MAE 5905: Introdução à Ciência de Dados Gabarito Lista 2

Alex Monito Nhancololo

2026-05-11

Critérios de avaliação

Independentemente do software empregado, aplicam-se os seguintes critérios de avaliação:

- Justificar sempre as escolhas metodológicas, não bastando apresentar apenas o código e/ou sua saída.
- Manter o documento limpo, sem # desnecessários, warnings, erros ou descrições triviais.
- Incluir tabelas e gráficos bem formatados, com títulos, legendas e rótulos claros para interpretação autônoma.
- Organizar todo o código (caso haja) em apêndice/anexo , mantendo coerência, consistência e redação técnica objetiva. Os códigos devem constar no trabalho, mas como apêndices, e não no corpo do texto.
- Se feito em grupo, deve constar a contribuição de cada membro.
- Não pode haver evidências de cópia de colegas deste curso, de anos anteriores, ChatGPT, etc.
- Excluindo os códigos, o trabalho não pode exceder 20 páginas, sendo necessário incluir apenas o essencial. Não existe número mínimo, quanto menor melhor.
- Comentários sobre funções básicas (como `set.seed()` ou outras funções triviais) não devem ser incluídos.

Dica

- O capítulo 1, seções 1.6 a 2.5 das notas de aula (Nhancololo, Scalon e Alencar, 2026; link) mostra como as tabelas, equações, graficos, etc., devem ser formatadas, seja em `Word, Overleaf/LaTeX, Markdown ou Quarto`.
- A seção 1.6 do mesmo material mostra como deve ser a saída do modelo caso haja ajuste; não é permitido apenas copiar o `summary` não formatado e colocá-lo no trabalho.
- Use `echo = TRUE, message = FALSE, warning = FALSE, comment = ""` para mostrar código e saída, ocultar mensagens automáticas e warnings, e remover o símbolo padrão (##) antes da saída do código.

> **!** Códigos no Apêndice
>
> Os códigos para reproduzir tudo que foi feito estão na Seção do apêndice.

EXERCÍCIO 1

Item (a)

Tabela 1: Variáveis simuladas (6 das n = 100).

X	erros
-0.08378436	0.07901367
-0.98294375	-0.10924157
-1.87506732	0.90552492
-0.18614466	0.97537461
-0.63348570	0.94308295
1.09079746	-0.01697289

> **i** Nota
>
> A função `rnorm()` por padrão simula valores de uma distribuição normal, com média 0 e desvio padrão 1. Caso precise usar média e desvio padrão diferentes, altere os parâmetros `mean` e `sd` em `rnorm(n, mean = 0, sd = 1)`. Você pode rodar `?rnorm()` para ver a documentação.

Item (b)

Sejam $\beta_0 = 30$, $\beta_1 = 7.3$, $\beta_2 = -3.6$ e $\beta_3 = 0.65$. O vetor Y, fica: $Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot X^3 + \text{erros}$

```
[1] 29.4417342 18.6197225  0.2752451 29.4875868 24.7086992 34.5060456
[7] 19.1066976 34.6538984 25.5669753 27.1148154
```

item c) Considere os β_i desconhecidos, e estime-os usando a função Linear Models (`lm`).

$$\hat{\text{Y}} = 30.1 + 7.29(\text{X}) - 3.49(\text{X\^2}) + 0.64(\text{X\^3}) \tag{1}$$

> **i** Nota
>
> No `R`, ao usar `lm()` para ajustar modelos de regressão, os operadores em fórmulas têm significados especiais. Especificamente, o operador `^` indica interações `(:)`, não potências. Por exemplo, `X^2` em uma fórmula é interpretado como `X + X:X`, o que não gera o termo quadrático.
> Para incluir potências, usa-se `I()`, para operações matemáticas dentro de fórmulas no R, garantindo que potências e outras transformações sejam corretamente interpretadas. Assim:
>
> - **Errado**: `lm(Y ~ X + X^2 + X^3)`, trata `X^2` e `X^3` como interações.
> - **Correto**: `lm(Y ~ X + I(X^2) + I(X^3))`,calcula de fato os quadrados e cubos de `X`.
>
> Você pode utilizar alternativamente, a função `poly(X, 3, raw = TRUE)`.

