**

- Cria novas colunas, mas não adiciona na base de dados.

- A diferença de **transmute()** para **mutate()** é que em **mutate()** acrescenta novas colunas aos dados originais, enquanto que em **transmute()** criamos novas colunas a partir dos dados originais.

- Exemplo:

```
library(magrittr)
library(dplyr)

dados1 %>%
  transmute(vmci = round(valor_compra/n_itens))
```

- `summarise()`
 - Permite sumarizar variáveis, ou seja, produzir tabelas resumidas do banco de dados.
 - Pode ser usado em conjunto com o comando `group_by()` para obter o resumo de grupos.
 - Sintaxe:
`summarise(nome_da_coluna = função_summarise(coluna))`
 - Exemplo:

```
dados1 %>%
  select(filial) %>%
  summarise(item_total = sum(n_itens))

ou

dados1 %>%
  group_by(filial) %>%
  summarise(cupons_distintos = n_distinct(cupom))
```

- Principais funções de sumarização:

Table 8: Principais funções de summarise

Funções	Descrição
<code>n()</code>	Conta o número de elementos da coluna x
<code>n_distinct(x)</code>	Conta os elementos distintos da coluna x
<code>sum(x)</code>	Soma os valores da coluna x, retorna o acumulado
<code>mean(x)</code>	Cálcula a média da coluna x
<code>median(x)</code>	Cálcula a mediana da coluna x
<code>quantile(x,k)</code>	Cálcula o percentil de ordem $0 \leq k \leq 1$ dos valores da coluna x
<code>min(x)</code>	Retorna o valor mínimo da coluna
<code>max(x)</code>	Retorna o valor máximo da coluna
<code>var(x) ou var(x,y)</code>	Cálcula a variância da coluna x, ou a covariância da coluna x em relação a coluna y
<code>sd(x)</code>	Cálcula o desvio-padrão da coluna x
<code>prod(x)</code>	Multiplica os valores da coluna x

- `group_by()`
 - Permite operações por grupo. Agrupa dados de determinadas colunas.
 - Agrupa as colunas priorizando a ordem em que aparecem na função.
 - Exemplo:
`group_by(coluna1,coluna2,...)`

- **rename()**

- Renomeia uma coluna.

```
rename(novo_nome = antigo_nome)
```

- Pode renomear várias colunas de uma vez.

```
dados1 %>%  
  rename(x1 = coluna1, x2 = coluna2, ...)
```

7.6.2 combinando tabelas de dados:

- `bind_cols()`
 - Une duas tabelas lado a lado, sobrepostas. Ou seja, soma o número de colunas das duas tabelas.
 - Acrescenta numeração as colunas repetidas. Ou seja, se houver a mesma coluna nas duas tabelas, será acrescido ao nome das colunas repetidas um valor.
 - É necessário que tenha o mesmo número de linhas nas duas tabelas para fazer essa combinação.
 - Dentro da função, a ordem de chamada de cada tabela determina a ordem das colunas.
 - Exemplo:

```
library(tibble)
library(magrittr)
library(dplyr)
x <- dados1 %>%
  select(cupom,filial,valor_compra)
y <- dados1 %>%
  select(cupom,n_itens)
z <- bind_cols(x,y)
colnames(z)
[1] "cupom...1" "filial" "valor_compra" "cupom...4" "n_itens"
```

- `bind_rows()`
 - Une duas tabelas pelas linhas.
 - Não é necessário que o número de linhas, ou colunas, seja igual nas duas tabelas. Nesse ponto é diferente do comando `bind_cols()`.
 - As colunas das duas tabelas são combinadas, porém das colunas repetidas é mantida apenas uma.
 - Quando não há correspondência entre as colunas o comando retorna **NA**, no valor específico da linha.
 - Dentro da função, a ordem de chamada de cada tabela determina a ordem das colunas.
 - Exemplo:

```
library(tibble)
library(magrittr)
library(dplyr)
x <- dados1 %>%
  select(cupom,filial,valor_compra)
y <- dados1 %>%
  select(cupom,n_itens)
z <- bind_rows(x,y)
```

```
colnames(z)
[1] "cupom" "filial" "valor_compra" "n_itens"
```

- **inner_join()**

- A tabela final será o resultado da intersecção das colunas de x e y, que possuem pelo menos uma coluna em comum, a coluna *chave*.
- Junta duas colunas pela interseção.
- Ao juntar as duas tabelas pela função `inner_join()`, apenas os registros que existam nas duas tabelas (pela coluna chave) são unidos, os demais registros de cada tabela não são agregados.
- Os filtros (`filter()`) aplicados a cada tabela são somados.
- Exemplo:

```
x = dados1 %>%
  select(cupom, filial, valor_compra) %>%
  filter(valor_compra >500)
x
y = dados1 %>%
  select(filial,n_itens) %>%
  filter(n_itens < 8)
y
inner_join(x,y)
```

- **left_join()**

- Une duas tabelas, definindo qual será a tabela principal (tabela da **esquerda**).
- Apresenta e prioriza os registros da tabela principal (tabela da **esquerda**).
- O resultado final reúne todos os registros da tabela principal e os correspondentes na outra tabela.
- É necessário que tenha pelo menos uma coluna em comum, a coluna chave.
- Exemplo:
`left_join(tabela_principal, tabela_secundaria)`

- **right_join()**

- Une duas tabelas, definindo qual será a tabela principal (tabela da **direita**).
- Apresenta e prioriza os registros da tabela principal (tabela da **direita**).
- O resultado final reúne todos os registros da tabela principal e os correspondentes na outra tabela.
- É necessário que tenha pelo menos uma coluna em comum, a coluna chave.
- Exemplo:
`right_join(tabela_secundaria, tabela_principal)`

- **full_join()**

- Une duas tabelas.
- Mantem todos os registros.

- Prestar atenção na junção das linhas/registros que formam novas informações, através da combinação de correspondentes.
- Os registros sem correspondentes na outra tabela são preenchidos com valor **NA**.
- É necessário que tenha pelo menos uma coluna em comum, uma coluna chave.
- Exemplo:

```
x <- dados1 %>%
  select(cupom,filial,valor_compra) %>%
  filter(valor_compra > 500)

y <- dados1 %>%
  select(filial,n_itens) %>%
  filter(n_itens < 8)

full_join(x,y)
```

- **intersect()**
Retorna a interseção entre tabelas.
`intersect(x,y)`
- **union()**
 - Retorna a união de tabelas.
 - Não repete registros iguais nas duas tabelas.
 - Monta a nova tabela na ordem em que as tabelas foram inseridas na função.
 - Exemplo:
`union(x,y)`
- **setdiff()**
 - Retorna a diferença entre tabelas.
 - A ordem das tabelas na função interfere na saída:
 - * `setdiff(x,y)`
Retorna tudo que está em x e não está em y.
 - * `setdiff(y,x)`
Retorna tudo que está em y e não está em x.
- **setequal()**
 - Esse comando verifica se duas tabelas de dados possuem linhas com os mesmos valores, independentemente da ordem em que tais valores se apresentem.
 - Retorna **TRUE**, se os registros forem iguais, ou **FALSE**, se os registros forem diferentes.
 - Sintaxe:
`setequal(tabela_1,tabela_2)`

7.7 Organizando dados com o tidyverse

- `pivot_longer()` ou `gather()`
 - Converte a tabela de dados do formato larga para o formato longo. (larga -> longo)
 - A função `pivot_longer()` substituiu a função `gather()`, após o ano de 2019.
 - Transformação:
 - * Converte várias colunas do dataframe original em duas colunas e várias linhas/registros.
 - * Uma coluna recebe o nome das variáveis em colunas e a outra recebe os valores dessas variáveis.
 - * Ao final o número de linhas do dataframe é ampliado e o número de colunas diminuiu.
 - Condição:
 - * As colunas originais devem ter em comum a mesma variável.
 - * Pelo menos duas colunas contendo os nomes das categorias de uma variável separados por colunas.
 - Sintaxe:

```
tbla_longa <- tabela_larga %>% pivot_longer(cols = c(coluna_4,coluna_5)),  
names_to = "nova_coluna_1", values_to = "nova_coluna_2")
```

 - * `cols`
Recebe as colunas que vão ser transformadas em linhas.
 - * `names_to`
Nome da nova coluna que vai receber como variável o nome das colunas originais.
 - * `values_to`
Nome da nova coluna que vai receber os valores contidos nas colunas originais.
 - Exemplo:

Table 9: Tabela de chegada de turistas no formato larga

Estado	Terrestre	Aéreo
SP	3900	4200
RS	2800	3800
RJ	2600	3950

Table 10: Tabela de chegada de turistas no formato longo

Estado	Meio	Chegada
SP	Terrestre	3900
SP	Aereo	4200
RS	Terrestre	2800

Estado	Meio	Chegada
RS	Aereo	3800
RJ	Terrestre	2600
RJ	Aereo	3950

- **pivot_wider()** ou **spread()**
 - Converte a tabela de dados do formato longo para o formato larga. (longo -> larga)
 - A função **pivot_wider()** substituiu a função **spread()**, após o ano de 2019.
 - As funções **pivot_wider()** e **spread()**, faz o inverso das funções **pivot_longer()** e **gather()**, ou seja, espalha os dados das linhas por colunas.
 - Transformação:
 - * Converte várias linhas do dataframe original em colunas.
 - Sintaxe:
`tabela_larga <- tabela_longa %>% pivot_wider(names_to = "coluna_4", values_to = "coluna_5")`
 - * **names_to**
Determina qual coluna terá seus valores transformados em novas colunas.
 - * **values_to**
Determina qual coluna terá seus valores distribuidos entre as novas colunas.
 - Exemplo:

Table 11: Tabela em formato longo dieta de pacientes

Pacientes	Tempo	Sexo	dieta	Peso
1	4	Homem	Antes	150
2	4	Homem	Antes	160
3	3	Mulher	Antes	90
4	3	Mulher	Antes	95
5	6	Mulher	Antes	110
1	4	Homem	Depois	140
2	4	Homem	Depois	110
3	3	Mulher	Depois	80
4	3	Mulher	Depois	80
5	6	Mulher	Depois	82

Table 12: Tabela em formato larga dieta de pacientes

Pacientes	Tempo	Sexo	Antes	Depois
1	4	Homem	150	140
2	4	Homem	160	110
3	3	Mulher	90	80
4	3	Mulher	95	80
5	6	Mulher	110	82

- **separate()**
 - Separa os dados contidos numa mesma coluna para diversas colunas.
 - Transforma um campo vetorizado em diversas colunas separadas.
 - É necessário determinar o separador, o caracter que separa os dados dentro do campo.
 - Argumentos:
 - * Coluna que vai ter seus dados separados.
 - * Novas colunas que vão receber os dados.
 - * Caracter que separa os dados na coluna original.
 - Exemplo:

```
resposta <- dados %>% separate(cor, c("cor1","cor2"), sep = ",")
```
- **unite()**
 - O comando **unite()** é utilizado para unir duas ou mais colunas em uma unica coluna.
 - Argumentos:
 - * Nova coluna que vai receber os dados unidos.
 - * Colunas originais que vão ceder os dados.
 - * Caracter separador usados para separar os dados na nova coluna.
 - Exemplo:

```
resposta_unida <- dados %>% unite("cor", c("cor1","cor2"), sep = ",")
```

- `complete()`
 - Gera todas as combinações possíveis entre as colunas, ou tabelas, selecionadas.
`dados %>% complete(coluna1,coluna2,coluna3,...)`
`dados %>% complete(dt1,dt2,...)`
 - Completa as combinações de colunas, se não houver valor, com **NA**.
 - O comando `nesting()`, que pode ser usado dentro da função `complete()`, cruza todos os valores de determinado grupo (tabela) com os pares únicos dos valores das colunas selecionadas em `nesting()`.
`dados1 %>% complete(dt,nesting(coluna1,coluna3))`
- `drop_na()`
 - Elimina as linhas, especificadas ou não, com valor **NA**.
 - Eliminando linhas com **NA**, de colunas especificadas:
`dados %>% drop_na(c(coluna1,coluna2))`
 - Eliminando todas as linhas com valor **NA**:
`dados %>% drop_na()`
- `replace_na()`
 - Substitui os valores **NA**, de determinada coluna, por outro valor especificado.
 - Especifica a coluna, ou as colunas através de `list()`, e define o valor caso **NA**.
 - Exemplo:
`dados %>% replace_na(list(paciente = "ausente", antes = 0, depois = 0))`

8 SINCRONIZAÇÃO COM BANCO DE DADOS

8.1 Pacotes de banco de dados

- Sincronização com banco de dados:
 - **DBI**
Conecta R ao sistema de gerenciamento de banco de dados (SGBD).
 - **odbc**
Componentes de conexão/interação com bancos de dados.
 - **RSQLite**
Conexão com SQLite.
 - **RMySQL**
Conexão com MySQL e MariaDB.
 - **RPostgres**
Conexão com Postgres.
 - **RPostgreSQL**
Alternativo Conexão com PostgreSQL.
 - **bigrquery**
Conexão Google's BigQuery.
- Manipulação de dados:
 - **dbplyr**
Tradução de dplyr em dbplyr (SQL).
 - **sqldf**
Permite manipular data.frame em R com instruções SQL.

8.2 Sincronização com banco de dados

- Referência para estudo de pacotes R de banco de dados:
<http://db.rstudio.com/>

8.2.1 DBI

- O pacote **DBI** ajuda a conectar **R** a sistemas de gerenciamento de banco de dados (SGBD).
- Ele suporta as seguintes operações:
 - Conectar/desconectar do SGBD.
 - Criar e executar instruções no SGBD.
 - Extrair resultados de declarações.
 - Tratamento de erros e exceções.
 - Informações (metadados) de objetos de banco de dados.
 - Gerenciamento de transação.
- Ele é instalado automaticamente quando instalado um dos *backends* do banco de dados:
 - **odbc**
 - **RSQLite**
 - **RMariaDB** ou **RMySQL**
 - **RPostgres**
 - **bigrquery**
 - ...

- Principais funções:

- `dbConnect(backends::SGBD(), dbname = "nome_banco_de_dados")`
Conecta a determinado banco de dados.
Ex.: `con <- dbConnect(RSQLite::SQLite(), dbname = "memory")`
- `dbListTables(con)`
Lista as tabelas.
- `dbWriteTable(con, "nome_tabela", valores)`
Cria uma tabela.
Ex.:
`dbWriteTable(con, "mtcars", mtcars)`
- `dbListFields(con, "nome_tabela")`
Lista os campos de uma tabela.
Ex.: `dbListFields(con, "mtcars")`
- `dbReadTable(con, "nome_tabela")`
Mostra determinada tabela contida no banco de dados.
Ex.:
`dbReadTable(con, "mtcars")`
- `dbSendQuery(con, "query")`
Salva uma query (SQL) numa variável (não executa).
Ex.:
`res <- dbSendQuery(con, "SELECT * FROM mtcars WHERE cyl = 4")`
- `dbFetch(variável)`
Executa determinada query (SQL) salva numa variável.
Ex.:
`dbFetch(res)`
- `dbClearResult(variável)`
Limpa determinada variável que contém uma query.
Ex.:
`dbClearResult(res)`
- `dbDisconnect(con)`
Desconecta do banco de dados.

8.2.2 odbc

- O pacote `odbc` oferece uma interface compativel com DBI para *drives open database connectivity* (ODBC).
- Permite de uma maneira fácil e eficiente de configurar a conexão com banco de dados usando um drive ODBC, incluindo:
 - SQLite
 - MySQL
 - PostgreSQL
 - SQL Server
 - Oracle
 - ...
- Os pacotes dos *drives* podem ser baixados separadamente.
- O pacote `odbc` funciona com em uso conjunto com o pacote DBI.

- Principais funções:

- Explorando objetos e colunas da banco de dados:

- * `odbcListObjects(con)`
Lista objetos.
- * `odbcListObjects(con, catalog="mydb", schema="dbo")`
Tabelas em determinado schema.
- * `odbcListColumns(con, catalog="mydb", schema="dbo", table="cars")`
colunas em uma tabela.
- * `odbcListObjectTypes(con)`
Estruturas de objetos contidas num banco de dados.

- Dados do sistema:

- * `odbcListDrivers()`
Todos os *drives* do sistema.
- * `odbcListDataSources()`
Todas as fontes do sistema.

- Queries (consultas) e declarações:

- * `dbGetQuery(con, "SELECT speed, dist FROM cars")`
Para enviar consultar interativas. Envia uma consulta e busca os resultados.
Faz as duas coisas “consultar” e “buscar” o resultado, é equivalente ao `dbSendQuery() + dbFetch()`
- * `dbSendQuery(con, "SELECT speed, dist FROM cars")`
Salva uma query (SQL) numa variável (não executa).
Ex.:`query <- dbSendQuery(con, "SELECT speed, dist FROM cars")`
- * `dbFetch(query, n = 10)`
Executa determinada query (SQL) salva numa variável.
`n = valor` limita o resultado (resultado parcial).
Ex.:`dbFetch(query)`
- * `dbClearResult(variável)`
Limpa determinada variável que contém uma query.
Ex.:`dbClearResult(query)`
- * `dbExecute(con, "INSERT INTO cars (speed, dist) VALUES (88, 30)")`
Força uma instrução SQL direta.

- Conexão com banco de dados usando pacotes **DBI** + **odbc** (modelo):

```
#Bibliotecas
library(DBI)
library(odbc)

#Conexão com banco de dado
con <- dbConnect(odbc::odbc(),
  driver = "PostgreSQL Driver",
  database = "test_db",
  uid = "postgres",
  pwd = "password",
  host = "localhost",
  port = 5432)
```

8.2.3 Drives

8.2.3.1 RSQLite

- Pacote **RSQLite** incorpora o mecanismo de banco de dados **SQLite** em **R** (**RSQLite**), fornece uma interface compatível com DBI.

- Por *default* as extensões SQLite3 estão habilitadas.

- Conectar ao banco de dados **SQLite**:

```
con <- dbConnect(RSQLite::SQLite(), dbname = "nome_database")
```

- Criar um nova *database* no **SQLite** pelo **R**:

```
con <- dbConnect(RSQLite::SQLite(), "my-db.sqlite")
```

- Criar um nova *database* temporaria no **SQLite** pelo **R**:

```
con <- dbConnect(RSQLite::SQLite(), ":memory:")
```

Este banco de dados será excluído automaticamente quando for desconectado.

- Desconectar do banco de dados **SQLite**:

```
dbDisconnect(con)
```

8.2.3.2 RMySQL ou RMariaDB

- O pacote `odbc`, em combinação com um drive (`RMySQL` e/ou `RMariaDB`), fornece suporte ao DBI e uma conexão ODBC.

- Configurações de conexão pelo pacote `RMySQL`:

- `dbname`

Nome da *database* que deseja se conectar dentro do **MySQL**.

- `host`

“localhost”.

- `port`

Para o **MySQL** é 3306.

- `user`

Login de usuário para se conectar ao **MySQL**.

Normalmente é “root”.

- `password`

Senha para loggar no banco de dados **MySQL**.

- Conexão com banco de dados **MySQL** usando pacotes DBI + `odbc` + `RMySQL` (modelo):

```
con = dbConnect(RMySQL::MySQL(),
                 dbname='Database name',
                 host='localhost',
                 port=3306,
                 user='Database user',
                 password='Database password')
```

8.2.3.3 RPostgres

- O pacote `odbc`, em combinação com um drive (`RPostgres`), fornece suporte ao DBI e uma conexão ODBC.
- Configurações de conexão pelo pacote `RPostgres`:

- `dbname`
Nome da *database* que deseja se conectar dentro do **PostgreSQL**.
- `host`
“localhost”.
- `port`
Para o **PostgreSQL** é 5432.
- `user`
Login de usuário para se conectar ao **PostgreSQL**.
Normalmente é “postgres”.
- `password`
Senha para loggar no banco de dados **PostgreSQL**.

- Conexão com banco de dados **PostgreSQL** usando pacotes DBI + `odbc` + `RPostgres` (modelo):

```
con <- dbConnect(RPostgres::Postgres(),
                  dbname = 'Database name',
                  host = 'localhost',
                  port = 5432,
                  user = 'Database user',
                  password = 'Database password')

ou,

con <- DBI::dbConnect(odbc::odbc(),
                      Driver = "PostgreSQL Driver",
                      Server = "localhost",
                      Database = "Database name",
                      UID = "rstudioapi::askForPassword("Database user")",
                      PWD = "rstudioapi::askForPassword("Database password")",
                      Port = 5432)
```

8.3 Importação de tabelas

8.3.1 Remoto

- Os principais pacotes para importar tabelas do banco de dados são DBI em conjunto com dplyr (do tidyverse).
- A partir dos pacotes DBI e dplyr é disponibilizado a função `tbl()` que possibilita a importação de determinada tabela do banco de dados para o R, via remota.
Ex.:`nome_tabela_no_R <- tbl(con, "nome_tabela_db")`

- A tabela ainda se encontra remota, locada dentro do banco de dados, ou seja, as manipulações são instruções enviadas para o banco de dados executar.

8.3.2 Local

- É importante observer que apesar da importação da tabela para o R, a tabela ainda se encontra remota, ou seja, estamos manipulando ela no banco de dados.

8.3.2.1 Remoto para local - `collect()`

- Para passar a tabela de remoto para local (para um `data.frame` de variável dentro do R) é necessário usar a função `collect()`.
Ex.:`tabela_local <- tabela_remoto %>% collect()`

8.3.2.2 Local para remoto (exportar dados para o banco de dados) - `copy_to()`

- A função `copy_to()` é o oposta da função `collect()`, pega um `data.frame` local e o carrega para a fonte remota.
Ex.:`copy_to(dest, df, name = deparse(substitute(df)), overwrite = FALSE)`
- Principais argumentos da função `copy_to()`:

– `dest`

Fonte de dados remota (con).

– `df`

`data.frame` local.

– `name`

Nome para a nova tabela remota.

– `overwrite`

Caso TRUE, substituirá uma tabela existente com o mesmo nome.

Caso FALSO, gerará um erro se o nome já existir.

8.4 Manipulação de tabelas

8.4.1 Manipular dados direto no banco de dados com R - dbplyr

- Para usar `dplyr` (manipulação de dados) direto no banco de dados (remotamente) é necessário instalar o pacote `dbplyr`.
- Existem duas situações em que é útil manipular dados remotamente pelo **R**:
 - Os dados já estão num banco de dados.
 - São tantos dados que não cabem na memória simultaneamente.
- Caso os dados caibam na memória, vale a pena trabalhar com eles localmente (importar para o **R**). Trabalhar com dados de forma remota deixa o processo mais lento.
- O pacote `dbplyr` funciona em conjunto com os pacotes `DBI`, `odbc`, pacote do backend específico do banco de dados (`SQLite` ou `RMySQL` ou `RPostgreSQL` ou ...) e `dplyr` (do pacote `tidyverse`).
- O objetivo do `dbplyr` é gerar instruções **SELECT** do **SQL** a partir do `dplyr`, para que não seja necessário usar o **SQL**.
- Funcionamento do `dbplyr`:
 - Traduz os comando `dplyr` para **SQL**.
 - Executa a instrução **SQL** no banco de dados.
 - Reune tudo que foi solicitado e envia tudo em uma única etapa para ser executado no banco de dados.
- Para visualizar o código da instrução **SQL** que está sendo gerada a partir do `dplyr` usamos a função `show_query()` em conjunto da variável salva com as instruções em `dplyr`.
Ex.: `Consulta_resultado_remoto %>% show_query()`

- Exemplo consulta usando `dplyr` (`dbplyr`):

```
# Importando a tabela
db_funcionarios <- tbl(con, "funcionarios")

#Consulta remota (dplyr tidyverse) enviar para banco de dados
Consulta_resultado_remoto <- db_funcionarios %>%
  select(idfuncionario, nome, sexo) %>%
  filter(sexo == "Masculino")
Consulta_resultado_remoto

# Source: SQL [?? x 3]
# Database: postgres [postgres@localhost:5432/data_science]
idfuncionario nome sexo
<int> <chr> <chr>
1 2 Armstrong Masculino
2 3 Carr Masculino
3 6 Phillips Masculino
4 7 Williamson Masculino
5 9 James Masculino
6 10 Sanchez Masculino
7 12 Black Masculino
8 13 Schmidt Masculino
9 18 Nguyen Masculino
10 19 Day Masculino
# i more rows
# i Use print(n = ...) to see more rows

#Mostrando a query gerada a partir do tidyverse
Consulta_resultado_remoto %>%
  show_query()

<SQL>
SELECT "idfuncionario", "nome", "sexo"
FROM "funcionarios"
WHERE ("sexo" = 'Masculino')

#Extraindo resultado remoto para local - R
Consulta_resultado_local <- Consulta_resultado_remoto %>%
  collect()
Consulta_resultado_local
```

8.4.2 Manipular dados localmente com SQL - sqldf

- O pacote `sqldf` permite manipular objetos `data.frame` de **R** através de instruções **SQL** simples.
- Usamos a função `sqldf()` para passar a instrução **SQL**.
Sintaxe: `sqldf("SELECT * FROM nome_data.frame;")`
O nome das tabelas são os nomes de objetos `data.frame`.
- Outra forma de executar a função, é passar a instrução **SQL** como texto para uma variável e executar a variável com da função `sqldf()`.

```
sql_inst = "SELECT * FROM dados;"  
sqldf(sql_inst)
```

- Exemplo - Manipulação de dados usando instruções **SQL**, com o pacote `sqldf`:

```
dados <- read_csv2(file = "/home/...")  
dados <- data.frame(dados) #Colocando no formato data.frame  
dados  
  
#Manipulando data.frame através de instruções SQL  
sql_instruction = "SELECT cupom, filial FROM dados;"  
df <- sqldf(sql_instruction)  
df
```

9 CAP. 6 - PACOTE DATA.TABLE

9.1 Teoria

- Manipula dados, porém usa uma filosofia diferente de **tidyverse**.
- Chega nos mesmos resultados que **tidyverse**.
- Apresenta um ganho em desempenho no tempo, em relação ao **tidyverse**.
- Não necessita de tantos pacotes para executar as tarefas.

9.2 Estrutura

- A estrutura básica do **data.table**:

– Sintaxe:

DT[i,j,by]

onde,

* **DT**

É o nome do **data.frame**.

* **i**

Corresponde a(s) linha(s) selecionadas, ou uma operação sobre a(s) linha(s).

* **j**

Corresponde a(s) coluna(s) selecionadas, ou uma operação sobre a(s) coluna(s).

* **by**

Agrupa os dados em torno de determinada(s) coluna(s) (semelhante a **group_by**).

– Exemplo:

dt[, .N, by = filial]

Obs.: A função **.N** conta número de registros.

- O **data.table** pode receber mais argumentos (como no **magrittr**, semelhante ao operador *pipe*):

– Sintaxe:

DT[i,j,by][...] ...

– Exemplo:

DT[c(1,7,9)][order(-valor_compra)]

9.3 Transformando **data.frame** em **data.table**

- Para transformar **data.frame** em **data.table** aplicamos a função **data.table()**.
- Exemplo:

```
# Biblioteca  
library(data.table)
```

```
# Transformando data.frame dados em data.table dt
dt <- data.table(dados)
```

9.4 data.table

9.4.1 Manipulando linhas

Table 13: 5 formas de manipulação de linhas no data.table

Comando	O que faz?
DT[condições sobre as colunas]	Seleciona as linhas de DT que satisfazem as condições.
DT[1:k]	Seleciona as linhas de 1 a k.
DT[order(j1,j2)]	Ordena os dados em ordem ascendente do vetor 1, seguido por vetor 2. Para ordem descendente use sinal de menos antes do nome do vetor. Ex.: DT[order(-j)]
unique(DT) ou unique(DT,by = colunas selecionadas)	Seleciona as linhas distintas (elimina as repetidas) considerando as colunas selecionadas.
na.omit(DT,cols = colunas selecionadas)	Elimina as linhas com valores faltantes, considerando as colunas selecionadas.

9.4.2 Manipulando colunas

Table 14: 8 formas de manipulação de colunas no data.table

Comando	O que faz?
DT[,j] ou DT[[j]]	Seleciona a coluna j e retorna um vetor.
DT[,list(j)] ou DT[,(j)]	Seleciona a coluna j e retorna um data.table
DT[,-c(j1,j2,...,jn)]	Exclui as colunas listadas j1, j2, ..., jn
DT[,(j1,j2,...,jn)]	Retorna as colunas listadas j1, j2, ..., jn
DT[,(nome_escolhido = função(j))]	Aplica a função especificada à coluna j e retorna um data.table.
DT[,(nome_1 = f(j1), nome_2 = f(j2), ..., nome_n = f(jn))]	Aplica várias funções a várias colunas e retorna um data.table.
DT[,novacol := vetor]	Adiciona uma nova coluna.
DT[,c('col1', 'col2', ..., 'coln') := c(vetor1, vetor2, ..., vetorn)]	Adiciona várias novas colunas.

9.4.3 Sumarizando dados

- Realiza operações para apuração de valores sobre linhas de um `data.table`.
- Argumentos de operações de sumariação de dados:

Table 15: Argumentos para operações em um DT aplicados a uma ou mais colunas

Comandos	O que faz?
<code>.N</code>	Conta o número de linhas.
<code>DT[.,N, by = c(j1,...,jn)]</code>	Conta o número de linhas de acordo com os agrupamentos das colunas j1, ..., jn
<code>DT[.,(f1(j1),...,fn(jn)), by = j]</code>	Aplica diversas funções nas colunas especificadas, de acordo com o agrupamento da coluna j em by.
<code>DT[.,(f1(j1),...,fn(jn)), keyby = .(j1,...,jn)]</code>	Aplica diversas funções nas colunas especificadas, de acordo com o agrupamento das colunas listadas j1, ..., jn em keyby.

9.4.4 Operando um subconjunto de dados

- O pacote possui um símbolo especial denotado por `.SD` para realizar operações em um subconjunto de dados do `data.table DT`, de acordo com um grupo definido por `by` (agrupa em torno de determinadas colunas, igual a `group_by`).

– `DT[, .SD]`

Subconjunto completo dos dados.

– `DT[, .SD, by = .(j)]`

É o subconjunto completo dos dados exeto pela coluna `j`, formando subconjuntos para cada grupo da coluna `j`.

– `DT[, .SD, by = .(j,k)]`

Podemos agrupar entorno de mais de uma coluna, definidas por `by`.

- É possível ainda definir (**selecionar**) as colunas do conjunto completo que deverão ser consideradas em `.SD` através do operador `.SDcols`. São as colunas que vão receber as funções.

```
DT[, lapply(.SD, mean), .SDcols = c("coluna_1", "coluna_2"), by = .(coluna_3)]
```

- `lapply`

– Aplica a função determinada no subconjunto.

Sintaxe: `lapply(.SD, função)`

– É comum que apareça dentro de `data.table` quando realizando operação de subconjuntos. É fundamental para as operações.

```
Ex.: DT[, lapply(.SD, mean), .SDcols = "coluna_1", by = .(coluna_2)]
```

– Podem ser aplicadas várias funções no subconjunto.

```
Ex.: DT[, c(lapply(.SD, mean), lapply(.SD, sum)), .SDcols = "coluna_1", by = .(coluna_2)]
```

- Exemplo:

```
DT[, c(lapply(.SD, mean), lapply(.SD, sum)), .SDcols = c("coluna_1", "coluna_2"), by = .(coluna_3, coluna_4)]
```

– Aplica as funções `média(mean)` e `soma(sum)` sobre as colunas selecionadas `coluna_1` e `coluna_2`.

– Agrupando os dados (`by`) entorno das colunas selecionadas `coluna_3` e `coluna_4`.

9.4.5 Modificando dados com set

- As funções **set** são para modificação de dados do **data.table**.
- São as principais funções **set**:

Table 16: Funções set para modificação de dados no formato data.table

Funções		Descrição	Fórmula
set		Modifica o valor da linha e coluna.	set(dt, i=1, j=3, value=999)
setnames		Modifica o nome da coluna.	setnames(dt, old='nome_coluna', new='novo_nome_coluna')
setorder	Reordena linhas de forma de forma decrescente ou crescente.		setorder(dt, -vendas, n_itens)
setcolorder	Reordena colunas.		setcolorder(dt, c(1,3,2))

10 CAP. 7 - GRÁFICOS PACOTE BÁSICO E PACOTE `ggplot2`

- Objetivo é obter o grafico ideal, com o qual se consiga visualizar os dados e analiza-los.
- Os principais passos:
 - Identificação dos tipos de variáveis.
 - Construção dos gráficos.
 - Ajustes.
 - Refinamento.

10.1 Gráficos com o pacote básico

- Principais funções de gráfico do pacote básico:

Table 17: Nome das principais funções para construção de gráficos do pacote base.

Função	Tipo de Gráfico
<code>barplot(x)</code>	Produz um gráfico de colunas do vetor x.
<code>boxplot(x)</code>	Produz o boxplot de x.
<code>coplot(y~x z)</code>	Produz um gráfico de dispersão entre x e y condicionado por z.
<code>curve(expressão)</code>	Produz um gráfico a partir da expressão de certa função de x.
<code>dotplot(x)</code>	Produz um gráfico de pontos.
<code>hist(x)</code>	Produz um histograma do vetor x.
<code>mosaicplot</code>	Produz um mosaico para tabelas de contigência.
<code>pairs(x)</code>	Produz uma grande de gráficos de dispersão entre variáveis quantitativas de uma tabela.
<code>pie(x)</code>	Produz um gráfico circular (pizza).
<code>plot(x)</code>	Produz um gráfico de dispersão entre x e y
<code>qqnorm(x)</code>	Plota os quantis de x usando como base a curva normal.
<code>stem(x)</code>	Produz um ramo e folha.
<code>stripchart</code>	Produz um gráfico de dispersão unidimensional.

- Principais argumentos das funções de gráfico do pacote básico:

Table 18: Nome dos argumentos para adicionar efeito em gráficos.

Função	Efeito no gráfico
adj=	Controla a formatação do texto (0 formatação à esquerda; 0.5 centralizada; 1 à direita).
main=','	Adiciona um título principal ao gráfico de acordo com texto entre aspas.
col=	Comando para colorir diversos itens do gráfico, pode ser valores como 1,2,..., ou por nome como 'red', 'blue', etc (consulte nomes com o comando colors() ou sistemas como rgb(), hsv(), gray() e rainbow()). Para cor das fontes use: col.lab, col.main, col.sub
border=	Especifica a cor da borda de uma coluna no gráfico.
font=	Controla o estilo da fonte de: 0-normal, 1-italíco, 2-negrito, 3-italíco e negrito.
cex=	Controla o tamanho da fonte de textos, o valor padrão é 1.(cex.axis, cex.lab, cex.main, cex.sub)
lty=	Especifica o tipo de linha (1-sólida, 2-tracejadas, etc).
lwd=	Especifica a espessura da linha (1, 2, ...).
pch=	Controla o tipo de símbolo (1 a 25 ou especificando entre aspas).
type=	Especifica o tipo de plotagem: 'p' (Pontos); 'l' (Linhas); 'b' (Pontos conectados por linhas); 'o' (idem a b com pontos sobrepostos à linhas); 'h' (Linhas verticais); 's' (degrau no qual o dado é representado no topo da linha vertical); 'S' (idem ao s porém o dado é representado na base da linha vertical).
xlim=(inicio,fim)	Controla os limites do eixo X.
ylim=(inicio,fim)	Controla os limites do eixo Y.
xlab=','	Adiciona rótulo para o eixo X de acordo com texto entre aspas.
ylab=','	Adiciona rótulo para o eixo Y de acordo com texto entre aspas.
las=','	Controla a orientação dos rótulos dos eixos. 0 - paralelo ao eixo; 1 - horizontal; 2 - perpendicular; 3 - vertical.
xaxt ou yaxt=''	Se xaxt = 'n', o eixo X é definido porém não é desenhado; Se yaxt = 'n', o eixo Y é definido porém não é desenhado.
text (x,y,'texto',cex,col)	Adiciona texto ao gráfico na coordenada (x,y) podendo ser diminuído o tamanho da fonte na proporção desejada em relação ao tamanho padrão 1 e com a cor especificada.
legend(x,y,legenda) locator	Adiciona uma legenda no ponto (x,y) com símbolos dados no campo legenda
(n,type='n',...)	Retorna as coordenadas correspondentes pedidas pelo usuário ao clicar (n vezes) no gráfico. Também desenha símbolos (type = 'p') ou linhas (type = 'l') Respeitando os parâmetros do gráfico. Por padrão type = 'n'.
segments(x0,y0,x1,y1)	Desenha segmentos de linha a partir do ponto (x0,y0) até (x1,y1).

- Observações sobre visualização:

Podemos usar o comando `par(mfrow = c(i,j))` que prepara uma janela gráfica para receber vários gráficos.

- Dois gráficos, lado a lado.

```
par(mfrow=c(1,2))
```

- Dois gráficos, um abaixo do outro.

```
par(mfrow=c(2,1))
```

- Quatro gráficos, sendo dois em cada linha.

```
par(mfrow=c(2,2))
```

- Um gráfico na janela gráfica.

Basta omitir o comando.

- Redefinir o número de linhas a partir das margens da janela gráfica (**default**).

```
par(mar=c(5,4,4,2))
```

Sendo na ordem: abaixo, esquerda, acima e direita.

Esses valores tem impacto no espaço dos títulos dos gráficos.

Mudar esses valores reajusta o gráfico.

- Define a medida das margens.

```
par(mai=c(x1,x2,x3,x4))
```

Sendo na ordem: abaixo, esquerda, acima e direita.

- Fecha a janela gráfica (*devices*).

```
dev.off()
```

10.1.1 Gráfico de barras (barplot)

- A função `barplot()` gera um gráfico de barras.

10.1.1.1 Pré-requisitos

- Necessita que os dados estejam preparados para gerar os gráficos, em formato *tabulado*.
- Para preparação dos dados é necessário o uso das funções dos pacotes `magrittr`, `dplyr` (ou `data.table`), e `tidyverse`.
- Uma coluna com os dados **númericos** (frequências e/ou valores).
- Uma coluna com os dados **string**, ou **factor**.

10.1.1.2 Preparação dos dados

- Organização dos dados das colunas, colocando uma coluna em função da outra. As principais funções nesse caso são:

– `order`

Retorna uma permutação que reorganiza seu primeiro argumento em ordem crescente ou decrescente, quebrando laços por argumentos adicionais.

```
x <- tabula_Estado$Estado[order(tabula_Estado$cheg_2012)]
```

– `sort`

Ordena um vetor em ordem crescente ou decrescente.

```
y <- sort(tabula_Estado$cheg_2012)/1000
```

- Definindo parâmetros para a janela gráfica (`par`):

– `mar`

Vetor numérico que oferece o número de linhas a partir das margens da janela gráfica.

No formato `c(inferior, esquerda, superior, direita)`.

```
mar = c(9,5,4,2)
```

– `mai`

Vetor numérico que oferece o tamanho da margem, especificado em polegadas.

No formato `c(inferior, esquerda, superior, direita)`.

```
mai = c(1.8,1,0.8,0.4)
```

– Exemplo:

```
par(mar = c(9,5,4,2),mai = c(1.8,1,0.8,0.4))
```

10.1.1.3 Plotagem gráfico de barras (barplot)

- Principais argumentos do gráfico de barras (`barplot()`):
 - **y**
Vetor do eixo Y.
 - **names.arg**
Vetor com os nomes das barras do eixo X.
 - **main**
Título principal do gráfico.
 - **cex.main**
Tamanho da fonte de textos (título do gráfico).
 - **xlab**
Rótulo do eixo X.
 - **ylab**
Rótulo do eixo Y.
 - **cex.names**
Tamanho da fonte de textos (nomes das barras do eixo X, vetor x).
 - **axisnames**
Inclui os nomes das categorias no eixo x.
 - **las**
Controla a orientação dos rótulos dos eixos.
 - **ylim**
Controla os limites do eixo Y.
 - **text**
Neste caso, adiciona valores de Y no topo de cada barra de **xbar** (variável com gráfico). Como consequência dessa função, é necessário colocar o gráfico dentro de uma variável (**xbar**), se não for desejado este artifício não é necessário colocar **barplot** dentro de uma variável.
 - Para plotar gráfico de barras na horizontal basta adicionar o argumento **horizon = TRUE**.

- Exemplo - Gráfico de barras (`barplot()`):

```
xbar = barplot(
  y, names.arg = x,
  main = "Titulo do gráfico.",
  cex.main = 1.5,
  xlab = "Rótulo do eixo X.",
  ylab = "Rótulo do eixo Y.",
  cex.names = 1,
  axisnames = T,
  las = 2,
  ylim = c(0,1.2*max(y))
)
text(xbar, y, label = round(y,2), pos = 3, cex = 0.8, col = "black")
```



Figure 4: Gráfico de barras - Vertical (`barplot()`).

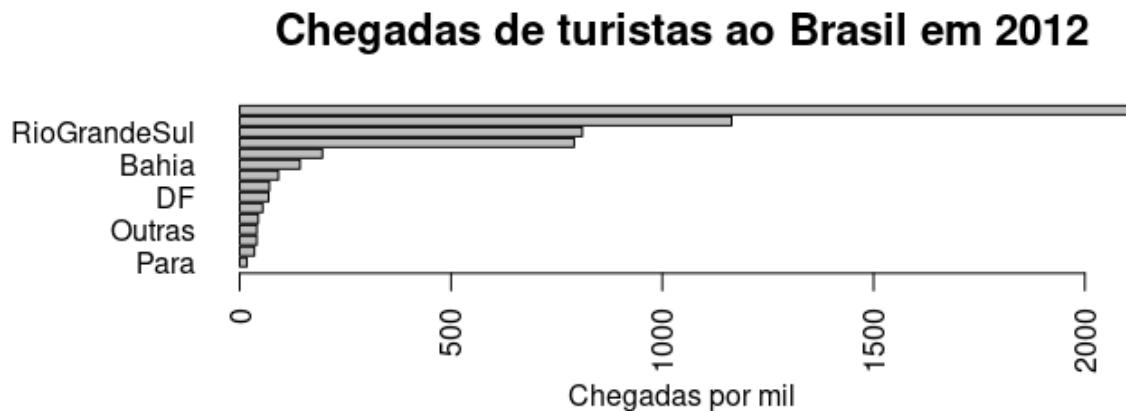


Figure 5: Gráfico de barras - Horizontal (`barplot(horizon = TRUE)`).

10.1.2 Gráfico circular/pizza (pie)

- A função `pie` gera um gráfico circular/pizza.
- Essa forma de visualização serve para analisar a frequencia de variáveis categóricas.

10.1.2.1 Pré-requisitos

- Utilize somente em casos de a variável possuir poucas categorias (em torno de cinco).
- Com quantidades diferentes entre si.
- Caso não se enquadrar nos pré-requisitos o ideal é optar por gráfico de barras.
- Os dados devem estar organizados em formato tabular.

10.1.2.2 Preparação dos dados

- Organização dos dados das colunas, colocando uma coluna em função da outra. As principais funções nesse caso são:

- `order`

Retorna uma permutação que reorganiza seu primeiro argumento em ordem crescente ou decrescente, quebrando laços por argumentos adicionais.

```
x <- tabula_Estado$Estado[order(tabula_Estado$cheg_2012)]
```

- `sort`

Ordena um vetor em ordem crescente ou decrescente.

```
y <- sort(tabula_Estado$cheg_2012)/1000
```

- Definindo parâmetros para a janela gráfica (`par`):

- `mar`

Vetor numérico que oferece o número de linhas a partir das margens da janela gráfica.
No formato `c(inferior, esquerda, superior, direita)`.

```
mar = c(9,5,4,2)
```

- `mai`

Vetor numérico que oferece o tamanho da margem, especificado em polegadas.
No formato `c(inferior, esquerda, superior, direita)`.

```
mai = c(1.8,1,0.8,0.4)
```

- Exemplo:

```
par(mar = c(9,5,4,2),mai = c(1.8,1,0.8,0.4))
```

10.1.2.3 Plotagem gráfico circular/pizza (pie)

- Principais argumentos:
 - **y**
O vetor do eixo Y, contendo os valores das categorias.
 - **main**
Título principal do gráfico.
 - **labels**
Rótulo com cada categoria do gráfico.
 - **cex.main**
Tamanho da fonte de textos (título do gráfico).
 - **cex**
Tamanho do texto dos rótulos (**labels**).
 - **col**
Comando para colorir diversos itens do gráfico, pode ser valores como 1,2,..., ou por nome como ‘red’, ‘blue’, etc.
Neste caso para colorir os pedaços do gráfico circular/pizza.
 - **text**
Adiciona texto ao final do gráfico, neste caso o texto é a fonte usada para elaboração do gráfico.
Dentro do **text** apresenta parâmetros para localizar o texto na janela gráfica, o texto e tamanho da fonte.
 - **rotulos**
Variável que recebe o texto, a partir da função **paste** de concatenação de texto e valores, com os rótulos de cada categoria, contendo nome de cada categoria (**x**) e porcentagem(**porc**).

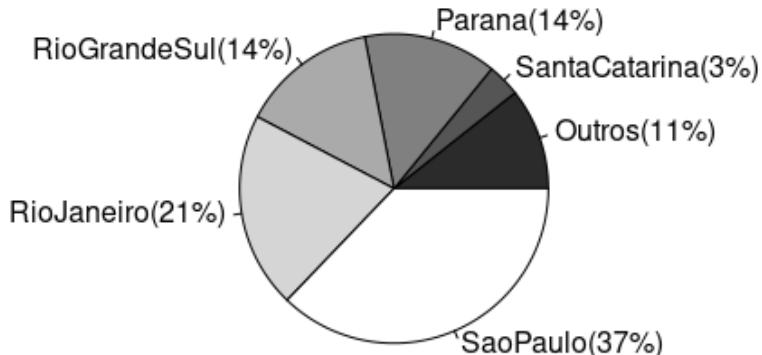
- Exemplo - Gráfico circular “pizza” (`pie()`):

```
#Juntando categorias com baixa proporção na categoria outros.
y <- c(sum(y[1:10]), y[11:15])
x <- c("Outros", as.character(x[11:15]))

#Gráfico circular/pizza
#Variável rótulo
porc = 100*round(y/sum(y),2) #calcula a %
rotulos = paste(x,"(",porc,"%)",sep = "") #texto para rotulo

#plotando gráfico pizza
par(mar = c(4,0,2,0), mai = c(0.8,0,0.4,0))
pie(
  y,
  main = "Título do gráfico",
  labels = rotulos,
  cex.main = 2,
  cex = 1.5,
  col = gray(1:length(x)/length(x))
)
text(0, -1, "Fonte: Elaborado com pacote graphics version 3.6.1 do R.", cex = 1)
```

Proporção de chegadas de turistas por Estado em 2012



Fonte: Elaborado com pacote graphics version 3.6.1 do R.

Figure 6: Gráfico circular “pizza” (`pie()`).

10.1.3 Gráfico de linhas (plot lines)

- O gráfico de linhas (`plot`) é utilizado para vizualizar uma ou mais variáveis numéricas que podem ser plotadas ao longo do tempo (não somente tempo) no eixo x.
- Podemos adicionar mais linhas (variáveis) no gráfico através do comando `lines`.

10.1.3.1 Pré-requisitos

- Os dados devem estar organizados em formato tabular.

```
dados_SP <- dados %>%
  select(Mes, Estado, cheg_2012, cheg_2013, cheg_2014, cheg_2015) %>%
  filter(Estado == "SaoPaulo")
dados_SP
```

- Separar as colunas em vetores.

```
# Definindo os valores dos eixos
x <- dados_SP$Mes
y1 <- dados_SP$cheg_2012/1000
y2 <- dados_SP$cheg_2013/1000
y3 <- dados_SP$cheg_2014/1000
y4 <- dados_SP$cheg_2015/1000
```

10.1.3.2 Preparação dos dados

- Definir os limites do eixo y:

– **li**

Variável que recebe o limite inferior do eixo y, com base no menor valor do eixo y dos vetores de cada coluna.

– **ls**

Variável que recebe o limite superior do eixo y, com base no maior valor do eixo y dos vetores de cada coluna.

– Exemplo:

```
# Definindo os limites do eixo y
li <- min(y1,y2,y3,y4)
ls <- max(y1,y2,y3,y4)
```

10.1.3.3 Plotagem gráfico plot

- Principais argumentos:
 - **x**
Variável do tipo vetor que recebe valores de tempo.
 - **y1, y2, y3 e y4**
Variáveis do tipo vetor em que cada uma representa os valores de uma linha.
 - **lty**
Especifica o tipo de linha.
 - **lw**
Especifica a espessura da linha.
 - **type**
Especifica o tipo de plotagem, ‘b’ (pontos conectados por linhas).
 - **ylim**
Controla os limites do eixo Y.
 - **xlab**
Rótulo do eixo X.
Quando o gráfico apresentar diversas linhas, o rótulo deve ser inicializado e zerado dentro do **plot** e adicionado seu valor definitivo na função **title**.
 - **ylab**
Rótulo do eixo Y.
Quando o gráfico apresentar diversas linhas, o rótulo deve ser inicializado e zerado dentro do **plot** e adicionado seu valor definitivo na função **title**.
 - **lines**
Adiciona novas linhas ao gráfico de linhas, cada uma com suas particularidades.
 - **title**
Função que adiciona títulos, rótulos e texto ao gráfico de linhas.
 - **main**
Título principal do gráfico.
 - **sub**
Adiciona texto ao final do gráfico.
 - **cex.sub**
Tamanho da fonte do texto.
 - **col**
Comando para colorir diversos itens do gráfico, pode ser valores como 1,2,..., ou por nome como ‘red’, ‘blue’, etc.
Neste caso para colorir as linhas do gráfico de linhas.
 - **legend**
Adiciona um quadro de legenda a janela gráfica.

- Exemplo - Gráfico de linhas (`plot()` `lines()`):

```
plot(x, y1, lty = 1, lwd = 1, type = "b", ylim = c(0.8*li,ls*1.2),xlab ="",
      ylab = "", col = "red")
lines(x, y2, lty = 2, lwd = 1, type = "b", col = "yellow")
lines(x, y3, lty = 3, lwd = 2, type = "b", col = "blue")
lines(x, y4, lty = 4, lwd = 1, type = "b", col = "green")
title(main = "Chegada de turistas em São Paulo",
      xlab ="Mês",
      ylab = "Chegadas por mil",
      sub = "Fonte: Elaborado com pacote graphics version 3.6.1 do R.",
      cex.sub = 0.8)
legend(9,400,c("2012","2013","2014","2015"), col = c("red","yellow","blue","green"),
      lty = 1:4, cex = 0.5)
```

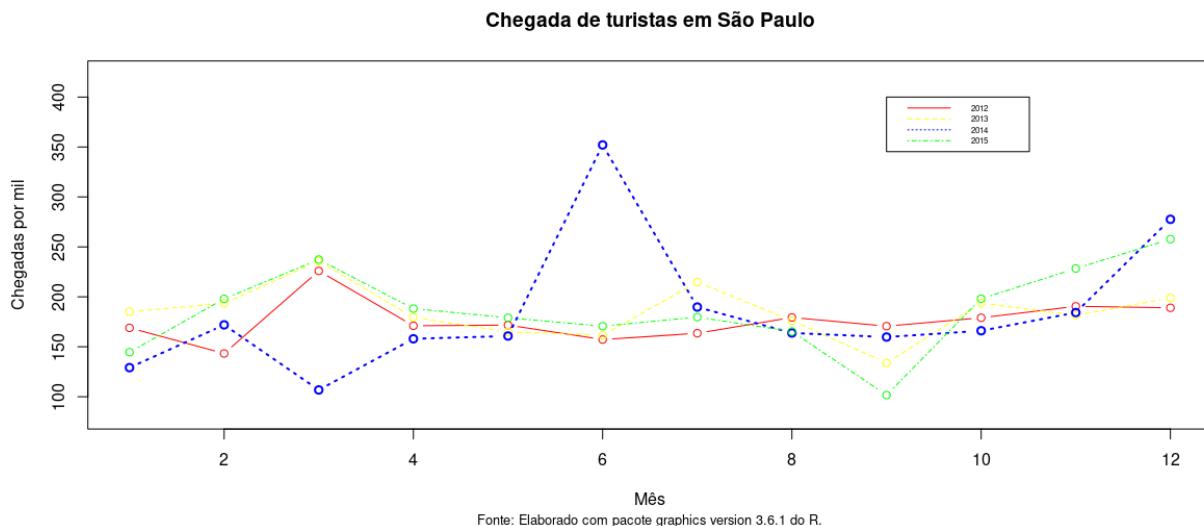


Figure 7: Gráfico de linhas (`plot()` `lines()`).

10.1.3.4 Comparando séries de gráficos de linhas

10.1.3.4.1 O que é comparar séries

- O intuito é plotar mais de um gráfico na mesma janela gráfica, para facilitar a comparação das diferentes séries de dados.
- Sendo possível preparar a janela gráfica assim como uma matriz, para receber gráficos em linha, ou em coluna, ou em linhas e colunas ao mesmo tempo.

10.1.3.4.2 Preparação da janela gráfica

- Existem duas formas de preparar a janela gráfica para receber os gráficos:

– `par(mfrow = c(nº_linhas, nº_colunas))`
Prepara a janela gráfica como uma matriz, sendo definido o número de linhas e colunas a receber os gráficos.
Necessita ajeitar os parâmetros de margem da janela gráfica(`mai` e `mar`), por conta disto não é o mais aconselhado de usar.

– `layout(matrix(nº_linhas,nº_colunas), nº_linhas, nº_colunas)`
Prepara a janela gráfica como uma matriz, sendo definido o número de linhas e colunas a receber os gráficos.
Mais recomendado a utilização.

10.1.3.4.3 Plotagem de gráficos de linhas comparando séries

```
#script para dois graficos de linha
#preparando a janela grafica para receber dois graficos

#metodo 1 - não recomendado
#par(mfrow = c(2,1)) #necessidade de configurar margens
#par(mar = c(6,4,1,1), mai = c(0.9,0.9,0.3,0.1))

#Metodo 2 - Recomendado
layout(matrix(c(1,2), 1, 2)) #tende a funcionar melhor que par(mfrow())
#numero de linhas, numero de colunas

#grafico 1

#define os limites do eixo y
li1 <- min(y1,y2,y3,y4)
ls1 <- max(y1,y2,y3,y4)

#script para o grafico de linha
plot(x1, y1, lty = 1, lwd = 1, type = "b", ylim = c(0.8*li1,ls1*1.2),xlab ="",
      ylab = "") #xlab e ylab vazios some com os rotulos x e y, para que possa colocar a no titulo (title)
lines(x1, y2, lty = 2, lwd = 1, type = "b") #acrescenta y2
lines(x1, y3, lty = 3, lwd = 2, type = "b") #acrescenta y3
lines(x1, y4, lty = 4, lwd = 1, type = "b") #acrescenta y4
#lty = especifica o tipo de linha
#lwd = especifica a espessura da linha
#type = especifica o tipo de plotagem, 'b' (pontos conectados por linhas)
```

```

title(main = "Chegada de turistas em São Paulo",
      xlab ="Mês",
      ylab = "Chegadas por mil")
legend(9,400,c("2012","2013","2014","2015"), lty = 1:4, cex = 0.5) #os dois primeiros valores são a pos
#grafico 2

#define os limites do eixo y
li2 <- min(z1,z2,z3,z4)
ls2 <- max(z1,z2,z3,z4)

#script para o grafico de linha
plot(x2, z1, lty = 1, lwd = 1, type = "b", ylim = c(0.8*li2,ls2*1.2), xlab ="",
      ylab = "") #xlab e ylab vazios some com os rotulos x e y, para que possa colocar a no titulo (
lines(x2, z2, lty = 2, lwd = 1, type = "b") #acrescenta y2
lines(x2, z3, lty = 3, lwd = 2, type = "b") #acrescenta y3
lines(x2, z4, lty = 4, lwd = 1, type = "b") #acrescenta y4
#lty = especifica o tipo de linha
#lwd = especifica a espessura da linha
#type = especifica o tipo de plotagem, 'b' (pontos conectados por linhas)
title(main = "Chegada de turistas em Rio de Janeiro",
      xlab ="Mês",
      ylab = "Chegadas por mil",
      sub = "Fonte: Elaborado com pacote graphics version 3.6.1 do R.", cex.sub = 0.8)
legend(9,300,c("2012","2013","2014","2015"), lty = 1:4, cex = 0.5) #os dois primeiros valores são a pos

```

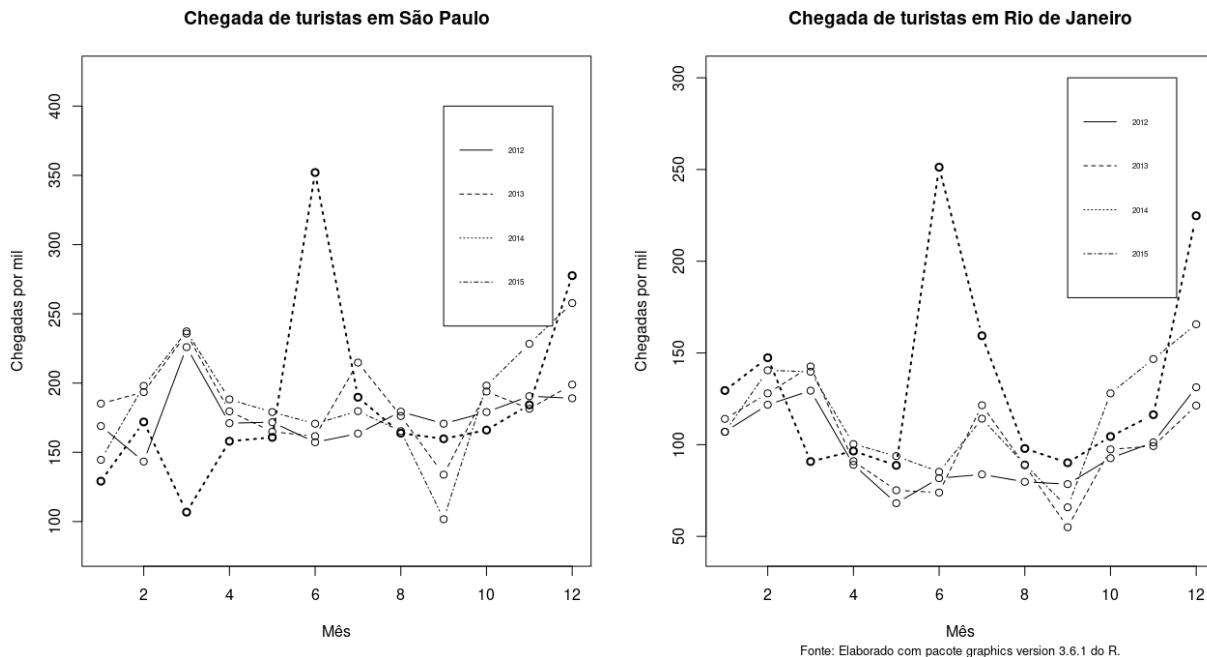


Figure 8: Gráfico de linha comparando séries

10.1.4 Gráfico de dispersão (plot abline)

- O gráfico de dispersão é usado para observar a relação entre duas variáveis quantitativas (que podem ser contadas).
- O pesquisar a princípio busca uma relação linear, logo a visão do gráfico deve ser acompanhado do **coeficiente de correlação linear**, que mede matematicamente a intensidade dessa relação.
- **Coeficiente de correlação linear:**
 - A relação entre duas variáveis é chamada **regressão linear simples**.
 - O coeficiente de **correlação linear** é uma medida que varia entre 1 e -1.
 - Espera-se encontrar valores próximos de -1 e 1 no caso de presença de relacionamento linear.
 - no caso de 1.
O coeficiente é positivo ou dito crescente, uma variável cresce acompanhando o crescimento da outra.
 - no caso de -1.
O coeficiente é negativo ou dito decrescente, uma variável cresce com o decaimento da outra.

10.1.4.1 Pré-requisitos

- Os dados devem estar em formato tabular, ou as variáveis em formato de vetor.

10.1.4.2 Preparação dos dados

- Coeficiente de Correlação linear:
 - Teoria
 - * O coeficiente de correlação tem o objetivo de entender como uma variável se comporta num cenário onde a outra variável variando. E se existe alguma relação entre a variabilidade de ambas as variáveis.
 - * Os coeficientes variam de 1 a -1. Quanto mais próximo dos extremos, mais forte é a relação entre as variáveis. Quanto mais próximo do centro 0, menor é a relação entre as variáveis. Em 0 não existe relação entre as variáveis.
 - * A correlação próximo do valor 1, significa que a relação é positiva, ou seja, a reta de regressão é ascendente. Quando uma variável aumenta a outra aumenta também.
 - * A correlação próximo do valor -1, significa que a relação é Negativa, ou seja, a reta de regressão é descendente. Quando uma variável diminui a outra aumenta.
 - Cálculo de correlação linear:

$$cor_{x,y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Onde,

n é o número de registros/linhas.

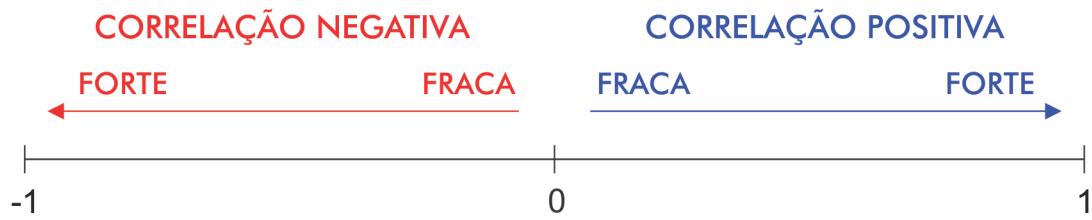


Figure 9: Correlações fortes e fracas

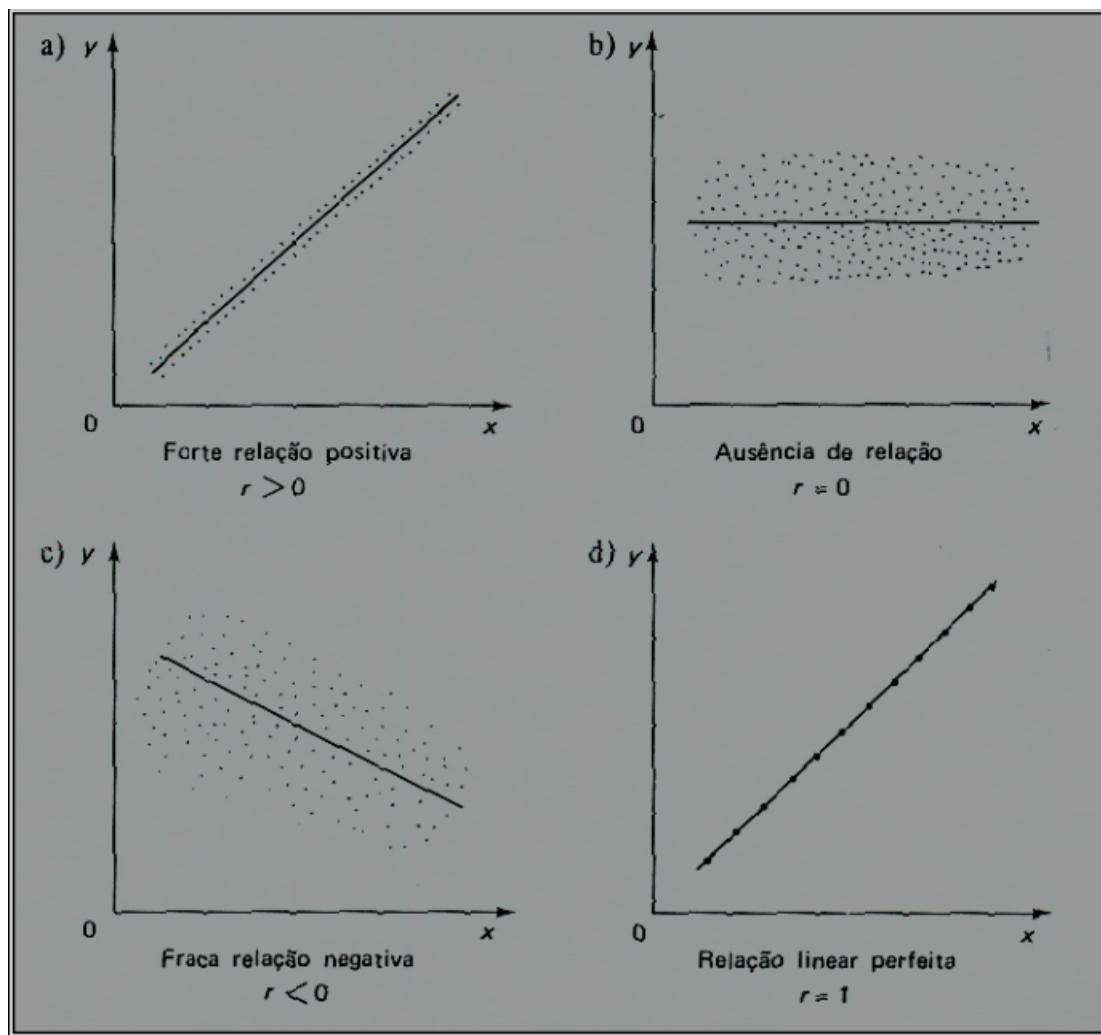


Figure 10: Tipos de Correlação

x_i é o vetor x.

y_i é o vetor y.

xy é x vezes y.

Uma forma rápida e simples de resolver o cálculo é preencher a tabela de correlação linear com as informações:

	x	y	xy	x^2	y^2
n					
	Σ				

Figure 11: Tabela de correlação linear

– `cor(x,y)`

Função do R que calcula a correlação linear das variáveis vetor x e y.

- Coeficiente de reta de regressão:

Tenta traçar uma reta que melhor aproxime todos os pontos dispersos.

$$y = A + Bx$$

Onde,

A é o intercepto

B é o coeficiente angular.

– Coeficiente angular:

$$B = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

– Intercepto:

$$A = \frac{\sum y - B \sum x}{n}$$

– `lm(y ~ x)$coef`

Esta função do R retorna os coeficientes da reta de regressão (**intercepto** e **coeficiente angular**).

A parte da função `$coef` apenas retorna de maneira mais direta os coeficientes separados, assim deixando claro em cada coluna o que é **intercepto** e o que é **coeficiente angular**.

10.1.4.3 Plotagem gráfico plot abline

- Funções usadas:
 - **paste()**
Concatena as strings e valores.
 - **expression()**
Salva numa variável o desenho de texto no formato **expression**, ou seja, uma expressão matemática, uma equação.
As Expressões (**expressions**) podem ser usadas como título, subtítulo e rótulos de eixos.
 - **eval()**
Avalia, se possível resolve, uma **expression**.
- Principais argumentos do gráfico de dispersão:
 - **x**
Vetor com variável quantitativa x.
 - **y**
Vetor com variável quantitativa y.
 - **main**
Título principal do gráfico.
 - **xlab**
Rótulo do eixo x.
 - **ylab**
Rótulo do eixo y.
 - **text**
Adiciona texto ao gráfico.
Neste caso adiciona, pela posição do texto inserido, o nome a reta e a equação/expressão da reta.
 - **abline**
Adiciona a reta tracejada.
 - **lty**
Especifica o tipo de linha.
 - **lwd**
Especifica a espessura da linha.

- Exemplo - Gráfico de dispersão “plot abline”(plot() abline()):

```
#Observando a correlação entre as chegadas de São Paulo e Rio de Janeiro
x <- dados_RJ$cheg_2014/1000
y <- dados_SP$cheg_2014/1000
x
y

#Obtendo a correlação
cor(x,y)

#Obtendo os coeficientes da reta de regressão
lm(y ~ x)$coef
#lm = é usado para ajustar modelos lineares. Ele pode ser usado para realizar regressão, análise de variâncias, etc.
#coef = é uma função genérica que extrai coeficientes de modelo de objetos retornados por funções de modelagem.

#Gráfico de dispersão
plot(x, y,
      main = paste("Gráfico de Dispersão entre as chegadas de turistas de 2014",
                   "\n", "São Paulo x Rio de Janeiro"),
      xlab = "Chegadas no Rio de Janeiro/1000",
      ylab = "Chegadas em São Paulo/1000"
)
abline(lm(y ~ x), lty = 2, lwd = 2) #adiciona a reta tracejada
#lty = especifica o tipo de linha
#lwd = especifica a espessura da linha
text(130,230,"reta de regressão") #adiciona texto na posição (130,230)
text(130,210,paste("y = ",eval(expression(round(lm(y ~ x)$coef[[2]],2))),",
[[j]] seleciona a coluna j.
" x + ",eval(expression(round(lm(y ~ x)$coef[[1]],2))))) #adiciona equação na posição (130,210)
#paste = Concatenar vetores após a conversão em caractere.
#eval = Avalie uma expressão R em um ambiente especificado.
#expression = Cria ou testa objetos do modo "expressão".
```

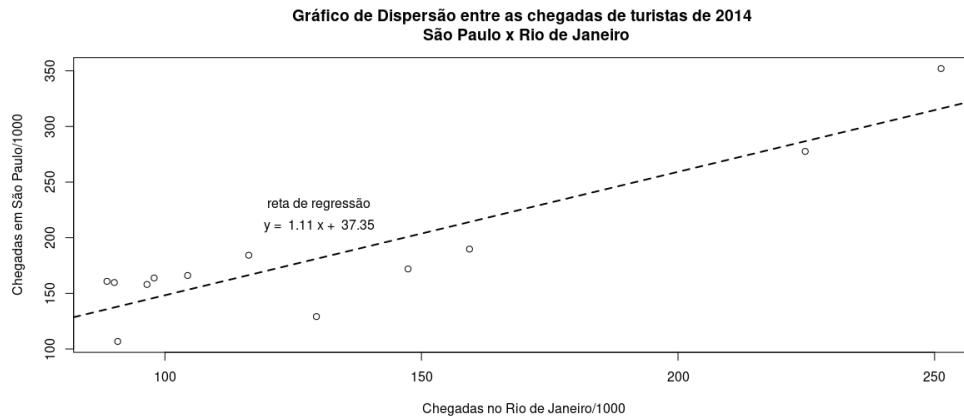


Figure 12: Gráfico de dispersão “plot abline” (plot() abline()).

10.1.5 Diagrama de caixa (boxplot)

- O **Diagrama de caixa** serve para compreensão da forma e amplitude dos dados.
- É importante para fazer o **diagrama de caixa** conhecer a fórmula das **separatrizes**.
- O **diagrama de caixa** usa em sua construção os conceitos de **quartis (Q1, Q2, e Q3)**.

10.1.5.1 Separatrizes

- Quartis:
 - Q1 (25%)
 - Q2 (50% ou mediana)
 - Q3 (75%)
- Tabela de distribuição de frequências:

Salário	N. º de prof. (<i>fi</i>)	fac
01 -- 03	20	20
03 -- 05	40	60
05 -- 07	60	120
07 -- 09	30	150
09 -- 11	10	160
Total	160	-----

Figure 13: Exemplo de tabela de distribuição de frequências

- Como achar o intervalo de classe que corresponde a separatriz calculada.
Ex.: Se Q1 (25%), achar na tabela de classes e frequências, na coluna frequência acumulada, a classe que contém o valor que corresponde a 25% da frequência acumulada total.
Esse intervalo de classe será a classe selecionada para aplicação da fórmula.

- Fórmula da separatriz:

$$P_k = Li + \frac{k \cdot \sum f_i - F_{anterior}}{f_{intervalo}} \cdot h$$

onde,

P_k é o percentil (separatriz),

Li é o limite inferior do intervalo de classe selecionada,

k é o número em fração do percentil,

f_i é a frequência,

$\sum f_i$ é a frequência acumulada total,

$F_{anterior}$ é a frequencia acumulada do intervalo de classe anterior (ao selecionado) do qual se esta calculando,

$f_{intervalo}$ é a frequênciia do intervalo de classe selecionada,

h é a amplitude de classe ($Ls - Li$).

10.1.5.2 boxplot

- Montando a box:
A box contém como limite superior Q_3 , limite inferior Q_1 e linha interna a mediana (Q_2).
- Intervalo interquartil:

$$IQR = Q_3 - Q_1$$

- Limites:

– Máximo

$$L_{\max} = Q_3 + 1,5 \cdot (IQR)$$

– Mínimo

$$L_{\min} = Q_1 - 1,5 \cdot (IQR)$$

- Valores discrepantes (**Outliers**):

– Possíveis erros (arredondamento ou observação).
– Alguma condição especial que deve ser observada separadamente.

- Exemplo explicativo de boxplot:

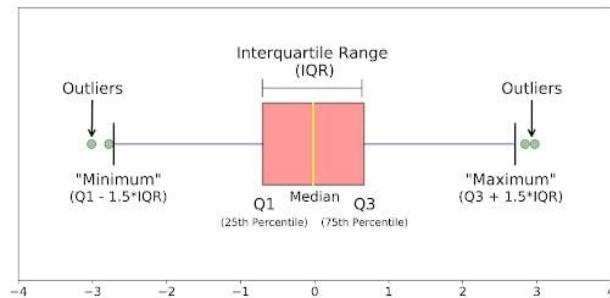


Figure 14: Exemplo explicativo de boxplot

10.1.5.3 Pré-requisitos

- Os dados devem estar em formato tabular.

10.1.5.4 Preparação dos dados

- A variável em formato de vetor.

Ex.: `x <- turismo$cheg_2012/1000`

10.1.5.5 Plotagem gráfico boxplot

- Principais argumentos do gráfico de dispersão:

– **x**

Variável em formato de vetor.

– **main**

Título principal do gráfico.

– **xlab**

Rótulo do eixo x.

– **ylab**

Rótulo do eixo y.

- Exemplo - Diagrama de caixa “boxplot” (`boxplot()`):

```
#Variável x
x <- turismo$cheg_2012/1000

#Plotando o diagrama de caixa - boxplot
boxplot(x,
         main ="Boxplot das chegadas de Turistas ao Brasil em 2012",
         xlab ="Ano de 2012",
         ylab ="Chegadas de turistas em 2012 por mil")
```

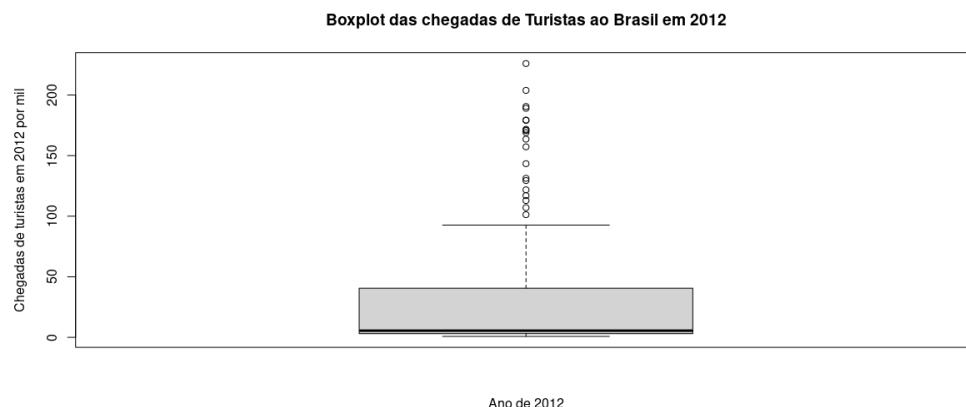


Figure 15: Gráfico de caixa (`boxplot()`)

10.1.6 Histograma (hist)

- Histograma é um tipo de gráficos de barras.
- É usado para variáveis quantitativas continuas.
- Este tipo de gráfico é muito usado para observar:

– **Distribuição de frequências**

– **Simetria**

– **Desvio**

Presença de valores discrepantes (Outliers).

– **Amplitude da variável**

- A diferença entre gráficos de barras e histograma:

– **Gráfico de barras**

É aplicado a variáveis categóricas, apenas um eixo representando variável numérica e o outro eixo representando um variável categórica.

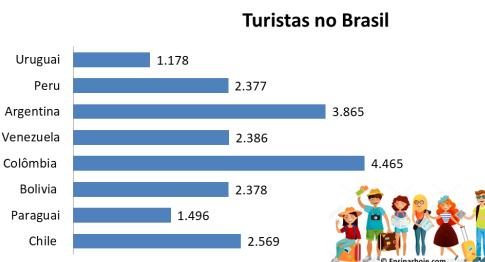


Figure 16: Exemplo gráfico de barras

– **Histograma**

É aplicado a variáveis numéricas e possui dois eixos numéricos (x representando a variável e y representando a frequência da variável).

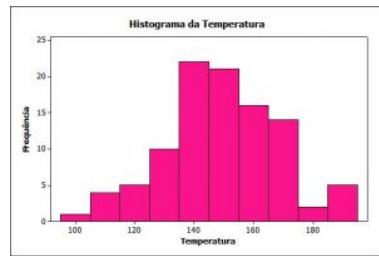


Figure 17: Exemplo histograma

10.1.6.1 Pré-requisitos

- Os dados devem estar em formato tabular.

10.1.6.2 Preparação dos dados

- A variável em formato de vetor.

Ex.: `x <- dados$cheg_2012/1000`

- A frequência é calculada automaticamente pela função `hist` (histograma), basta informar a função se ela deve calcular a frequência absoluta (**T**) ou a frequência relativa (**F**).

Ex.: `hist(... , freq = T | F, ...)`

10.1.6.3 Plotagem histograma

- Principais argumentos da função histograma (`hist()`):

– **x**

Variável em formato de vetor.

– **freq**

É a frequência.

Caso queira a frequência absoluta = **T**.

Caso queira a frequência relativa = **F**.

– **col**

Comando para colorir diversos itens do gráfico, pode ser valores como 1,2,..., ou por nome como ‘red’, ‘blue’, etc.

– **main**

Título principal do gráfico.

– **xlab**

Rótulo do eixo x.

– **ylab**

Rótulo do eixo y.

– **sub**

Adiciona texto ao final do gráfico.

- Exemplo - Histograma:

```
#Compreendendo a distribuição frequênciade chegadas de turistas do Brasil em
#2012
x <- dados$cheg_2012/1000

#histograma
hist(x,
      freq = T, #se T fornece a frequencia absoluta, se F fornece a frequencia relativa
      main = "Histograma das chegadas de turistas ao Brasil em 2012",
      xlab = "Chegadas de turistas em 2012 por mil",
      ylab = "Frequencia Absoluta das chegadas",
      sub = "Fonte: elaboração propria") #legenda
```

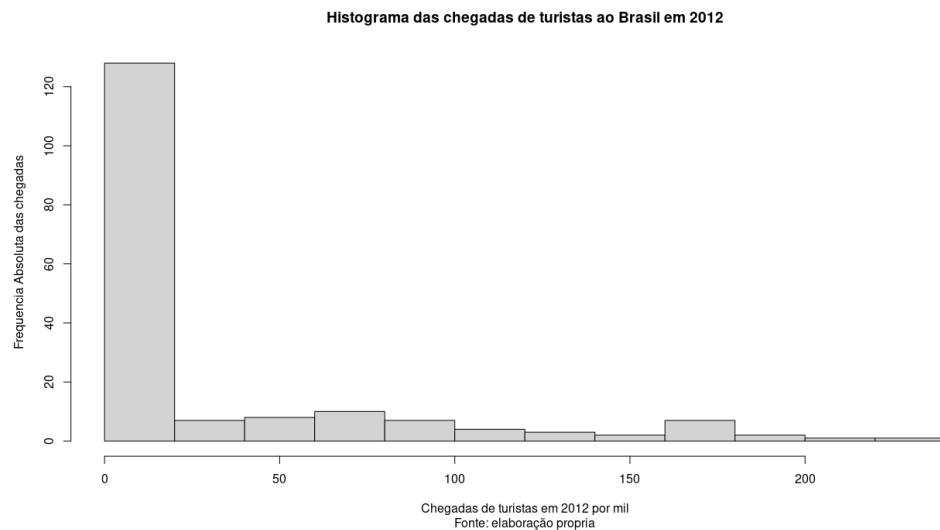


Figure 18: Histograma (hist())

10.2 Pacote ggplot2

O pacote `ggplot2` constroi diversos tipos de graficos a partir da mesma estrutura de componentes:

- `data`: referente ao banco de dados.
- `geom_forma`: um rol de tipos possiveis de representação dos dados.
- `coord_system`: referente ao sistema de coordenadas, que podem ser cartesianas, polares e projeção de mapas.

10.2.1 O que precisa para fazer o gráfico?

A. Um nome de objeto para guardar o grafico (uma variavel).

B. A base de dados que será utilizada para a plotagem.

`ggplot(data=nome_da_base)`

C. Descrever como as variaveis serão utilizadas na plotagem:

`aes(x=..., y=..., ...)`

D. Especificar o tipo de gráfico:

`geom_forma(...)`

E. Utilizar o operador “+” para adicionar camadas (*layers*) ao objeto `ggplot` criado.

F. Pacotes auxiliares como `ggthemes` e `grid`, dentre outros.

10.2.2 Quais formatos podemos utilizar no ggplot2 (*geom_forma*)?

Table 19: Nome das principais formas geométricas para construção de gráficos do pacote ggplot2

Forma	Tipo de gráfico
<code>geom_area</code> ou <code>geom_ribbon</code>	Produz um grafico para visualizar área sob a curva ou entre curvas.
<code>geom_bar</code> ou <code>geom_col</code>	Produz um grafico de colunas do vetor x.
<code>geom_bar+coord_polar</code>	Produz um grafico circular (Pizza).
<code>geom_boxplot</code>	Produz o boxplot de x.
<code>geom_curve</code>	Produz um grafico em curva.
<code>geom_density</code>	Produz um grafico da densidade de x.
<code>geom_dotplot</code>	Produz um grafico de pontos.
<code>geom_histogram</code>	Produz um histograma do vetor x.
<code>geom_line</code> , <code>geom_abline</code> , <code>geom_hline</code> , <code>geom_vline</code>	Produz um grafico de linhas
<code>geom_point</code>	Produz um grafico de dispersão entre x e y.
<code>geom_qq</code> ou <code>geom_qq_line</code>	plota os quantis de x usando como base a curva normal.
<code>geom_tile</code> , <code>geom_rect</code> ou <code>geom_raster</code>	Produz uma grade de retangulos.
<code>geom_violin</code>	Produz um grafico em forma de violino.

10.2.3 Nome dos argumentos para adicionar efeito em gráficos do pacote `ggplot2`

Table 20: Nome dos argumentos para adicionar efeito em gráficos do pacote `ggplot2`.

Funções	Efeitos no gráfico
<code>autoplot</code>	Produz um grafico apropriado para o tipo de variavel.
<code>coord_cartesian</code>	Coordenada cartesiana.
<code>coord_fixed</code>	Coordenada cartesiana com razão entre eixo x e y fixada.
<code>coord_flip</code>	Inverte a posição dos eixos x e y.
<code>coord_polar</code>	Coordenada polar.
<code>geom_blank</code>	Janela em branco.
<code>geom_jitter</code>	Produz um efeito jitter.
<code>geom_smooth</code>	Produz uma curva suavizada.
<code>geom_text</code>	Aplica texto a janela grafica.
<code>scale_fill_(=brewer ou grey ou gradient)</code>	Define a escala de cores.
<code>scale_*_continuos</code>	Define parametros para o eixo x ou y continuos.
<code>scale_*_discrete</code>	Define parametros para o eixo x ou y discreto.
<code>scale_*_manual</code>	Define parametros para os eixos manualmente.

10.2.4 Definindo um tema para o grafico `ggplot`

- `theme_gray`
Fundo cinza e linhas grandes brancas.
- `theme_bw`
O classico preto e branco. Otimo para projetor.
- `theme_linedraw`
Linhas pretas de varias larguras num fundo branco. semelhante ao `theme_bw`.
- `theme_light`
Semelhante ao `theme_linedraw`, porem com as linhas mais cinza claro, para dar atenção aos dados.
- `theme_dark`
Versão escura do `theme_light`, com o fundo escuro, util para criar linhas finas coloridas.
- `theme_minimal`
Um tema minimalistico sem anotações de fundo.
- `theme_classic`
Tema classico, com linhas do eixo x e y, sem linhas de grade.
- `theme_void`
Um tema completamente vazio.

10.2.5 Pacote ggthemes

Table 21: Temas do pacote ggthemes

Tema	Semelhanças
theme_base	Tema do pacote básico do R
theme_calc	Semelhante aos gráficos produzidos pelo Calc do LibreOffice B
theme_economist	Semelhante ao The Economist
theme_economist_white	Semelhante ao The Economist com fundo branco
theme_excel	Semelhante aos gráficos produzidos pelo Excel
theme_few	Baseado nas regras de Stephen Few sobre regras práticas para o uso de cores nos gráficos
theme_fivethirtyeight	Baseado nos gráficos do site fivethirtyeight.com
theme.foundation	Tema de fundação, para produzir novos temas
theme_gdocs	Semelhante aos gráficos do Google Docs
theme_hc	Baseado em Highcharts JS
theme_igray	Inverte o tema gray
theme_map	Limpa o tema para incluir mapas
theme_pander	Baseado no pacote pander
theme_par	Baseado nos parâmetros definidos em par() do pacote base
theme_solarized	Baseado na paleta Solarized
theme_solarized_2	Baseado na paleta Solarized
theme_solid	Elimina todas as linhas e textos, mantendo somente os objetos geométricos
theme_stata	Semelhante aos gráficos do Stata
theme_tufte	Baseado no designer de Edward Tufte
theme_wsj	Semelhante aos gráficos do Wall Street Journal

Exemplo:

```
#Bibliotecas
library(ggthemes)

#Plotando gráficos
f <- ggplot(dados,aes(cheg_2012/1000,cheg_2013/1000)) +
  geom_blank() +
  labs(x="",y="")

#Tema
p1 <- f +
  theme_gdocs(base_size = 18) +
  ggtitle("theme_gdocs")

p1
```

10.2.6 Inserindo títulos, subtítulos e rótulos aos eixos de um ggplot

- Existem duas formas de inserir textos no gráfico no **ggplot2**.

10.2.6.1 Primeira forma

- Podemos adicionar texto ao gráfico **ggplot2** através do comando **labs()** e seus parâmetros:

- **title**
Adicionar um título ao gráfico **ggplot**.
- **x**
Adiciona um rótulo ao eixo x no gráfico **ggplot**.
- **y**
Adiciona um rótulo ao eixo y no gráfico **ggplot**.
- **subtitle**
Adicionar um subtítulo ao gráfico **ggplot**.
- **caption**
Adiciona texto ao final do gráfico **ggplot**.

- Exemplo:

```
#Plotando gráfico
p <- ggplot(data = Turismo, aes(x=cheg_2012/1000,y=cheg_2013/1000)) #Salva gráfico em variável

#Aplicando elementos de texto na forma janela em branco
p +
  geom_blank() + #Produz efeito janela em branco
  labs(title = "Título", #Adiciona texto ao gráfico
       x = "Eixo x", #Adiciona rótulo ao eixo x
       y = "Eixo y", #Adiciona rótulo ao eixo y
       subtitle = "Subtítulo", #Adiciona subtítulo ao gráfico
       caption = "Elaborado por ...") + #Adiciona texto ao final do gráfico
  theme_bw(base_size = 18)
```

10.2.6.2 Segunda forma

- Podemos adicionar texto ao **ggplot2** através dos comandos:

- **ggtitle("", subtitle = "")**
O comando **ggtitle** adiciona título ao gráfico **ggplot**.
Podemos adicionar o parâmetro **subtitle** para adicionar um subtítulo ao gráfico **ggplot**.
- **xlab("")**
Adiciona um rótulo ao eixo x no gráfico **ggplot**.
- **ylab("")**
Adiciona um rótulo ao eixo y no gráfico **ggplot**.
- **labs(caption = "")**
O comando **labs()** acompanhado do parâmetro **caption**, adiciona texto ao final do gráfico **ggplot**.