Tabela 2: Estimativas dos parâmetros

Characteristic	**Beta**	**95% CI**	**p-value**
(Intercept)	30.105	29.868, 30.341	**<0.001**
X	7.285	6.952, 7.618	**<0.001**
I(X^2)	-3.490	-3.646, -3.333	**<0.001**
I(X^3)	0.639	0.539, 0.740	**<0.001**

Abbreviation: CI = Confidence Interval
R^2 = 0.990; Adjusted R^2 = 0.990; Sigma = 0.940; Statistic = 3,157; p-value = <0.001; df = 3; Log-likelihood = -134; AIC = 277; BIC = 290; Deviance = 84.8; Residual df = 96; No. Obs. = 100

As estimativas mostradas na Tabela 2 possuem valores próximos àqueles que foram usados para simular (Tabela 1). Tanto o valor de R^2 quanto o $Adj.R^2$ foi de 0,99, indicando que a curva estimada pelo modelo consegue explicar grande parte da variabilidade dos dados.

Para avaliar a adequação do modelo, realizou-se uma análise de resíduos (Figura 1). A Verificação Preditiva Posterior confirma que o modelo replica com precisão a distribuição dos dados observados. As premissas de Linearidade e Homocedasticidade são satisfeitas, com linhas de referência estáveis e ausência de padrões de funil. A Normalidade dos Resíduos é verificada, com os pontos seguindo a linha teórica, apesar de um leve desvio a partir da abscissa 1,5. No gráfico de Observações Influentes, todos os pontos situam-se dentro dos limites da Distância de Cook, embora os casos 30 (alto leverage), 32, 52, 90 e 94 sejam destacados. Por fim, a Colinearidade (VIF < 5) indica baixa incerteza nos parâmetros, mesmo diante da natureza polinomial do modelo.