- Exemplo:

```
#Plotando gráfico
p <- ggplot(data = Turismo, aes(x=cheg_2012/1000,y=cheg_2013/1000)) #Salva gráfico em variável

#Aplicando elementos de texto na forma janela em branco
p +
  geom_blank() + #Produz efeito janela em branco
  ggtitle("Título",subtitle = "Subtítulo") + #Adiciona título e subtítulo ao gráfico
  xlab("Eixo x") + #Adiciona rótulo ao eixo x
  ylab("Eixo y") + #Adiciona rótulo ao eixo y
  labs(caption = "Elaborado por ...") + #Adiciona texto ao final do gráfico
  theme_bw(base_size = 18)
```

10.2.7 Escalas no ggplot2

- Podemos definir a escala dos eixos utilizando uma camada específica para esse fim:
 - Variáveis Discretas
`scale_x_discrete()` ou `scale_y_discrete()`
 - Variáveis Continuas
`scale_x_continuous()` ou `scale_y_continuous()`
- Principais argumentos das funções `scale_(x|y)_discrete()` e `scale_(x|y)_continuous()`:
 - `drop`
T omite do gráfico os níveis de um fator que não aparecem nos dados; F usa todos os níveis de um fator.
 - `na.translate`
F remove valores faltantes da escala.
 - `labels`
NULL (nenhum nome `ticks`) ou um vetor com nome (caracteres) dos `ticks`.
`labels = c("One Hundred Fifty", "Three Hundred", "Four Hundred Fifity")`
O comando `abbreviate`, abrevia o nome dos vetores nos `ticks`.
`labels = abbreviate`
 - `limits`
Vetor de caracteres, ou números, com os possíveis limites dos valores de escala e sua ordem.
`limits = c("caracter_1", "caracter_2")`
`limits = c(0, 600)`
 - `name`
Nome da escala que aparece na legenda (rótulos de x).
 - `breaks`
O argumento `breaks` nos permite especificar onde os `ticks` aparecem.
Podemos dar nomes aos `ticks` especificados para aparecerem, usando o argumento `labels`.
`breaks = c(150, 300, 450),`
`labels = c("One Hundred Fifty", "Three Hundred", "Four Hundred Fifity")`
 - `expand`
Expande a escala por adição de x aos limites da escala.
`expand = expand_scale(add = x)` ou `expand = expansion(add = x)`
Expande a escala por multiplicação de x aos limites da escala.
`expand = expand_scale(mult = x)` ou `expand = expansion(mult = x)`
 - `position`
Posição da escala no eixo x (*top* ou *bottom*) e no eixo y (*left* ou *right*).
`position = 'top'`
 - `trans`
Transforma a escala continua.
`trans = "reverse"`
Principais transformações:

- * ans
- * atanh
- * boxcox
- * date
- * exp
- * hms
- * identity
- * log
- * log10
- * log1p
- * log2
- * logit
- * modulus
- * probability
- * probit
- * pseudo_log
- * reciproval
- * reverse
- * sqrt
- * time

- Exemplos:

```
#Plotando gráfico
p = ggplot(data = Turismo, aes(x=Estado, y=cheg_2012))
p+
  geom_blank()+
  labs(title = "Título",
       x = "Eixo x",
       y = "Eixo y",
       subtitle = "Subtítulo",
       caption = "Elaborado por ...")+
  theme_bw(base_size = 18)+ #Tema
  scale_x_discrete(limits=c("Amazonas","RioJaneiro"))
#Vetor de caracteres com os possíveis valores de escala e sua ordem.
```

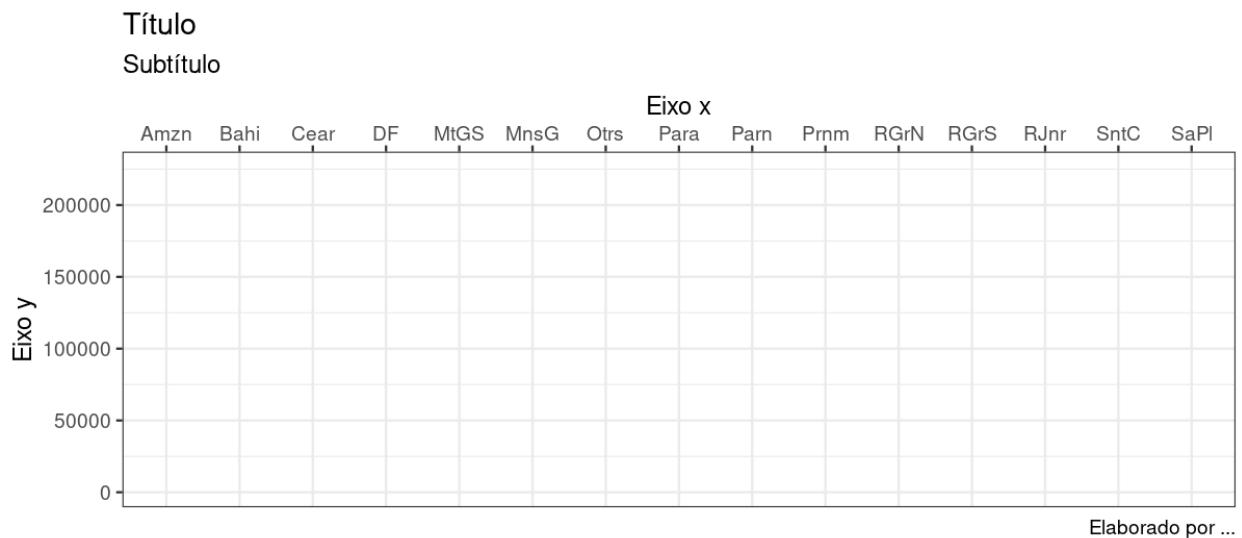


Figure 19: Exemplo 1 - scale_x_discrete

```

#Plotando gráfico
p = ggplot(data = Turismo, aes(x=cheg_2012/1000, y=cheg_2013/1000))
p+
  geom_blank()+
  labs(title = "Título",
       x = "Eixo x",
       y = "Eixo y",
       subtitle = "Subtítulo",
       caption = "Elaborado por ...")+
  theme_bw(base_size = 18)+ #Tema
  scale_y_continuous(
    breaks = c(75,150,225),
    labels = c("75 mil","150 mil","225 mil"), #nome dos ticks do eixo y
    position = "right",
    trans = "reverse")+
  scale_x_continuous( limits = c(50,150)) #limites do eixo x

```

Título

Subtítulo

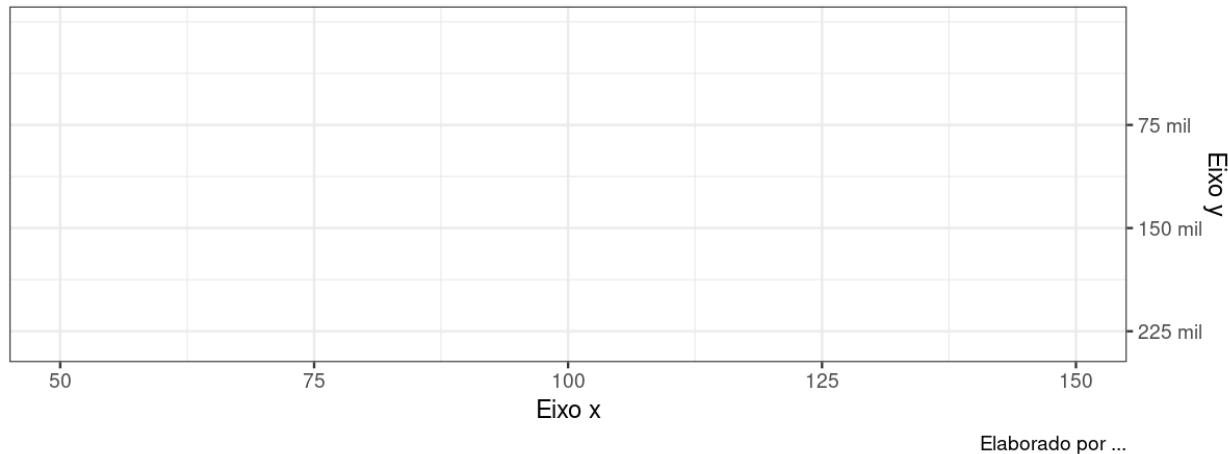


Figure 20: Exemplo 2 - scale_(x|y)_continuous

10.2.8 Cores nos gráficos ggplot2

- As cores podem ser aplicadas em diversos elementos do gráfico:
 - Linhas
 - Preenchimentos da forma gráfica
 - Texto
- Principais parâmetros:
 - **fill**
Controla o preenchimento de um gráfico.
 - **colour** ou **color**
Cor de linha ou contorno do gráfico.
 - **alpha**
Controla o grau de transparência da cor, valores entre 0 e 1 (0 sendo muito transparente e 1 sendo opaco).
Ex.: `alpha <- 1`

10.2.8.1 Método para obter cores em R

- Pelo número

`col = x`, sendo algum número. `x = 1, 2, ...`

- Pelo nome

Há 657 nomes de cores disponíveis no R.

Através da função `colors()` diretamente na linha de comando será exibido o nome das 657 cores.

Ademais se colocar `colors()` [x] será exibido o nome na posição x.

Ex.: `colors()` [657] = “grey”.

- Pelo sistema **RGB** (*Red, Green, Blue*)

rgb(0,0,0)

- Pelo sistema hexa decimal

#ff0000

Figure 21: 657 cores e seus respectivos nomes.

10.2.8.2 Principais pacotes de paletas de cores do R

- R básico
- Pacote **RColorBrewer**

10.2.8.3 Tipos de paletas de cores

- *sequencial*
Cores que variam em sequência da mais clara para mais escura.
- *divergente*
O centro da paleta é mais claro e os extremos mais escuros em ambas as direções.
- *qualitativa*
Não possui um ordenamento nas variações das cores.

10.2.8.4 5 funções básicas do R que geram paletas de cores sequenciais

- `rainbow(n,alpha)`
- `heat.colors(n,alpha)`
- `terrain.colors(n,alpha)`
- `topo.colors(n,alpha)`
- `cm.colors(n,alpha)`

10.2.8.5 RColorBrewer paletas de cores disponíveis

- *sequencial*

Table 22: **Pacote RColorBrewer:** Nome das paletas sequencial.

Nome das paletas
Blues
BuGn
BuPu
GnBu
Greens
Greys
Oranges
OrRd
PuBu
PuBuGn
PuRd
Purples
RdPu
Reds
YlGn
YlGnBu
YlOrBr
YlOrRd

As variações de cores vão de três a nove valores possíveis em cada paleta.

- *divergente*

Table 23: **Pacote RColorBrewer:** Nome das paletas divergente.

Nome das paletas
BrBG
PiYG
PRGn
PuOr
RdBu
RdGy
RdYlBu
RdYlGn
Spectral

As variações de cores vão de três a onze valores possíveis em cada paleta.

- *qualitativa*

Table 24: **Pacote RColorBrewer:** Nome das paletas qualitativas e número de cores possíveis em cada paleta.

Nome das paletas	Número de cores
Accent	8
Dark2	8
Paired	12
Pastel1	9
Pastel2	8
Set1	9
Set2	8
Set3	12

10.2.8.6 Aplicando escala de cinza ao gráfico

- `scale_fill_grey(..., start = x, end = x)`
x é um valor entre 0 e 1, sendo 0 mais escuro e 1 mais claro.

Aplica-se a gráficos que possuem preenchimento interno na forma como: **boxplot**, **histograma**, **violino** e **barras**.

- `scale_color_grey(..., start = x, end = x)`
x é um valor entre 0 e 1, sendo 0 mais escuro e 1 mais claro.
Aplica-se a gráficos como **dispersão** ou **linhas**.

10.2.9 Ajustando parâmetro de textos de um ggplot

- Os temas possuem formatações padronizadas para todos os elementos textuais de um gráfico como título, subtítulo ou rótulos dos eixos.
- É possível realizar ajustes através da camada `theme()`, utilizando-se dos argumentos de `element_text()`.
- Em `element_text()` podemos ajustar os seguintes parâmetros:
 - `family`
Tipo de fonte, o padrão é “sans”. No sistema **Windows** é possível consultar as famílias disponíveis através do comando `windowsFonts()`. Para mais opções de fontes utilize o pacote `extrafont` ou `showtext`.
 - `face`
`plain`, `italic`, `bold`, `bold.italic` para ajustar a fonte em **plana**, **ítálico**, **negrito** ou **negrito-ítálico** respectivamente.
 - `colour`
Cor da linha.
 - `size`
Tamanho do texto em pontos. Pode usar um valor ou proporcional ao padrão, fazendo `rel(1.5)` para o aumento de 50% ou `rel(0.5)` para diminuir 50%.
 - `hjust`
Alinhamento horizontal entre [0,1], `hjust = 0.5` centraliza.
 - `vjust`
Alinhamento vertical entre [0,1], `vjust = 0.5` centraliza.
 - `angle`
De 0 a 360.
 - `lineheight`
Altura da linha.
- Podemos aplicar os elementos de texto de forma global ou especificando o elemento que pode ser só o título, só um dos eixos, etc.
- Para maiores detalhes utilize o comando `??theme`.

- Exemplo:

```
#Plot
p = ggplot(data = dados, aes(x = cheg_2012/1000, y = cheg_2013/1000))

#ajustando parâmetros de texto
p+
  geom_blank()+
  labs(title = "Título",
       x = "Eixo x",
       y = "Eixo y",
       subtitle = "Subtítulo")+
  theme_bw(base_size = 18)+
  theme(text = element_text(family = "mono"))+
#Altera a fonte de todos os textos
  theme(axis.text.x = element_text(size = rel(1.2)))+
#Aumenta a fonte só do eixo x em 20%
  theme(axis.text.y = element_text(angle = 45))+
#Muda o angulo do texto do eixo y em 45 graus
  theme(axis.title.y = element_text(face = "bold.italic"))+
#Muda o rótulo do eixo y para negrito-itálico
  theme(plot.title = element_text(hjust = 0.5))+
#Centraliza o título
  theme(plot.subtitle = element_text(hjust = 1))
#Subtítulo a direita

#fechando dispositivo gráfico
dev.off()
```

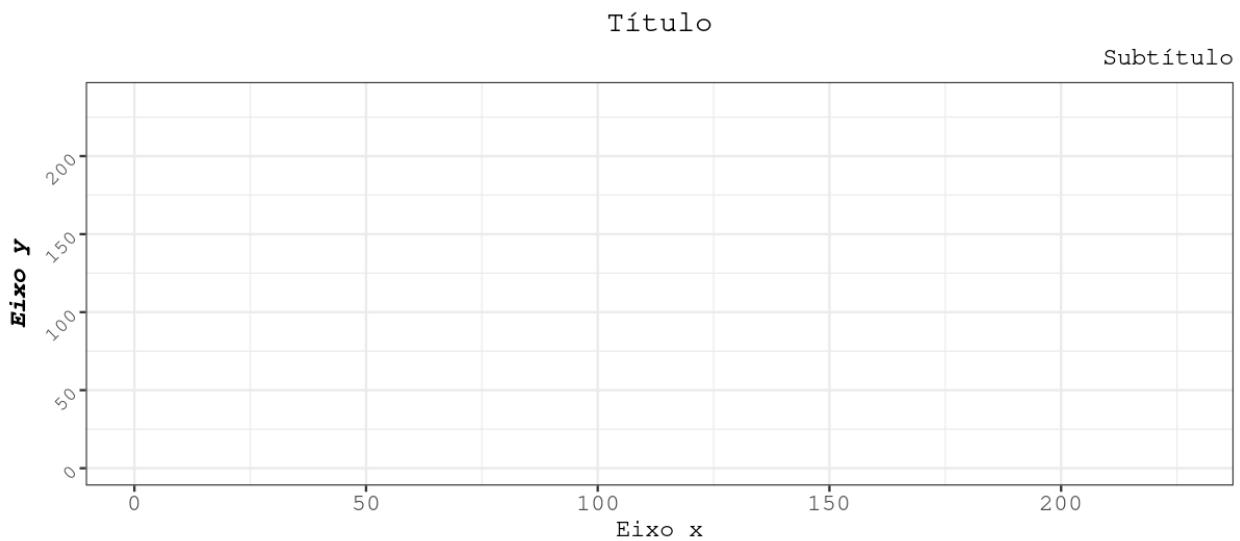


Figure 22: Gráfico com ajustes de texto.

10.2.10 Layout da janela gráfica e plotagem de vários gráficos em uma janela

10.2.10.1 Principais pacotes para configurar layout da janela gráfica

- `grid`
- `patchwork`

10.2.10.2 Pacote grid

- O pacote `grid` implementa funções gráficas no sistema de plotagem `ggplot2` e também é um pacote gráfico independente, apesar de ser pouco usada essa última função.
- O pacote `grid` oferece uma ampla variedade de funções que personalizam elementos de plotagem do `ggplot2`, como:
 - Temas
 - Cores
 - Grafos (pequenos e múltiplos)
 - Incluir anotações matemáticas em objetos `ggplot`
<https://github.com/tidyverse/ggplot2/wiki/Plotmath>
 - Alterar sistemas de coordenadas
 - *Layout* da janela gráfica
- Sobre o *layout* da janela gráfica, podemos configurar para adicionar vários gráficos numa mesma janela gráfica. O passo a passo:
 - Criar e configurar os gráficos.
 - Definir quais gráficos devem aparecer e configuração do *layout* da janela gráfica (em tabela).
 - *Print* dos gráficos em suas devidas posições no *layout* (em tabela) na janela gráfica.
- Exemplo:

```
#ggplot
p = ggplot(data = dados, aes(x = cheg_2012/1000,y = cheg_2013/1000))

#Gráfico 1
g1 <- p+
  geom_blank()+
  theme_bw(base_size = 18)

#Gráfico 2
g2 <- p+
  geom_point()+
  theme_minimal(base_size = 18)

#Layout para 1 linha e 2 colunas (g1 ao lado de g2)
pushViewport(viewport(layout = grid.layout(1,2)))
```

```

#Atribuindo g1
print(g1, vp=viewport(layout.pos.row = 1, layout.pos.col = 1))

#Atribuindo g2
print(g2, vp=viewport(layout.pos.row = 1, layout.pos.col = 2))

#fechando dispositivo grafico
dev.off()

```

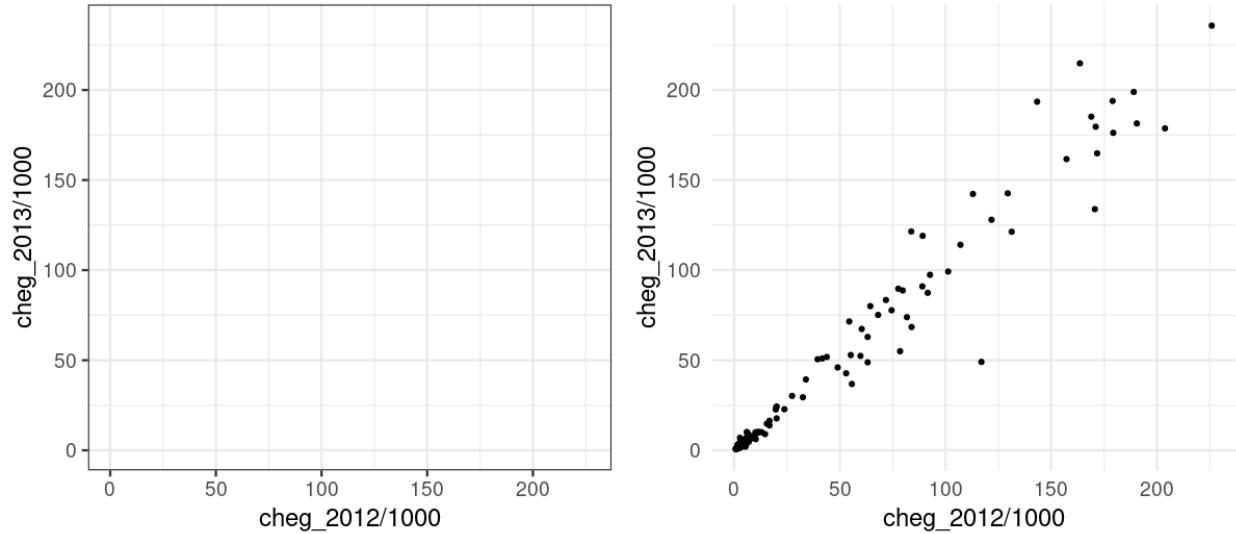


Figure 23: *Layout* da janela gráfica com dois gráficos, usando biblioteca grid do **R**.

10.2.10.3 Pacote patchwork

- O objetivo do `patchwork` é tornar simples juntar `ggplot` separados no mesmo gráfico.
- O `patchwork` usa uma API que incita a exploração, iteração e escala para *layout* arbitrariamente complexos.
- Formas de agrupar gráficos num *layout* usando `patchwork`:
 - `g1 + g2`
Agrupa em linha.
 - `(g1|g2|g3)/g4`
Agrupa em linhas e colunas.
- Exemplo:

```
#ggplot
p = ggplot(data = dados, aes(x = cheg_2012/1000,y = cheg_2013/1000))

#Gráfico 1
g1 <- p+
  geom_blank()+
  theme_bw(base_size = 18)

#Gráfico 2
g2 <- p+
  geom_point()+
  theme_minimal(base_size = 18)

#Gráfico 3
g3 <- p+
  geom_boxplot()+
  theme_grey(base_size = 18)

#Gráfico 4
g4 <- p+
  geom_point()+
  theme_minimal(base_size = 18)

#Layout 1 para 1 linha e 2 colunas (g1 ao lado de g2)
g1 + g2 + plot_layout(ncol = 2) + plot_annotation(title = "Dois gráficos com patchwork",
                                                    tag_levels = "1")
                                                    #Título geral e número em cada gráfico

#fechando dispositivo grafico
dev.off()

#Layout 2
(g1 | g2 | g3) / g4

#fechando dispositivo grafico
dev.off()
```

Dois gráficos com patchwork

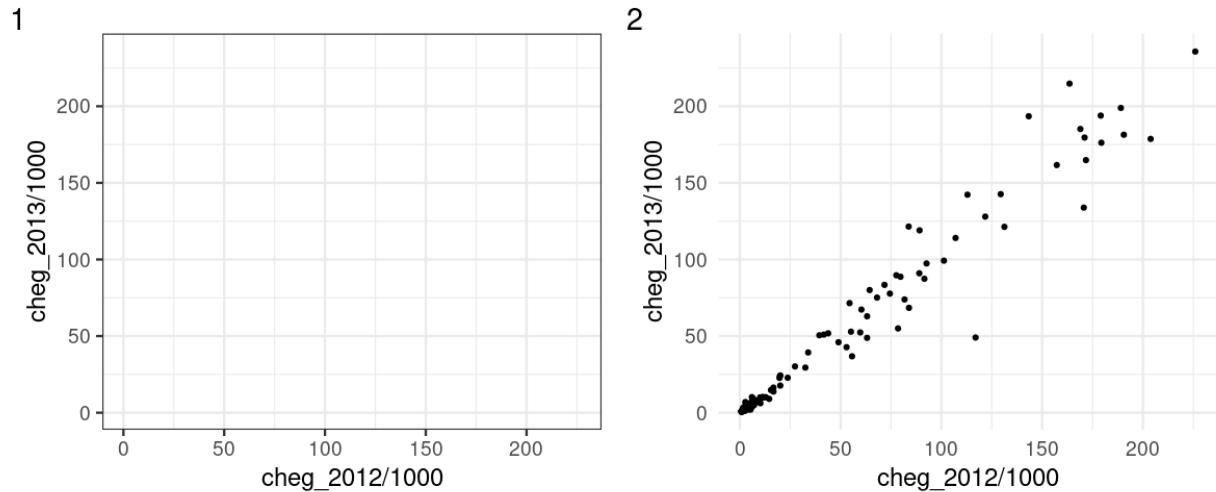


Figure 24: *Layout 1* da janela gráfica com dois gráficos, usando biblioteca patchwork do R.

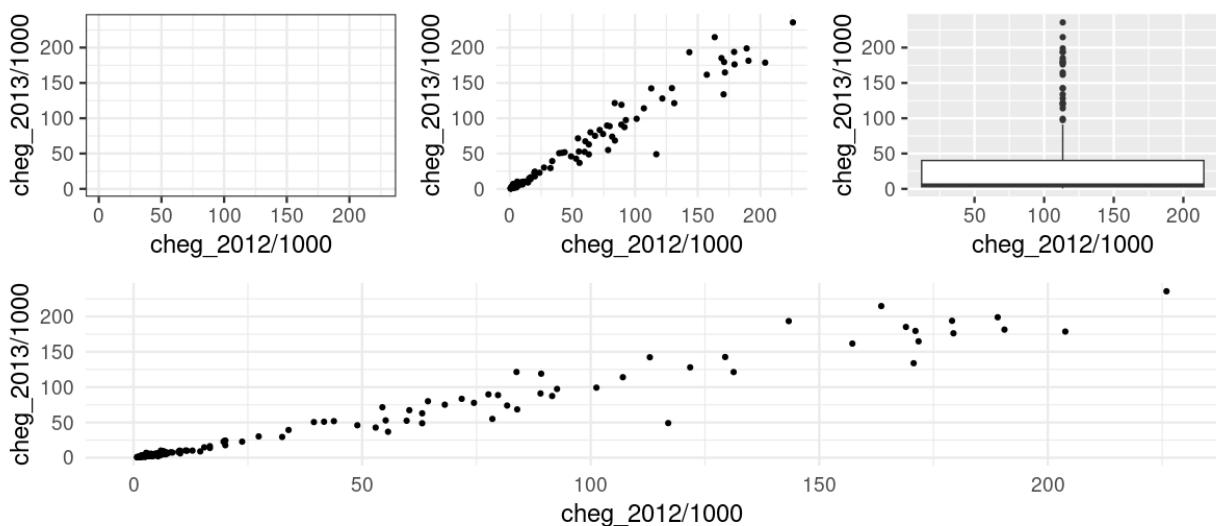


Figure 25: *Layout 2* da janela gráfica com quatro gráficos, usando biblioteca patchwork do R.

10.2.11 Gráficos usando pacote `ggplot2`

- Passo a passo (principais pacotes):

- Importa dados

- `readr`

- Organizar os dados

- `tidyR`

- `dplyr`

- `magrittr`

- Preparar gráfico

- `ggplot2`

- Adicionar camadas ao gráfico

- `grid`

- `ggthemes`

- `RColorBrewer`

- `extrafont`

- `showtext`

- Plotar gráfico e definir layout

- `grid`

- `patchwork`

10.2.11.1 Gráfico de barras (geom_bar) com ggplot2

- Tipos de gráficos de barras do ggplot2:

- **geom_bar**

Para plotar um gráfico de barras no ggplot2, definimos o tipo de gráfico **geom_bar**.

A altura das barras é proporcional ao número de casos em cada grupo.

- **geom_col**

Outra forma de gerar gráficos de barras no ggplot2 é o tipo de gráfico **geom_col**.

A altura das barras representão os valores dos dados.

- Observações:

- A função **reorder** serve para reordenar uma variável em função de outra. Muito útil para organizar as barras em formato crescente, ou decrescente, dos grupos.

- Quando temos mais de uma variável numérica associada a variável categorica (grupo) podemos ter no gráfico de barras com barras lado a lado ou empilhadas.(Ver exemplo 2)

- * **fill = variável**

Categorias secundarias agrupadas.

Gera legenda automática na lateral.

- * **geom_bar(stat = "identity", position = "dodge")**

barras agrupadas da mesma categoria ficam lado a lado.

- * **geom_bar(stat = "identity")**

Sem **position = "dodge"**, por default é barras agrupadas da mesma categoria na forma empilhadas.

- Os argumentos:

- **stat = identity**

Não altera o gráfico de barras.

- **stat = count**

Conta o número de casos em cada posição x.

- Exemplo 1 - Gráfico de Barras:

```
#Plotando gráfico de barras (geom_bar)
p <- ggplot(data = dt, aes(x = reorder(Estado,y), y))+ #Mapeamento das variáveis
  geom_bar(stat = "identity")#Forma de barras
  labs(title = "Chegada de Turistas ao Brasil em 2013", #Título
       x = "Estados", #Texto do eixo x
       y = "Número de chegadas por mil")+ #Texto do eixo y
  geom_text(aes(label = round(y,2)), hjust=0.5, vjust=0)+ #Insere valores sobre as barras
  theme_bw(base_size = 18)+ #Define o tema
  theme(plot.title = element_text(hjust = 0.5))+ #Centraliza o texto do título
  theme(axis.text.x = element_text(angle = 90)) #Muda o ângulo do texto do eixo x em 90 graus
p

#Fechando dispositivo gráfico
dev.off()
```

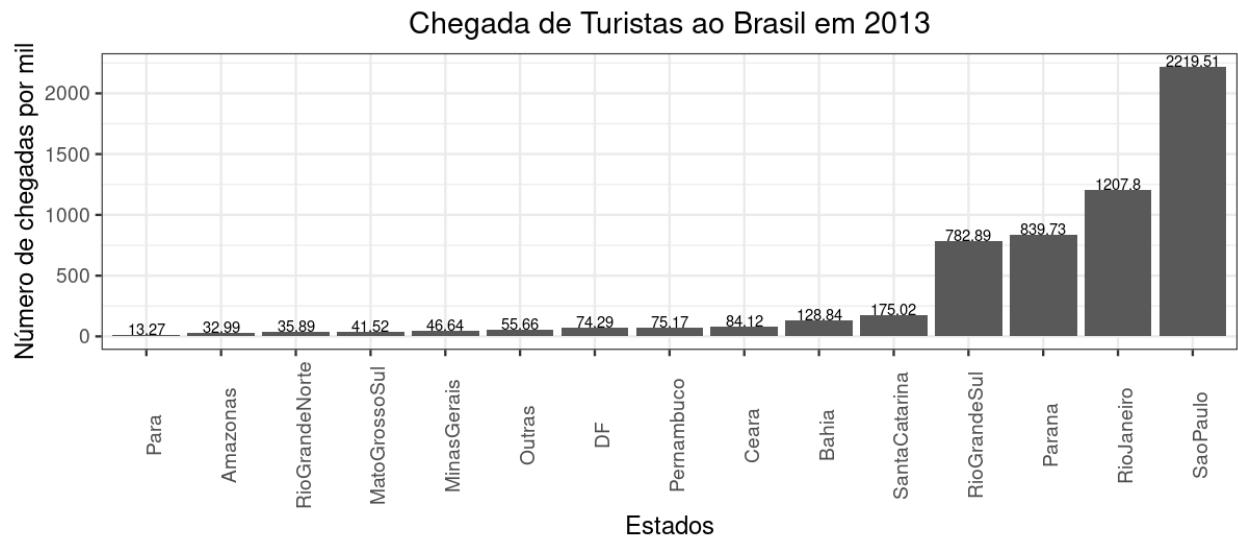


Figure 26: Gráfico de barras (geom_bar)

- Exemplo 2 - Gráfico de Barras com mais de uma categoria e *layout* com dois gráficos:

```
#Preparação dos dados
dt <- dados %>%
  filter(Estado == "SaoPaulo" | Estado == "RioJaneiro") %>%
  rename('2012' = cheg_2012, '2013' = cheg_2013, '2014' = cheg_2014, '2015' = cheg_2015) %>%
  gather(ano, chegada, '2012':'2015') %>%
  select(Estado, ano, chegada) %>%
  group_by(Estado,ano) %>%
  summarize(chegada=sum(chegada)) %>%
  ungroup()

#Convertendo ano para fator
dt$ano = factor(dt$ano)

#Visualizando dados
dt
View(dt)

#Plotagem gráfico de barras com duas ou mais categorias
p = ggplot(dt)+
  aes(x=reorder(Estado,chegada), y=chegada/1000, fill = ano)+ 
  geom_bar(stat = "identity",position = "dodge")+
  geom_text(aes(label = round(chegada/1000,2)),
            position = position_dodge(width = 0.9),
            vjust=-0.25)

#Adicionando camadas a p
p1 = p +
  labs(title = "Chegada de Turistas ao Brasil - Versão barras lado a lado",
       x = "Estados",
       y = "Número de chegadas por mil")+
  theme_bw(base_size = 18)+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_fill_grey(start = 0, end = 0.9)

p1

#Plotando versão barras empilhadas
p = ggplot(dt)+
  aes(x=reorder(Estado,chegada), y=chegada/1000, fill = ano)+ 
  geom_bar(stat = "identity")+
  geom_text(aes(label = round(chegada/1000,2)),
            position = position_stack(vjust=1))

#Adicionando camadas a p
p2 = p+
  labs(title = "Chegada de Turistas ao Brasil - Versão barras empilhadas",
       x = "Estados",
       y = "Número de chegadas por mil")+
  theme_bw(base_size = 18)+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_fill_grey(start = 0.4, end = 1)
```

p2

```
#Layout (patchwork)
pp = p1 + p2 + plot_layout(ncol = 1) +
  plot_annotation(title = "Gráfico de Barras com duas ou mais categorias",
                 tag_levels = "1")
```

pp

```
#Fechando dispositivo gráfico
dev.off()
```

Gráfico de Barras com duas ou mais categorias

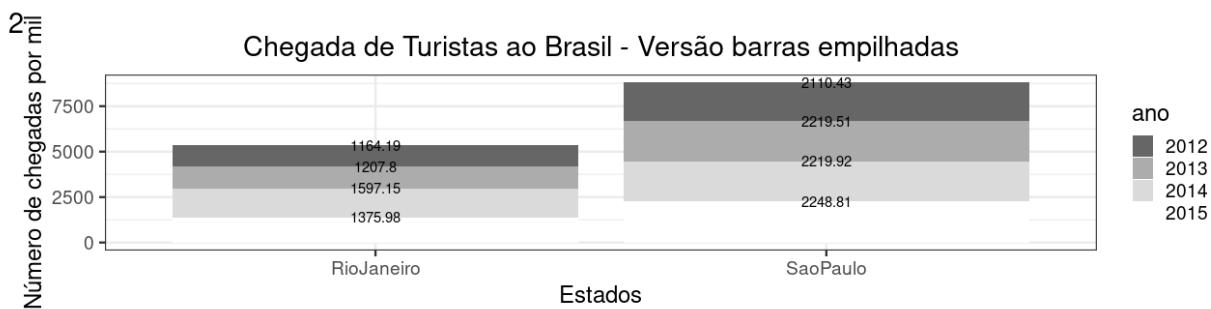
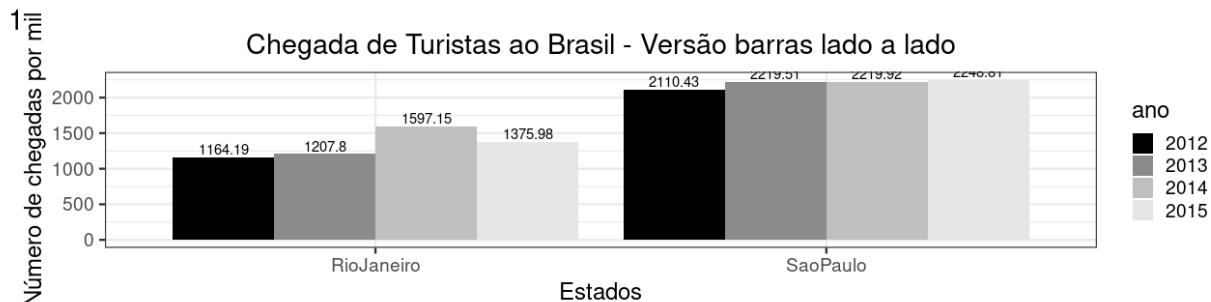


Figure 27: Gráfico de Barras (geom_bar) com duas ou mais categorias e *layout* com dois gráficos.

10.2.11.2 Histograma com ggplot2

10.2.11.2.1 Teoria histograma

- Histograma é um tipo de gráficos de barras.
- É usado para variáveis quantitativas continuas.
- Para elaborar um histograma é necessário uma variável quantitativa.
- No eixo x teremos os valores da variável e no eixo y sua frequência que pode ser absoluta ou relativa.
- Este tipo de gráfico é muito usado para observar:

– **Distribuição de frequências**

– **Simetria**

– **Desvio**

Presença de valores discrepantes (Outliers).

– **Amplitude da variável**

- A diferença entre gráficos de barras e histograma:

– **Gráfico de barras**

É aplicado a variáveis categóricas, apenas um eixo representando variável numérica e o outro eixo representando um variável categórica.

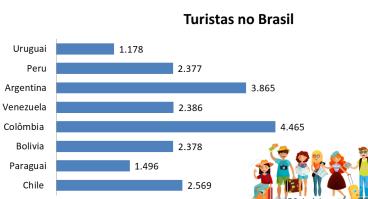


Figure 28: Exemplo gráfico de barras

– **Histograma**

É aplicado a variáveis numéricas e possui dois eixos numéricos (x representando a variável e y representando a frequência da variável).

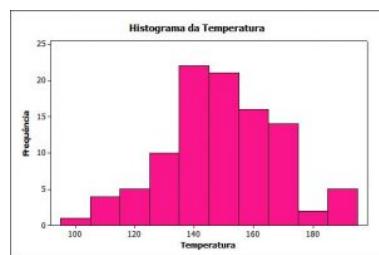


Figure 29: Exemplo histograma

10.2.11.2.2 Histograma

- Principais argumentos da função `geom_histogram()`:
 - `binwidth`
A largura das caixas (barras).
 - `color = "nome_cor", fill = "nome_cor"`
Altera a cor da linha e do preenchimento.
 - `color = variável_factor, fill = variável_factor`
Controla a variação de cor em função da variável **fator(factor)** do histograma.
 - `linetype="dashed"`
Altera o tipo da linha de contorno das caixas (barras) para pontilhado.
 - `position="identity"`
Histogramas sobrepostos.
 - `position="dodge"`
Histogramas intercaladas.
`-alpha=0.6`
Controlar transparencia da cor do preenchimento.
- Alterar a posição da legenda:
 - `theme(legend.position="top")`
Legenda posicionada na parte superior da janela gráfica.
 - `theme(legend.position="bottom")`
Legenda posicionada na parte inferior da janela gráfica.
 - `theme(legend.position="none")`
Remove a legenda.

- Exemplo - Histograma com eixo x logarítmico no ggplot2:

```
#Plotagem
p = ggplot(dados,aes(x=cheg_2013/1000))

#Adição de camadas
p+
  geom_histogram(aes(y= ..count.., fill = factor(Regiao)), #fill = variável factor
                 position = "identity", #Histogramas sobrepostos
                 alpha = 0.6, #Densidade das cores
                 binwidth = 0.1)+ #Largura das caixas (barras)
  scale_x_log10()#Eixo x em escala logarítmica
  labs(x = "Chegadas em escala logarítmica",
       y = "Frequência Absoluta",
       title = "Histograma do número de chegadas de turistas ao Brasil \nAno de 2013")+
  theme_bw(base_size = 18)+ #Tipo de tema
  scale_fill_discrete(name = "Região")#Escala de cores dos dados discretos
  scale_fill_grey(start = 0.2,end = 0.8) #Escala de cinza

#Fechando dispositivo gráfico
dev.off()
```

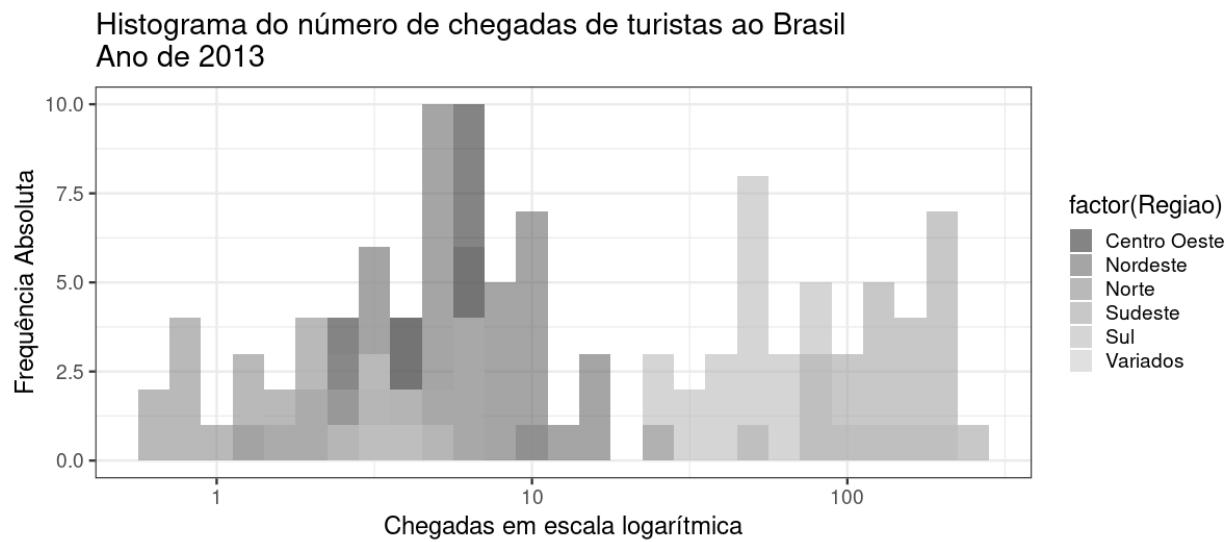


Figure 30: Histograma (geom_histogram) com eixo x logarítmico.

10.2.11.3 boxplot (diagrama de caixa) com ggplot2

10.2.11.3.1 Teoria boxplot

- O **Diagrama de caixa** serve para compreensão da forma e amplitude dos dados.
- Para elaborar um boxplot é necessário pelo menos uma variável quantitativa.
- Se a variável pode ser agrupada por fatores(factor), temos um boxplot comparativo $y \sim x$, isto é, a variável numérica y agrupada pelas categorias da variável do tipo fator x .
- É importante para fazer o **diagrama de caixa** conhecer a fórmula das **separatrizes**.
- O **diagrama de caixa** usa em sua construção os conceitos de **quartis (Q1, Q2, e Q3)**.

10.2.11.3.2 Separatrizes

- Quartis:
 - Q1 (25%)
 - Q2 (50% ou mediana)
 - Q3 (75%)
- Tabela de distribuição de frequências:

Salário	N. º de prof. (<i>fi</i>)	fac
01 -- 03	20	20
03 -- 05	40	60
05 -- 07	60	120
07 -- 09	30	150
09 -- 11	10	160
Total	160	-----

Figure 31: Exemplo de tabela de distribuição de frequências

- Como achar o intervalo de classe que corresponde a separatrix calculada.
Ex.: Se Q1 (25%), achar na tabela de classes e frequências, na coluna frequência acumulada, a classe que contém o valor que corresponde a 25% da frequência acumulada total.
Esse intervalo de classe será a classe selecionada para aplicação da fórmula.

- Fórmula da separatrix:

$$P_k = Li + \frac{k \cdot \sum f_i - F_{anterior}}{f_{intervalo}} \cdot h$$

onde,

P_k é o percentil (separatrix),

Li é o limite inferior do intervalo de classe selecionada,

k é o número em fração do percentil,

f_i é a frequência,

$\sum f_i$ é a frequência acumulada total,

$F_{anterior}$ é a frequencia acumulada do intervalo de classe anterior (ao selecionado) do qual se esta calculando,

$f_{intervalo}$ é a frequência do intervalo de classe selecionada,

h é a amplitude de classe ($Ls - Li$).

10.2.11.3.3 boxplot

- Montando a box:
A box contém como limite superior Q_3 , limite inferior Q_1 e linha interna a mediana (Q_2).
- Intervalo interquartil:

$$IQR = Q_3 - Q_1$$

- Limites:

- Máximo

$$L_{\max} = Q_3 + 1,5 \cdot (IQR)$$

- Mínimo

$$L_{\min} = Q_1 - 1,5 \cdot (IQR)$$

- Valores discrepantes (**Outliers**):

- Possíveis erros (arredondamento ou observação).
 - Alguma condição especial que deve ser observada separadamente.

- Exemplo explicativo de boxplot:

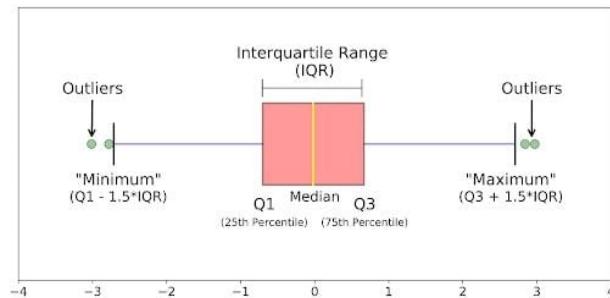


Figure 32: Exemplo explicativo de boxplot

- Exemplo - boxplot com eixos invertidos no ggplot2:

```
#Plotagem
p <- dados %>%
  ggplot(aes(x=as.factor(Mes),y=cheg_2012/1000))

#Adição de camadas
p +
  geom_boxplot() +
  labs(title = "Visualizando a variabilidade de chegadas de turistas ao Brasil no ano de 2012",
       x= "Mês",
       y="Número de chegadas") +
  theme_bw(base_size = 18) + #Adiciona tema "black and white"
  coord_flip() #Inverte posição do eixo x

#Fechando dispositivo gráfico
dev.off()
```

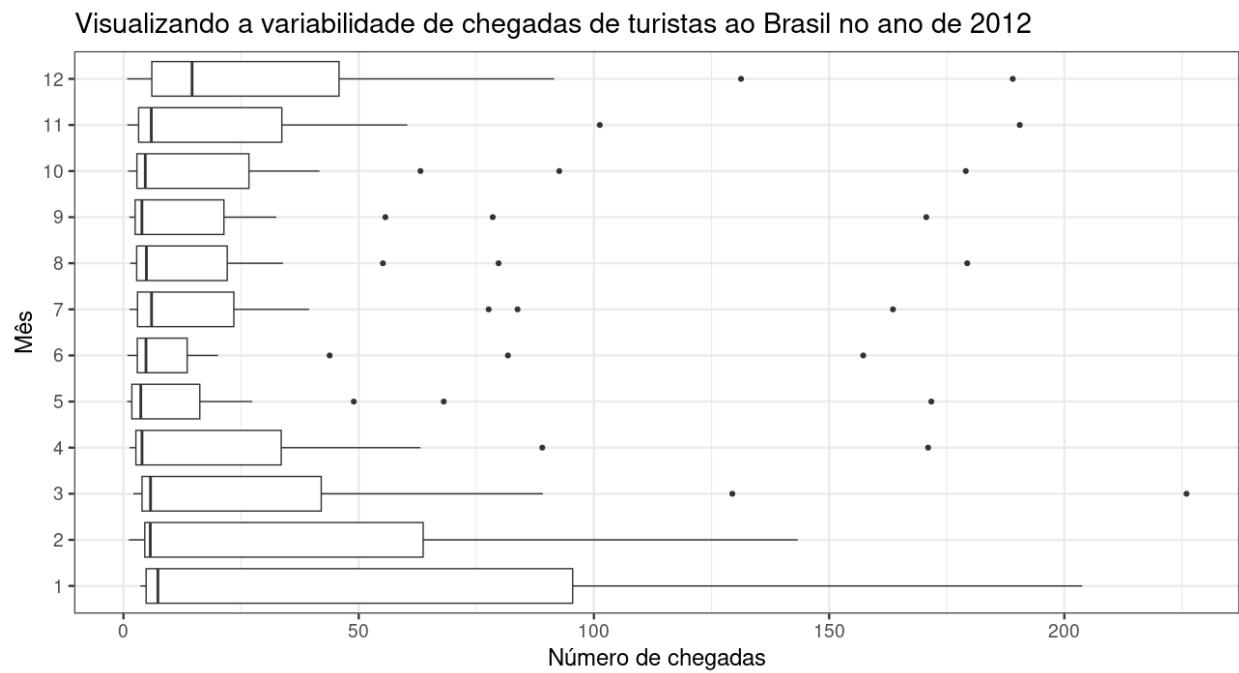


Figure 33: boxplot (diagrama de caixa) com eixos invertidos.

- Exemplo - boxplot com efeito jitter:

```
#Plotagem
p <- dados %>%
  ggplot(aes(x=as.factor(Mes),y=cheg_2012/1000))

#Adição de camadas
p +
  geom_boxplot() +
  geom_jitter() + #Adiciona o feito jitter
  labs(title = "Efeito jitter",
       x= "Mês",
       y="Número de chegadas") +
  theme_bw(base_size = 18) #Adiciona tema "black and white"

#Fechando dispositivo gráfico
dev.off()
```

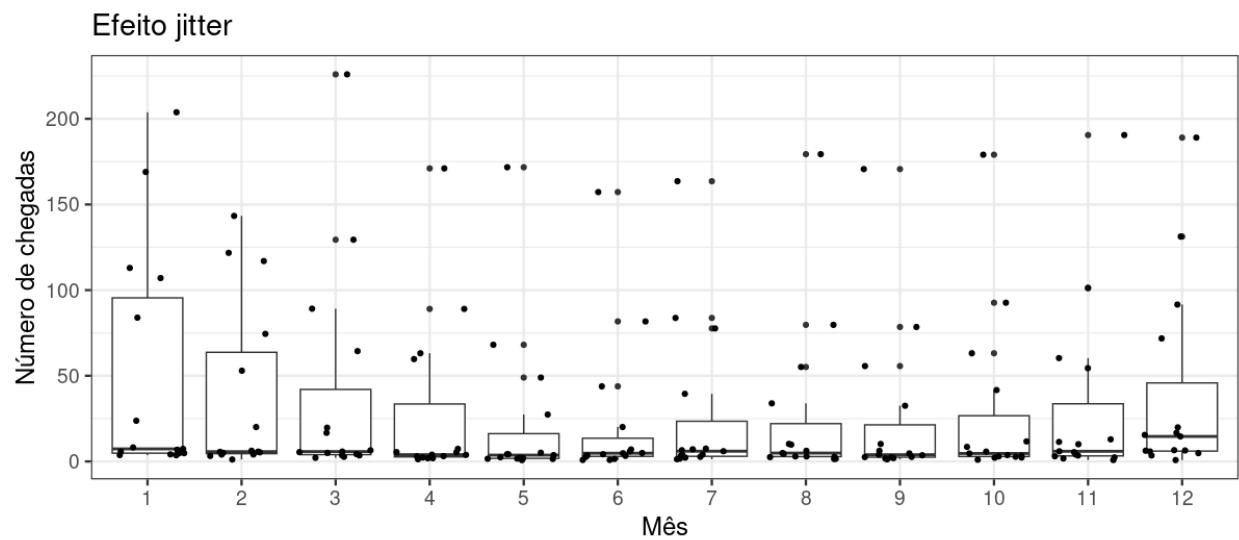


Figure 34: Boxplot com efeito jitter.

- Exemplos - boxplot dividindo gráfico por facetas (`facet_grid()` e `facet_wrap()`):

```
#Organizando dados
dt = dados %>%
  filter(Regiao == "Norte" | Regiao == "Nordeste")

#Plotagem
p <- dt %>%
  ggplot(aes(x=as.factor(Mes),y=cheg_2012/1000))

#Adição de camadas
p +
  geom_boxplot() +
  labs(title = "Chegada de turistas ao Brasil em 2012: Regiões Norte e Nordeste",
       x= "Mês",
       y="Número de chegadas") +
  theme_bw(base_size = 18)+ #Adiciona tema "black and white"
  facet_grid(Regiao ~.,scales = "free_y", space = "free")

#Fechando dispositivo gráfico
dev.off()
```

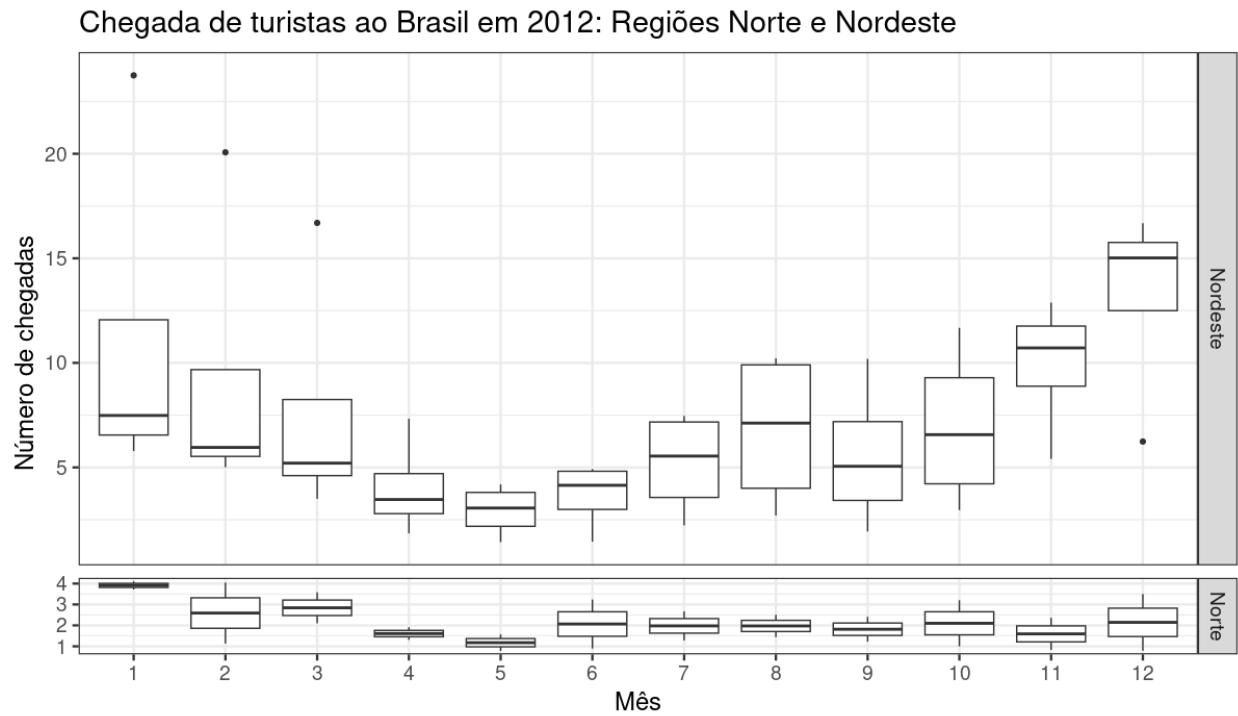


Figure 35: boxplot divido por facetas das regiões norte e nordeste, usando `facet_grid()`.

```

#Organizando dados
dt = dados %>%
  filter(Regiao != "Variados")

#Plotagem
p <- dt %>%
  ggplot(aes(x=as.factor(Mes),y=cheg_2012/1000))

#Adição de camadas
p +
  geom_boxplot() +
  labs(title = "Chegada de turistas ao Brasil em 2012 por Regiões",
       x= "Mês",
       y="Número de chegadas") +
  theme_bw(base_size = 18)+ #Adiciona tema "black and white"
  facet_wrap(~Regiao, scale = "free_y", nrow=2)

#Fechando dispositivo gráfico
dev.off()

```

Chegada de turistas ao Brasil em 2012 por Regiões

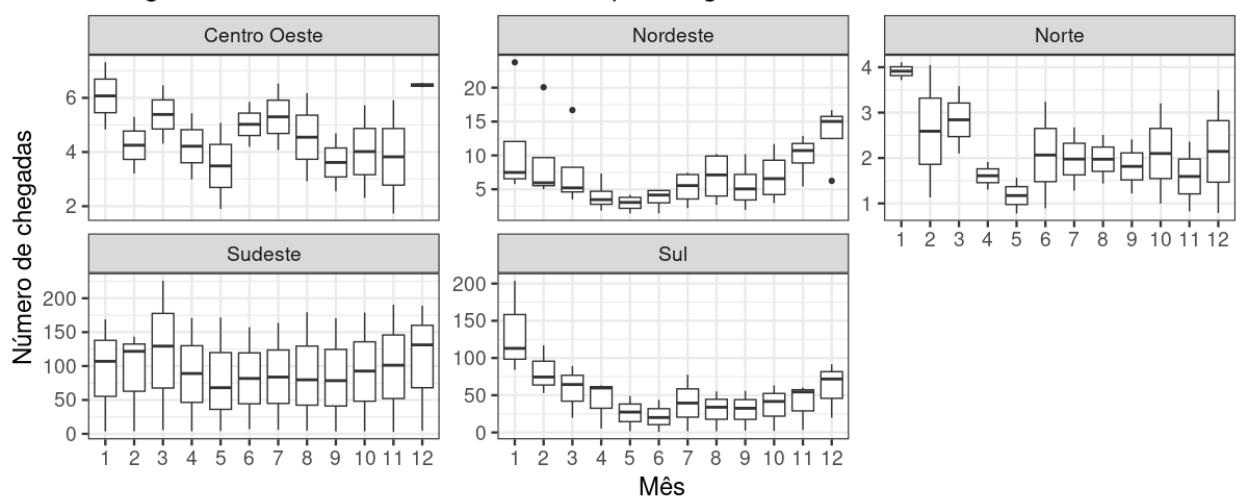


Figure 36: Gráfico dividido por facetas das regiões, usando `facet_wrap()`.

10.2.11.4 Gráfico circular (pizza) com ggplot2

- O gráfico circular é montado a partir do gráfico de barras.
- O gráfico circular é produzido a partir do gráfico de barras (`geom_bar()`), com uma única coluna e empilhado, e a variável `y` transforma em coordenada polar (`coord_polar()`).
`geom_bar(aes(x=1, weight=variável_y, fill=variável_de_agrupamento))+coord_polar(theta="y")`
- Passo a passo:
 - Produzir um gráfico de barras.
`geom_bar()`
 - As barras devem ser em formato empilhado, com uma única coluna.
`aes(x = 1, weight = variável_y, fill = variável_de_agrupamento)`
 - Aplicar coordenadas polar para transformar as barras em formato circular.
`coord_polar(theta = "y")`
 - Adicionar textos.
* `geom_text(x = 1.3, aes(y = Calculo_porcentagem, label = paste0(), "%"))`
label adiciona texto as fatias do gráfico circular.
`paste0` concatena texto.
 - Adicionar camadas:
 - * Retirar escalas
`scale_x_continuous(breaks = NULL)`
`scale_y_continuous(breaks = NULL)`
 - * Adicionar título e rótulos.
`labs(title = "Título", x = "", y = "Rótulo de y")`
x não recebe rótulo no gráfico circular.
 - * Atribuir tema e tamanho de fonte.
`theme_bw(base_size = 18)`
 - * Retirar grades.
`theme(panel.grid = element_blank())`

- Exemplo - Gráfico circular (pizza):

```
#Organizando dados
d <- data.frame(orcamento = c(10,20,30,40),
                 Empresa = c(paste("E",1:4)))

#Plotando gráfico
d %>%
  ggplot(aes(x = 1, weight = orcamento,
             fill = Empresa))+ #Gráfico de barras empilhado e coluna única, com variável de agrupamento
  geom_bar(color = "black")+
  coord_polar(theta = "y")+ #Transforma variável y em polar e o gráfico de barras em circular.
  geom_text(x = 1.3,
            aes(y = cumsum(orcamento[4:1]) - orcamento[4:1]/2,
                label = paste0(100*round(orcamento[4:1]/sum(orcamento[4:1]),3), "%")))+ 
  scale_x_continuous(breaks = NULL)+ #Retirar a escala x
  scale_y_continuous(breaks = NULL)+ #Retirar a escala y
  labs(title = "Gráfico circular com ggplot2",
       x = "",
       y = "Fatia do orçamento a ser pago")+
  theme_bw(base_size = 18)+ #Tema e tamanho da fonte base.
  theme(panel.grid = element_blank())+ #Não atribui nenhum valor as grades (sem desenho).
  scale_fill_grey(start = 0.6, end = 1) #Escala de cores cinza.

#Fechando dispositivo gráfico
dev.off()
```

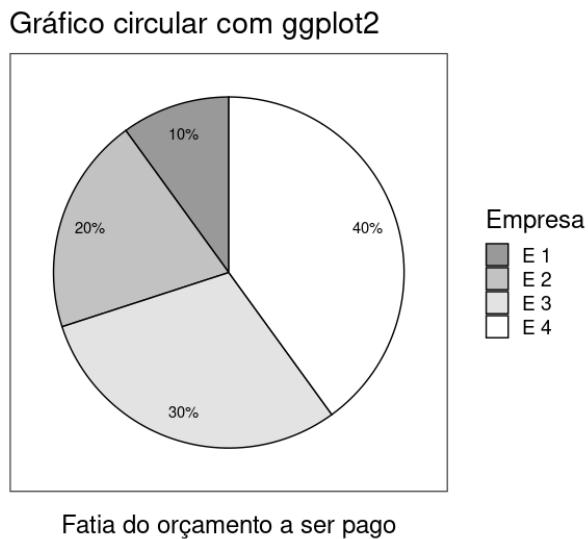


Figure 37: Gráfico circular por ggplot2, construído a partir das funções gráficas `geom_bar()` + `coord_polar()`.

10.2.11.5 Gráfico de pontos com ggplot2

- No gráfico de pontos temos dois eixos numéricos produzindo um gráfico de dispersão.
- Comando para aplicar gráfico de pontos:
`geom_point()`
- É possível agrupar os pontos por grupos e atribuir cores e formas distintas.
- Principais argumentos:
 - `subset`
Subconjunto de dados de um `data.frame`, aplica um filtragem nos dados.
`subset(data, variável %in% c(valor1,valor2,...))`
 - `shape`
Altera as formas de determinados grupos a partir de variável.
 - `color`
Altera as cores de determinados grupos a partir de variável.
 - `scale_x_continuous`
Configuração do eixo x (para dados numéricos).
 - * `limits`
Define os limites do eixo x.
`limits = c(limite_inferior, limite_superior)`
 - * `breaks`
Define o espaçamento da escala do eixo x.
`breaks = seq(limite_inferior, limite_superior, valor_espaçamento)`
 - `geom_point(size = valor)`
`size` altera o tamanho dos ícones do gráfico (os pontos).

- Exemplo - Gráfico de pontos (`geom_point()`) com subgrupos:

```
#Plotagem
ggplot(subset(dados, Regiao %in% c("Sul", "Sudeste")), #Subconjunto de dados de um data.frame
       #O comando filtra dos dados do data.frame
       aes(x = Mes, y = cheg_2013/1000,
           shape = Estado, #Alterar formas desse grupo de variáveis
           color = Regiao))+ #Diferencia os grupos das variáveis regiao por cor
       scale_x_continuous(limits = c(1,12), #Limites do eixo x
                           breaks = seq(1,12,1))+ #Espaçamento do eixo x
       geom_point(size = 3)+ #Tamanho dos ícones (pontos)
       labs(title = "Gráfico de Dispersão: Mês x chegadas em 2013",
            x = "Meses",
            y = "Chegadas por mil")+
       theme_bw(base_size = 18)+ #Adiciona tema "black and white" e tamanho da fonte
       theme(plot.title = element_text(hjust = 0.5))+ #Título centralizado
       scale_color_grey(start = 0.8, end = 0.2) #Aplica escalas de cinza

#Fechando dispositivo gráfico
dev.off()
```

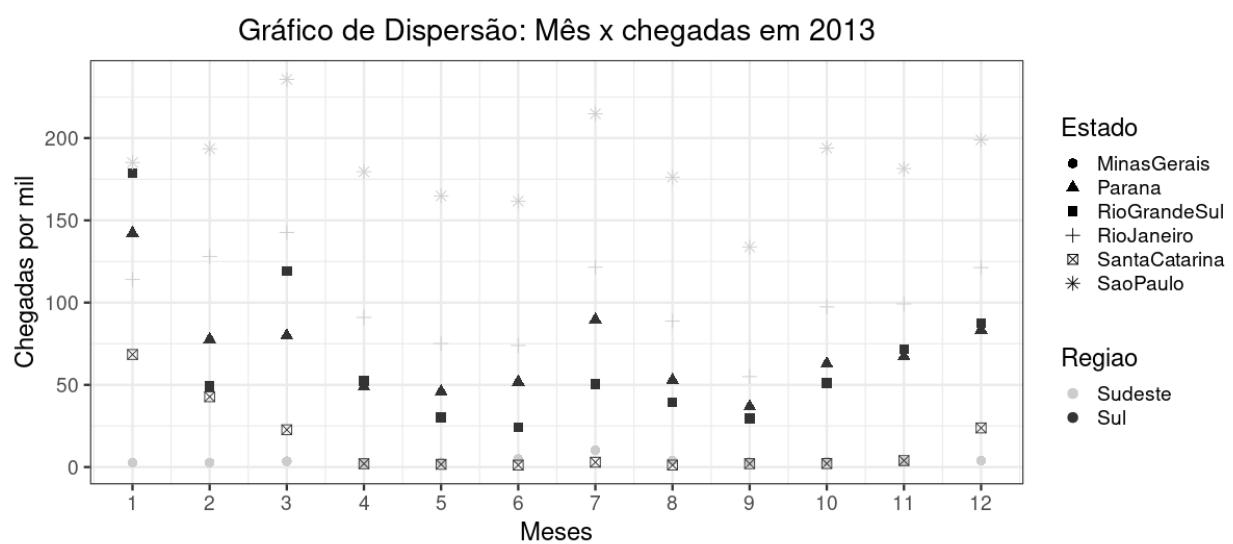


Figure 38: Gráfico de pontos, construído a partir da biblioteca `ggplot2` usando a função `geom_point()`.

- Exemplo - Gráfico de pontos com efeito jitter (`geom_jitter()`):

```
#Plotagem
ggplot(subset(dados, Regiao %in% c("Sul", "Sudeste")), #Subconjunto de dados de um data.frame
       #O comando filtra dos dados do data.frame
       aes(x = Mes, y = cheg_2013/1000,
           shape = Regiao))+ #Diferencia os grupos das variáveis regiao por forma
       scale_x_continuous(limits = c(1,12), #Limites do eixo x
                           breaks = seq(1,12,1))+ #Espaçamento do eixo x
       geom_jitter(size = 3, #Tamanho dos ícones (pontos)
                   aes(colour = Estado), #Camada color ao agrupamento por Estado (legenda)
                   width = 0.25)+ #Controla a largura do espalhamento jitter
       labs(title = "Gráfico de Dispersão com efeito jitter (espalhamento)",
            subtitle = "Mês x chegadas em 2013",
            x = "Meses",
            y = "Chegadas por mil")+
       theme_bw(base_size = 18)+ #Adiciona tema "black and white" e tamanho da fonte
       theme(plot.title = element_text(hjust = 0.5))+ #Título centralizado
       theme(plot.subtitle = element_text(hjust = 0.5))+ #Subtítulo centralizado
       scale_color_grey(start = 0.8, end = 0.2) #Aplica escalas de cinza

#Fechando dispositivo gráfico
dev.off()
```

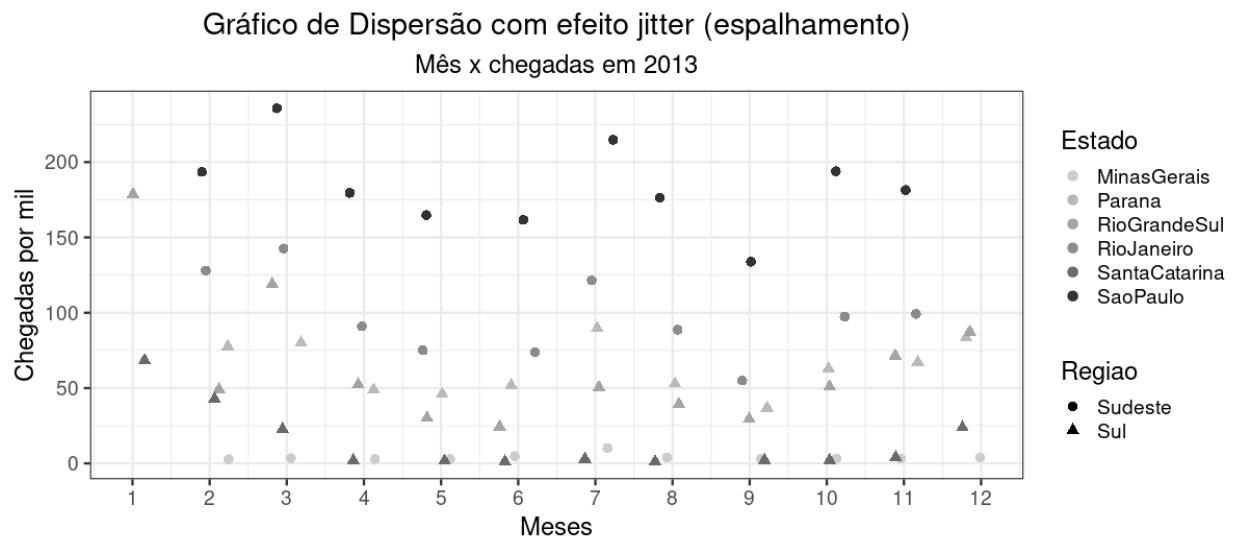


Figure 39: Gráfico de pontos com efeito jitter, substitui `geom_point()` por `geom_jitter()`.

- Exemplo - Gráficos de pontos com e sem efeito jitter (`geom_point()` e `geom_jitter()`):

```
#Gráfico 1
g1 <- ggplot(subset(dados, Regiao %in% c("Sul", "Sudeste"))), #Subconjunto de dados de um data.frame
#O comando filtra dos dados do data.frame
aes(x = Mes, y = cheg_2013/1000,
    shape = Estado, #Alterar formas desse grupo de variáveis
    color = Regiao))+ #Diferencia os grupos das variáveis regiao por cor
scale_x_continuous(limits = c(1,12), #Limites do eixo x
                    breaks = seq(1,12,1))+ #Espaçamento do eixo x
geom_point(size = 3)+ #Tamanho dos ícones (pontos)
labs(title = "Gráfico de Dispersão: Mês x chegadas em 2013",
     x = "Meses",
     y = "Chegadas por mil")+
theme_bw(base_size = 18)+ #Adiciona tema "black and white" e tamanho da fonte
theme(plot.title = element_text(hjust = 0.5))+ #Título centralizado
scale_color_grey(start = 0.8, end = 0.2) #Aplica escalas de cinza

#Gráfico 2
g2 <- ggplot(subset(dados, Regiao %in% c("Sul", "Sudeste"))), #Subconjunto de dados de um data.frame
#O comando filtra dos dados do data.frame
aes(x = Mes, y = cheg_2013/1000,
    shape = Estado))+ #Diferencia os grupos das variáveis regiao por forma
scale_x_continuous(limits = c(1,12), #Limites do eixo x
                    breaks = seq(1,12,1))+ #Espaçamento do eixo x
geom_jitter(size = 3, #Tamanho dos ícones (pontos)
            aes(colour = Regiao), #Camada color ao agrupamento por Estado (legenda)
            width = 0.25)+ #Controla a largura do espalhamento jitter
labs(title = "Gráfico de Dispersão com efeito jitter (espalhamento)",
     subtitle = "Mês x chegadas em 2013",
     x = "Meses",
     y = "Chegadas por mil")+
theme_bw(base_size = 18)+ #Adiciona tema "black and white" e tamanho da fonte
theme(plot.title = element_text(hjust = 0.5))+ #Título centralizado
theme(plot.subtitle = element_text(hjust = 0.5))+ #Subtítulo centralizado
scale_color_grey(start = 0.8, end = 0.2) #Aplica escalas de cinza

#Layout
g1 / g2

#fechando dispositivo grafico
dev.off()
```

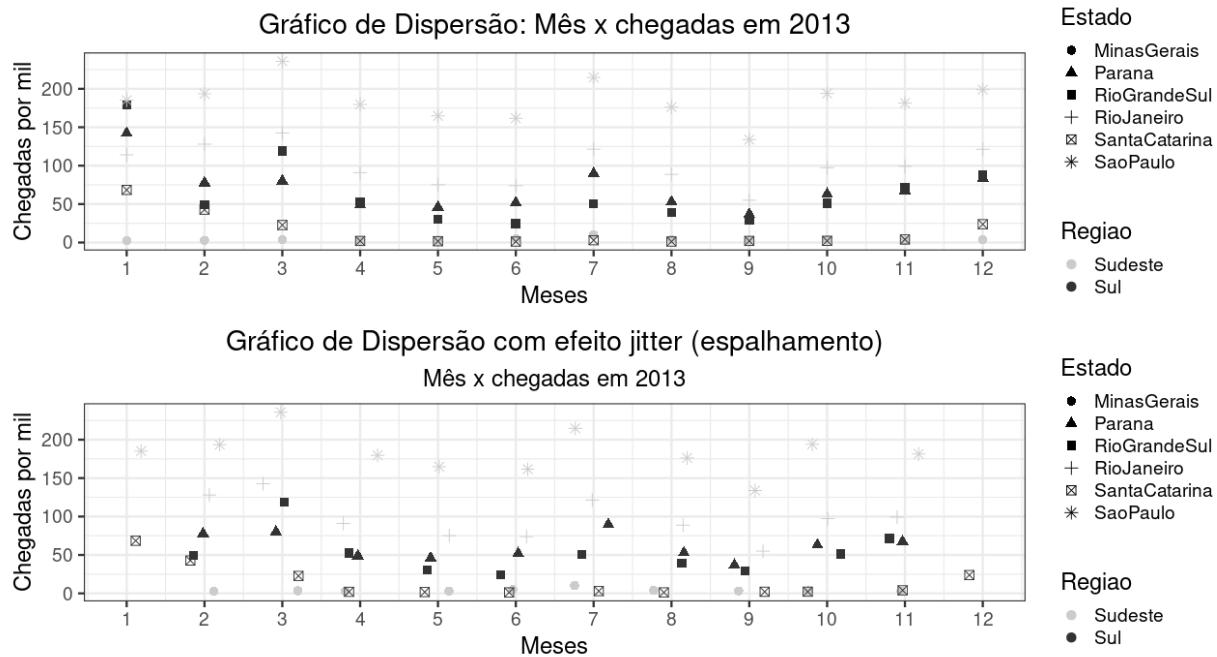


Figure 40: Gráfico de pontos comparando sem efeito jitter (`geom_point()`) e com efeito jitter (`geom_jitter()`).

10.2.11.6 Gráfico de linhas com ggplot2

- No gráfico de linhas temos dois eixos numéricos.
- O gráfico de linhas é produzido ligando os pontos do gráfico de pontos por linhas. Na biblioteca `ggplot2` o gráfico de linhas é construído através das funções `geom_point()`+`geom_line()`.
- É o gráfico adequado para representar séries de valores no tempo.
- Principais argumentos do gráfico de linhas na biblioteca `ggplot2`:

– `ggplot()`

Essa função recebe um `data.frame` e cria a camada básica do gráfico.

* `subset(data, variavel_grupos %in% c(valor_1, valor_2, ...))`

Subconjunto de dados de um `data.frame`.

O comando filtra dos dados do `data.frame`.

Ex.: `subset(dt, Regiao %in% c("Suldeste", "Sul"))`

* `aes()`

A função `aes()` mapeia os aspectos visuais do gráfico.

· `x`

Define qual vai ser a variável e por consequência os valores do eixo x.

· `y`

Define qual vai ser a variável e por consequência os valores do eixo y.

· `color`

Adiciona diferentes cores aos diversos grupos da variável selecionada.

`color = variável_agrupamento`

· `shape`

Adicionar diferentes formatos aos pontos dos diversos grupos da variável selecionada.

`shape = variável_agrupamento`

– `scale_x_continuous()`

Configuração do eixo x (para dados numéricos).

* `limits`

Define os limites do eixo x.

`limits = c(limite_inferior, limite_superior)`

* `breaks`

Define o espaçamento da escala do eixo x.

`breaks = seq(limite_inferior, limite_superior, valor_espaçamento)`

– `geom_point()`

Adiciona a camada de gráfico de pontos (os pontos).

* `size`

`geom_point(size = número)` altera o tamanho dos ícones do gráfico (os pontos).

– `geom_line()`

Adiciona a camada do gráfico de linhas (as linhas que liga os pontos).

* **size**
goem_line(size = número) altera a espessura das linhas.

– **theme_bw()**
Adiciona o tema “black and white”.

* **base_size**
Define o tamanho da fonte dos texto do gráfico.

– **scale_color_gray()**
Aplica escalas de cinza.

- Exemplo - Gráfico de linhas (`geom_point()`+`geom_line()`):

```
#Plotagem
ggplot(subset(dados, Regiao %in% c("Sul", "Sudeste")), #Subconjunto de dados de um data.frame
       #0 comando filtra dos dados do data.frame
       aes(x = Mes,
           y = cheg_2013/1000,
           shape = Estado, #Alterar formas desse grupo de variáveis
           color = Regiao))+ #Diferencia os grupos das variáveis regiao por cor
       scale_x_continuous(limits = c(1,12), #Limites do eixo x
                           breaks = seq(1,12,1))+ #Espaçamento do eixo x
       geom_point(size = 3)+ #Tamanho dos ícones (pontos)
       geom_line(size = 1.0)+ #Espessura da linha
       labs(title = "Gráfico de Linhas: Mês x chegadas em 2013",
            x = "Meses",
            y = "Chegadas por mil")+
       theme_bw(base_size = 18)+ #Adiciona tema "black and white" e tamanho da fonte
       scale_color_grey() #Aplica escalas de cinza

#Fechando dispositivo gráfico
dev.off()
```

Gráfico de Linhas: Mês x chegadas em 2013

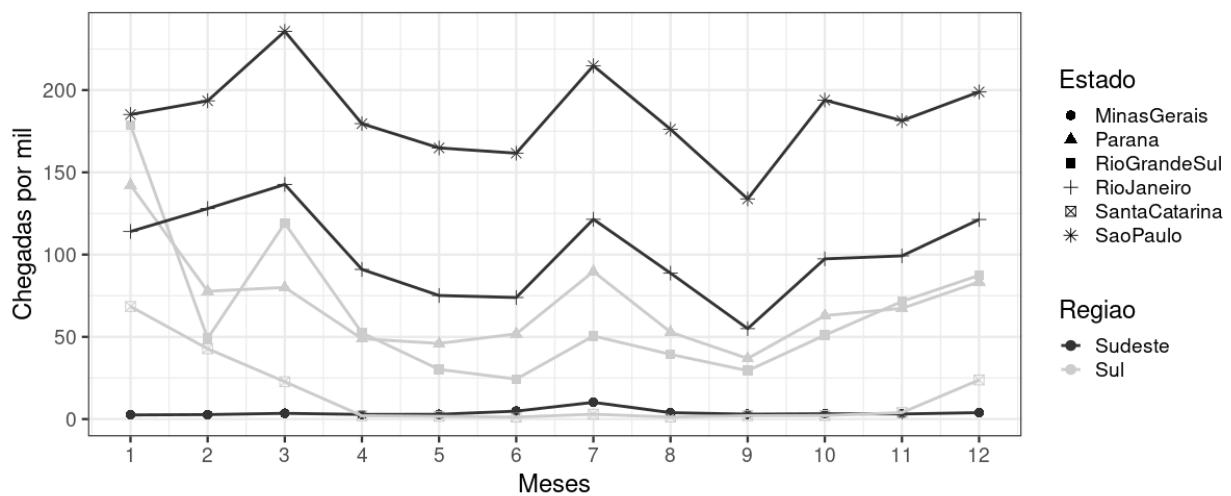


Figure 41: Gráfico de linhas, construído a partir da biblioteca `ggplot2` usando as funções `geom_point()` + `geom_line()`.

- Exemplo - Gráfico de linhas com efeito de suavização smooth(geom_smooth()):

```
#Plotagem
ggplot(subset(dados, Estado %in% c("SaoPaulo")), #Subconjunto de dados de um data.frame
       #O comando filtra dos dados do data.frame
       aes(x = cheg_2012/1000,
           y = cheg_2013/1000,
           color = Estado))+ #Diferencia os grupos das variáveis regiao por cor
       scale_x_continuous(limits = c(140,230), #Limites do eixo x
                           breaks = seq(140,230,10))+ #Espaçamento do eixo x
       geom_point(size = 1.5)+ #Tamanho dos ícones (pontos)
       geom_smooth(size = 1.0)+ #Espessura da curva
       labs(title = "Gráfico com ajuste de curva de tendência: 2012 x 2013",
            x = "Chegadas por mil em 2012",
            y = "Chegadas por mil em 2013")+
       theme_bw(base_size = 18)+ #Adiciona tema "black and white" e tamanho da fonte
       scale_color_grey() #Aplica escalas de cinza

#Fechando dispositivo gráfico
dev.off()
```

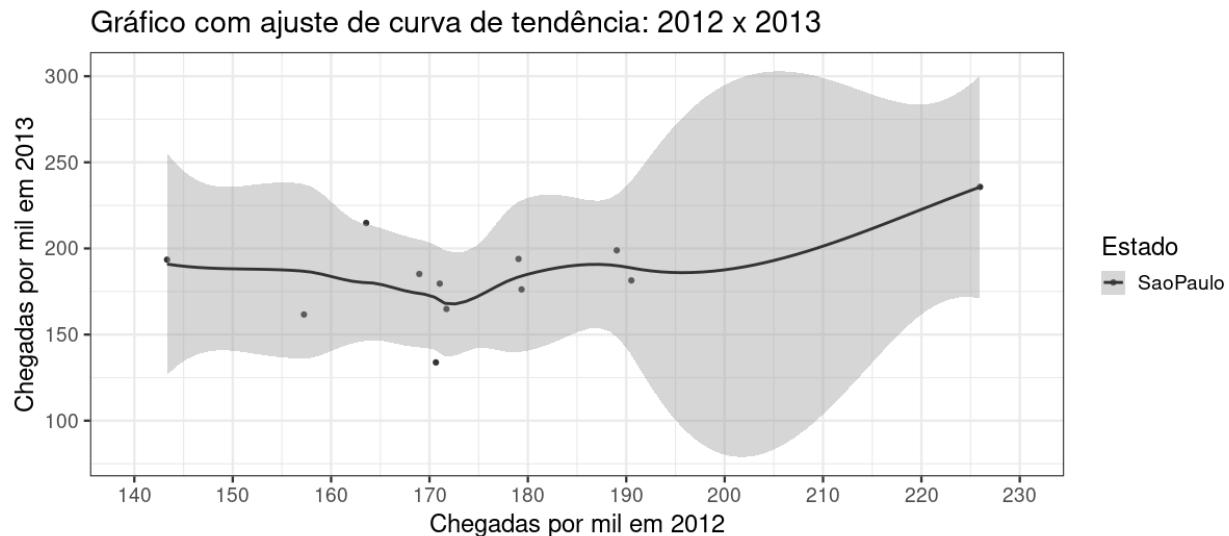


Figure 42: Gráfico de linha com ajuste por curva suavizada (geom_point() + geom_smooth()).

10.2.11.7 Gráfico de pontos com ajuste por curva de tendência com ggplot2

- Trata-se de um ajuste por curva de tendência entre duas variáveis numéricas.
- Usa a função `geom_smooth()` em conjunto com a função `geom_point()`, ambas da biblioteca `ggplot2`, para traçar uma curva suavizada que melhor se ajuste aos pontos, criada através do modelo de **regressão local**.

`geom_point() + geom_smooth()`

- A função `geom_smooth()` aceita outros métodos para geração de tendência, através do argumento `method`.

Ex.: `geom_smooth(method = lm)`

- A suavização de smooth descreve uma tendência (uma variabilidade), apresentando maior ou menor confiabilidade (área cinza do gráfico, argumento `se` da função `geom_smooth()`) da estimativa de tendência, dependendo da quantidade de pontos no local.

Ex.: `geom_smooth(se = FALSE)`

- O argumento `span` controla a “ondulação”. Números pequenos fazem a curva mais sinuosa ($0 > x > 1$), enquanto números maiores (~ 1) fazem a curva mais suave.

Ex.: `span = 0.3`

- Gráfico de linha x gráfico de tendência:

– Os gráficos se assemelham ao passo que são gerados a partir do gráfico de pontos (`geom_point`).

`geom_point() + geom_line()`
`geom_point() + geom_smooth()`

– Diferente do gráfico de linhas, o gráfico de tendência não liga os pontos. O gráfico de tendência gera uma curva suavizada que melhor se ajuste aos pontos.

- Exemplo - Gráfico de pontos com ajuste por curva de tendência por ggplot2 (`geom_point()`+`geom_smooth()`):

```
#Plotagem
ggplot(subset(dados, Estado %in% c("SaoPaulo")), #Subconjunto de dados de um data.frame
       #O comando filtra dos dados do data.frame
       aes(x = cheg_2012/1000,
           y = cheg_2013/1000,
           color = Estado))+ #Diferencia os grupos das variáveis regiao por cor
       scale_x_continuous(limits = c(140,230), #Limites do eixo x
                           breaks = seq(140,230,10))+ #Espaçamento do eixo x
       geom_point(size = 1.5)+ #Tamanho dos ícones (Pontos)
       geom_smooth(size = 1.0, span = 0.7, se = TRUE)+ #Espessura da curva (size = 1.0)
       #Curva mais suave (span = 0.7)
       #Apresenta confiabilidade estimada (se = TRUE)
       labs(title = "Gráfico com ajuste de curva de tendência: 2012 x 2013",
            x = "Chegadas por mil em 2012",
            y = "Chegadas por mil em 2013")+
       theme_bw(base_size = 18)+ #Adiciona tema "black and white" e tamanho da fonte
       scale_color_grey() #Aplica escalas de cinza

#Fechando dispositivo gráfico
dev.off()
```

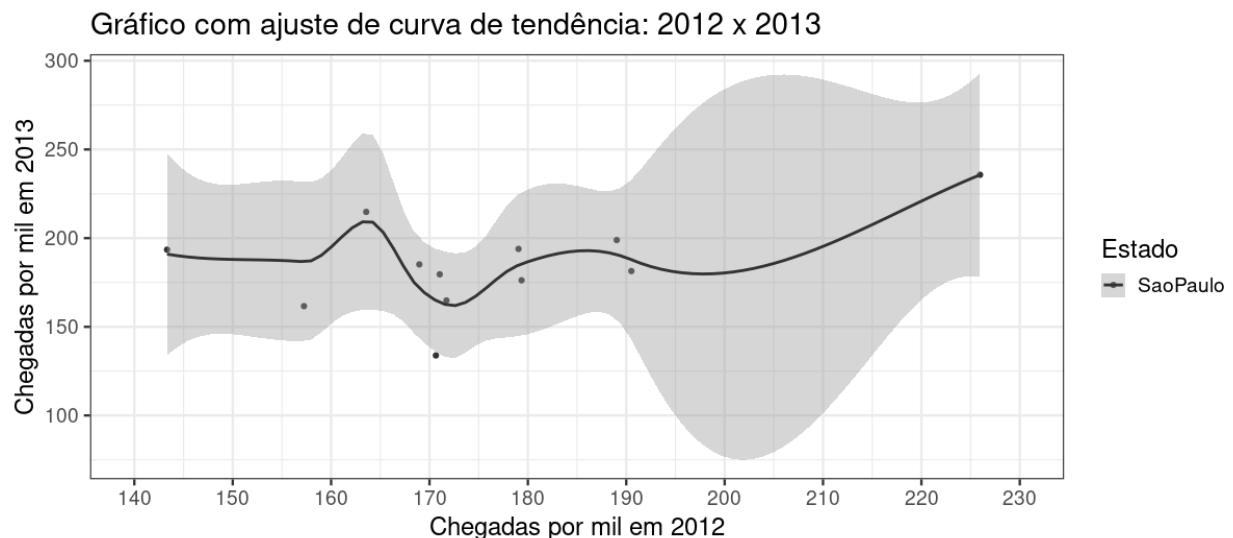


Figure 43: Gráfico de pontos com ajuste com curva de tendência suavizada smooth, com `span = 0.7`.

10.2.11.8 Gráfico de dispersão com linha de tendência com ggplot2

- A regressão linear calcula uma equação que minimiza a distância entre a linha ajustada e todos os pontos dos dados.
- Quando realizamos um ajuste por regressão linear observamos o gráfico de dispersão, o coeficiente de correlação e o valor de R^2 (coeficiente de determinação).

10.2.11.8.1 Coeficiente de reta de regressão

- Regressão linear tenta traçar uma reta que melhor aproxime todos os pontos dispersos.

$$y = A + Bx$$

Onde,

A é o intercepto

B é o coeficiente angular.

- Coeficiente angular:

$$B = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

- Intercepto:

$$A = \frac{\sum y - B \sum x}{n}$$

- `lm(y ~ x)$coef`

Esta função do **R** retorna os coeficientes da reta de regressão (**intercepto** e **coeficiente angular**). A parte da função `$coef` apenas retorna de maneira mais direta os coeficientes separados, assim deixando claro em cada coluna o que é **intercepto** e o que é **coeficiente angular**.

10.2.11.8.2 Coeficiente de Correlação linear

- O coeficiente de correlação tem o objetivo de entender como uma variável se comporta num cenário onde a outra variável variando. E se existe alguma relação entre a variabilidade de ambas as variáveis.
- Os coeficientes variam de 1 a -1. Quanto mais próximo dos extremos, mais forte é a relação entre as variáveis. Quanto mais próximo do centro 0, menor é a relação entre as variáveis. Em 0 não existe relação entre as variáveis.
- A correlação próximo do valor 1, significa que a relação é positiva, ou seja, a reta de regressão é ascendente. Quando uma variável aumenta a outra aumenta também.
- A correlação próximo do valor -1, significa que a relação é negativa, ou seja, a reta de regressão é descendente. Quando uma variável diminui a outra aumenta.
- Cálculo de correlação linear:

$$\text{cor}_{x,y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

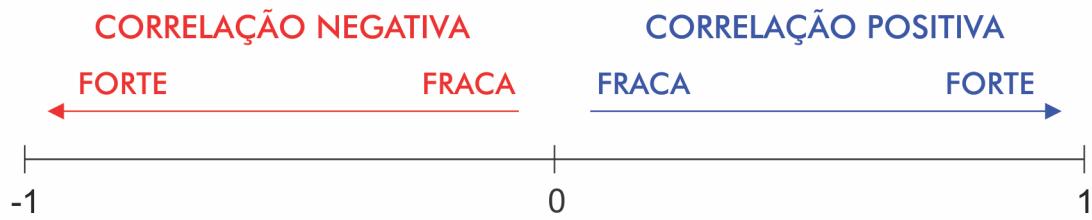


Figure 44: Correlações fortes e fracas

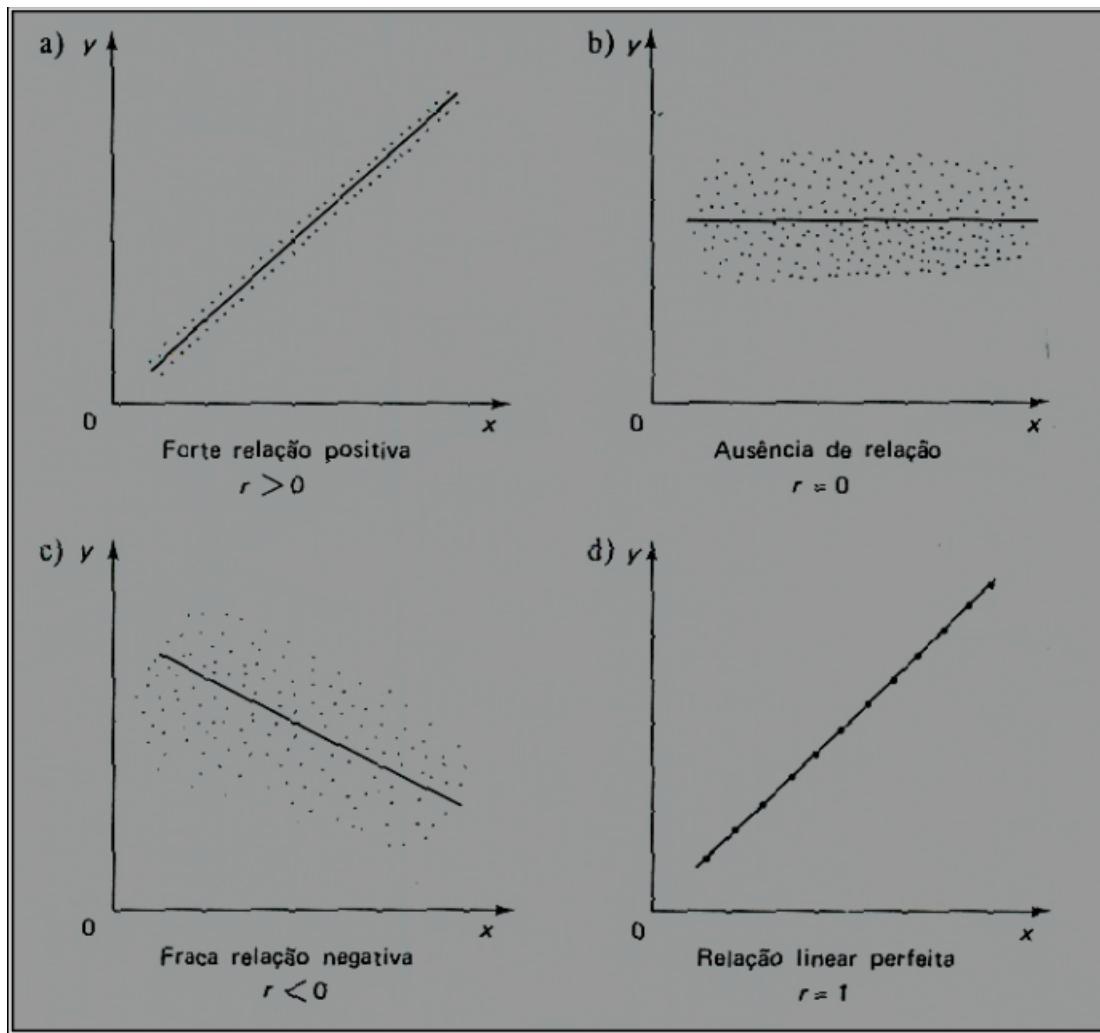


Figure 45: Tipos de Correlação

Onde,

n é o número de registros/linhas.

x_i é o vetor x.

y_i é o vetor y.

xy é x vezes y.

Uma forma rápida e simples de resolver o cálculo é preencher a tabela de correlação linear com as informações:

	x	y	xy	x^2	y^2
n					
Σ					

Figure 46: Tabela de correlação linear

- $\text{cor}(x,y)$

Função do **R** que calcula a correlação linear das variáveis vetor x e y.