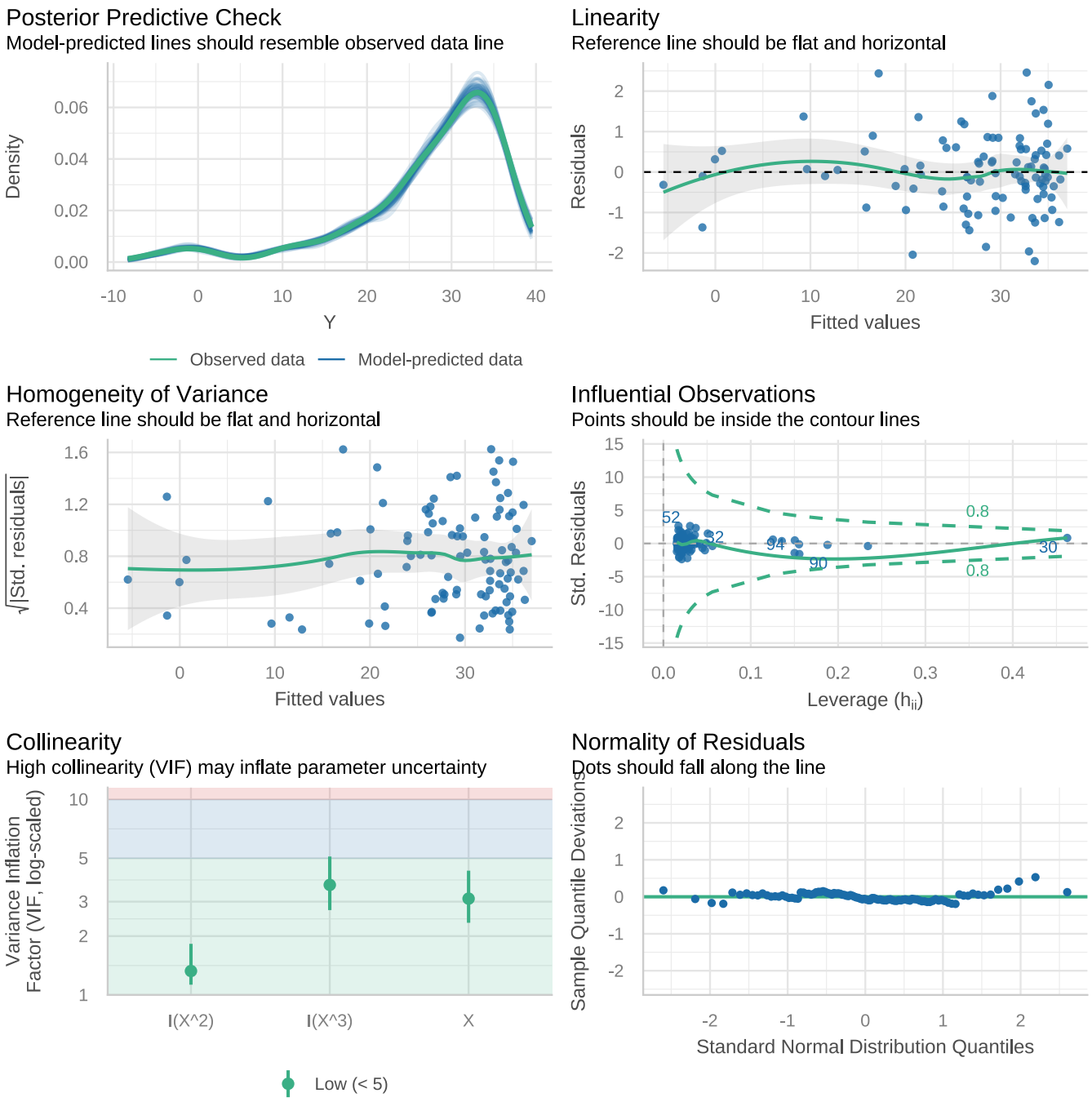

Figura 1: Diagnóstico do modelo ajustado

Item d)

O modelo Ridge ($\alpha = 0$) mantém a estrutura original ao preservar todos os preditores (covariaveis), aplicando uma contração proporcional nos coeficientes para reduzir a variância e evitar instabilidades. Com a penalidade λ baixa ($2,21 \times 10^{-3}$), as estimativas obtidas ($\beta_1 = 7,4208$; $\beta_2 = -3,5135$; $\beta_3 = 0,5898$) permanecem próximas aos considerados no item b), garantindo um ajuste fiel que distribui o peso entre os termos sem descartar informação (Tabela 3).

O Lasso ($\alpha = 1$) atua como um seletor de variáveis, sendo capaz de zerar termos redundantes para favorecer a parcimônia e a interpretabilidade. Nesta análise, o método preservou todas as covariaveis

por identificar a força da tendência cúbica, resultando em coeficientes ($\beta_1 = 7,3712$; $\beta_2 = -3,4984$; $\beta_3 = 0,5847$) ligeiramente menores que os do Ridge sob um λ de $1,25 \times 10^{-2}$ (Tabela 3).

A qualidade do ajuste é confirmada pelas métricas de performance, com erro médio (RMSE) de $0,9416$ no Ridge e $0,9421$ no Lasso, além de um elevado R^2 ($0,9856$). A semelhança entre os resultados confirma que a estrutura simulada foi capturada, exigindo pouca intervenção da regularização (Tabela 3).

> **Dica**
>
> Para uma base teórica sólida, recomenda-se o livro de Morettin e Singer (2023). É fundamental, ainda, o estudo dos artigos de Tibshirani (1996), que introduz a regressão LASSO e discute sua relação com a regressão Ridge (Hoerl e Kennard, 1970) e os métodos de seleção de subconjuntos. Complementarmente, o artigo de Zou e Hastie (2005) apresenta o Elastic Net, técnica que combina as vantagens das abordagens anteriores.

Tabela 3: Estimativas do modelo Ridge e LASSO

Parâmetro	Ridge (L_2)	Lasso (L_1)
Intercepto (β_0)	29.8331	30.0667
X	6.3869	7.2958
X^2	-3.1801	-3.4454
X^3	0.7260	0.6203
Alpha (α)	0	1
Lambda (λ)	7.64×10^{-1}	2.54×10^{-2}
RMSE	**1.1985**	**0.9232**
R^2	**0.9885**	**0.9900**

EXERCÍCIO 2

O conjunto de dados Weekly compreende os retornos semanais do índice S&P 500 entre 1990 e 2010. As variáveis de defasagem (Lag1 a Lag5), que representam o desempenho de semanas anteriores, exibem um comportamento estatístico uniforme: as médias entorno de zero, indicando que não há uma tendência de alta ou baixa persistente apenas com base no passado imediato (Tabela 4). A maior parte dos retornos varia entre -1,15% e 1,41%, embora a presença de valores extremos possa ser um indicativo de forte oscilação e mudanças bruscas no mercado financeiro.

Tabela 4: Estatísticas descritivas das variáveis de retornos defasados e volume.