10.2.11.8.3 Coeficiente de determinação (R^2)

- O R^2 (coeficiente de determinação) é um indicador da qualidade do ajuste, que varia de [0,1], e indica a porcentagem da variabilidade de y é explicada pela variabilidade de x.
 - 0%
Indica que o modelo não explica nada da variabilidade dos dados de resposta ao redor de sua média.
 - 100%
Indica que o modelo explica toda variabilidade dos dados de resposta ao redor de sua média.
Teoricamente se um modelo pudesse explicar 100% da variância, os valores ajustados seriam sempre iguais aos valores observados e, portanto, todos os pontos de dados cairiam na linha de regressão ajustada.
- Valores baixos de R^2 não são necessariamente ruins, em algumas áreas é esperado que seus valores sejam baixos.
Ex.: Áreas de previsão de comportamento humano, normalmente $R^2 < 50\%$.

10.2.11.8.4 Gráfico de dispersão com linha de tendência

- Principais argumentos do gráfico de dispersão com linhas de tendência:
 - `geom_point()`
Adiciona a camada de gráfico de pontos (os pontos).
 - `geom_smooth()`
 - * `method = lm`
Adiciona uma função de modelagem.
Uma comum usada é de regressão linear `lm`.
 - `sprintf()`
Chama uma função de C `printf()`, que retorna um vetor de caracteres, contendo uma combinação formatada de texto e valores de variáveis.
 - `summary(model)`
Retorna resumos matemáticos da classe do argumento.
 - * `summary(model)$coefficients[2]`
Retorna o **intercepto** (A) da reta de regressão linear ($y = A + Bx$).
 - * `summary(model)$coefficients[1]`
Retorna o **coeficiente angular** (B) da reta de regressão linear ($y = A + Bx$).
 - * `summary(model)$r.squared`
Retorna o **coeficiente de determinação** (R^2) da regressão linear.
 - * `cor(y,x)`
Retorna o **coeficiente de correlação** da regressão linear.

- Exemplo - Gráfico de dispersão com linha de tendência (**regressão linear**) por `ggplot2` (`geom_point() + geom_smooth(method = "lm")`):

```
#Plotagem
ggplot(df, aes(x, y, color = ""))+
  geom_point()+
  geom_smooth(method = "lm") #Método lm, Regressão linear
  labs(title =
    sprintf("Regressão linear\nR-quadrado = %1.3f\nEquação: %1.2fX+%1.2f\nCoeficiente de correlação = %1.2f",
           summary(model)$r.squared, #R-quadrado
           summary(model)$coefficients[2], #coeficiente A de Y=AX+B
           summary(model)$coefficients[1], #coeficiente B de Y=AX+B
           cor(y,x)),#Coeficiente de correlação linear
  color = #legenda color (reta)
  sprintf("Equação: %1.2fX+%1.2f",
         summary(model)$coefficients[2], #coeficiente A de Y=AX+B
         summary(model)$coefficients[1]))+ #coeficiente B de Y=AX+B
  #Função de C printf, que retorna um vetor de caracteres,
  #contendo uma combinação de formatada de texto e valores de variáveis
  #"\n" = Próxima linha
  #"%1.2f" = "%" imprimir argumentos,
  # "1" tamanho mínimo do campo,
  # "2" número de casas decimais e
  # "f" ponto flutuante.
  scale_color_grey() #Adiciona escala de cinza na variável color
  scale_fill_grey() #Adiciona escala de cinza na variável fill
  theme_bw(base_size = 18) #Adiciona tema "black and white" e tamanho da fonte
  theme(plot.title = element_text(hjust = 0.5)) #Alinhamento horizontal do título

#Fechando dispositivo gráfico
dev.off()
```

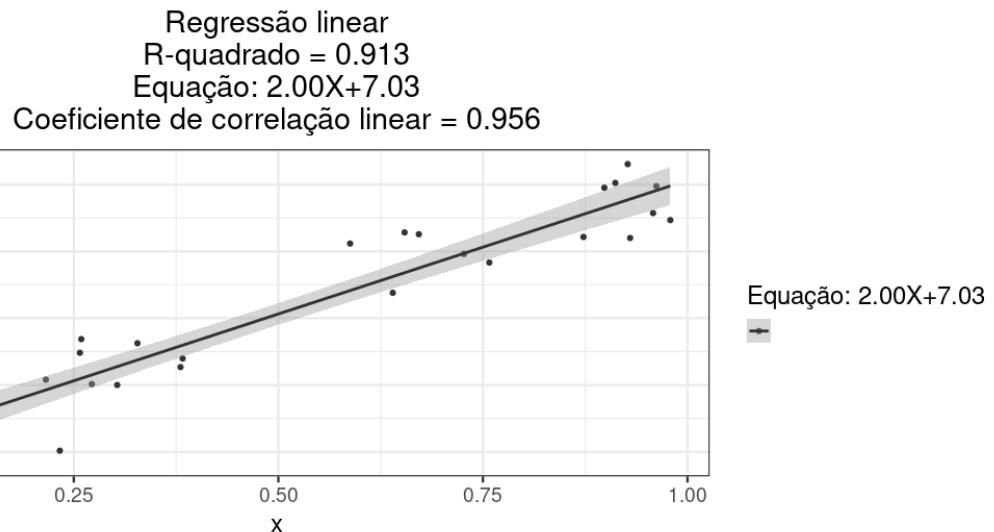


Figure 47: Gráfico de dispersão com linha de tendência (**regressão linear**).

10.2.11.9 Efeitos

10.2.11.9.1 O efeito jitter

- No gráfico de pontos ou dispersão, quando diversas observações (pontos) apresentam o mesmo valor, na visualização convencional, não é possível perceber esse fato.
- O feito jitter estabelece uma forma de evidenciar essas repetições. É um efeito que mostra os pontos sobrepostos num resultado de espalhamento em torno do ponto de sobreposição, permitindo que se visualize melhor a quantidade de ocorrências.
- Adiciona uma pequena variação aleatória à localização de cada ponto, é uma maneira útil de lidar com *overplotting* causada pela discrepância em conjuntos de dados menores.
- *Overplotting* é quando os dados ou rótulos em uma visualização de dados se sobrepõem, dificultando a visualização de pontos de dados individuais em uma visualização de dados.
- Comando para adicionar o efeito jitter no gráfico, usando a biblioteca **ggplot2**: `geom_jitter()`

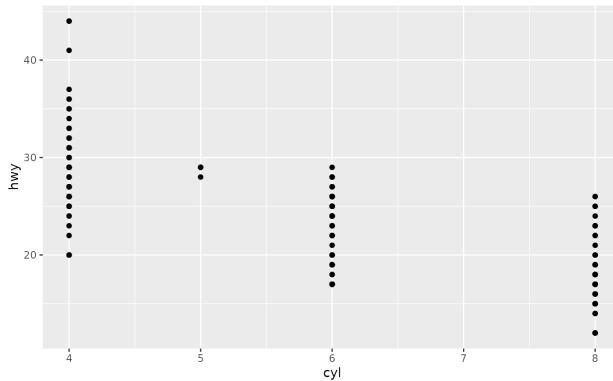


Figure 48: Gráfico sem efeito jitter.

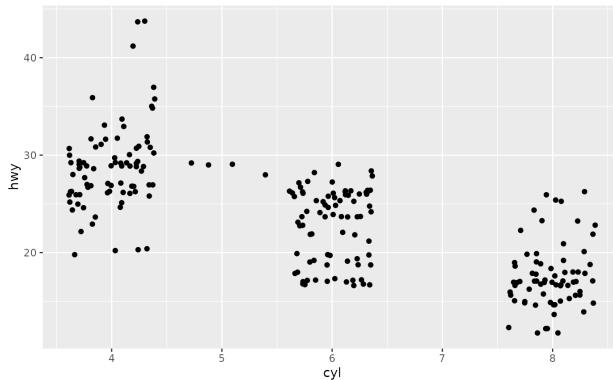


Figure 49: Gráfico com feito jitter.

10.2.11.9.2 Facetas

- Divide o gráfico em vários painéis.
- O `facet_grid()` forma uma matriz de painéis definidos por variáveis de facetação de linha e coluna.
- É mais útil quando se tem duas variáveis discretas e todas as combinações das variáveis existem nos dados.
- Caso tenha apenas uma variável com muitos níveis, vale tentar `facet_wrap()`.

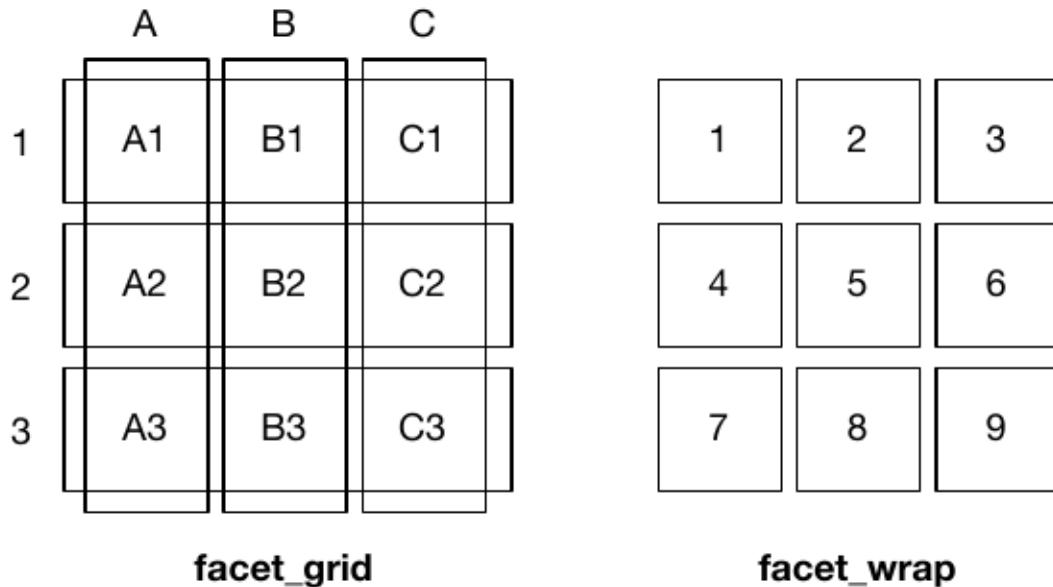


Figure 50: Diferença entre `facet_grid()` e `facet_wrap()`.

- Principais argumentos do `facet_grid()`:

- **rows ou cols**

O conjunto de variáveis `var()`. Definição de grupo de facetas na dimensão de linhas e colunas.

Exemplo: `row = var(variável_de_agrupamento)`

Outra forma de escrever é através da expressão `variável_linha ~ variável_coluna`, onde qualquer variável pode ser substituída por ponto (`variável_linha ~ .` ou `. ~ variável_coluna`), que é usado para indicar não haver lapidação nessa dimensão.

Exemplo: `Regiao ~ .`

- **nrow ou ncol**

Funciona para o `facet_wrap()`, controla como a feixa de opções é agrupada em uma grade, ou seja, controla o número de colunas e linhas, assim o como as facetas estão dispostas numa tabela. Exemplo: `nrow = 3`

- **scale**

Por padrão, todos os painéis tem as mesmas escalas (`scale = "fixed"`). Eles podem se tornar independente, definindo as escalas como `free_x` (variando no eixo x), `free_y` (variando no eixo y) ou `free` (variando em ambos os eixos).

Exemplo: `scale = "free"`

- **space**

Se `fixed`, o padrão, todos os painéis tem o mesmo tamanho.

Se `free_y` sua altura será proporcional ao comprimento da escala y.

Se `free_x` sua largura será proporcional ao comprimento da escala x.

Ou se `free` tanto a altura quanto a largura vão variar em função das escalas x e y.

Exemplo: `space = "free"`

- **labeller**

O argumento `labeller` pode ser usado para controlar o rótulo dos painéis.

Exemplo: `labeller=label_both`

Adiciona o nome das variáveis nos rótulos, não somente os valores.

- Exemplos - `facet_grid()` alterando parâmetros:

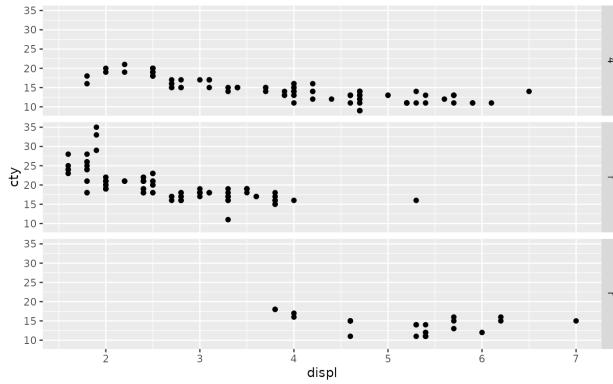


Figure 51: Gráfico por facetas com `facet_grid(rows = variável)` ou `facet_grid(variável ~ .)`.

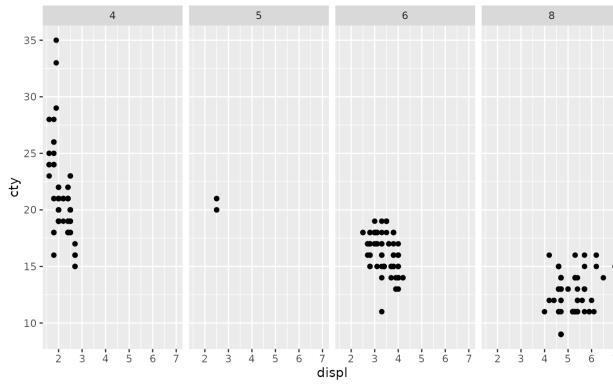


Figure 52: Gráfico por facetas com `facet_grid(cols = variável)` ou `facet_grid(.~ variável)`.

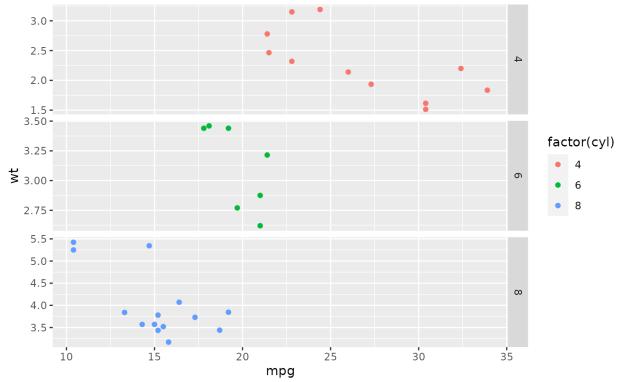


Figure 53: Gráfico por facetas com escala do eixo y livre. `facet_grid(scale = "free_y")`

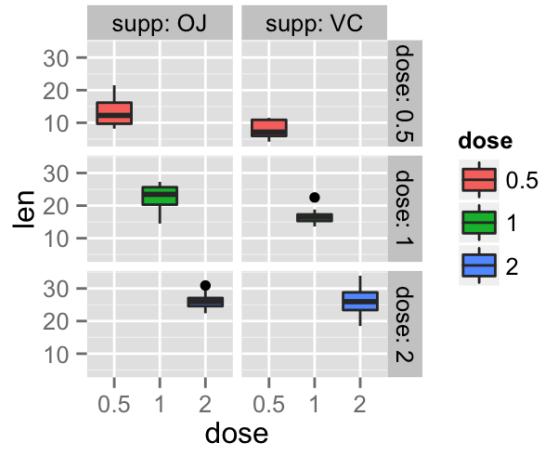


Figure 54: Gráfico por facetas com rótulos com nome da variável. `facet_grid(labeler = label_both)`

10.2.11.9.3 O efeito de suavização smooth

- Trata-se de um ajuste de curva de tendência entre duas variáveis numéricas.
- Usa o modelo de regressão local para gerar uma curva suavizada que melhor se ajuste aos pontos.
- Na regressão local, estima-se uma função na vizinhança de cada ponto de interesse.
- O efeito de suavização smooth é gerado na biblioteca `ggplot2` pelas funções (`geom_point() + geom_smooth()`) com alguma semelhança ao gráfico de linhas, porém ao invés de ligar os pontos gera uma curva que se ajuste de maneira aproximada aos pontos.
- A suavização de smooth descreve uma tendência (uma variabilidade), apresentando maior ou menor confiabilidade (área cinza do gráfico, argumento `se` da função `geom_smooth()`) da estimativa de tendência, dependendo da quantidade de pontos no local.
- Podemos adicionar outros métodos de modelagem para uma curva de suavização, através do argumento `method`. Um método comum de se utilizar é a **regressão linear** (`lm`).
Ex.: `geom_smooth(method = lm, se = FALSE)`
- Principais argumentos da função `geom_smooth()`:
 - `orientation`
Define a orientação, define se o ajuste deve ser feito ao longo do eixo y ao invés do eixo x.
Ex: `orientation = "y"`
 - `span`
Use `span` para controlar a “ondulação”.
Números pequenos fazem a curva mais sinuosa ($0 > x > 1$), enquanto números maiores (~ 1) fazem a curva mais suave.
Ex.: `span = 0.3`
 - `method`
Adiciona uma função de modelagem.
Uma comum usada é de regressão linear `lm`.
Ex.: `method = lm, se = FALSE`
 - `se`
Especifica se mostra uma área de confiança em torno da linha suave (área cinza).
Por default é `True`, mostrando a área.
Caso `False`, omite a área.
Ex.: `se = FALSE`

- Exemplos - Gráficos de tendência usando a função de efeito de suavização smooth (`geom_point() + geom_smooth()`):

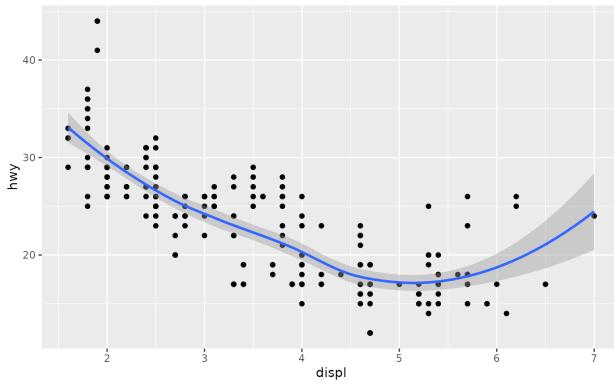


Figure 55: Gráfico de tendência com curva suavizada smooth, área cinza indicador de confiabilidade.

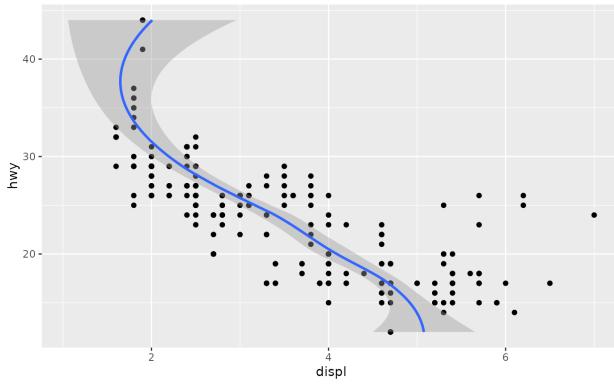


Figure 56: Gráfico de tendência com curva suavizada smooth, orientada no eixo y `orientation = "y"`.

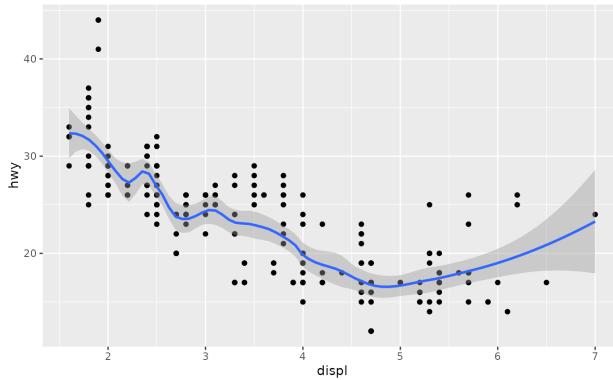


Figure 57: Gráfico de tendência com curva suavizada smooth, com `span = 0.3` (mais sinuoso).

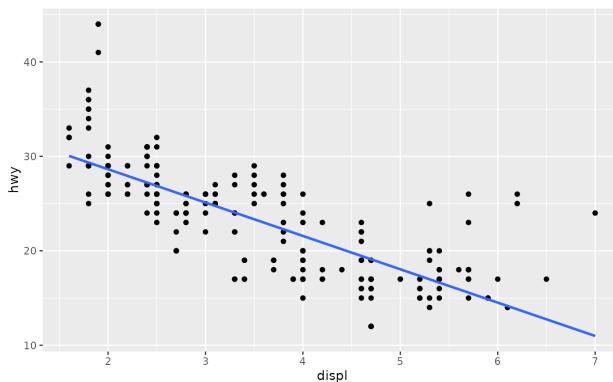


Figure 58: Gráfico de tendência com curva suavizada smooth, com `method = lm`, `se = FALSE` (regressão linear com área de confiança omitida).

10.2.12 Assistentes para ggplot2 - esquisse

10.2.12.1 Pacotes auxiliares ao ggplot2

- **ggThemeAssist**

Fornece uma interface gráfica (Addins) para editar os elementos do tema ggplot2.

- **esquisse**

- Pacote para criação de gráficos (ggplot2) de maneira *point and click*.

- Assistente gráfico, para criação de gráficos rápidos e simples.

- Não tem todo o poderio do ggplot2, porém apresenta facilidades.

- **hrbrthemes**

Uma compilação de temas, escalas e utilitários extras de ggplot2, incluindo uma função de verificação ortográfica para campos de rótulos de plotagem.

- **ggthemes**

Temas adicionais para gráficos ggplot2.

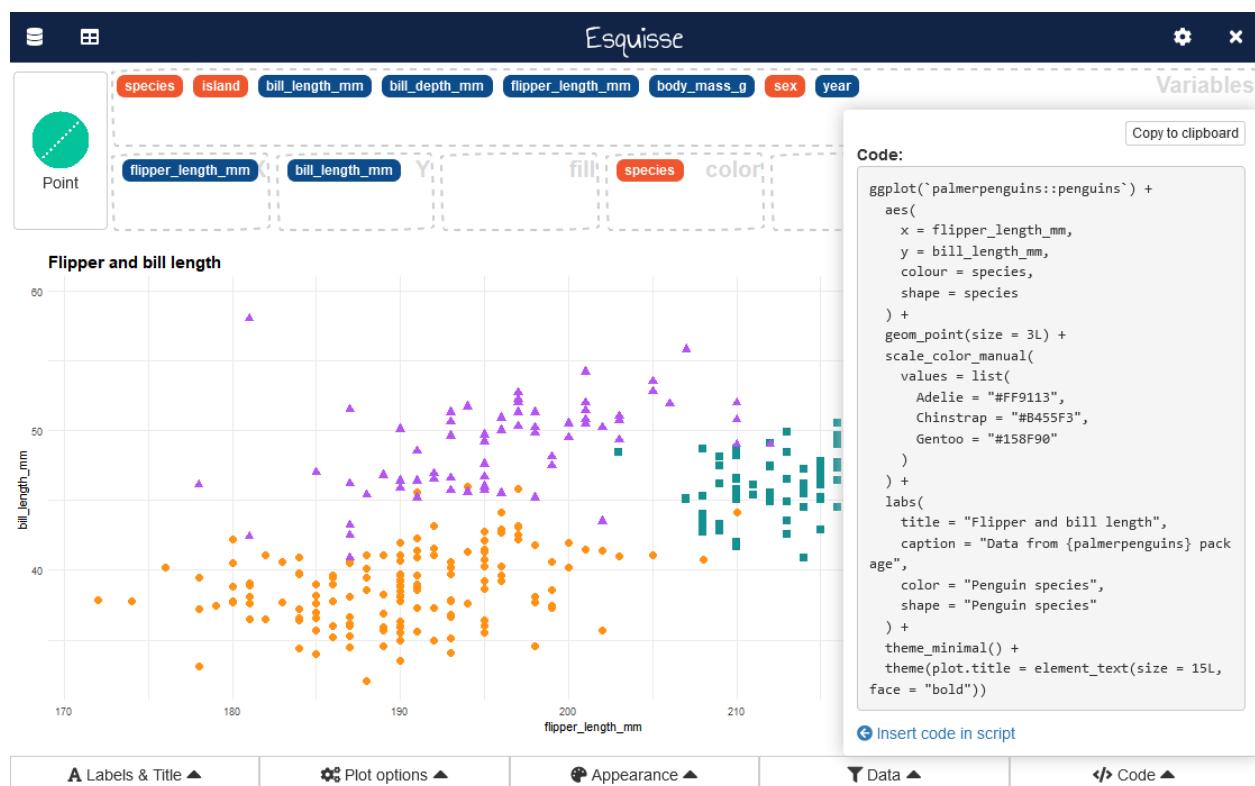


Figure 59: Pacotes auxiliares (ggThemeAssist e esquisse) de construção de gráfico - ggplot2 builder

10.2.12.2 Referências e modelos de assistente gráfico Modelos do pacote ggplot2:

<https://exts.ggplot2.tidyverse.org/gallery/>

10.2.12.3 Funções do assistente de gráficos `ggplot2` - `esquisse`

- Abrir assistente gráfico:
 - `esquisse()`
Chama o assistente gráfico, abre no menu do **RStudio**.
 - `esquisse(viewer = "browser")`
Abrir no browser.
 - `esquisse(viewer = "pane")`
Abrir no viewer.
- `esquisse(dados)`
Passa dados para o assistente gráfico.
Mesmo sem passar os dados para o assistente gráfico, o programa pega os dados na memória do sistema.

10.2.12.4 Assistente gráfico esquisse

- Ao instalar o pacote **esquisse**, ele fica disponível no addins do **RStudio**.

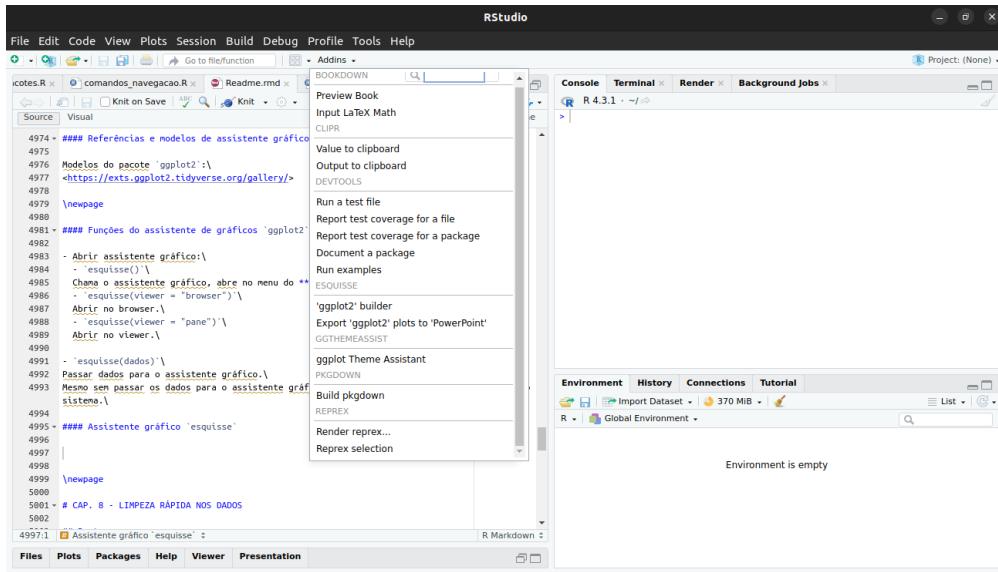


Figure 60: Abrir o **esquisse** a partir do addins no **RStudio**.

- Dentro do assistente gráfico **esquisse**, a primeira aba é relativa a inserir textos no gráfico (título, subtítulo, legenda e rótulos).

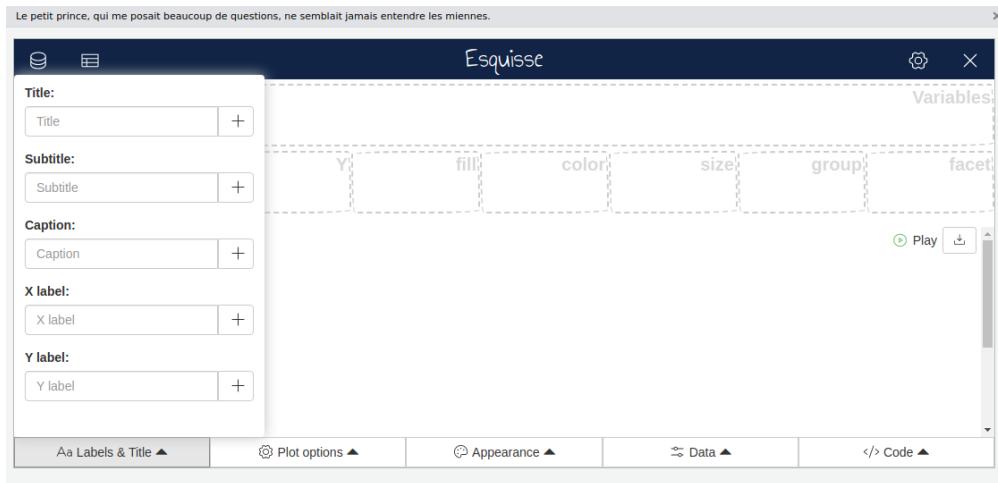


Figure 61: No **esquisse**, aba para editar textos do gráfico (título, subtítulo, legenda e rótulos).

- A segunda aba são opções de personalização do gráfico, variando de gráfico para gráfico (número de linhas).

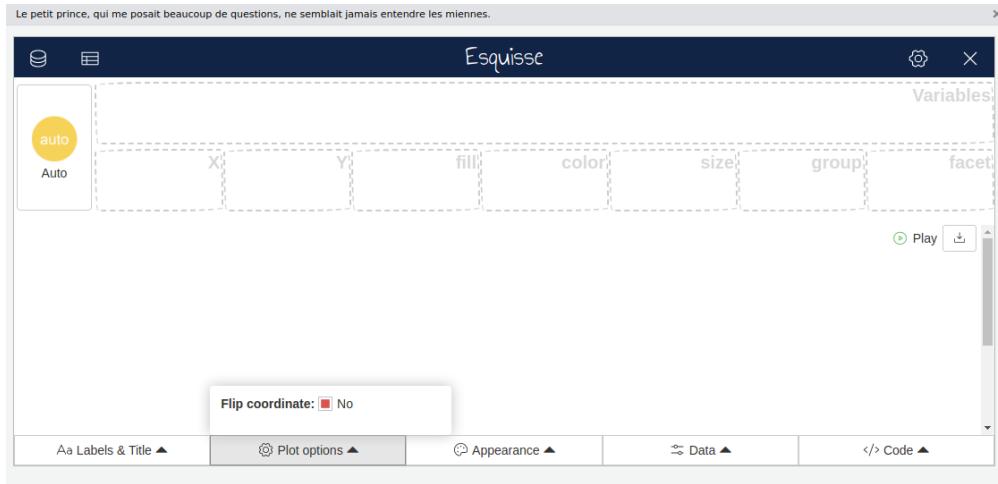


Figure 62: No **esquisse**, aba para editar opções do gráfico.

- Na aba appearance podemos mudar os temas e cores do gráfico.

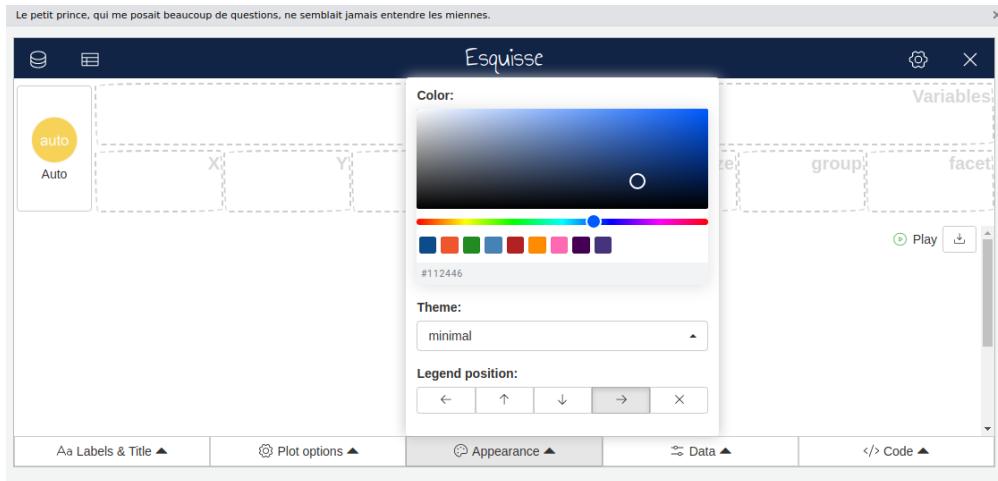


Figure 63: No **esquisse**, aba para editar aparência do gráfico.

- Na aba *data*, podemos aplicar filtros no *dataset*.

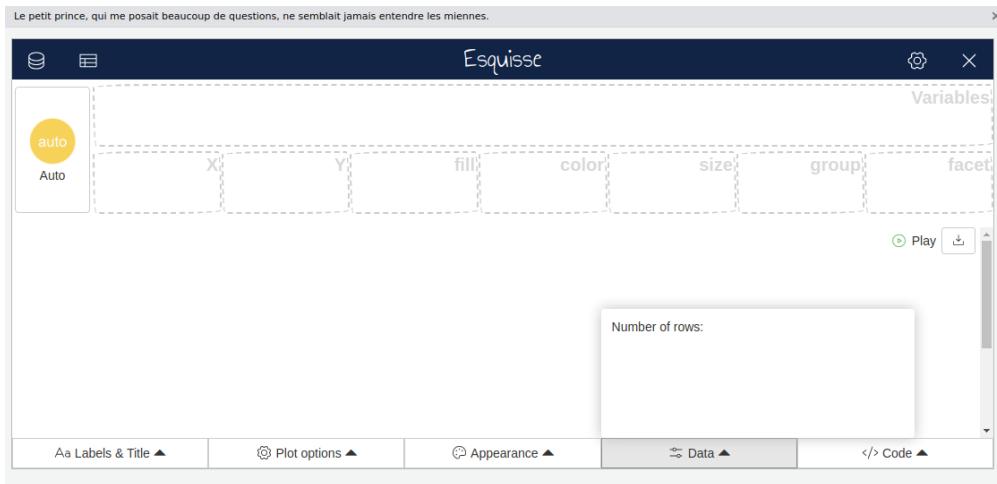


Figure 64: No **esquisse**, aba para exportar o gráfico (seja o código ou o .png).

- Por fim, na aba *code*, podemos exportar o gráfico, tanto o código de criação dele para replicar num *script*, ou o .png.

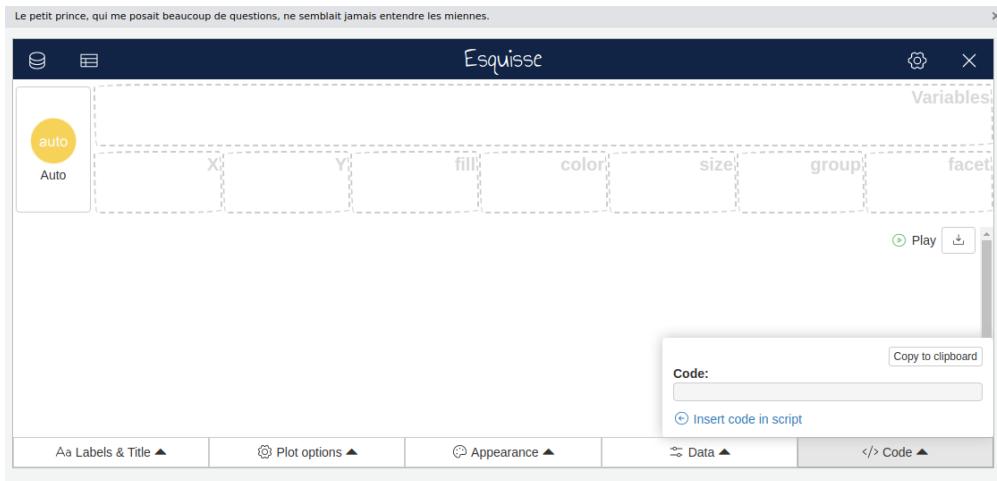


Figure 65: No **esquisse**, aba para exportar o gráfico (seja o código ou o .png).

11 CAP. 8 - LIMPEZA RÁPIDA NOS DADOS

11.1 Pactoes

- `janitor`
Projetado para inspeção e limpeza de dados “sujos”.

11.2 Dados “sujos”

- Dados que podem apresentar diversos problemas ao utilizar dados abertos ou quando várias pessoas digitaram os dados.
- Registros que necessitam de ajustes antes de sua análise.

11.3 Principais funções janitor

11.3.1 Limpando nomes do `data.frame` - `clean_names()`

- Manipulação de nomes problemáticos de variáveis (`clean_names()`).
 - O que a função `clean.names()` faz?
 - Retorna nomes somente com letras em caixa baixa e com “_” como separador.
 - Manipula caracteres especiais e espaços.
 - Inclui números para nomes duplicados.
 - Converte o símbolo “%” para “percent” preservando o sentido.
 - Existe um função no pacote básico do R que também faz limpeza de nomes (corrigem nomes) `make.names()`. Porém ela é muito básica, não sendo assim a solução ideal:
 - Elimina espaços e substitui por pontos.
 - Substitui símbolos por pontos.
 - Exemplo - Limpando nomes do `data.frame`:

```
#Dados
dfp <- as.data.frame(matrix(ncol = 6))
names(dfp) <- c("OriGem",
                 "REPETE",
                 "REPETE",
                 "% de acertos",
                 "R!$@$$*&*",
                 "")

dfp
#OriGem REPETE REPETE % de acertos R!$@$$*&
#1      NA      NA      NA          NA      NA NA

#Limpando nomes
clean_names(dfp)
#ori_gem repete repete_2 percent_de_acertos r  x
#1      NA      NA      NA          NA NA NA

#Comparando com a função básica do R make.names
make.names(names(dfp))
#[1] "OriGem"           "REPETE"           "REPETE"           "X..de.acertos"
#[5] "R....."          "X"
```

11.3.2 Remova colunas ou linhas inútes

- Funções:

- `remove_constant()`

A função `remove_constant()` remove as colunas constantes.

- `remove_empty()`

A função `remove_empty()` remove valores vazios (*NA*), tanto colunas quanto linhas.
Sem especificar remove colunas e linhas.

- * `which = "rows"`

remover linhas vazias.

- * `which = "cols"`

remover colunas vazias.

- Exemplos - Eliminando colunas e/ou linhas com `remove_constant()` e `remove_empty()`:

```
#Dados
x <- c("b","a","b","c","c",NA,"a","a",NA,"a")
y <- rep("Brasil",10)
z <- c(NA,1:7,NA,NA)
vazia <- rep(NA,10)

#Criando dataframe
dt <- data.frame(x,y,z,vazia)
dt

#Elimina coluna com valores constantes - remove_constant()
#Remove colunas y e vazia
dt_clean1 <- remove_constant(dt)
dt_clean1
      x   z
1     b NA
2     a  1
3     b  2
4     c  3
5     c  4
6 <NA>  5
7     a  6
8     a  7
9 <NA> NA
10    a NA

#Eliminando as linhas vazias - remove_empty()
#Remove linha 9, sem valores
dt_clean2 <- remove_empty(dt_clean1)
dt_clean2
      x   z
1     b NA
2     a  1
3     b  2
4     c  3
5     c  4
6 <NA>  5
7     a  6
8     a  7
10    a NA
```

11.3.3 Substitua valores perdidos/faltantes - `mice()`

- Apesar do `janitor` auxiliar a eliminar linhas e colunas com valores perdidos, caso necessite substituir tais valores, o pacote `mice` ajuda nessa tarefas, usando técnica de imputação de valores (*Multivariate Imputation by Chained Equations*).

```
install.packages("mice")
```

- A técnica `complete(mice())`, para substituição de valores, levam em conta o tipo de variável, produzindo a substituição dos valores pedidos.

```
complete(mice(data.frame))
```

– O tipo de valores `character` não são apropriados para serem substituidos. O ideal, se possível, é passar as variáveis `character` para `factor`.

– Os tipos de valores mais apropriados para serem substituidos são `numeric` e `factor`.

- Boas práticas ao aplicar substituição de valores perdidos/faltantes (`complete(mice())`):

– Antes aplicar a técnica de substituição de valores perdidos, é interessante fazer uma análise do `data.frame` para entender profundamente a situação:

* `str()`

Para entender a estrutura dos dados envolvidos (tipos de dados).

* `summary()`

Obter um resumo estatístico para conhecer de modo geral melhor o `data.frame`.

– Ao final da aplicação das funções `complete(mice())`, outra boa prática é obter o resumo estatístico (`summary()`) para observar e comparar a extensão dos impactos das substituições dos valores perdidos/faltantes para o conjunto dos dados.

- Exemplo - Substituição de valores perdidos/faltantes (`complete(mice())`):

```
#mice para substituição de valores perdidos/faltantes
#Analise estatística
dt
str(dt)
summary(dt)

#Substituindo valores perdidos com mice
dt_ajustado = complete(mice(dt))
dt_ajustado

#Nova analise estatística
summary(dt_ajustado)
```

11.3.4 Produzindo tabelas de frequência para uma variável - `taby()`

- A função `taby()`, do pacote `janitor`, é uma versão melhorada da função `table()`, do pacote base do R.
- Diferenças:
 - Retorna `data.frame` que pode ser melhorado e impresso com `kable()`, do pacote `knitr`.
 - Calcula porcentagens automaticamente (porcentagem total e porcentagem valida, sem **NA**).
 - Pode opcionalmente exibir valores **NA**.
 - Quando **NA** ocorre, uma coluna adicional `valid_percent` (porcentagem valida) é adicionada.
 - Pode opcionalmente ordernar as contagens (frequência).
 - Pode ser usado com operador pipe `%>%`, do pacote `magrittr`.
 - Quando a variável for do tipo categórica, os valores perdidos são contabilizados na tabela.

Km rodados por veículo	Frequência Absoluta f_i	Frequência relativa f_{ri} (%)	Frequência acumulada F_{ac}	Frequência Acumulada Relativa F_{Ri} (%)
20 -> 29	3	14%	3,3	14%
29 -> 38	6	25%	9,3	38%
38 -> 47	6	25%	15,3	63%
47 -> 56	4	16%	19,3	79%
56 -> 65	5	21%	24,3	100%
Total	24	100%	---	---

Figure 66: Exemplo tabela de frequência de uma variável.

- Exemplo - Produzindo tabelas de frequência para uma variável (`taby()`):

```
#Dados
x <- c("b","a","b","c","c",NA,"a","a",NA,"a")
y <- rep("Brasil",10)
z <- c(NA,1:7,NA,NA)
vazia <- rep(NA,10)

#Tabela de frequência da variável x
tabyl(x, sort = TRUE)
x n percent valid_percent
a 4      0.4          0.50
b 2      0.2          0.25
c 2      0.2          0.25
<NA> 2      0.2          NA

tabyl(x) %>%
  select(x, valid_percent)
x valid_percent
a          0.50
b          0.25
c          0.25
<NA>        NA
```

11.3.5 Tabulação cruzada - tabyl()

11.3.5.1 Tabulação cruzada

- As tabelas de tabulação cruzada (tabelas de contigência) exibem o relacionamento entre duas variáveis categóricas (nominais ou ordinais). O tamanho da tabela é determinado pelo número de valores distintos para cada variável, com cada célula na tabela representando uma combinação exclusiva de valores.
- Exemplo didático:
Numa tabela onde temos duas colunas (x e y), a tabulação cruzada é a incidência (frequência) de ocorrências de x em y.
Logo, para tornar uma tabela em tabulação cruzada o x mantém-se coluna e os valores de y passam a ser colunas, os valores da tabulação cruzada são as frequências das combinações.

	CA	AZ	NY	Azul	Verde	Vermelho
Alto lucro	1	0	1	2	0	0
Médio lucro	1	0	1	0	1	1
Baixo lucro	0	2	0	0	1	1

Tabela 1 – Exemplo de Tabela Cruzada.

Figure 67: Exemplo de tabulação cruzada

11.3.5.2 Função tabyl() para tabulação cruzada

- Uma tabulação cruzada é gerada com a função `tabyl()`.
- Propriedades:
 - Retorna um `dataframe`.
 - Calcula frequências absolutas, mas é possível incluir frequências relativas por linha ou coluna.
 - Também é possível incluir as frequências em forma de porcentagem.
 - Pode (opcionalmente) mostrar os valores `NA` (“`show_na = F`” exclui `NA`).
`Ex.: tabyl(dt,x,y, show_na = F)`
- A função `tabyl()`, do pacote `janitor`, produz resultado que só seria possível através de um conjunto de funções do sistema `tidyverse` (`dplyr` e `tidyr`).
`Ex.: group_by %>% summarise %>% mutate %>% spread`

- O pacote `janitor` tem um conjunto de funções para *adornar* a tabulação:

- `adorn_totals`

Adiciona o total por linha (`where = "row"`) ou coluna (`where = "col"`) ou por ambos.
Ex.: `tabyl(dt,col1,col2) %>% adorn_totals(.)`

- `adorn_percentages`

Calcula porcentagens com base nos totais de linha (“`row`”), de coluna (“`col`”) ou sobre o total geral da tabela (“`all`”).
Ex.: `adorn_percentages("row")`

- `adorn_pct_formatting`

Formata as porcentagens, controlando o número de dígitos a serem exibidos e incluindo o símbolo de “%”.

- `adorn_rounding`

Produz arredondamento nos números da tabela, não deve ser usado em conjunto com `adorn_pct_formatting`.

Ex.: `adorn_rounding(digits = 0, rounding = "half up")`

Métodos de arredondamento (`rounding`):

- * `half up`

Arredonda para cima.

Ex.: $10.5 \rightarrow 11$

- * `half to even`

Arredonda para baixo.

Ex.: $10.5 \rightarrow 10$

- `adorn_ns`

Adiciona as contagens (frequência absoluta).

- `adorn_title`

Adiciona título a tabela, podendo ser no topo da tabela ou combinado na primeira linha/coluna. A opção “`top`” adiciona o nome da coluna numa linha vazia (dificulta manipulação posterior). Outra opção é “`combined`” anexando a variável de nome na linha já presente.

Ex.: `adorn_title("combined", row_name = "Cor de selo", col_name = "Categorias")`

- Exemplos - Pacote janitor, função tabyl() para tabulação cruzada com adornos:

```
#Tabulação cruzada entre x e y
#frequência de y em x
tabyl(dt,x,y)
  x A B C NA_
  azul 2 0 0   1
  preto 1 2 0   1
  vermelho 1 1 1   0
  <NA> 0 0 1   0

#Tabulação cruzada entre x e y com adornos
tabyl(dt,x,y) %>%
  adorn_percentages("row") %>% #Inclui porcentagens com base na linha
  adorn_pct_formatting() #Formata para %
    x      A      B      C      NA_
    azul 66.7% 0.0% 0.0% 33.3%
    preto 25.0% 50.0% 0.0% 25.0%
    vermelho 33.3% 33.3% 33.3% 0.0%
    <NA> 0.0% 0.0% 100.0% 0.0%

tabyl(dt,x,y) %>%
  adorn_percentages("col") %>% #Inclui porcentagens com base na coluna
  adorn_pct_formatting() #Formata para %
    x      A      B      C      NA_
    azul 50.0% 0.0% 0.0% 50.0%
    preto 25.0% 66.7% 0.0% 50.0%
    vermelho 25.0% 33.3% 50.0% 0.0%
    <NA> 0.0% 0.0% 50.0% 0.0%

tabyl(dt,x,y) %>%
  adorn_percentages("all") %>% #Inclui porcentagens com base no total geral
  adorn_pct_formatting() #Formata para %
    x      A      B      C      NA_
    azul 18.2% 0.0% 0.0% 9.1%
    preto 9.1% 18.2% 0.0% 9.1%
    vermelho 9.1% 9.1% 9.1% 0.0%
    <NA> 0.0% 0.0% 9.1% 0.0%
```

- Exemplos - Pacote `janitor`, função `tabyl()` para tabulação cruzada com adorno `adorn_title` para títulos e exclusão de **NA**:

```
tabyl(dt,x,y, show_na = F) %>% #Exclui NA
  adorn_percentages("all") %>% #Inclui porcentagens com base no total geral
  adorn_rounding(2) %>% #Arredonda com 2 casas decimais
  adorn_ns() %>% #Inclui contagens
  adorn_title("combined",
    row_name = "Cor de selo",
    col_name = "Categorias") #Adiciona títulos

Cor de selo/Categorias      A      B      C
  azul 0.25 (2) 0.00 (0) 0.00 (0)
  preto 0.12 (1) 0.25 (2) 0.00 (0)
  vermelho 0.12 (1) 0.12 (1) 0.12 (1)
```

- Exemplos - Pacote `janitor`, função `tabyl()` para tabulação cruzada com adorno `adorn_rounding` para arredondamentos:

```
#Arredondamentos
#half up - arredonda para cima
dt %>%
  tabyl(x,y) %>%
  adorn_percentages() %>%
  adorn_rounding(digits = 0,rounding = "half up")
  x A B C NA_
  azul 1 0 0 0
  preto 0 1 0 0
  vermelho 0 0 0 0
  <NA> 0 0 1 0

#half to even - arredonda para baixo
dt %>%
  tabyl(x,y) %>%
  adorn_percentages() %>%
  adorn_rounding(digits = 0,rounding = "half to even")
  x A B C NA_
  azul 1 0 0 0
  preto 0 0 0 0
  vermelho 0 0 0 0
  <NA> 0 0 1 0
```

11.3.6 Teste qui-quadrado para tabela cruzada - `chisq.test()`

- O teste qui-quadrado pode ser aplicado em dados tabelados de forma cruzada.
- O teste qui-quadrado é um teste não paramétrico, não depende de parâmetros populacionais (média e variância).
- É aplicado em variáveis qualitativas nominais, categóricas.
- O objetivo é testar as hipóteses:
 - H_0 :
A variável linha é independente da variável coluna, ou seja, a proporção no total das linhas deve ser a mesma quando considerada em cada categoria da variável linha.
 - H_1 :
A variável linha é dependente da variável coluna, ou seja, a proporção no total das linhas é diferente da proporção em cada categoria da variável linha.
- As variáveis devem ser categóricas.
- Para concluir o teste, observamos o resultado do `p-valor` e comparamos com o nível de significância (α) adotado na pesquisa, que em geral costuma ser de 0,05(ou seja, 5%).
 - Se $p\text{-valor} > 0,05 = H_0$
Não é possível rejeitar a hipótese nula (H_0).
Concluímos que as variáveis não são dependentes, são independentes, ao nível de significância de 5%.
 - Se $p\text{-valor} \leq 0,05 = H_1$
Rejeitamos a hipótese conhecida como nula (H_0).
Concluímos que as variáveis são dependentes ao nível de significância de 5%.

- Exemplo 1 - Hipótese H_1 :

```
#Suponha dois grupos que receberam tratamento
grupo <- c(rep("A",15),rep("B",15))
set.seed(20)
resposta <- c(sample(0:1,16,replace = T),rep(1,14))
dt = data.frame(grupo,resposta)

tab = tabyl(dt, grupo, resposta, show_na= F)
tab
grupo 0 1
  A 9 6
  B 1 14

#Aplicando o teste qui-quadrado
chisq.test(tab)

Pearson's Chi-squared test with Yates' continuity correction

data: tab
X-squared = 7.35, df = 1, p-value = 0.006706

#Conclusão
##p-value = 0.006706
##Logo, p-value (0.006706) <= alpha (0.05)
##Assim, rejeitamos a hipótese nula ( $H_0$ ),
##há uma associação estatisticamente significativa entre as variáveis.
##É a hipótese  $H_1$ .
```

- Exemplo 2 - Hipótese H_0 :

```
#Suponha dois grupos que receberam tratamento
grupo <- c(rep("A",15),rep("B",15))
set.seed(20)
resposta <- c(sample(0:1,25,replace = T),rep(1,5))
dt = data.frame(grupo,resposta)

tab = tabyl(dt, grupo, resposta, show_na= F)
tab
grupo 0 1
  A 9 6
  B 6 9

#Aplicando o teste qui-quadrado
chisq.test(tab)

Pearson's Chi-squared test with Yates' continuity correction

data: tab
X-squared = 0.53333, df = 1, p-value = 0.4652

#Conclusão
##p-value = 0.4652
##Logo, p-value (0.4652) > alpha (0.05)
##Assim, não podemos rejeitar a hipótese nula ( $H_0$ ),
##Não há evidências suficientes para concluir que as variáveis estão associadas.
##É a hipótese  $H_0$ .
```

11.3.7 Caça aos registros com valores duplicados - `get_dupes()`

- A função `get_dupes()`, do pacote `janitor`, realiza a tarefa de retornar os registros duplicados do conjunto de dados em análise (exibe uma coluna com a contagem de duplicatas), para que seja possível detectar os casos problemáticos.
- É uma função útil para casos em que não deveriam aparecer registros duplicados.
Ex.: ID, registros de notas fiscais, conjuntos únicos (ID, registro de faturamento),...
- Exemplo - Uso da função `get_dupes()`, do pacote `janitor`, para identificar registros duplicados:

```
#data.frame
df = data.frame(ID = c(1000,1001,1000,1002),
                 FAT = c(2098.60,345.00,2098.60,1332.44),
                 ANO = c(2016,2016,2016,2017))

#Identificando registros duplicados - get_dupes()
##Metodo 1
get_dupes(df, ID, FAT)

  ID      FAT dupe_count  ANO
1 1000  2098.6          2 2016
2 1000  2098.6          2 2016

##Metodo 2 - com uso de magrittr
df %>%
  get_dupes(ID,FAT)

  ID      FAT dupe_count  ANO
1 1000  2098.6          2 2016
2 1000  2098.6          2 2016
```

11.3.8 Corrija número para data com a função `excel_numeric_to_date()`

- A função `excel_numeric_to_date()`, do pacote `janitor`, é para consertar data em uma arquivo importado de *Excel*, se uma data veio em formato número ao invés de data.
- A função `excel_numeric_to_date()` converte número para a classe *Date*.
- Exemplo - Conversão de número para data, usando a função `excel_numeric_to_date()` de um arquivo importado do excel:

```
#Aplicação da função excel_numeric_to_date
#Metódo 1
excel_numeric_to_date(51503)
[1] "2041-01-02"

#Metódo 2 - MAC
excel_numeric_to_date(51503, date_system = "mac pre-2011")
[1] "2045-01-03"

#Metódo 3 - com magrittr
dt <- 51503
dt %>%
  excel_numeric_to_date()
[1] "2041-01-02"
```

11.3.9 Conte os níveis dos fatores - escala de *Likert*

11.3.9.1 Escala *Likert*

- A escala *Likert* é utilizada para mensurar sentimentos numa escala que pode variar entre um e cinco níveis (a mais usada é de cinco níveis).
- Sendo um o menor nível e cinco o maior nível de concordância ou discordância sobre uma pergunta ou afirmação.
- Os níveis dos fatores geralmente se apresentam como:
 - Concordo totalmente
 - Concordo parcialmente
 - Neutro
 - Discordo parcialmente
 - Discordo totalmente

Medição de 5 pontos



Figure 68: Exemplo de escala *Likert*.

11.3.9.2 A função `top_levels()`

- A função `top_levels()`, do pacote `janitor`, realiza a contagem dos níveis da escala do tipo *Likert*.
- A função `top_levels()` fornece uma tabela com as contagens e percentuais dos níveis agrupados em três grupos:
 - Alto
 - Médio
 - Baixo
- O argumento `n` da função estabelece quantos níveis serão incluídos no grupo alto e baixo da escala.
 - `n = 1`
Um nível no grupo alto, três níveis no grupo médio e um nível no grupo baixo.
 - `n = 2`
Dois níveis no grupo alto, um nível no grupo médio e dois níveis no grupo baixo.
- As características da variável para ser avaliada pela função `top_levels()`:
 - Deve ser um fator (classe `factor`)
 - Conter as respostas e os níveis (`levels`)
 - Respostas que não correspondem aos níveis (`levels`) são ignoradas.
 - Exemplo:

```
f <- factor(c("neutro","concordo parcialmente",
              "discordo parcialmente","concordo",
              "concordo","concordo totalmente",
              "concordo totalmente","concordo","discordo totalmente"),
              levels = c("concordo totalmente",
                        "concordo parcialmente",
                        "neutro",
                        "discordo parcialmente",
                        "discordo totalmente"))
```

- Exemplo - Contagem dos níveis da escala do tipo *Likert*, usando a função `top_levels()`:

```
#Variável
f <- factor(c("neutro","concordo parcialmente",
             "discordo parcialmente","concordo",
             "concordo","concordo totalmente",
             "concordo totalmente","concordo","discordo totalmente"),
             levels = c("concordo totalmente",
                       "concordo parcialmente",
                       "neutro",
                       "discordo parcialmente",
                       "discordo totalmente"))

#Contagem levels
#n = nº de níveis no grupo alto e baixo
top_levels(f,n=1)
#Um nível no grupo alto, três níveis no grupo médio e um nível no grupo baixo.
      f   n    percent
      concordo totalmente 2 0.3333333
      <<< Middle Group (3 categories) >>> 3 0.5000000
      discordo totalmente 1 0.1666667

top_levels(f,n=2)
#Dois níveis no grupo alto, um nível no grupo médio e dois níveis no grupo baixo.
      f   n    percent
      concordo totalmente, concor... 3 0.5000000
      neutro 1 0.1666667
      discordo parcialmente, disc... 2 0.3333333
```

11.3.9.3 Plotagem de escala *Likert*

- Podemos utilizar o pacote *likert*, para obter um resumo e formas de visualização da análise de respostas na escala *Likert*.

```
install.packages("likert")
```

- O pacote *likert* é usado em conjunto com os pacotes gráficos (*ggplot2* por exemplo).

- Exemplo - Visualização de respostas na escala *Likert*:

```
#Bibliotecas
library(janitor) #Limpeza de dados
library(ggplot2) #Elaboração de gráficos
library(likert) #Opções de figuras no ggplot para itens com escala likert
library(RColorBrewer) #Pacote com paleta de cores para gráficos

#Criando uma base de dados simulada com 3 questões
niveis <- c("concordo totalmente",
           "concordo parcialmente",
           "neutro",
           "discordo parcialmente",
           "discordo totalmente")

#Cria variáveis com 40 valores semialeatórios de 5 fatores
set.seed(30); q1 = factor(sample(1:5,40,replace = T))
levels(q1) = niveis #Substitui os 5 levels de q1 pelos 5 levels da variável niveis
set.seed(31); q2 = factor(sample(1:5,40,replace = T))
levels(q2) = niveis #Substitui os 5 levels de q2 pelos 5 levels da variável niveis
set.seed(32); q3 = factor(sample(1:5,40,replace = T))
levels(q3) = niveis #Substitui os 5 levels de q3 pelos 5 levels da variável niveis

#As respostas na escala de Likert
respostas <- data.frame(q1,q2,q3)

#Preparo para plotagem
tb_likert <- likert(respostas) #Transforma data.frame na classe likert

#Resumo das respostas
summary(tb_likert)

#Gráfico das respostas em escala de Likert
plot(tb_likert,
      colors = c("gray30","gray65","gray95","gray75","gray55"))+
      ggtitle("Gráfico das respostas em escala de Likert")+
      #Título
      labs(x = NULL,
            y = "Porcentagem de respostas")+
      #Rótulos dos eixos
      theme(plot.title = element_text(hjust = 0.5)) #Centraliza o título

#Fechando dispositivo gráfico
dev.off()
```

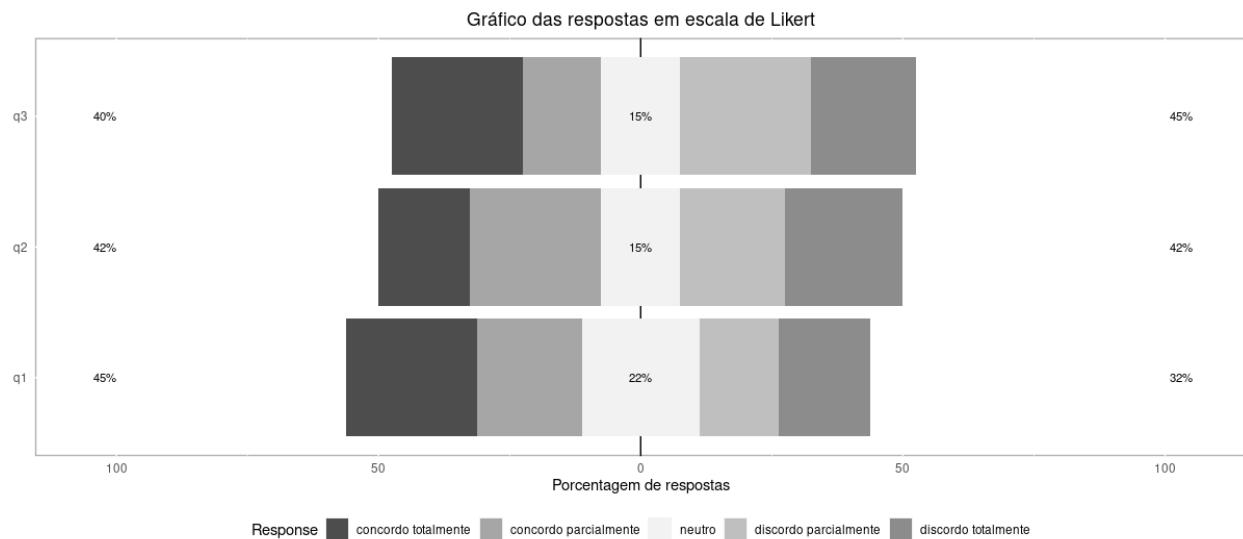


Figure 69: Gráfico das respostas em escala de *Likert*

- Exemplo - Visualização de respostas na escala *Likert* no formato type = heat:

```
#Bibliotecas
library(janitor) #Limpeza de dados
library(ggplot2) #Elaboração de gráficos
library(likert) #Opções de figuras no ggplot para itens com escala likert
library(RColorBrewer) #Pacote com paleta de cores para gráficos

#Criando uma base de dados simulada com 3 questões
niveis <- c("concordo totalmente",
           "concordo parcialmente",
           "neutro",
           "discordo parcialmente",
           "discordo totalmente")

#Cria variáveis com 40 valores semialeatórios de 5 fatores
set.seed(30); q1 = factor(sample(1:5,40,replace = T))
levels(q1) = niveis #Substitui os 5 levels de q1 pelos 5 levels da variável niveis
set.seed(31); q2 = factor(sample(1:5,40,replace = T))
levels(q2) = niveis #Substitui os 5 levels de q2 pelos 5 levels da variável niveis
set.seed(32); q3 = factor(sample(1:5,40,replace = T))
levels(q3) = niveis #Substitui os 5 levels de q3 pelos 5 levels da variável niveis

#As respostas na escala de Likert
respostas <- data.frame(q1,q2,q3)

#Resumo das respostas
tb_likert <- likert(respostas) #Transforma data.frame na classe likert, prepara para plotagem
summary(tb_likert)

#Gráfico das respostas em escala de Likert
plot(tb_likert,
      colors = c("gray30","gray65","gray95","gray75","gray55"),
      type = "heat")+
  ggtitle("Gráfico das respostas em escala de Likert")+ #Título
  labs(x = NULL,
       y = "Porcentagem de respostas")+ #Rótulos dos eixos
  theme(plot.title = element_text(hjust = 0.5)) #Centraliza o título

#Fechando dispositivo gráfico
dev.off()
```

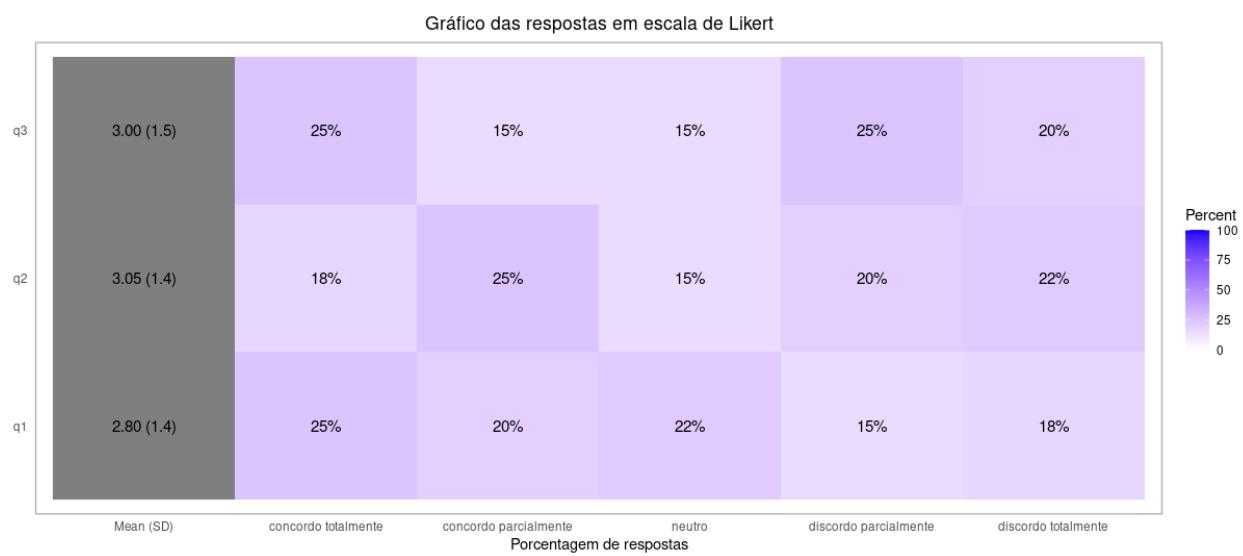


Figure 70: Gráfico das respostas em escala de *Likert* - type = "heat".

- Exemplos de modelos de visualização da escala *Likert*:

```
p1 = likert(items = pptqc, nlevels = 3)
plot(p1)
```

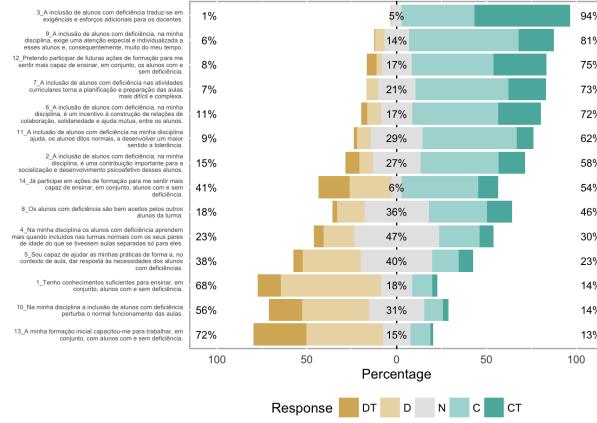


Figure 71: Modelo 1 de visualização de escala *Likert*.

```
plot(p1, type = "heat")
```

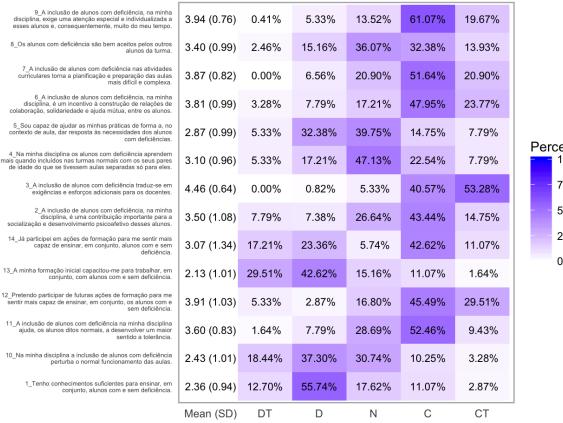


Figure 72: Modelo 2 de visualização de escala *Likert* - type = "heat".

```

lik2 <- likert(as.data.frame(bd[, 3:5]), grouping = bd$categ)
plot(lik2, wrap = 60, text.size=3) + theme(axis.text.y = element_text(size="6"))

```

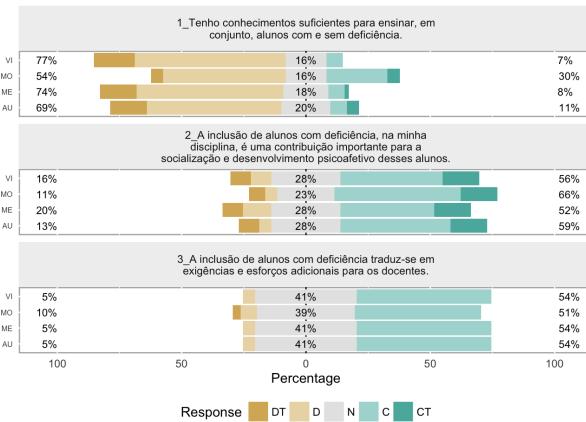


Figure 73: Modelo 3 de visualização de escala *Likert* - grouping = bd\$categ.

12 CAP. 9 - ANÁLISE DESCRIPTIVA DOS DADOS

12.1 Teoria

- Objetivo do capítulo é fazer uma análise descritiva dos dados através da tabulação das variáveis e cálculo de medidas descritivas (média, desvio-padrão, etc).
- Análise descritiva dos dados (Informações preliminares):
 - Contagem dos resultados observados em cada variável do conjunto de dados.
 - Natureza descritiva dos dados, tipo de variáveis (categórica ou numérica).
 - Três objetivos principais:
 - * Verificar erros e anomalias.
 - * Compreender a distribuição de cada uma das variáveis isoladamente.
 - * Compreender a natureza e a força das relações entre as variáveis.
- Após essas etapas, estabelecer um modelo estatístico formal e relatar suas conclusões.

12.2 Tipos de variáveis

- Variável numérica:
 - Contínua
Se seus valores pertencer ao conjunto dos números reais.
Ex.: Temperatura corporal, saldo em caixa, peso da carga de um caminhão, etc.
 - Discreta
Se seus valores pertencer ao conjunto dos números inteiros.
Ex.: Número de pessoas com febre, número de empresas, número de caminhões, etc.
- Variável categórica:
 - Ordinal
Se seus valores podem ser ordenados do menor para o maior.
Ex.: Temperatura (baixa, média ou alta), saldo em caixa (negativo, nulo ou positivo), etc.
 - Nominal
Quando não for possível estabelecer ordenamento.
Ex.: Sexo do indivíduo, atividade fim da empresa, marca/modelo do caminhão, etc.
- Podemos usar, no **R**, a função **str()** (structure) para conhecer o tipo dos dados.
Ex.: **str(variavel)**

12.3 Tabulação dos dados

- Na etapa de tabulação, o pesquisador prepara as tabelas de frequência com o intuito de entender o comportamento das variáveis.
- Para construir as tabelas de frequência utilizamos o pacote `janitor`, a função `tabyl()` e os argumentos `adorn_`.
- Variáveis numéricas contínuas:
 - No caso de variáveis numéricas contínuas, para dividir o intervalo de classes podemos usar a função `cut()`, junto com o argumento `b = nclass.Sturges(coluna)`.
Ex.:
`intervalo=(cut(dados$valor_compra,b=nclass.Sturges(dados$valor_compra)))`
 - Por *default* os intervalos de classe da função `cut()`, com argumento `b = nclass.Sturges(coluna)`, as classes são formadas aberta na esquerda e fechada na direita.
Ex.: [11.4,153]
 - Podemos adicionar o argumento `right = FALSE` para inverter a forma das classes, ficando fechada na esquerda e aberta na direita.
Ex.: [11.4,153)
 - Outra forma de formar os intervalos de classe é inserindo os valores manualmente dos intervalos de classe, na função `cut()`, no argumento `b = c(valores_do_intervalo)`.
Ex.:
`intervalo3 = (cut(dados$valor_compra, b = c(12,182,352,522,692,862)))`

- Exemplo - Tabela de frequência para variável categórica:

```
#Análise descritiva dos dados
#Tabulação dos dados - variável categórica

#Bibliotecas
library(knitr) #Interpretação e compilação do documento rmd, formato tabela kable
library(magrittr) #Operador pipe "%>%", concatena linhas de comando
library(readr) #Leitura de dados
library(janitor) #Limpeza de dados

#Leitura da base de dados
dados <- read.csv2(file = "~/Programacao/R/Dados/Dados_de_importacao/vendas.csv")
dados <- data.frame(dados)

#Exibindo as 6 primeiras linhas da base de dados
head(dados)

  cupom filial valor_compra n_itens desconto_perc quinzena
1   101     A      100.22      5           2         1
2   102     A      80.89       20          0         1
3   103     A      75.44       7           0         1
4   104     A      305.33      3           10        2
5   105     A      120.99      1           2         2
6   106     A      27.89       1           0         2

#Exibindo a estrutura dos dados
#Tipo das variáveis
str(dados)

'data.frame': 23 obs. of 6 variables:
 $ cupom      : int  101 102 103 104 105 106 201 202 203 204 ...
 $ filial     : chr "A" "A" "A" "A" ...
 $ valor_compra : num 100.2 80.9 75.4 305.3 121 ...
 $ n_itens    : int 5 20 7 3 1 1 20 30 17 14 ...
 $ desconto_perc: int 2 0 0 10 2 0 0 12 10 0 ...
 $ quinzena   : int 1 1 1 2 2 2 2 2 2 1 ...

#Tabela de frequência variável categórica
tb_filial <- tabyl(dados,filial) %>%
  adorn_totals() %>%
  adorn_rounding(2)

tb_filial

  filial  n percent
    A    6    0.26
    B   12    0.52
    C    5    0.22
  Total 23    1.00

#Plotar tabela
kable(tb_filial, caption = "Tabela de frequência para variável categórica", align = "ccc")
```

Table 25: Tabela de frequênci para variável categórica

Filial	n	Porcentagem
A	6	0.26
B	12	0.52
C	5	0.22
Total	23	1.00

- Exemplo - Tabela de frequência para variável numérica (contínua), usando o método de separação de classes `nclass.Sturges()`:

```
#Análise descritiva dos dados
#Tabulação dos dados - Variável numérica (continua)

#Bibliotecas
library(knitr) #Interpretação e compilação do documento rmd, formato tabela kable
library(magrittr) #Operador pipe "%>%", concatena linhas de comando
library(readr) #Leitura de dados
library(janitor) #Limpeza de dados

#Leitura da base de dados
dados <- read.csv2(file = "~/Programacao/R/Dados/Dados_de_importacao/vendas.csv")
dados <- data.frame(dados)

#Exibindo as 6 primeiras linhas da base de dados
head(dados)

  cupom filial valor_compra n_itens desconto_perc quinzena
1   101     A      100.22      5          2         1
2   102     A      80.89       20         0         1
3   103     A      75.44       7          0         1
4   104     A      305.33      3          10        2
5   105     A      120.99      1          2         2
6   106     A      27.89       1          0         2

#Exibindo a estrutura dos dados
#Tipo das variáveis
str(dados)

'data.frame': 23 obs. of 6 variables:
 $ cupom      : int  101 102 103 104 105 106 201 202 203 204 ...
 $ filial     : chr "A" "A" "A" "A" ...
 $ valor_compra: num  100.2 80.9 75.4 305.3 121 ...
 $ n_itens    : int  5 20 7 3 1 1 20 30 17 14 ...
 $ desconto_perc: int  2 0 0 10 2 0 0 12 10 0 ...
 $ quinzena   : int  1 1 1 2 2 2 2 2 2 1 ...

#Cut para categorizar valor_compra em b intervalos
#Usar o metodo cut(dados, nclass.Sturges()) para separar as classes
intervalo = (cut(dados$valor_compra, b = nclass.Sturges(dados$valor_compra)))
intervalo

[1] (11.4,153] (11.4,153] (11.4,153] (294,434] (11.4,153] (11.4,153]
[7] (11.4,153] (434,575] (153,294] (11.4,153] (11.4,153] (715,857]
[13] (11.4,153] (153,294] (434,575] (11.4,153] (11.4,153] (153,294]
[19] (11.4,153] (153,294] (11.4,153] (294,434] (715,857]
Levels: (11.4,153] (153,294] (294,434] (434,575] (575,715] (715,857]

#Tabela de frequência da variável valor_compra
#Tabela frequência da uma variável numérica continua
tb_valor = tabyl(intervalo) %>%
```

```

adorn_totals() %>%
adorn_rounding(2)

tb_valor

intervalo  n percent
(11.4,153] 13    0.57
(153,294]   4    0.17
(294,434]   2    0.09
(434,575]   2    0.09
(575,715]   0    0.00
(715,857]   2    0.09
      Total 23    1.00

#Plotar tabela
kable(tb_valor, align = "ccc",
      caption = "Tabela de frequênci para variável numérica contínua")

```

Table: Tabela de frequênci para variável numérica contínua

intervalo	n	percent
(11.4,153]	13	0.57
(153,294]	4	0.17
(294,434]	2	0.09
(434,575]	2	0.09
(575,715]	0	0.00
(715,857]	2	0.09
Total	23	1.00

Table 26: Tabela de frequência para variável numérica contínua, usando o método de separação de classes `nclass.Sturges()`

Intervalo	n	Porcentagem
(11.4,153]	13	0.57
(153,294]	4	0.17
(294,434]	2	0.09
(434,575]	2	0.09
(575,715]	0	0.00
(715,857]	2	0.09
Total	23	1.00

- Exemplo - Tabela de frequência para variável numérica (contínua), com separação de classes inserido manualmente e limites com aberturas invertidas usando `right = FALSE`:

```
#Cut para categorizar valor_compra em b intervalos
#Entrando com os intervalos de classe
#O comando "right = FALSE", inverte o intervalo de classe, esquerdo fechado e direito aberto [,).
intervalo = (cut(dados$valor_compra, b = c(12,182,352,522,692,862),
                  right = FALSE))
intervalo

[1] [12,182)  [12,182)  [12,182)  [182,352) [12,182)  [12,182)  [12,182)
[8] [352,522) [182,352) [12,182)  [12,182)  [692,862) [12,182)  [182,352)
[15] [352,522) [12,182)  [12,182)  [182,352) [12,182)  [182,352) [12,182)
[22] [352,522) [692,862)
Levels: [12,182) [182,352) [352,522) [522,692) [692,862)

#Tabela de frequência da variável valor_compra
#Tabela frequência da uma variável numérica continua
tb_valor = tabyl(intervalo) %>%
  adorn_totals() %>%
  adorn_rounding(2)

tb_valor

intervalo n percent
[12,182) 13    0.57
[182,352) 5     0.22
[352,522) 3     0.13
[522,692) 0     0.00
[692,862) 2     0.09
  Total 23    1.00

#Plotar tabela
kable(tb_valor, align = "ccc",
      caption = "Tabela de frequência para variável numérica continua")

Table: Tabela de frequência para variável numérica continua

| intervalo | n   | percent |
|:-----:|:---:|:-----:|
| [12,182) | 13 | 0.57   |
| [182,352) | 5  | 0.22   |
| [352,522) | 3  | 0.13   |
| [522,692) | 0  | 0.00   |
| [692,862) | 2  | 0.09   |
|  Total    | 23 | 1.00   |
```

Table 27: Tabela de frequênci para variável numérica continua, com separação de classes inserido manualmente e limites com aberturas invertidas usando `right = FALSE`

Intervalo	n	Porcentagem
[12,182)	13	0.57
[182,352)	5	0.22
[352,522)	3	0.13
[522,692)	0	0.00
[692,862)	2	0.09
Total	23	1.00

12.4 Estatística descritiva com o pacote DescTools

12.4.1 Teoria

- O pacote **DescTools** foi desenvolvido com o objetivo de fornecer uma análise descritiva de forma rápida e completa.
- A principal função do pacote é **Desc()**, descreve as variáveis de acordo com sua natureza, produzindo medidas estatísticas descritivas e uma representação gráfica adequada.
- Tipos de variáveis:
 - Lógicas
 - Fatores (ordenados e não ordenados)
 - Inteiros
 - Numéricos
 - Datas
 - Tabelas
 - Matrizes
- Principais saídas da função **Desc()**:

Table 28: Principais saídas da função Desc para variáveis numéricas.

Saídas da função Desc	Descrição
length	O comprimento do vetor.
n	O número de observações validas.
NAs	O número de observações faltantes.
unique	O número de observações distintas entre si.
0s	O número total de valores nulos.
mean	A média aritmética do vetor.
meanSE	Fornece um intervalo de 95% de confiança para a média, com base no erro padrão da média.
.05, ..., .95	Percentil de x, iniciando em 5%, 10%, 1 quartil, mediana, ...
range	A amplitude do vetor x.
sd	O desvio-padrão do vetor x.
vcoef	O coeficiente de variação de x.
mad	O desvio médio absoluto.
IQR	A amplitude interquartil, definida por 3 quartil - 1 quartil.
skew	O coeficiente de assimetria do vetor x.
kurt	O coeficiente de curtose do vetor x.
lowest	Os cinco menores valores do vetor x.
highest	Os cinco maiores valores do vetor x.

- Principais parâmetros da função `Desc()`:

– `plotit = F`
 Omitir os gráficos.
 Ex.: `Desc(dados, plotit = F)`

- Mapeamento de dados faltantes:

```
PlotMiss(dados, main="Dados Faltantes", clust = TRUE)
```

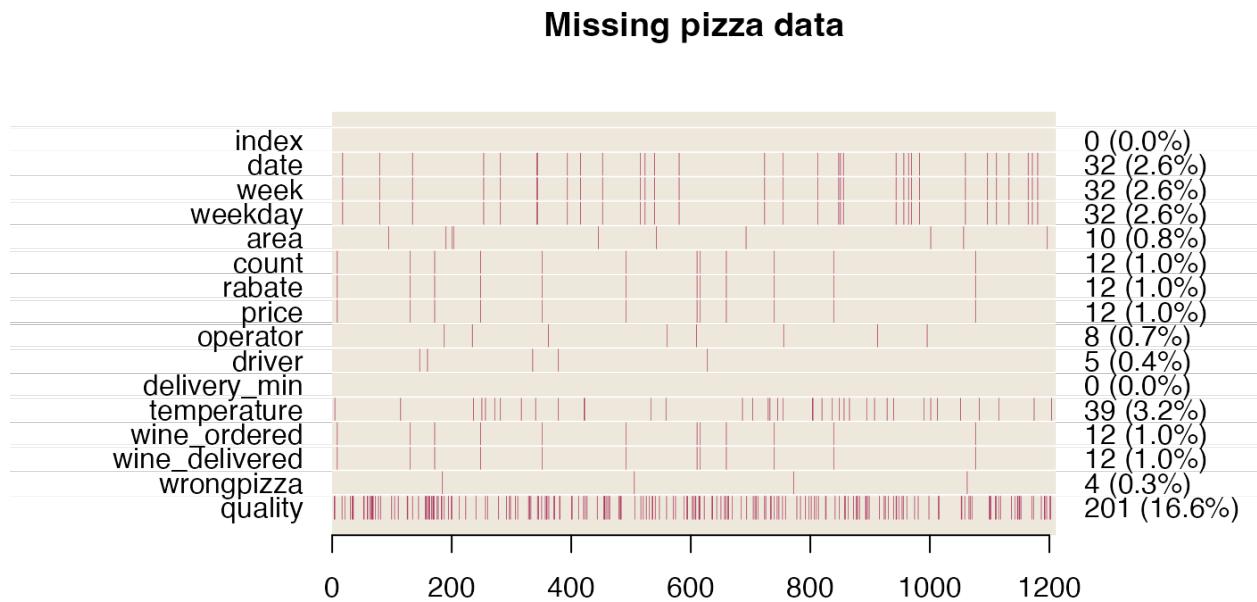


Figure 74: Exemplo da função `PlotMiss()` para mapemaento de dados faltantes.

12.4.2 Customizar os gráficos

- É possível plotar determinada coluna a partir da função `plot()` e `Desc()`, obtendo assim as principais informações da variável envolvida.
- Os gráficos plotados (pode ser mais de um), vão depender do tipo da variável envolvida:
 - Numérica
 - Histograma** sobreposto com curva de densidade;
 - Boxplot**;
 - Frequência acumulada** para cada intervalo da variável.
 - Inteira
 - Categórica
 - Dicotômica** (ate dois níveis), intervalos de confiança de 90, 95, 99% (assemelha-se a um boxplot).
 - Politômica** (mais de dois níveis), gráfico de barras tanto para frequência absoluta quanto para frequência relativa.
- Principais argumentos da função `plot()`:
 - `Desc(dados$coluna)`
Coluna/variável da qual serão plotados os gráficos, a partir de suas principais medidas estatísticas descritivas.
 - `main = "Título"/NULL`
Insere um título ao gráfico.
 - `maxlablen = 25`
Controla o número de caracteres máximo m impresso nos rótulos do gráfico.
 - `type = c("bar", "dot")`
Customização do tipo de plotagem.
 - `col = "red"/NULL`
Adiciona cor aos pontos.
 - `xlim = c(150,200)/NULL`
Limites do eixo x.
 - `ecdf = TRUE`
Exibe (TRUE), ou não (FALSE), as barras acumuladas do gráfico de barras.
 - Exemplo:

```
plot(Desc(dados$coluna), main= NULL,
      maxlablen = 25,
      type = c("bar", "dot"),
      col = NULL,
      border = NULL,
      xlim = NULL,
      ecdf = TRUE)
```

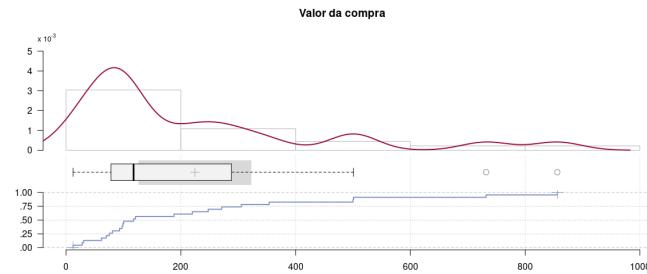


Figure 75: Gráficos de variável quantitativa contínua (numérica) a partir de medidas estatística descritiva. `plot(Desc(dados$variavel_numerica))`

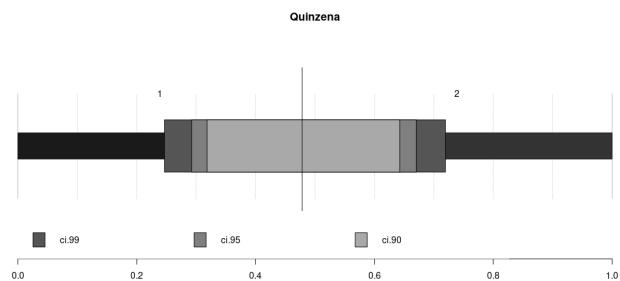


Figure 76: Gráfico de variável quantitativa discreta (inteiro), dicotômica (dois níveis), a partir de medidas estatística descritiva. `plot(Desc(dados$variavel_inteiro))`

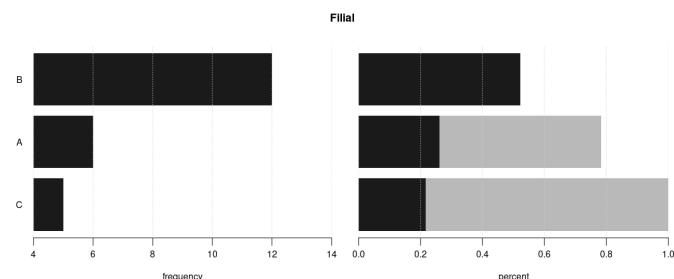


Figure 77: Gráficos de variável qualitativa (categórica), politômica (mais de dois níveis), a partir de medidas estatística descritiva. `plot(Desc(dados$variavel_categorica))`

12.4.3 Interpretação dos coeficientes

- $CV | vcoef$ (Coeficiente de variação)

– O coeficiente de variação é uma medida de dispersão, quanto menor a porcentagem mais próximos os dados estão da média.

– Calculando Coeficiente de variação CV :

$$CV = \frac{dp}{\bar{x}} \times 100$$

onde,

$$\text{Média}(\bar{x}) = \frac{\sum_{i=1}^n x_i}{n}$$

$$\text{Desvio}(dm) = \sum_{i=1}^n (|x_i - \bar{x}|)$$

$$\text{Variância}(var) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

$$\text{Desvio-padrão}(dp) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

– Análise do coeficiente de variação:

- * $CV \leq 15\%$
Baixa dispersão.
- * $15\% < CV \leq 30\%$
Média dispersão.
- * $CV > 30\%$
Alta dispersão.

- As | **shew** (Coeficiente de assimetria)

– Ajuda a definir assimetria dos dados.

– Casos:

* Simetrica: $\bar{x} = md = m_o$.

* Assimetrica a esquerda: $\bar{x} - md = negativo$.

* Assimetrica a direita: $\bar{x} - md = positivo$.

– Calculando coeficiente de assimetria:

$$As = \frac{3 \times (\bar{x} - md)}{dp}$$

Onde,

\bar{x} é média;

md é a mediana;

m_o é a moda;

dp é o desvio-padrão.

– Análise do coeficiente de assimetria:

* $|skew| \leq 0.15$

Distribuição praticamente simétrica ($\bar{x} = md = m_o$).

* $0.15 < |skew| \leq 1$

Assimetria moderada.

* $|skew| > 1$

Assimetria Forte.

– Casos de assimetria:

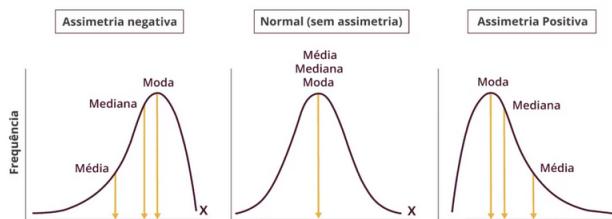


Figure 78: Casos de assimetria.

- C | kurt (Coeficiente de Curtose)
 - Curtose é o grau de achatamento (ou afilamento) de uma distribuição em comparação com a curva normal.
 - Calculando o coeficiente de Curtose:

$$C = \frac{Q_3 - Q_1}{2 \times (P_{90} - P_{10})}$$

$$* = L_i + \frac{k \cdot \sum f_i - F_{anterior}}{f_{intervalo}} \times h$$

$$P_k = L_i + \frac{\frac{k}{100} \cdot \sum f_i - F_{anterior}}{f_{intervalo}} \times h$$

Onde,

Q_3 é o terceiro quartil;

Q_1 é o primeiro quartil;

P_{90} é o percentil 90;

P_{10} é o percentil 10.

- Análisando o coeficiente de Curtose:

* $C \cong 0.263$
A distribuição é mesocúrtica.

* $C < 0.263$
A distribuição é leptocúrtica (em cume).

* $C > 0.263$
A distribuição é platicúrtica (plana).

- Tipos de achatamento da curva de distribuição:

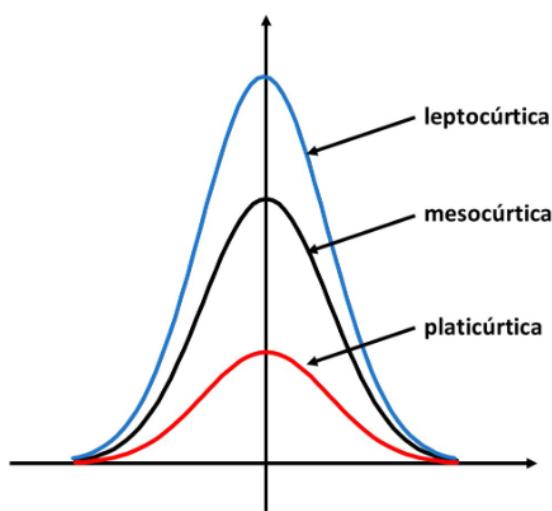


Figure 79: Tipos de distribuição normal.