Variável	**Média ($\bar{x}$)**	**Desvios-padrão (σ)**	**IQR**	**Assimetria**	**Curtose**	Q_1	**Mediana (Q_2)**	Q_3
Lag1	0.15	2.36	2.56	-0.48	5.72	-1.15	0.24	1.41
Lag2	0.15	2.36	2.56	-0.48	5.71	-1.15	0.24	1.41
Lag3	0.15	2.36	2.57	-0.48	5.67	-1.16	0.24	1.41
Lag4	0.15	2.36	2.57	-0.48	5.67	-1.16	0.24	1.41
Lag5	0.14	2.36	2.57	-0.48	5.66	-1.17	0.23	1.41
Volume	1.57	1.69	1.72	1.62	2.09	0.33	1.00	2.05
Today	0.15	2.36	2.56	-0.48	5.72	-1.15	0.24	1.41

O Volume de negociações apresenta média de 1,57, maior concentração entre 0,33 e 2,05 e assimetria, indicando que, embora o fluxo de transações seja moderado na maior parte do tempo, ocorrem picos esporádicos de alta atividade. Diferente dos retornos, que oscilam em torno de uma média fixa, o volume exibe uma tendência de crescimento consistente ao longo dos anos, refletindo o aumento da liquidez e da participação no mercado nas últimas duas décadas (Figura 2).

Essa semelhança entre os "Lags" sugere que as propriedades do mercado (como o nível de risco e a amplitude das variações) são persistentes ao longo do tempo. No entanto, essa uniformidade reforça a dificuldade de prever a direção do índice baseando-se apenas em resultados passados, confirmando a natureza incerta e instável das bolsas de valores, onde o volume sinaliza a intensidade da atividade, mas os retornos permanecem essencialmente imprevisíveis.