12.5 Mapeamento de dados faltantes - PlotMiss()

- Ao analisar a base de dados com dados faltantes, devemos observar como estão distribuídos os dados faltantes.
- Ao aplicarmos o comando `PlotMiss()` podemos analisar a distribuição dos dados faltantes, identificando quais variáveis apresentam o maior número de dados faltantes.
- Através do mapeamento de dados faltantes, podemos observar quais variáveis apresentam perda de dados e as variáveis, que pela maior ausência de dados, exigem maior atenção.
- A partir do mapeamento podemos tomar duas decisões em relação as variáveis que apresentam ausência de dados:
 - Realizar limpeza dos dados faltantes.
 - * Procurar/remover colunas ou linhas inúteis.
`remove_empty()`
 - * Registros duplicados.
`get_dupes()`
 - Aplicar um algoritmo de imputação para esses dados.
`complete(mice())`
- Exemplo - Mapeamento de dados faltantes:

```
#Análise de distribuição de dados faltantes
PlotMiss(dados, col = colorRampPalette(c("gray10","gray90"))(1))

#Fechando dispositivo gráfico
dev.off()
```

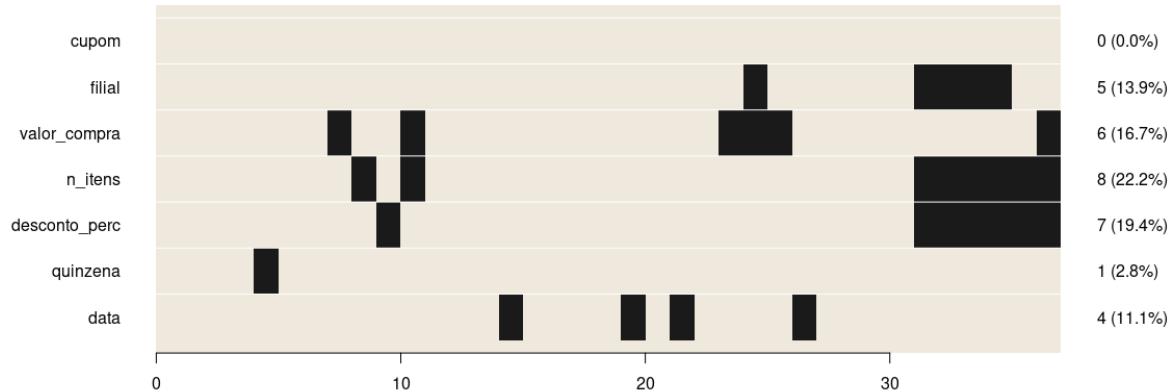


Figure 80: Mapeamento de dados faltantes - `PlotMiss()`

12.6 Analisando datas com o pacote DescTools

- Na análise descritiva de datas (`Desc(variável_data)`) é feito uma análise das frequências (frequência, porcentagem da frequência, frequência acumulada e porcentagem da frequência acumulada).
- O pacote `lubridate` adiciona a classe `date` e funções para trabalhar com datas.

12.6.1 Principais funções (`date`) do pacote `lubridate`

- Converter classes para `date`:

– `date(variável_chr)`

Converte variável do tipo `chr` em tipo `date`.
Ex.: `data_date <- date(data_string)`

– `as_date(variável_chr)`

– `as.Date(variável_chr)`

Função da base do **R** para converter variável do tipo `chr` em tipo `date`.

- Formatos de data:

– `dmy(variável_chr)`

Converte variável `chr` no formato `date` (dia, mês e ano).
Ex. formato: 21-10-2015.

– `mdy(variável_chr)`

Converte variável `chr` no formato `date` (mês, dia e ano).
Ex. formato: 10-21-2015.

– `myd(variável_chr)`

Converte variável `chr` no formato `date` (mês, ano e dia).
Ex. formato: 10-2015-21.

– `ymd(variável_chr)`

Converte variável `chr` no formato `date` (ano, mês e dia).
Ex. formato: 2015-10-21.

– `ydm(variável_chr)`

Converte variável `chr` no formato `date` (ano, dia e mês).
Ex. formato: 2015-21-10.

- Extrair componentes da classe **date**:

- **second()**

Extrai os segundos.

- **minute()**

Extrai os minutos.

- **hour()**

Extrai a hora.

- **wday()**

Extrai o dia da semana.

- **mday**

Extrai o dia do mês.

- **month()**

Extrai mês.

- **year()**

Extrai ano.

- **today()**

Extrai o dia em que foi feita a ultima execução.

- **now()**

Extrai o dia em que foi feita a ultima execução, data completa.

- Fusos horários:

O argumento `tz` ou `tzone` define o fuso.

– `with_tz(data, tzzone = "novo_fuso")`
Retorna qual seria a data em outro fuso.

– `force_tz(data, tzzone = "novo_fuso")`
Altera o fuso sem mudar a hora.

– Exemplo:

```
estreia_GoT <- ymd_hms("2017-07-16 22:00:00", tz = "America/Sao_Paulo")
estreia_GoT
## [1] "2017-07-16 22:00:00 -03"

# Devolve qual seria a data em outro fuso

with_tz(estreia_GoT, tzzone = "GMT")
## [1] "2017-07-17 01:00:00 GMT"
with_tz(estreia_GoT, tzzone = "US/Alaska")
## [1] "2017-07-16 17:00:00 AKDT"

# Altera o fuso sem mudar a hora

force_tz(estreia_GoT, tzzone = "GMT")
## [1] "2017-07-16 22:00:00 GMT"
```

12.6.2 Operações com datas

- Intervalos:

– Intervalos podem ser salvos em objetos da classe `interval`.

```
inicio <- dmy("01-04-1991")
evento <- dmy("31-10-1993")

sobrevo <- interval(inicio, evento)
sobrevo
## [1] 1991-04-01 UTC--1993-10-31 UTC
class(sobrevo)
## [1] "Interval"
```

– Outra forma de definir um intervalo é usar o operador `%--%`.

– Verificar se dois intervalos tem intersecção entre si, usando a função `int_overlaps()`.

```
intervalo_1 <- dmy("01-02-2003") %--% dmy("02-03-2005")
intervalo_2 <- dmy("04-05-2004") %--% dmy("12-03-2012")
int_overlaps(intervalo_1, intervalo_2)
## [1] TRUE
```

- Aritmética com datas:

– Somando datas:

```
* data + dday(1)
Adiciona um dia a variável da classe date.
```

```
* data + dyear(1)
Adiciona um ano a variável da classe date.
```

– Criando datas recorrentes:

Cria um evento que se repete no mesmo dia da semana, durante 1 (evento inicial) + 10 semanas.

```
reuniao <- dmy("18-03-2017")
reunioes <- reuniao + weeks(0:10)
reunioes
## [1] "2017-03-18" "2017-03-25" "2017-04-01" "2017-04-08" "2017-04-15"
## [6] "2017-04-22" "2017-04-29" "2017-05-06" "2017-05-13" "2017-05-20"
## [11] "2017-05-27"
```

– Duração de intervalos:

```
intervalo <- dmy("01-03-2003") %--% dmy("31-03-2003")
```

```
* intervalo / ddays(1)
Retorna o número de dias.
```

```
* intervalo / dminutes(1)
Retorna o número de minutos.
```

```
* as.period(intervalo)
Retorna o tempo do intervalo.
```

12.6.3 Análise DescTools com datas

- A análise foca em três componentes:
 - Semana
Frequências distribuídas nos dias da semana.
 - Mês
Frequências distribuídas pelos meses.
 - Dia
Frequências distribuídas pelos dias.
- É possível fazer analise de data, com e sem gráfico:
 - `Desc(data, plotit = F)`
Sem gráficos.
 - `Desc(data)`
Com gráficos.

- Exemplo - Análise descritiva de data com gráficos:

```
#Transformando a coluna dados$data no formato data ano, mês e dia
str(dados$data)
x <- dmy(as.character(dados$data))

#Analizando a variável data sem gráficos
Desc(x, plotit = F)
#Analise descritiva de datas, leva em consideração três componentes
## semana
## mês
## dia

#Analise descritiva de datas com gráficos
Desc(x)
# Plota três gráficos
## semana
## mês
## dia
```

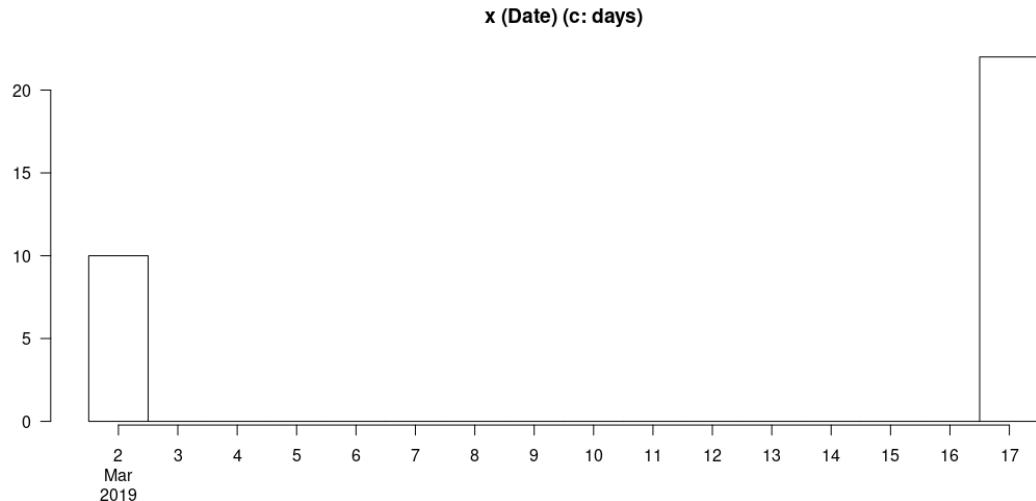


Figure 81: Analise de data, gráfico de dias.

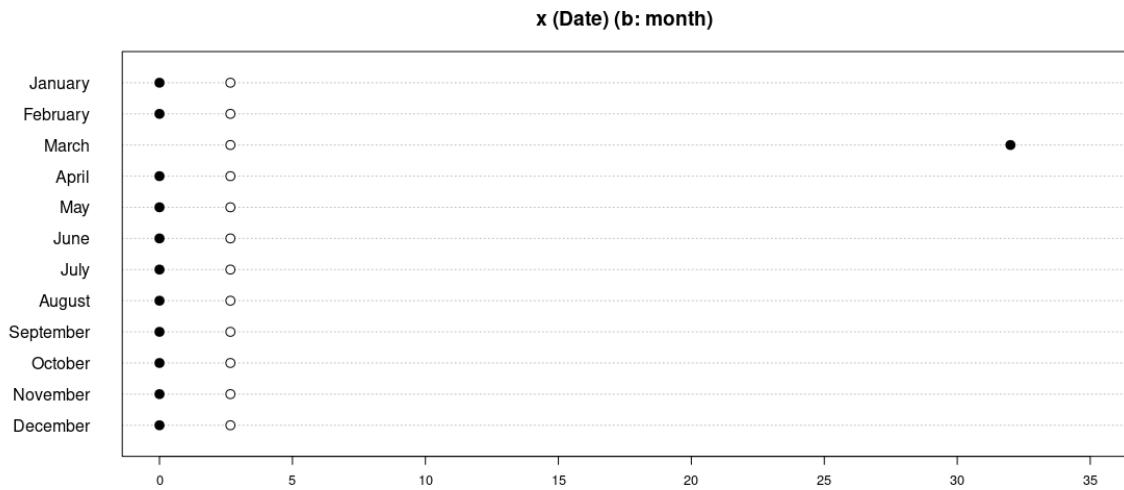


Figure 82: Analise de data, gráfico de meses.

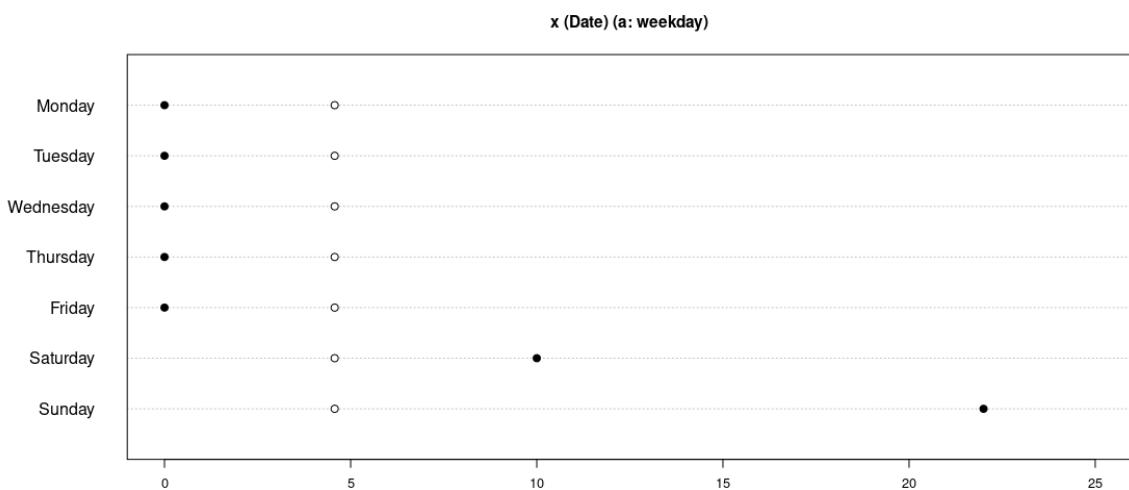


Figure 83: Analise de data, gráfico de dias da semana.

13 CAP. 10 - DISTRIBUIÇÕES DE PROBABILIDADES

13.1 Teoria e introdução

- É possível realizar comparações gráficas com uma distribuição teórica de probabilidades.
- As distribuições de probabilidade descrevem o comportamento de uma variável aleatória que pode ser **discreta** ou **contínua**.
- Principais distribuições:
 - Variável discreta
 - * Binomial
 - * Binomial negativa
 - * Geométrica
 - * Hipergeométrica
 - * Poisson
 - Variável contínua
 - * Beta
 - * Cauchy
 - * Exponencial
 - * F de Snedecor
 - * Gama
 - * Lognormal
 - * Normal
 - * Qui-quadrada
 - * T de Student
 - * Uniforme
 - * Weibull

Table 29: Nome das funções correspondentes às principais distribuições de probabilidade

Distribuição de probabilidade	Nome da função no R
Beta	beta
Binomial	binom
Binomial negativa	nbinom
Cauchy	cauchy
Qui-quadrada	chisq
Exponencial	exp
F Snedecor	f
Gamma	gamma
Geométrica	geom
Log-normal	lnorm
Logística	logis
Normal	norm
Poisson	poi
T Student	t
Weibull	weibull

- Modelos:
 - Em situações práticas, o pesquisador possui uma série de observações da variável de interesse, obtidas através de um processo de amostragem, e deseja ajustar uma distribuição de probabilidades que reproduza o comportamento geral dessa variável.
 - É importante conhecer diversos modelos teóricos para que se possa selecionar o mais adequado que reproduza o comportamento da variável em questão.
 - O pesquisador precisará estimar os valores dos parâmetros do modelo que ele identifica como adequado (ajustar parâmetros).
 - Ao final verificar a qualidade de ajuste.
- Comandos sobre as distribuições de probabilidade e pesquisa:
 - Em relação aos comandos sobre as distribuições de probabilidade, é possível avaliar a densidade (**d**), a probabilidade (**p**), os percentis (**q**) e gerar números aleatórios (**r**).
 - Cada um desses prefixos **d**, **p**, **q** e **r**, associados ao nome da função de distribuição, produz informações sobre essa função (prefixo + função).
Os paramêtros da função **norm** são a **média** (**mean**, símbolo μ) e o **desvio-padrão** (**sd**, símbolo σ).
Sintaxe: **ddist(x, parâmetros)**
Ex.: **pnorm(x, media, desvio-padrão)**
 - Para saber quais argumentos e a ordem com que devem ser especificados no argumento das funções, utilize o comando **?ddist**, substituindo **dist** pelo nome da função no **R**.
Ex.: **?dnorm**

13.2 Distribuição normal

- A distribuição normal é uma das mais importantes e conhecidas distribuições de probabilidade pela sua enorme aplicabilidade em diversas áreas do conhecimento representando o comportamento de diversas variáveis aleatórias, em especial, quando se trata de descrever o comportamento da média amostral.
- A função de densidade de probabilidade de uma variável X dessa natureza é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} x \in R$$

- Notação: $X \sim N(\mu, \sigma)$ que dizer que a variável aleatória X tem distribuição normal com os parâmetros especificados entre parênteses.
- Os parâmetros dessa distribuição são:

– μ

Representa a média da variável.

– σ

Representa o desvio-padrão da variável.

$$\sigma = \sqrt{\sigma^2}$$

– σ^2

Representa a variância da variável (grau de dispersão).

* Variância populacional:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

* Variância amostral:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n-1}$$

– $f(x)$

Representa a função densidade de probabilidade no ponto x .

- No caso contínua, avalia-se a probabilidade para intervalos de valores da variável X , que é dada pela área sob a curva da função de densidade restrita aos limites do intervalo. Em notação matemática representamos a probabilidade de X pertencer ao intervalo $[a, b]$ da seguinte forma:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Seja a área sob a curva no intervalo $[a, b]$ a probabilidade de algo ocorrer entre os valores de a até b . É importante salientar que o valor da densidade, ou seja os valores de $f_X(x)$ representam as densidades, enquanto que a área sob essas densidades é a probabilidade.

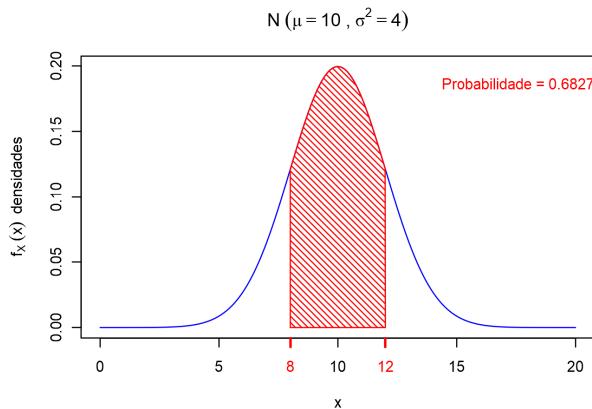


Figure 84: Probabilidade ou área sobre a curva.

- Funções `norm()`:

- Avaliamos a função de densidade de probabilidade no ponto x usando a função $dnorm(x, \mu, \sigma)$.
- Para avaliar a probabilidade acumulada no ponto x , isto é, $P(X \leq x)$, usamos a função $pnorm(x, \mu, \sigma)$.
- Probabilidade acumulada num intervalo ($P(a < X < b)$):

$$pnorm(final, \mu, \sigma) - pnorm(inicial, \mu, \sigma)$$

Ex.: $P(40 < X < 55)$, sendo $X \sim N(50, 1)$

```
pnorm(55, 50, 1) - pnorm(40, 50, 1)
[1] 0.9999997
```

- Para obter o percentil de ordem k da distribuição usamos $qnorm(k, \mu, \sigma)$.

Ex.: Percentil 97,5 de uma distribuição normal padronizada ($X \sim N(0, 1)$)

```
qnorm(0.975, 0, 1)
[1] 1.959964
```

- Para gerar k números aleatórios usamos $rnorm(k, \mu, \sigma)$.

Ex.: 20 números aleatórios de uma distribuição normal padronizada ($X \sim N(0, 1)$)

```
set.seed(30)
y <- rnorm(20, 0, 1)

[1] -1.28851820 -0.34768941 -0.52162885  1.27347316  1.82452060 -1.51130794
[7]  0.11050805 -0.76079623 -0.66989702  0.27451969 -1.02327202 -1.81939791
[13] -0.66778981 -0.05929799  0.88016591  0.26851292 -0.01957938 -0.52494697
[19] -1.40933143 -1.83398921
```

- O gráfico da distribuição normal é simétrico em relação ao parâmetro μ , em forma de sino e sua escala depende do parâmetro σ , quanto menor o valor de σ , maior será a área em torno de μ .
- A distribuição normal é dita padronizada quando $\mu = 0$, e $\sigma = 1$.

13.3 Aderência da distribuição normal

- Grande parte das ferramentas de inferência estatística passa pelo teste de aderência quanto a normalidade dos dados, isto é, analisa-se a partir da distribuição de densidade empírica da amostra se ela pode ser considerada proveniente de uma população com distribuição normal.
- O início dessa análise parte do princípio de que os dados seguem a distribuição normal e após análise gráficas (histograma, boxplot, densidade empírica) e testes de aderência (teste de shapiro, teste Kolmogorov-Smirnov, entre outros), conclui-se por aceitar, ou não, a distribuição normal dos dados.
 - Análise gráfica:
 - * Histograma
 - * Boxplot
 - * Densidade empírica
 - Testes de aderência:
 - * Teste de shapiro
 - * Teste de Kolmogorov-Smirnov
 - * ...
- A função `PlotFdist()` do pacote `DescTools` fornece uma representação gráfica univariada para um vetor numérico mostrando num único gráfico um histograma, a curva de densidade, um boxplot e uma distribuição acumulada empírica, sendo que a curva normal estimada pode ser sobreposta a esses gráficos baseados na amostra.
- Desse modo, para análise de qualidade de aproximação da distribuição estimada da variável com a distribuição normal, usamos a função `PlotFdist()` do pacote `DescTools`.
 - No primeiro argumento informamos o nome do vetor que contém os dados da variável.
 - No segundo argumento os dados da distribuição teórica (distribuição normal) e seus argumentos respectivos como média e desvio-padrão.
 - Para omitir o gráfico boxplot e a distribuição acumulada, acrescente:

```
args.boxplot = NA  
args.ecdf = NA
```

- Exemplo - Aderência de distribuição normal:

```
#Importação de dados
dados = read.csv2(file = "~/Programacao/R/Dados/Dados_de_importacao/vendas.csv")
dados
str(dados)
summary(dados)

#Ajuste da distribuição normal para vetor y
y <- dados$valor_compra
PlotFdist(y, #vetor dados da variável
          args.curve = list(expr = "dnorm(x,mean(y),sd(y))", #dados da distribuição normal
                             col = "black"), #Distribuição normal em preto
          args.dens = list(col = "gray"), #grafico densidade em cinza
          #args.boxplot = NA, #Omitir boxplot
          args.ecdf = NA, #Omitir distribuição acumulada
          xlim = c(-500, 1000))

#Organizando a legenda
legend(x = "topright",
       legend = c("kernel density", #Legenda 1
                  expression(X%~%N(mu==31.75, sigma==7.94))), #Legenda 2 - Expressão matemática
       fill = c("gray80","black"),
       text.width = 0.2) #Tamanho da caixa de legenda

#Fechando dispositivo gráfico
dev.off()
```

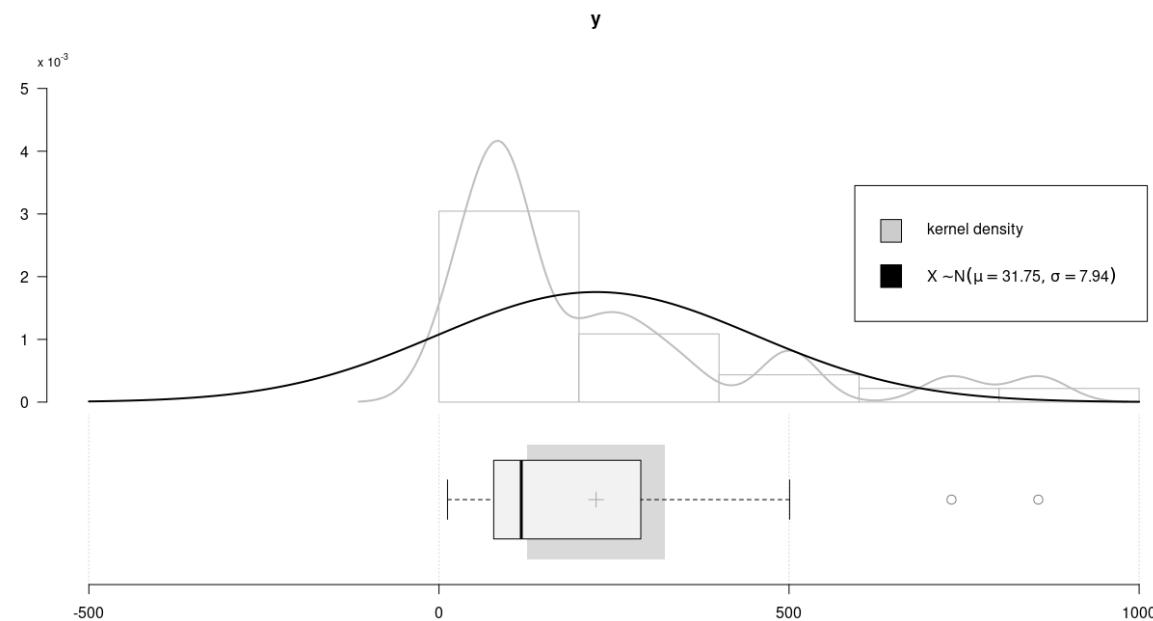


Figure 85: Análise gráfica (histograma, boxplot, densidade empírica) através da função `PlotFdist()`, do pacote `DescTools`.

Ao observar o gráfico, percebe-se que não é adequado a aproximação entre as curvas, a distribuição teórica está abaixo da distribuição empírica (kernel density) na maior parte do intervalo de valores do eixo x, há uma forte assimetria à direita, assim como presença de *outliers*, de acordo com o que vemos no boxplot.

A partir dessa análise gráfica da aproximação entre curvas de densidade, o próximo passo é realizar um teste de aderência da normalidade, como o teste Shapiro. Se o teste confirmar a aderência, torna-se possível avaliar probabilidades para essa variável, gerar números aleatórios simulando a obtenção de amostras dessa população bem como avaliar percentis para a variável, com base na distribuição normal.

13.4 As hipóteses de um teste estatístico

- Um teste estatístico é composto de duas hipóteses sobre parâmetros populacionais relacionados à pergunta da pesquisa:

– H_0 :

Hipótese nula, afirma valores para os parâmetros.

– H_1 :

Hipótese alternativa, nega o valor do parâmetro afirmado em H_0 .

- Em geral, a hipótese nula é normalmente a negação do que o pesquisador deseja provar.
- Exemplos:

1. O contrato de uma empresa estabelece que o percentual de certa substância num produto não pode ultrapassar certo valor, 1%. Há suspeitas de que esse item do contrato não esteja sendo cumprido, logo o pesquisador que provar que a proporção é maior do que 1%. O parâmetro a ser testado é a proporção p da substância no produto.

Nesta situação a formulação do teste é:

$$H_0: p = 0.01;$$

$$H_1: p > 0.01;$$

2. Caso o contrato estabeleça que o percentual de certa substância deva ser igual a (ou em torno de) certo valor, digamos, 1%, a formulação do teste é:

$$H_0: p = 0.01;$$

$$H_1: p \neq 0.01;$$

3. Caso o contrato estabeleça que o percentual de certa substância deva ser maior do que, digamos, 1% e que haja suspeita de que o contrato não esteja sendo cumprido, a formulação do teste é:

$$H_0: p = 0.01;$$

$$H_1: p < 0.01;$$

- Análise do caso 3:

A hipótese nula vai de encontro à afirmativa do contrato e a hipótese alternativa retrata a suspeita do pesquisador.

É usual que H_0 se apresente em termos de igualdade de parâmetros populacionais enquanto que H_1 em termos de desigualdade conforme os exemplos dos casos 1, 2 e 3.

O procedimento de um teste de hipótese faz uso de amostras da população de forma que uma das hipóteses será validade à luz das estatísticas estabelecidas pela amostra selecionada.

Assim, as informações de uma amostra aleatória da população de interesse é que servem de base para decisão de qual hipótese é a mais provável de ser verdadeira. Se a informação for consistente com a hipótese nula, então podemos concluir que ela é verdadeira; Caso contrário, concluímos que a hipótese é falsa.

Para testar uma hipótese, devemos extrair uma amostra aleatória, calcular uma estatística de teste apropriada, obter o *p-valor* dessa estatística e então tomar a decisão.

Podemos pensar este *p-valor* como a probabilidade que dá sustentação à manutenção da hipótese nula como verdadeira. Assim, se *p-valor* for muito pequeno, próximo de zero, não teremos motivos para acreditar que a hipótese nula é a verdadeira. Neste caso, rejeitamos a hipótese nula.

13.5 Erros possíveis no teste de hipótese

- Quando se toma uma decisão usando a informação de uma amostra aleatória, estamos sujeitos a **ERROS**.
- Os erros recebem convencionalmente os seguintes nomes:
 - Erro tipo I: rejeitar H_0 quando H_0 é verdadeira.
 - Erro tipo II: aceitar H_0 quando H_0 é falsa.
- Análise dos erros:
 - α é o **nível de significância do teste**, é a probabilidade do erro tipo I, em geral o valor adotado é de 5%.
 - $1 - \alpha$ é o **nível de confiança do teste** (é a probabilidade de aceitar H_0 quando de fato essa hipótese é verdadeira), em geral seu valor é de 95%.
 - β é a **probabilidade do erro tipo II**, seu valor não é fixado já que essa probabilidade depende do valor do parâmetro em H_1 , o qual pode assumir infinitas possibilidades.
 - $1 - \beta$ é conhecida como **poder do teste** (é a probabilidade de rejeitar H_0 quando de fato essa hipótese é falsa).

13.6 Regra de decisão para todos os testes de hipóteses

- A regra de decisão de um teste de hipóteses leva em conta o nível de significância adotado pelo pesquisador, o α , e o p -valor associado à estatística do teste.
 - p -valor $\geq \alpha$
Implica em aceitar H_0 como verdadeira.
 - p -valor $< \alpha$
Implica em rejeitar H_0 como verdadeira e consequentemente aceitar H_1 como verdadeira.
- Exemplo:
Considere que o teste seja formado sobre as hipóteses, $H_0:p = 0.20$ e $H_1:p > 0.20$ com nível de significância $\alpha = 0.05$; Após a seleção da amostra e sua análise estatística, obteve-se hipoteticamente p -valor = 0.39, a conclusão do teste é de que não há evidência suficiente para afirmar que $p > 0.20$, ou seja o teste leva a aceitar H_0 como verdadeira.
Caso, hipoteticamente p -valor = 0.001. Logo, a conclusão do teste é de que não há evidência suficiente para afirmar que $p = 0.20$, ou seja o teste leva a rejeitar H_0 como verdadeira e aceitar $H_1:p > 0.20$.

- Os principais testes estatísticos estão disponíveis no pacote base. A tabela a seguir mostra as funções para realização de diversos testes de hipóteses:

Table 30: Lista de funções para realização de teste de hipóteses do pacote **stats**.

Função	O que faz
<code>ansari.test(x, y, alternative = c('two.sided', 'less', 'greater'), exact = NULL, conf.int = FALSE, conf.level = 0.95)</code> <code>bartlett.test(x,y)</code>	Teste de Ansari-Bradley compara a diferença de escala em duas amostras. Teste de Bartlett compara se a variância entre os grupos são iguais.
<code>binom.test(x, n, p = 0.5, alternative = c('two.sided', 'less', 'greater'), conf.level = 0.95)</code>	Teste Binomial exato, testa a probabilidade de sucesso em um experimento binomial.
<code>Box.test(x, lag = 1, type = c('Box-Pierce', 'Ljung-Box'), fitdf = 0)</code>	Box-Pierce ou Ljung-Box teste para examinar a independência de uma série temporal.
<code>chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = False, simulate.p.value = FALSE, B = 2000)</code>	Teste Qui-quadrado para tabelas de contingência ou avaliar a qualidade do ajuste de duas distribuições.
<code>cor.test(x, y, alternative = c('two.sided', 'less', 'greater'), method = c('pearson', 'kendall', 'spearman'), exact = NULL, conf.level = 0.95, continuity = FALSE, ...)</code>	Teste para associação entre amostras pareadas usando o métodos de Pearson, Kendall ou Spearman.
<code>fisher.test(x, y = NULL, workspace = 200000, hybrid = FALSE, control = list(), or = 1, alternative = 'two.sided', conf.int = TRUE, conf.level = 0.95, simulate.p.value = FALSE, B = 2000)</code> <code>fligner.test(x, ...)</code>	Teste exato de Fisher para tabelas de contingência.
<code>friedman.test(y, ...)</code>	Teste de Fligner-Killeen da igualdade de variâncias.
<code>kruskal.test(x, ...)</code>	Teste de Friedman para a soma de postos com dados em blocos não duplicados.
<code>ks.test(x, y, alternative = c('two.sided', 'less', 'greater'), exact = NULL)</code>	Teste de Kruskal-Wallis para a soma de postos.
<code>mantelhaen.test(x, y = NULL, z = NULL, alternative = c('two.sided', 'less', 'greater'), correct = TRUE, exact = FALSE, conf.level = 0.95)</code>	Teste de Kolmogorov-Smirnov para comparação de duas ou mais distribuições.
<code>mauchly.test(objeto de classe mlm, ...)</code>	Teste Qui-quadrado de Cochran-Mantel-Haenszel para verificar a independência condicional em cada estrato de duas variáveis nominais.
<code>mcnemar.test(x, y = NULL, correct = TRUE)</code>	Teste de esfericidade de Mauchlys.
<code>mood.test(x, ...)</code>	Teste Qui-quadrado de Mcnemar para tabela de contingência.
<code>oneway.test(formula, data, subset, na.action, var.equal = FALSE)</code>	Teste de Mood da igualdade de variâncias.
<code>pairwise.prop.test(x, n, p.adjust.method = p.adjust.methods)</code>	Teste de igualdade da média entre duas ou mais variáveis.
<code>pairwise.t.test(x, g, p.adjust.method = p.adjust.methods, pool.sd = !paired, paired = FALSE, alternative = c('two.sided', 'less', 'greater'))</code>	Teste de igualdade entre pares de proporção.
<code>pairwise.wilcox.test(x, g, p.adjust.method = p.adjust.methods, paired = FALSE, alternative = c('two.sided', 'less', 'greater'))</code>	Teste t para igualdade entre pares de média de grupos com níveis.
	Teste de Wilcoxon para igualdade entre pares de média de grupos com níveis.

Função	O que faz
<pre>poisson.test(x, T = 1, r = 1, alternative = c('two.sided','less','greater'), conf.level = 0.95)</pre>	Teste de Poisson para a taxa média ou para igualdade entre duas taxas médias de distribuição de Poisson.
<pre>power.anova.test(groups = NULL, n = NULL, between.var = NULL, within.var = NULL, sig.level = 0.05, power = NULL) power.prop.test(n = NULL, p1 = NULL, p2 = NULL, sig.level = 0.05, power = NULL, alternative = c('two.sided','one.sided'), strict = FALSE, tol = .Machinedouble.eps^0.25) Calcula o poder do teste de proporção. power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05, power = NULL, type = c('two.sample','one.sample','paired'), alternative = c('two.sided','one.sided'), strict = FALSE, tol = .Machinedouble.eps^0.25) PP.test(x, lshort = TRUE)</pre>	Calcula o poder do teste anova. Calcula o poder do teste t. Calcula o poder do teste Phillips-Perron para raízes unitárias.
<pre>prop.test(x, n, p = NULL, alternative = c('two.sided','less','greater'), conf.level = 0.95, correct = TRUE) prop.trend.test(x, n, score = seq_along(x)) quade.test(y, ...)</pre>	Teste para a proporção de sucessos em vários grupos. Teste Qui-quadrado para a tendência da proporção. Teste Quade para dados em blocos não duplicados.
<pre>shapiro.test(x) t.test(x, y, paired = FALSE, alternative = c('two.sided','less','greater'), conf.level = 0.95)</pre>	Teste de Shapiro-Wilk para verificar a normalidade da variável. Teste t para comparação da média de uma amostra ou duas amostras (pareadas ou independentes).
<pre>TukeyHSD(x, which, ordered = FALSE, conf.level = 0.95, ...) var.test(x, y, alternative = c('two.sided','less','greater'), conf.level = = 0.95) wilcox.test(x, ...)</pre>	Teste de Tukey para comparar par a par a diferença entre médias. Teste F para comparação de igualdade de variâncias. Teste de Wilcoxon para o posto de sinais. Também conhecido como Teste de Mann-Whitney.

- Obs: O pacote **stats** (pacote de funções estatísticas do **R**) é um pacote básico e não precisa ser atualizado.

13.7 Teste de normalidade

- Os principais testes estatísticos têm como suposição a normalidade dos dados, que deve ser verificada antes da realização das análises principais.
- Um teste de normalidade pressupõe duas hipóteses estatísticas sobre a distribuição de probabilidade de uma variável aleatória X:
 - H_0
A distribuição de probabilidade de X é normal.
 - H_1
A distribuição de probabilidade de X não é normal.
- O teste de Shapiro-Wilk é um teste indicado para essa finalidade e sua aplicação no R é muito simples:

```
x <- dados$valor_compra

#Teste de Shapiro-Wilk
shapiro.test(x)

#  Shapiro-Wilk normality test
#
#data:  x
#W = 0.79892, p-value = 0.0003682
```

- Analisando *p-value*:
 - Como $p\text{-value} < 0.05$, conclui-se que pela rejeição de H_0 ao nível de significância de 5%, logo, assume-se que a distribuição de probabilidade da variável *valor_compra* não segue um modelo normal.
 - Caso *p-value* seja maior ou igual a 0.05, conclui-se pela aceitação de H_0 ao nível de significância de 5%.
 - Desse modo, ao aplicar o teste de Shapiro-Wilk, deve-se verificar se $p\text{-value} > 0.05$, nesse caso, aceita-se H_0 ; Caso contrário, rejeita-se H_0 .
 - * Se $p\text{-value} > 0.05$: H_0 ;
 - * Se $p\text{-value} \leq 0.05$: H_1 ;

13.8 Análise gráfica

- Juntamente com o teste de Shapiro, é importante realizar a inspeção gráfica através do histograma da variável que pode ser produzido de diversas formas e com diferentes pacotes:

1. Pacote básico:

```
hist(x)
```

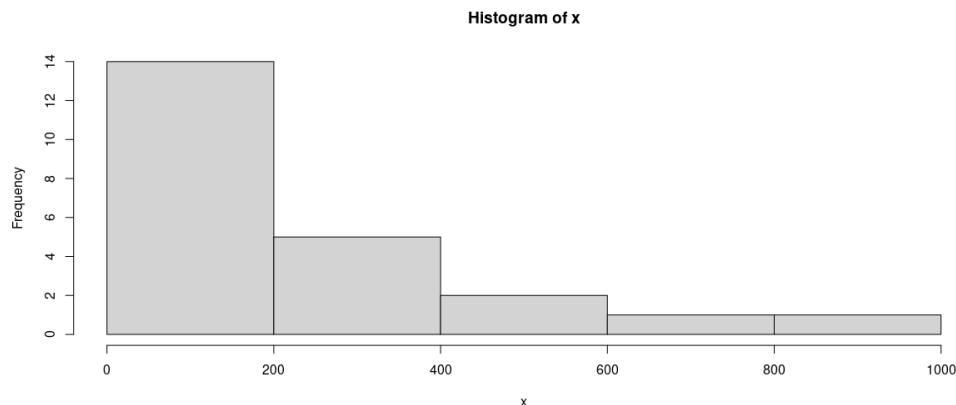


Figure 86: Histograma para análise de normalidade dos dados - pacote básico.

2. Pacote ggplot2:

```
dados %>%
ggplot(aes(x))+
geom_histogram(position = "identity", bins = 5) +
theme_bw()
```

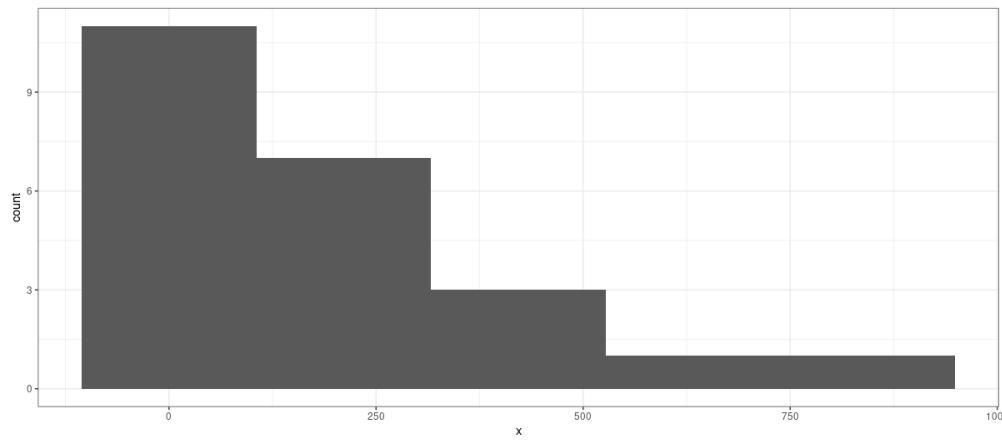


Figure 87: Histograma para análise de normalidade dos dados - pacote ggplot2.

- O gráfico `qqplot` produz uma análise de comparação entre percentil da amostra e percentil teórico da distribuição normal. Os pontos devem estar próximos da reta, indicando que os dados seguem a mesma distribuição normal.

1. `qqplot` com `ggplot2`:

```
ggplot(dados, aes(sample = valor_compra))+
  stat_qq()+
  stat_qq_line()+
  theme_bw()
```

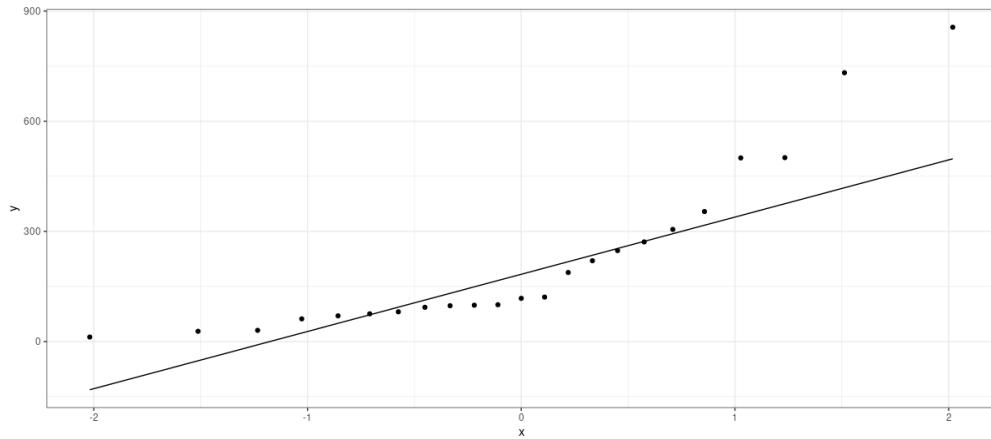


Figure 88: `qqplot` para análise da normalidade dos dados - pacote `ggplot2`.

2. `qqplot` com `DescTools`:

```
PlotQQ(x, args.cband = list(col = SetAlpha("grey", 0.25)))
```

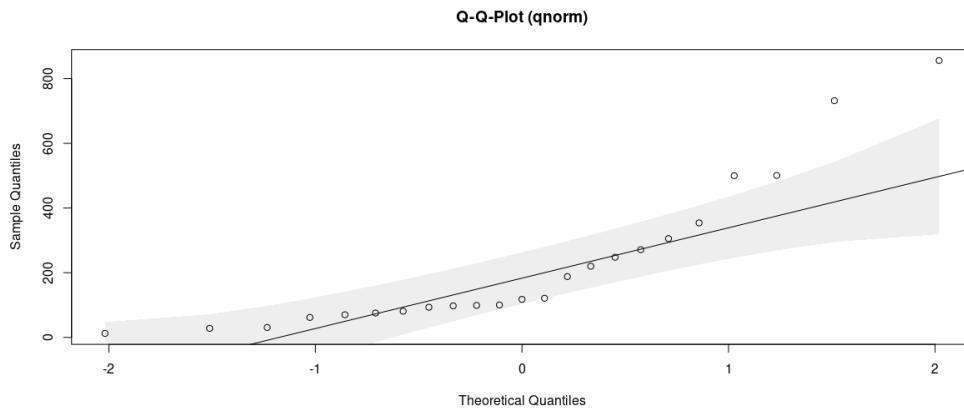


Figure 89: `qqplot` para análise da normalidade dos dados - pacote `DescTools`.

- Pode-se ainda usar a função `PlotFDist` para comparar as curvas teóricas e empíricas dos dados, observando a aderência das curvas.

13.9 E quando os dados não são normais?

13.9.1 O que são dados com distribuição normal?

- Quando a distribuição de probabilidade segue o seguinte modelo:

– Seja uma variável aleatória contínua com média (μ) em que $-\infty < x < \infty$ e o desvio-padrão (σ) é $\sigma > 0$.

– Função densidade de probabilidade:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} x \in R$$

– Cálculo da probabilidade:

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

– Podemos dizer que X tem uma distribuição normal e escrever na fórmula: $X \sim N(\mu, \sigma)$.

- E quando o vetor com a variável aleatória x passa no teste de *Shapiro-Wilk*, hipótese H_0 : $p-value > 0.05$ (5%).

`shapiro.test(x)`

13.9.2 Normalização de dados

- Dizemos que os dados não são normais quando eles não apresentam uma distribuição compatível com as características de uma distribuição normal.
- A normalidade dos dados é importante para atender os pressupostos de muitos testes de hipóteses, especialmente os testes conhecidos na estatística clássica como testes paramétricos.
- Quando o pressuposto da normalidade não é atendido, o analista pode tentar uma transformação da variável original como, por exemplo, aplicar uma transformação logarítmica ou transformação polinomial, ou ainda técnicas consagradas como é o caso da transformação **Box-Cox**.