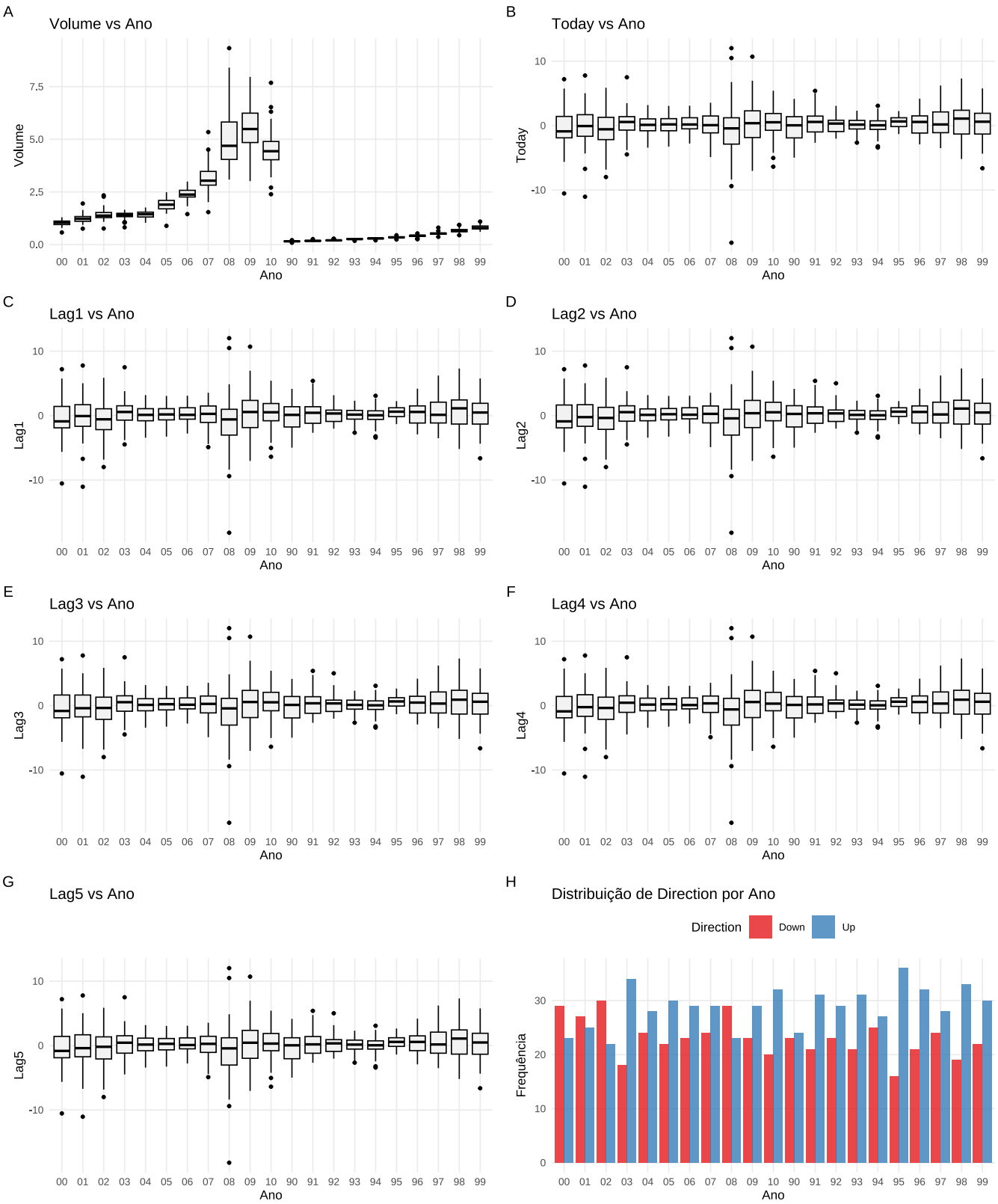

Figura 2: Análise exploratória temporal das variáveis do dataset Weekly.

Pela análise por direção (Direction) observa-se uma sobreposição quase entre as distribuições dos retornos passados (Tabela 5). As variações nas médias de Lag1 e Lag2 entre os grupos são mínimas, o que sugere

que o histórico recente possui baixo poder preditivo isolado sobre a tendência futura.

Tabela 5: Estatísticas descritivas segmentadas pela direção do mercado (Down vs. Up).

Variável	**Geral** (N = 1089) *	**Down** (N = 484) *	**Up** (N = 605) *
Lag 1	0.151 (2.357)	0.282 (2.315)	0.045 (2.387)
Lag 2	0.151 (2.357)	-0.040 (2.292)	0.304 (2.399)
Lag 3	0.147 (2.361)	0.208 (2.282)	0.099 (2.423)
Lag 4	0.146 (2.360)	0.200 (2.443)	0.102 (2.293)
Lag 5	0.140 (2.361)	0.188 (2.372)	0.102 (2.354)
Volume	1.575 (1.687)	1.609 (1.699)	1.547 (1.678)
Retorno Atual	0.150 (2.357)	-1.747 (1.760)	1.667 (1.531)

* Mean (SD)

Nota: Valores expressos como Média (Desvio-Padrão).

O Volume médio é idêntico em ambos os cenários, indicando que a intensidade das negociações, por si só, não sinaliza a direção do mercado. O retorno de hoje (Today) apenas confirma a construção da variável dependente (negativo em quedas e positivo em altas), enquanto as defasagens mais distantes (Lags 3 a 5) reforçam a ausência de padrões e a predominância do ruído na série (Figura 3).

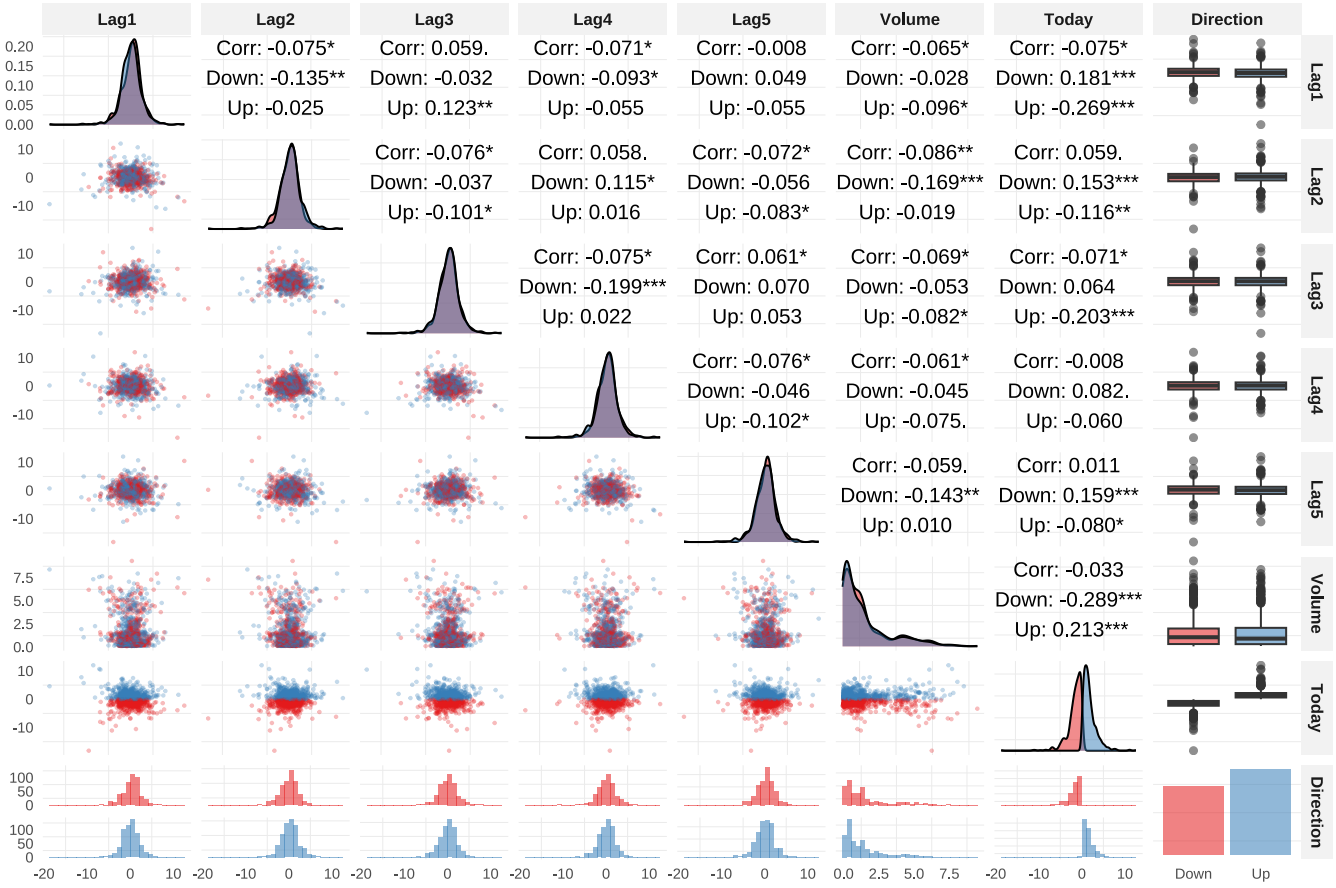

Figura 3: Matriz de dispersão e densidade das variáveis financeiras estratificadas por direção do mercado.

Item b)

Ao nível de significância de 5%, a associação entre a resposta Direction e Lag1 não é estatisticamente significante, sugerindo, assim, que Lag1 não é um bom preditor para a direção (Tabela 6), sugerindo que qualquer padrão observado pode ser fruto do acaso. A sua taxa de acerto (55,37%) é ligeiramente inferior à chance de acerto caso apenas apostássemos, sem critério algum, que o mercado subiria todas as semanas (55,56%).

Tabela 6: Estimativas do modelo de regressão logística para a direção do mercado em função do retorno da semana anterior (Lag1).

$$\log\left[\frac{P(\widehat{\text{Direction}} = \text{Up})}{1 - P(\widehat{\text{Direction}} = \text{Up})}\right] = 0.23 - 0.04(\text{Lag1}) \tag{2}$$

Characteristic	**log(OR)**	**95% CI**	**p-value**
(Intercept)	0.2302	0.1105, 0.3506	<0.001
Retorno da Semana Anterior (Lag1)	-0.0431	-0.0950, 0.0079	0.10

Abbreviations: CI = Confidence Interval, OR = Odds Ratio
Null deviance = 1,496; Null df = 1,088; Log-Likelihood = -747; AIC = 1,497; BIC = 1,507; Deviance = 1,493; Residual df = 1,087; No. Obs. = 1,089

Como observa-se na Tabela 7, o modelo sugere na subida do mercado em quase 98% das vezes. Embora ele acerte quase todas as semanas de alta (Sensibilidade), ele falha em identificar as quedas, errando mais de 98% dos períodos em que o mercado de fato caiu (Especificidade). Isso reforça a natureza imprevisível dos retornos financeiros, onde o histórico recente não consegue superar a estratégia de simplesmente seguir a tendência majoritária de alta.