13.9.2.1 Distribuição normal padronizada

- A distribuição normal padronizada é a distribuição normal sempre com os mesmos parâmetros, que são $\mu = 0$ e $\sigma^2 = 1$.
- Para transformar uma normal qualquer $X \sim N(\mu, \sigma^2)$ em uma normal padrão $X \sim N(\mu = 0, \sigma^2 = 1)$, basta seguir a seguinte forma:
 - Como passar uma normal com $\mu = 10$ (média) em 0:
Basta pegar todos os valores de X e subtrair 10.
 - Para o desvio-padrão(σ) se tornar 1:
Seja qual for o valor de σ , basta dividir por ele mesmo.
 - Assim qualquer distribuição normal pode ser padronizada, de forma que no processo de padronização dos valores da variável aleatória (X) os parâmetros se tornem $\mu = 0, \sigma^2 = 1$. Essa abordagem é dada pela definição de uma nova variável aleatória Z , chamada de **variável aleatória normal padronizada**, dada pela função linear Z :

$$Z = \frac{X - \mu}{\sigma}$$

- Exemplo - Distribuição normal padronizada:

```
#Distribuição normal
#Layout
par(mfrow = c(1,2))

x <- rnorm(50,100,5)
x

#mu = 100, sigma^2 = 25 (ou seja, sigma = 5) (plot 1,1)
curve(dnorm(x,100,5),
      xlim = c(80,120),
      main = "Gráfico 1")

#Padronizando a distribuição normal
transx <- (x - 100)/5
transx

#mu = 0, sigma = 1 (plot 1,2)
PlotFdist(transx, #vetor dados da variável
          args.curve = list(expr = "dnorm(x,mean(transx),sd(transx))", #dados da distribuição normal
                             col = "black"), #Distribuição normal em preto
          args.rug = NA,
          args.dens = NA, #grafico densidade em cinza
          args.boxplot = NA, #Omitir boxplot
          args.ecdf = NA, #Omitir distribuição acumulada
          xlim = c(-4, 4),
          main = "Gráfico 2")
```

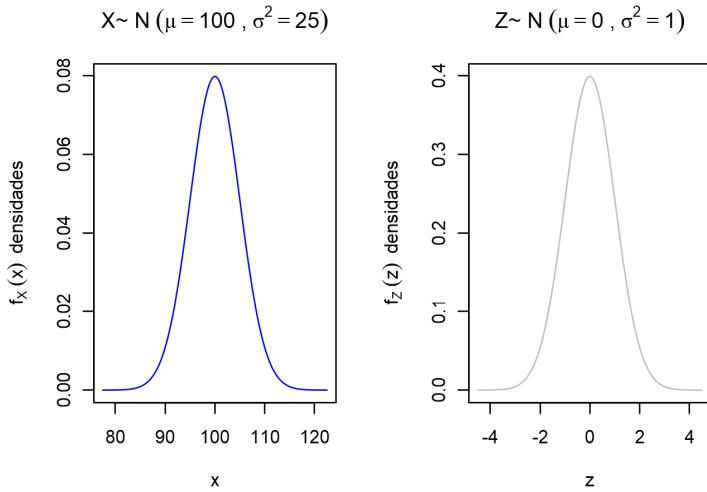


Figure 90: Distribuição normal padronizada, $X \sim N(\mu = 0, \sigma^2 = 1)$

13.9.2.2 Transformação logarítmica

```
#Bibliotecas
library(readr) #Leitura de dados
library(magrittr) #Operador pipe "%>%", concatena linhas de comando
library(DescTools) #Análise descritiva de forma rápida e completa

#Importação de dados
dados <- read.csv2(file = "~/Programacao/R/Dados/Dados_de_importacao/vendas.csv")
head(dados)
str(dados)
summary(dados)

#Tratamento de dados
dados$filial <- as.factor(dados$filial)
str(dados)

#Separação valor compra
x <- dados$valor_compra
head(x)

#Aplicando a transformação logarítmica
transx <- log(x)
head(transx)
#[1] 4.607368 4.393090 4.323338 5.721393 4.795708 3.328268

#Testando a normalidade após a transformação
shapiro.test(transx)

#Shapiro-Wilk normality test
#
#data: transx
#W = 0.9739, p-value = 0.7811
```

Ao aplicarmos a transformação logarítmica aos dados de valor de compra, obtivemos um $p\text{-}valor > 0.05$, o que nos leva a concluir que o logaritmo do valor da compra possui distribuição normal.

13.9.2.3 Transformação raiz quadrada

```
#Bibliotecas
library(readr) #Leitura de dados
library(magrittr) #Operador pipe "%>%", concatena linhas de comando
library(DescTools) #Análise descritiva de forma rápida e completa

#Importação de dados
dados <- read.csv2(file = "~/Programacao/R/Dados/Dados_de_importacao/vendas.csv")
head(dados)
str(dados)
summary(dados)

#Tratamento de dados
dados$filial <- as.factor(dados$filial)
str(dados)

#Separação valor compra
x <- dados$valor_compra
head(x)

#Aplicando a transformação raiz quadrada
transx <- sqrt(x)
head(transx)
#[1] 10.010994 8.993887 8.685620 17.473695 10.999545 5.281098

#Testando a normalidade após a transformação
shapiro.test(transx)

#Shapiro-Wilk normality test
#
#data: transx
#W = 0.92395, p-value = 0.08097
```

Ao aplicarmos a transformação raiz quadrada aos dados de valor de compra, obtivemos um $p\text{-}valor > 0.05$, o que nos leva a concluir que a raiz quadrada do valor da compra possui distribuição normal.

13.9.2.4 Transformação Box-Cox

- A transformação BoxCox consiste em transformar os dados de acordo com a expressão:

$$transx = \frac{x^\lambda - 1}{\lambda}, \lambda \neq 0$$

Onde, λ é um parâmetro a ser estimado pelos dados.

Se $\lambda = 0$ a equação acima se reduz a $transx = \log x$.

```
#Bibliotecas
library(readr) #Leitura de dados
library(magrittr) #Operador pipe "%>%", concatena linhas de comando
library(DescTools) #Análise descritiva de forma rápida e completa

#Importação de dados
dados <- read.csv2(file = "~/Programacao/R/Dados/Dados_de_importacao/vendas.csv")
head(dados)
str(dados)
summary(dados)

#Tratamento de dados
dados$filial <- as.factor(dados$filial)
str(dados)

#Separação valor compra
x <- dados$valor_compra
head(x)

#Lambda ótimo da transformação BoxCox
lambda <- BoxCoxLambda(x)
lambda
#[1] -0.05260669

#Aplicando a transformação ao vetor
transx <- BoxCox(x, lambda)
head(transx)
#[1] 4.091510 3.922403 3.866941 4.940628 4.238582 3.053183

#Testando a normalidade após a transformação
shapiro.test(transx)

#Shapiro-Wilk normality test
#
#data: transx
#W = 0.96909, p-value = 0.6672
```

- Neste exemplo o valor do parâmetro $lambda$ foi obtido pela função `BoxCoxLambda` do pacote `DescTools` que produz uma seleção automática para o parâmetro através do método de Guerrero. Para obter os detalhes dos métodos utilizados na obtenção do parâmetro consultar documentação (`?BoxCoxLambda`).
- Após a transformação BoxCox com $\lambda = -0.0526$ a variável valor de compra transformada apresenta distribuição normal com $p\text{-value} = 0.6672$.

13.10 Plot dois gráficos ou mais

- Para plotar dois ou mais gráficos ao mesmo tempo podemos usar a função `par(mfrow = c(1,c))`, onde os parâmetros `l` é referente ao número de linhas e o `c` ao número de colunas.
- Os gráficos devem ficar logo na sequência do comando `par`:

```
#Layout
par(mfrow = c(1,2))

#mu = 50, sigma = 2 (plot 1,1)
curve(dnorm(x,50,2),
      xlim = c(20,80),
      main = "Gráfico 1")

#mu = 50, sigma = 2 (plot 1,2)
curve(dnorm(x,50,2),
      xlim = c(20,80),
      main = "Gráfico 2")
```

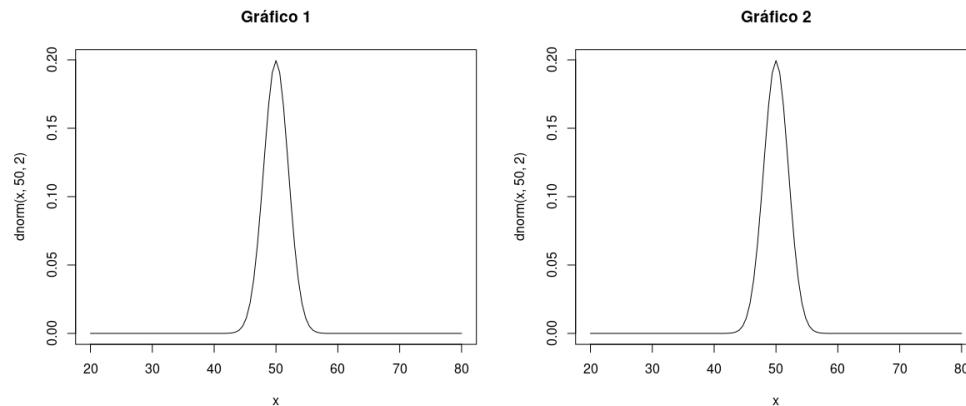


Figure 91: Plotando dois ou mais gráficos de distribuição de probabilidade ao mesmo tempo, usando a função `par(mfrow = c(1,2))`

14 CAP. 11 - MODELANDO A RELAÇÃO ENTRE DUAS VARIÁVEIS

14.1 Modelando a relação entre duas variáveis

- Modelos de relação entre duas variáveis e combinações possíveis de relação dada a natureza da variável, seja numérica ou categórica.

“Existe relação ou associação entre duas variáveis do conjunto de dados em análise?”

- Processo:

– Definição das variáveis (qualitativa/quantitativa)

– Exploração das variáveis individualmente

– **Processo de cruzamento das variáveis**

Procura-se estabelecer padrões de comportamento das variáveis correlacionadas e iniciar com as modelagens possíveis para cada caso.

14.2 Elaboração do modelo (casual)

- A primeira etapa é estabelecer um modelo casual, isto é, determinar qual das variáveis é tida como resposta (**Y**) e qual é tida como explicação (**X**).

– **Y** é a variável resposta.

– **X** é a variável explicação.

- Notação que usamos para representar a relação entre X e Y é $Y \sim X$.

- Combinações de entre as relações das variáveis:

– Categórica com categórica;

– Numérica com categórica;

– Categórica com numérica;

– Numérica com numérica.

14.3 Relações das variáveis

14.3.1 Numérica ~ categórica

14.3.1.1 Passos da criação do modelo numérica ~ categórica

- Definir o que é a variável resposta(Y) do que é variável explicação(X) ($Y \sim X$):
Variável resposta (numérica);
Variável explicação (categórica).
- Definir o **nível de significância**.
Normalmente é igual a 0.05 (ou seja, 5%).
- Análise rápida (função `Desc()`, do pacote `DescTools`) para formulação de modelos entre duas variáveis.
 - Informações relevantes:
 - * Número de grupos
 - * Número de pares
 - * Número de valores perdidos
 - * Média
 - * Mediana
 - * Desvio-padrão
 - * Amplitude
 - * Interquartil
 - * Quantidade
 - * Porcentagem
 - Sobre o gráfico, é gerado **boxplot** para cada categoria e gráfico de comparação entre as médias.
 - Exemplo:

```
Desc(Y ~ X, data, #Variáveis  
      digits = 1, #Digitos significativos  
      plotit = F, #Plotar gráfico TRUE(T) ou FALSE(F)  
      test = kruskal.test) #Teste impregnado na análise (Kruskal-Wallis ou Friedman)
```

- Definir o teste que será trabalhado (**Kruskal-Wallis** ou **Friedman**) e as Hipótese H_0 (afirmação) e H_1 (negação).

- **Kruskal-Wallis**

- * Teste não parametrico para avaliar a diferença entre k grupos.
 - * $k \geq 2$
 - * Neste teste comparamos se a **mediana** da variável resposta (Y) é igual ou não em cada grupo.
 - * Amostra aleatória simples.
 - * Os grupos devem ser independentes.
 - * Variável resposta(Y) deve ser numérica.
 - * Exemplo:
`Desc(..., test = kruskal.test)`

- **Friedman**

- * Para o caso em que se tem amostras pareadas, isto é, amostras cujas unidades foram avaliadas em dois momentos distintos, formam pares de valores para os mesmos elementos, deve-se utilizar o teste **Friedman**.
 - * Exemplo:
`Desc(..., test = friedman.test)`

- Exemplo - Análise relação entre duas variáveis numérico ~ categórica:

```
#A localidade de um estabelecimento fornece resposta quanto ao valor médio de
#compra dos clientes?

#Resposta (Y): valor_compra
#Explicação (X): filial
#Nível de significância: 0.05

#Hipóteses:
#H0: O valor de compra médio é igual em todas as filiais do estabelecimento.
#H1: O valor da compra média não é igual em todas as filiais do esbelecimento.

#Importando dados
dados <- read.csv2(file = "/home/sergio/Programacao/R/Dados/Dados_de_importacao/vendas.csv")
head(dados)
str(dados)

#Tratando os dados
dados$filial <- as.factor(dados$filial) #Transformando dados filial em tipo fator
dados$quinzena <- as.factor(dados$quinzena) #Transformando dados quinzena em tipo fator
str(dados)

#Análise rápida dos dados
Desc(valor_compra ~ filial, dados,
      digits = 1,
      plotit = TRUE,
      test = kruskal.test) #Teste de Kruskal-Wallis
-----
valor_compra ~ filial (dados)

Summary:
n pairs: 23, valid: 23 (100.0%), missings: 0 (0.0%), groups: 3

      A      B      C
mean 118.5 254.0 280.8
median 90.6 159.5 188.0
sd    96.7 248.7 281.2
IQR   39.0 241.0 236.4
n     6      12      5
np    26.1% 52.2% 21.7%
NAs   0      0      0
Os    0      0      0

Kruskal-Wallis rank sum test:
  Kruskal-Wallis chi-squared = 1.1822, df = 2, p-value = 0.5537
-----
#Pelo teste de Kruskal-Wallis: p-value = 0.5537 > 0.05 (nível de significância),
#logo, aceitamos a hipótese H0 como verdadeira.
#H0: O valor de compra médio é igual em todas as filiais do estabelecimento.
```

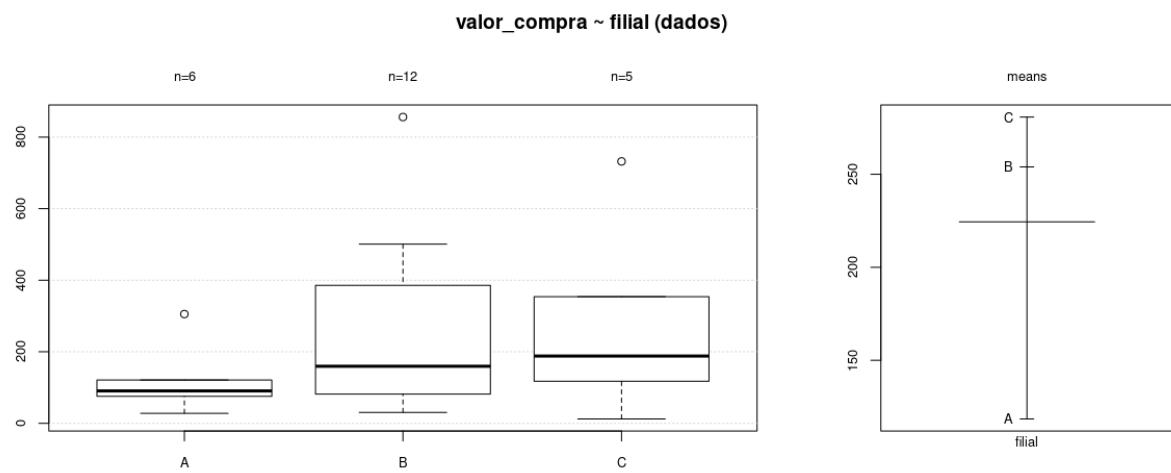


Figure 92: Gráfico referente ao teste de hipóteses variável numérica ~ variável categórica, do pacote **DescTools**.

14.3.1.2 Comparando gráficamente a distribuição de densidade de uma variável numérica condicionada a categorias de outra variável

- Uma forma de ver claramente a diferença de distribuição da variável numérica agrupada em cada categoria é a plotagem do gráfico de distribuição de desidate com as categórias sobrepostas, usando a função PlotMultiDens() do pacote DescTools.

```
PlotMultiDens(desconto_perc ~ quinzena, dados,
               xlab = "Percentual de desconto",
               main = "Percentual de desconto ~ quinzena",
               panel.first = grid(),
               col = c("black", SetAlpha("gray", 0.95), lwd = 2))
```

- Principais parâmetros:

– **$Y \sim X$**

– **data**

Base de dados de origem dos vetores (variáveis).

– **xlab**

Adiciona rótulo para o eixo X de acordo com o texto entre aspas.

– **main**

Adiciona um título principal ao gráfico de acordo com o texto entre aspas.

– **panel.first = grid()**

Adiciona grid ao gráfico.

– **col**

Comando para colorir diversos itens do gráfico, pode ser valores como 1,2,... , ou por nome como 'red', 'blue', etc (consulte nomes com o comando colors() ou sistemas como rgb(), hsv(), gray() e rainbow()). Para cor das fontes use: col.lab, col.main, col.sub.

– **lwd**

Especifica a espessura da linha (1,2,3,...).

- Exemplo - Comparação entre as distribuições de densidade de uma variável numérica entre categorias ($Y \sim X$):

```
#Comparando gráficamente a distribuição de densidade de uma variável numérica
#condicionada a categórias de outra variável.
```

```
#Análise variáveis
Desc(dados$desconto_perc, plotit = F) #Análise variável numérica
```

```
dados$desconto_perc (integer)
```

	length	n	NAs	unique	0s	mean	meanCI'
	23	23	0	5	10	4.39	2.07
					43.5%		6.71
	.05	.10	.25	median	.75	.90	.95
	0.00	0.00	0.00		2.00	10.00	12.00
	range	sd	vcoef	mad	IQR	skew	kurt
	15.00	5.37	1.22	2.97	10.00	0.65	-1.40
	value	freq	perc	cumfreq	cumperc		
1	0	10	43.5%		10	43.5%	
2	2	5	21.7%		15	65.2%	
3	10	4	17.4%		19	82.6%	
4	12	3	13.0%		22	95.7%	
5	15	1	4.3%		23	100.0%	

```
' 95%-CI (classic)
```

```
Desc(dados$quinzena, plotit = F) #Análise variável categórica
```

```
dados$quinzena (factor - dichotomous)
```

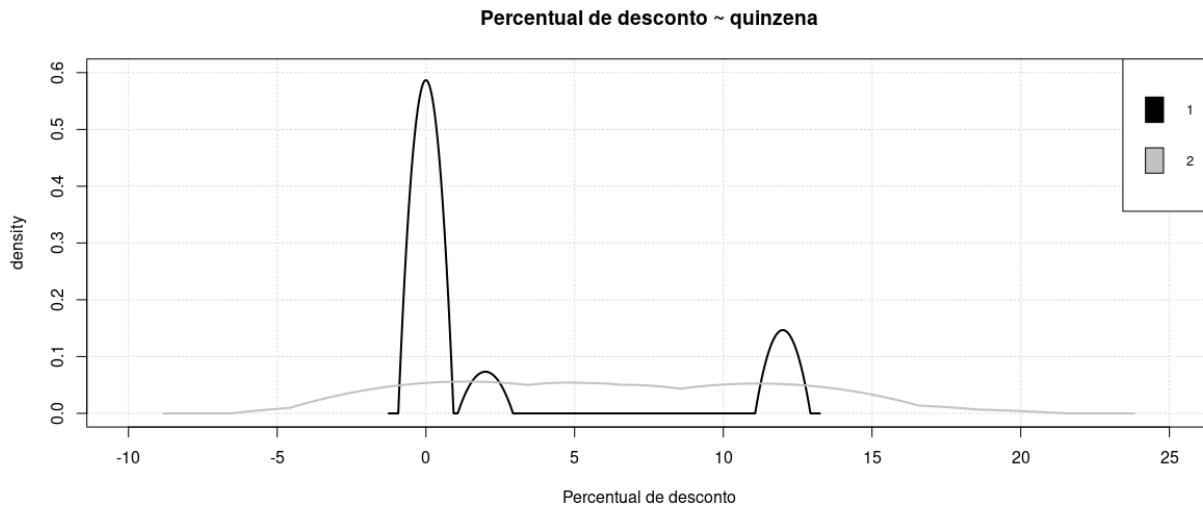
	length	n	NAs	unique
	23	23	0	2
				100.0% 0.0%
	freq	perc	lci.95	uci.95'
1	11	47.8%	29.2%	67.0%
2	12	52.2%	33.0%	70.8%

```
' 95%-CI (Wilson)
```

```
#Gráficos de densidade do desconto percentual sobrepostos
```

```
#para cada grupamento da quinzena
```

```
PlotMultiDens(desconto_perc ~ quinzena, dados,
               xlab = "Percentual de desconto",
               main = "Percentual de desconto ~ quinzena",
               panel.first = grid(),
               col = c("black", SetAlpha("gray", 0.95), lwd = 2))
```



Podemos observar claramente a diferença de distribuição da variável compra quando consideramos a primeira quinzena (vários picos) e a segunda quinzena (um único pico e grande dispersão).

Os dados mostram não haver uma normalidade na distribuição do percentual de desconto (variável numérica) em cada quinzena (variável categórica). O teste de Shapiro pode levar a comprovação dessa hipótese.

```
#Teste de Shapiro-Wilk
shapiro.test(dados$desconto_perc)
#Não há normalidade
#p-value = 7.375e-05 < 0.05
```

14.3.2 Categórica ~ numérica (regressão logística)

14.3.2.1 O que é regressão logística?

- Análise de **regressão** é usada para modelar a relação entre a variável dependente e uma ou mais variáveis independentes.
- **Regressão Logística** é um algoritmo que lida com questões e problemas de classificação.
- É importante ter em mente que a **regressão logística** não é apenas para classificar coisas entre duas categorias. A **regressão logística** trabalha com os conceitos de estatística e probabilidade.
- A **regressão logística** mede a relação entre a variável dependente categórica e uma ou mais variáveis independentes, estimando as probabilidades usando uma função logística (categórica ~ numérica).
- Quer dizer que esse tipo de algoritmo de *machine learning* analisa diferentes aspectos ou variáveis de um objeto para depois determinar uma classe na qual ele se encaixa melhor.

14.3.2.2 Tipos de regressão logística

- Tipos de regressão logística:
 - Para duas categorias:
 - * Binária ou binomial
 - Para mais de duas categorias:
 - * Ordinal
 - * Politômica ou multinomial
- Regressão logística binária ou binomial:

No modelo de regressão logística binomial, os objetos são classificados em dois grupos ou categorias. É quase um jogo do que é e o que não é.

Ex.:
A mensagem é spam ou não é.
A imagem é colorida ou não.
A célula é cancerígena ou não.
- Regressão logística ordinal:

Trabalha com o conceito de categorias ordenadas. Neste caso, os objetos são classificados em três ou mais classes que possuem uma ordem já determinada.

Ex.:
O desempenho do atleta é ruim, justo ou excelente.
O grau de satisfação do paciente com o tratamento é insatisfatório, satisfatório ou muito satisfatório.
- Regressão logística politômica ou multinomial:

No modelo de regressão logística multinomial, os objetos são classificados em três ou mais categorias que não possuem ordem entre si.

Ex.: O animal é gato, leão ou tigre.
Esta fruta é maçã, uma pera, uma manga ou um maracujá.

14.3.2.3 Modelo de regressão logística

- Passos (resumo):
 - Definir as variáveis
 - Formular as hipóteses (H_0 e H_1)
 - Definir o tipo de regressão logística
 - * Binomial
 - * Ordinal
 - * Multinomial
 - Definir o nível de referência da variável resposta (categórica)
 - Modelo
 - * Checagem de pressupostos (Modelo binomial e Modelo multinomial)
 - * Modelagem:
 - Gerar modelo (Binomial e Multinomial).
 - Obter coeficientes do modelo.
 - Cálculando a razão de chance ou *Odds Ratio (OR)*
 - Prevendo cada probabilidade

14.3.2.3.1 Definir as variáveis que serão trabalhadas

- Logicamente devem ser, para o caso do método de regressão logística, variável categórica ~ variável numérica ($Y \sim X$).
 - Variável resposta (Y): variável categórica.
 - Variável explicativa (X): variável numérica.
 - Nível de significância: normalmente 0.05 (5%)

14.3.2.3.2 Formulação de hipóteses

- Formular as hipóteses há serem respondidas pela inferição.
Ex.:
 - H_0 : Não há relação entre local de compra e desconto na compra.
 - H_1 : Há relação entre local de compra e desconto na compra.

14.3.2.3.3 Definir o tipo de regressão logística que será aplicada

- Com base na característica da variável categórica (resposta), fica claro qual o tipo de regressão logística devemos trabalhar (**binomial**, **ordinal** ou **multinomial**).
- Pacote **nnet** fornece modelo de regressão logística, a função **multinom()** é usada para geração do modelo e posteriormente a obtenção dos coeficientes do modelo.

14.3.2.3.4 Categoría de referência

- Ao preparar os dados, a variável resposta (categórica) deve ser um fator (factor) e é necessário definir o nível de referência desejado pelo analista, da variável resposta, através do comando **relevel** do pacote **stats**.

Exemplo definindo a categoria de referência como filial A:

```
dados$filial_f <- relevel(dados$filial, ref = "A")
```

- Ao interpretar a análise é preciso ter em mente qual a categoria usada como referência.
- Para checar a categoria de referência, basta usar a função **levels(data\$variável_categórica)**, a primeira na fila apresentada é a categoria de referência.

14.3.2.3.5 Checagem de pressupostos

- Checagem de pressupostos do modelo binário/binomial:
 1. A variável dependente deve ser dicotômica e mutuamente exclusivas.
A variável resposta (categórica) deve ter duas categorias (**dicotómica**) e uma variável classificada em uma categoria não pode estar em outra categoria também (**mutuamente exclusiva**).
 2. Independência das observações (sem medidas repetidas).
Não tem medidas repetidas, nem medidas ao longo do tempo.
 3. Ausência de *outliers*/Pontos de alavancagem (Testado no modelo).
 - Pontos de alavancagem é a medida dos valores de observação da variável independente, ou seja, valor desproporcional no próprio valor ajustado. São pontos que apresentam um perfil diferente dos demais no que diz respeito aos valores.
 - Plotando gráfico para identificar *outliers*/Pontos de alavancagem.
`plot(modelo, which = 5)`
 - * O modelo é uma variável com o modelo de regressão logística salvo dentro dela (será visto no próximo tópico).
 - * Os pontos *outliers*/Pontos de alavancagem aparecem longe da linha de regressão e fora dos limites da linha tracejada.
 - Verificar os resídos padronizado:
 - * Verificar o resumo dos resídos padronizado, os limites devem estar dentro da faixa $mínimo > -3$ e $máximo < 3$.
`summary(stdres(modelo))`
 - * Se os limites estiverem fora dessa faixa, existem *outliers*.
 - * A função `stdres()` é do pacote MASS.

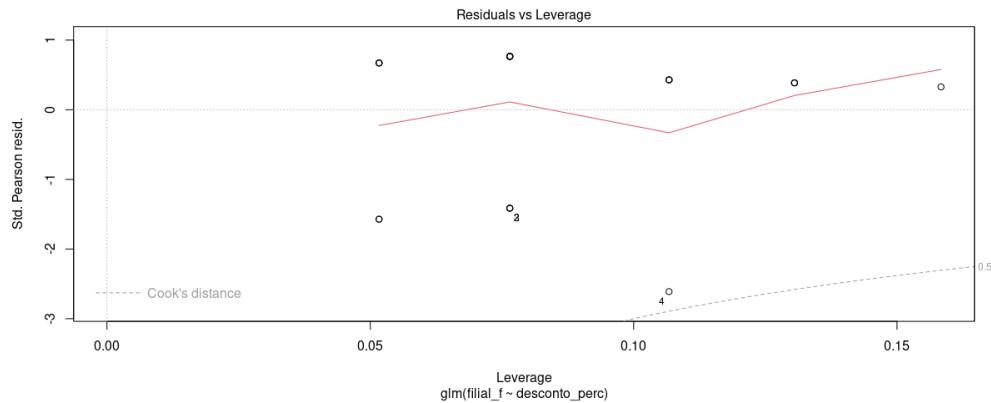


Figure 93: Gráfico residual, usando a função `plot(modelo, which = 5)`, para observar possíveis pontos *outliers*/pontos de alavancagem.

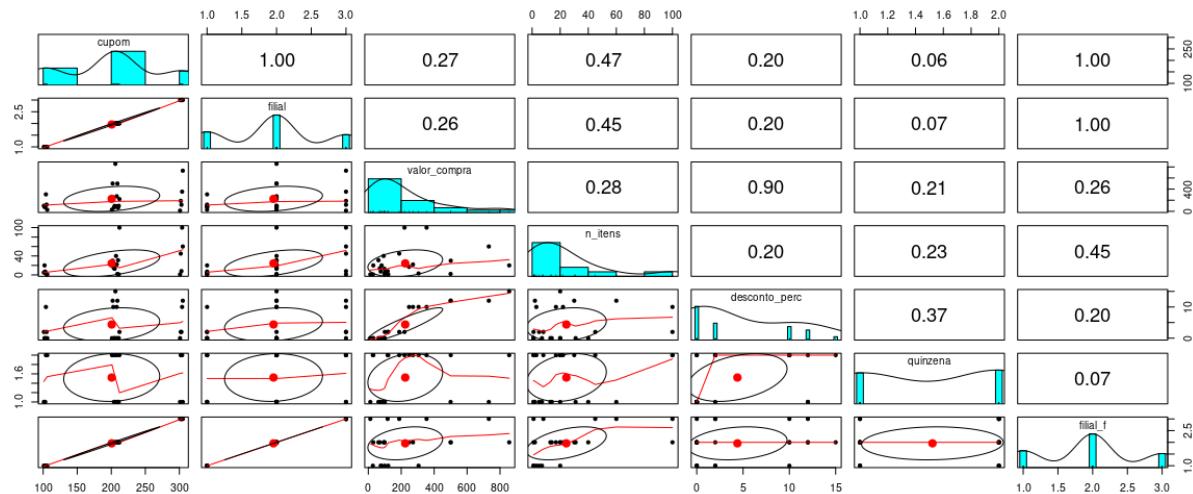


Figure 94: Painel para avaliar possíveis multicolinearidade, usando a função `pairs.panels()`, do pacote psych.

- Checagem de pressupostos do modelo ordinal:

- Checagem de pressupostos do modelo multinomial:
 1. A variável dependente seja nominal (categorias mutuamente exclusivas).
Uma variável classificada em uma categoria não pode estar em outra categoria também (**mutuamente exclusiva**).
 2. Independencia das observações (sem medidas repetidas).
Não tem medidas repetidas, nem medidas ao longo do tempo.
 3. Ausência de multicolinearidade.
 - Relação muito alta (acima de 0.8 ou 0.9) entre variáveis independentes (de explicação numérica).
 - Se a relação entre duas ou mais variáveis independentes for muito alta, torna a análise redundante.
 - Usamos a função `pairs.panels(dados)`, do pacote `psych`, que retorna um painel com as relações entre variáveis. Os valores de colinearidade entre as variáveis é a linha e coluna em que as variáveis se encontram.
Ex.: A correlação entre “Etnia” e “libEcon” é de 0.18.
 - Caso o valor for acima de 0.9 (ou 0.8 para alguns autores) existe a colinearidade entre as variáveis. Causando redundância na análise dessas duas variáveis.
 4. Ajuste do modelo.
 - Consiste em comparar dois modelos, o modelo construído do problema (`mod`) e um modelo nulo (`mod0`) (sem nenhum previsor, ou seja, que não prevê nada).
 - Salvamos esses modelos em variáveis (`mod` e `mod0`).


```
mod <- multinom(Voto ~ ConfPol + LibEcon + Etnia, data = dados, model = TRUE)
mod0 <- multinom(Voto ~ 1, data = dados, model = TRUE)
– Usamos a função anova() para comparar os modelos.
anova(mod, mod0)
```
 - As hipóteses são:
 - * H_0 : Ajuste do modelo construído = ajuste do modelo nulo, ou seja, $p > 0.05$.
Explicação H_0 : O modelo construído não prevê nada, assim como o modelo nulo.
 - * H_1 : Ajuste do modelo construído \neq ajuste do modelo nulo, ou seja, $p \leq 0.05$.
Explicação H_1 : Existe uma diferença entre os modelos, ou seja, o modelo construído é melhor que o modelo nulo, logo, consegue prever algo.
 - * O valor de p é dado pelo elemento `Pr(chi)` na função `anova()`.

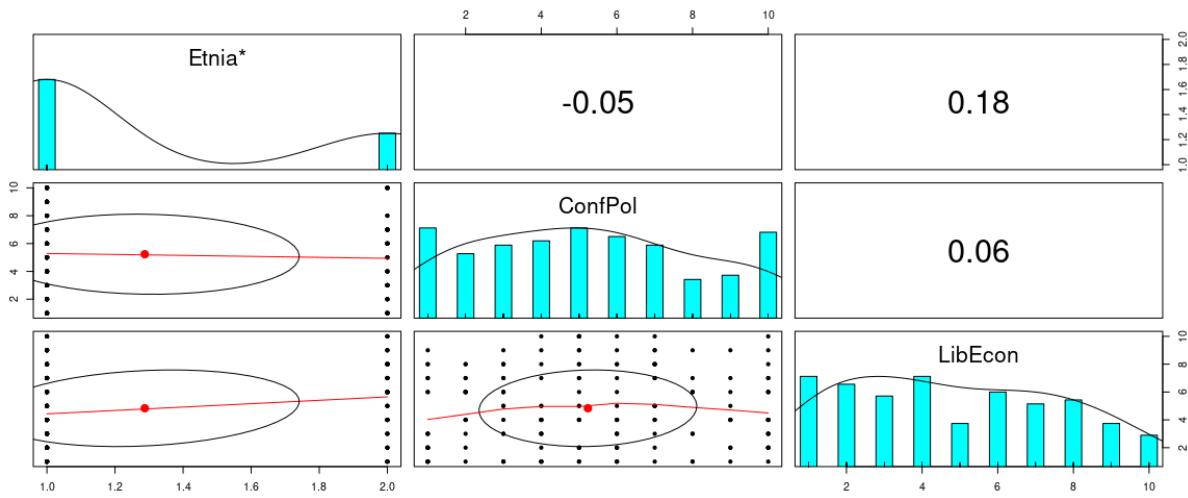


Figure 95: Painel para avaliar possíveis multicolinearidade, usando a função `pairs.panels()`, do pacote psych.

14.3.2.3.6 Modelagem

- Binomial:
 - Para gerar o modelo binomial, usamos a função `glm()`.
 - Sintaxe:

```
modelo_glm <- glm(Y ~ X1 + X2 + ..., family = binomial(link = 'logit'), data = dados)
```

Onde,
 - * `modelo_glm` é a variável onde será armazenada o modelo;
 - * `glm()` é a função de modelos lineares generalizados;
 - * Y é a variável dependente (categórica binomial, variável resposta);
 - * X_n são as variáveis independentes (numéricas, variáveis explicação);
 - * `family` é o tipo de modelo de regressão logística, no caso, binomial com link a função `logit`;
 - * `data` é o banco de dados.
 - Para extrair os coeficientes do modelo usamos a função `summary(modelo_glm)`.
Onde,
$$y(x) = \ln \left(\frac{P(\text{outro nível})}{P(\text{nível de referência})} \right) = intercept + valor_variável \cdot x$$
 - * x é o valor da variável numérica.
 - * É basicamente a equação da reta.
 - * *Estimate* são os coeficientes: `intercept` e `valor_variável`.
 - * Pr é o *p-valor*, neste caso *p-valor* deve ser **menor** que 0.05 para ser estatisticamente significativo para o modelo de previsão.
- Ordinal:
- Multinomial:

14.3.2.4 A razão de chance ou *Odds Ratio* (OR)

- *Odds*

É a razão de um evento de interesse ocorrer sobre os eventos de não interesse.

$$Odds = \frac{P(A)}{P(\bar{A})}$$

- Probabilidade

É a razão de um evento de interesse ocorrer sobre o total de eventos.

$$Probabilidade = \frac{P(A)}{Total}$$

- Razão de chance ou *Odds Ratio* (OR)

É a relação entre dois eventos ocorrerem, dada pela razão entre a probabilidade de um primeiro evento de interesse ocorrer, sobre o evento de não interesse desse primeiro evento, sobre um segundo evento de interesse de ocorrer, sobre o evento de não interesse desse segundo evento.

$$OR = \frac{\frac{P(A)}{P(\bar{A})}}{\frac{P(B)}{P(\bar{B})}}$$

ou seja,

$$OR = \frac{Odds(A)}{Odds(B)}$$

Para obtermos a razão de chance fornecida pelo modelo construído de regressão logística, utilizamos a função inversa ao logaritmo, ou seja, a função exponencial:

Se

$$y(x) = \ln \left(\frac{P(outro\ nível)}{P(nível\ de\ referência)} \right) = intercept + valor_variável \cdot x$$

então,

$$OR = exp^{y(x)}$$

$$OR = e^{(intercept+valor_variável \cdot x)}$$

$$OR = novo_intercept + novo_valor_variável^x$$

x é o valor da variável numérica.

Logo, para achar os coeficientes da equação de razão de chance (OR) no R:
`exp(coef(modelo))`

- Análise do *Odds*:
 - $Odds > 1$
Significa que a probabilidade do evento ocorrer é maior do que de não ocorrer.
 - $Odds < 1$
Significa que a probabilidade do evento ocorrer é menor do que de não ocorrer.
 - $Odds = 1$
Significa que a probabilidade do evento ocorrer ou não ocorrer é a mesma.
- Análise dos coeficientes do logaritmo da razão de chance ($y(x)$):
 - O aumento de uma unidade de valor da variável numérica (x) está associado a um aumento no logaritmo da razão de chance ($y(x)$) entre a variável categórica B versus a variável categórica de referência A, no valor de *valor_variável*.
$$y(x) = intercept + valor_variável \cdot x$$
 - Exemplo:
O desconto na compra fornece resposta quanto ao local de compra?
$$(Y \sim X)$$

y é a variável categórica = filial.
 x é a variável numérica = o valor de desconto na compra.

$$y(x) = 0.27 + 0.118x$$

Então,
o aumento de uma unidade de valor da variável desconto (x) está associado a um aumento no logaritmo da razão de chance ($y(x)$) entre comprar na filial **B** versus comprar na filial **A** (categoria de referência), no valor de 0.118.

 - Conclusão:
 - * A análise a partir do logaritmo da razão de chance ($y(x)$) nos oferece informação sobre o modelo, porém para uma análise mais conclusiva a razão de chance (**OR**) se mostra melhor.
 - * O logaritmo da razão de chance ($y(x)$) é um passo importante para obtenção da razão de chance (**OR**).

- Análise do modelo a partir da razão de chance (**OR**):
 - A partir da equação podemos avaliar a razão de chance em função dos valores de x (variável explicação, numérica).
 - Exemplo:
O desconto na compra fornece resposta quanto ao local de compra?

$$(Y \sim X)$$

y é a variável categórica = filial.

x é a variável numérica = o valor de desconto na compra.

`summary(modelo)`

$$y(x) = 0.27 + 0.118x$$

Logo,

`exp(coef(modelo))`

$$OR = e^{0.27+0.118x} = 1.31 + 1.126^x$$

Então,

Se $x = 0$ temos que a razão de chance (**OR**) entre filial B e a filial A é de 2.31 ($1.31 + 1.126^0$), isto é, a probabilidade de não obter desconto na compra é 2.31 vezes maior na filial B do que na filial A (nível de referência).

- Conclusão:

- * Através da análise de razão de chance (**OR**) podemos entender melhor a relação entre os níveis da variável categórica em relação ao nível de referência.
- * Obtemos a incidência de determinadas categorias.

14.3.2.5 Prevendo as probabilidades

- Usamos a função `predict()` para obter as probabilidades estimada a partir do modelo de regressão logística.

- Sintaxe:

A função `predict` tem como argumentos:

– `objeto = modelo`

O modelo é a variável salva com o modelo de regressão logística.

– `newdata = valor`

Dados de entrada para prever os valores.

Ex.:

```
valor <- data.frame(desconto_perc = seq(0,20,5))
```

Vetor de valores de 0 a 20 com o passo de 5.

– `type = "prob"`

Retorna uma probabilidade para cada valor da variável numérica distribuído pelos níveis.

Ex.:

Para o valor 0 (numérico) os níveis A, B e C são 35%, 46% e 19% respectivamente, no total somando os percentuais 100%.

- Exemplo:

```
#Prevendo a probabilidade de cada filial face aos descontos
valor <- data.frame(desconto_perc = seq(0,20,5))
p <- predict(modelo,newdata = valor,type = "prob")
tabela <- data.frame(valor,p)
tabela
#  desconto_perc      A          B          C
#1      0 0.35138307 0.4605713 0.1880456
#2      5 0.22975316 0.5437937 0.2264532
#3     10 0.14105843 0.6028764 0.2560651
#4     15 0.08291156 0.6398835 0.2772050
#5     20 0.04740715 0.6606725 0.2919203
```

14.3.2.6 Avaliando a significância dos parâmetros

- A **análise da significância estatística** é considerado um procedimento para verificar a discrepância de uma hipótese estatística em relação aos dados observados.
- Para avaliar a significância dos parâmetros estimados vamos obter a estatística de z e avaliar o *p-valor* de um teste bicaudal, ou seja, vamos testar se os parâmetros são nulos ou não.
- Padronizando os coeficientes do modelo:

– Em **R**:

```
z <- summary(modelo)$coefficients / summary(modelo)$standard.errors
```

– Modelo matemático:

$$z = \frac{\text{coeficientes do modelo}}{\text{erro padrão da função } lm()}$$

- Obtendo o p-valor, em **R**:

```
pval <- (1 - pnorm(abs(z), 0, 1))*2
```

Table 31: Exemplo de coeficiente estimados do modelo de regressão logística

	Z-intercept	Z-x	pval-intercept	pval-x
B	0.438	1.028	0.661	0.304
C	-0.803	0.933	0.422	0.351

- Análise dos parâmetros:

Como todos os p-valor > 0.05 , logo os coeficientes estimados são significativos.

14.3.2.7 Avaliando a acurácia do modelo

- Acurácia é a proporção de previsões corretas do modelo.
- Para obtê-las construímos a matriz de confusão do modelo em análise:

$$Acurácia = \frac{\text{total de acertos}}{\text{total de dados}}$$

- Gerando matriz de confusão no R:

```
mc <- table(dados$filial_f, predict(modelo))
```

– `predict`

Aplica o método de regressão lm ao objeto.

- Matriz de confusão

– Mostra o número de classificações corretas (em cada linha) versus o número de classificações preditivas (em cada coluna) para cada nível de variável sob análise.

– A diagonal da matriz de confusão são as previsões corretas.

Table 32: Matriz de confusão do modelo de regressão logística.

	A	B	C
A	0	6	0
B	0	12	0
C	0	5	0

- Análise da matriz de confusão:

- Classificou 6 vezes filial A como B (erro).
- Classificou 12 vezes filial B como B (acerto).
- Classificou 5 vezes filial C como B (erro).

- Cálculo de acurácia:

$$Acurácia = \frac{\text{Soma diagonal}}{\text{Soma de todos os valores da matriz}}$$

ou seja,

```
acuracia <- sum(diag(mc))/sum(mc)
```

O resultado é a porcentagem de acerto do modelo.

$$Acurácia = \frac{12}{23}$$

$$Acurácia \approx 52\%$$

- Exemplo - Avaliando a acurácia do modelo:

```
#Avaliando a acurácia do modelo

#Acurácia é a proporção de previsões corretas do modelo.
#Gerar uma matriz de confusão do modelo de análise.
#Acurácia = total de acertos / total de dados

mc <- table(dados$filial_f, predict(modelo)) #Gerar a matriz de confusão
kable(mc) #Apresentação da tabela em formato kable

#   |   | A| B| C|
# |:--|---|---|---|---|
# |A | 0| 6| 0|
# |B | 0| 12| 0|
# |C | 0| 5| 0

#Classificou 6 vezes filial A como B (erro)
#Classificou 12 vezes filial B como B (acerto)
#Classificou 5 vezes filial C como B (erro)

#Cálculo da acurácia
sum(diag(mc))/sum(mc) #A diagonal da matriz de confusão são as previsões corretas.
#[1] 0.5217391
#Acurácia = 52%
#Prever erroneamente = 48%
#Esse modelo possui baixa acurácia.
```

14.3.3 Categórica ~ categórica

14.3.4 Numérica ~ numérica

15 ANDAMENTO DOS ESTUDOS

Assunto em andamento:

Atualmente estou estudando Cap. 11 - Modelando a relação entre duas variáveis.

REFERÊNCIAS

ALCOFORADO, L. F. UTILIZANDO A LINGUAGEM R: conceitos, manipulação, visualização, modelagem e elaboração de relatórios. Rio de Janeiro: Departamento de estatística da UFF; Alta Books Editora, 2021.