Tabela 7: Matriz de Confusão e Métricas de Desempenho do Modelo Logístico.

Predito / Real	Down	Up
Down	8	10
Up	476	595

Estatísticas: Acurácia: 0.5537 | Sensibilidade: 0.9835 | Especificidade: 0.0165 | NIR: 0.5556

Item c) Ao nível de significância de 5%, a variável Lag1 continua não estatisticamente significantiva ($p = 0,14$), enquanto a variável Lag2 mostrou ser um preditor significantivo ($p = 0,023$) (Tabela 8).

Tabela 8: Estimativas do modelo de regressão logística com dois preditores (Lag1 e Lag2).

$$\log\left[\frac{P(\widehat{\text{Direction}} = \text{Up})}{1 - P(\widehat{\text{Direction}} = \text{Up})}\right] = 0.22 - 0.04(\text{Lag1}) + 0.06(\text{Lag2}) \tag{3}$$

Characteristic	**log(OR)**	**95% CI**	**p-value**
(Intercept)	0.2212	0.1010, 0.3420	<0.001
Retorno da Semana Anterior (Lag1)	-0.0387	-0.0906, 0.0124	0.14
Retorno de Duas Semanas Atrás (Lag2)	0.0602	0.0086, 0.1129	0.023

Abbreviations: CI = Confidence Interval, OR = Odds Ratio
Null deviance = 1,496; Null df = 1,088; Log-Likelihood = -744; AIC = 1,494; BIC = 1,509; Deviance = 1,488; Residual df = 1,086; No. Obs. = 1,089

Como observa-se na Tabela 9, a acurácia do modelo ($55,56\%$) é idêntica à taxa de não informação (NIR), evidenciando que o ajuste não oferece ganho preditivo. Embora a inclusão de Lag2 tenha elevado ligeiramente a capacidade de identificar quedas (Especificidade de $7,85\%$) em relação ao modelo simples, o viés em favor da alta persiste. O modelo captura a maioria dos movimentos de subida (Sensibilidade de $93,72\%$), mas falha em detectar mais de 92% das semanas de baixa.

Tabela 9: Matriz de Confusão e métricas para o modelo com Lag1 e Lag2.

Predito / Real	Down	Up
Down	38	38
Up	446	567

Estatísticas: Acurácia: 0.5556 | Sensibilidade: 0.9372 | Especificidade: 0.0785 | NIR: 0.5556

item d)

Ao nível de significância de 5%, o retorno de duas semanas atrás (Lag2) apresentou uma relação estatisticamente relevante (significativa) com a direção do mercado durante o período de treinamento ($p = 0,043$) (Tabela 10). Para os anos de teste (2009-2010), o modelo teve uma acurácia de $62,5\%$, superando a taxa básica de acerto esperada de $58,6\%$ (NIR).

Tabela 10: Estimativas do modelo treinado com dados de 1990 a 2008.

$$\log\left[\frac{P(\widehat{\text{Direction}} = \text{Up})}{1 - P(\widehat{\text{Direction}} = \text{Up})}\right] = 0.2 + 0.06(\text{Lag2}) \tag{4}$$

Characteristic	**log(OR)**	**95% CI**	**p-value**
(Intercept)	0.20	0.08, 0.33	0.002
Retorno de Duas Semanas Atrás (Lag2)	0.06	0.00, 0.12	0.043

Abbreviations: CI = Confidence Interval, OR = Odds Ratio
Null deviance = 1,355; Null df = 984; Log-Likelihood = -675; AIC = 1,355; BIC = 1,364; Deviance = 1,351; Residual df = 983; No. Obs. = 985

O modelo permanece é muito eficaz em capturar os movimentos de alta (acertando quase 92% das subidas), mas continua falhando em identificar as quedas (acertando apenas 21% das baixas) (Tabela 11).

Tabela 11: Matriz de Confusão para o período de teste (2009-2010).

Predito / Real	Down	Up
Down	9	5
Up	34	56

Estatísticas: Acurácia: 0.625 | Sensibilidade: 0.918 | Especificidade: 0.2093 | NIR: 0.5865

A curva ROC (Figura 4), com um valor de $0,454$, indica que a capacidade real do modelo em distinguir entre um período de alta e um de baixa é, na prática, inferior ao acaso, reforçando a dificuldade de prever o mercado financeiro com base apenas em resultados passados.

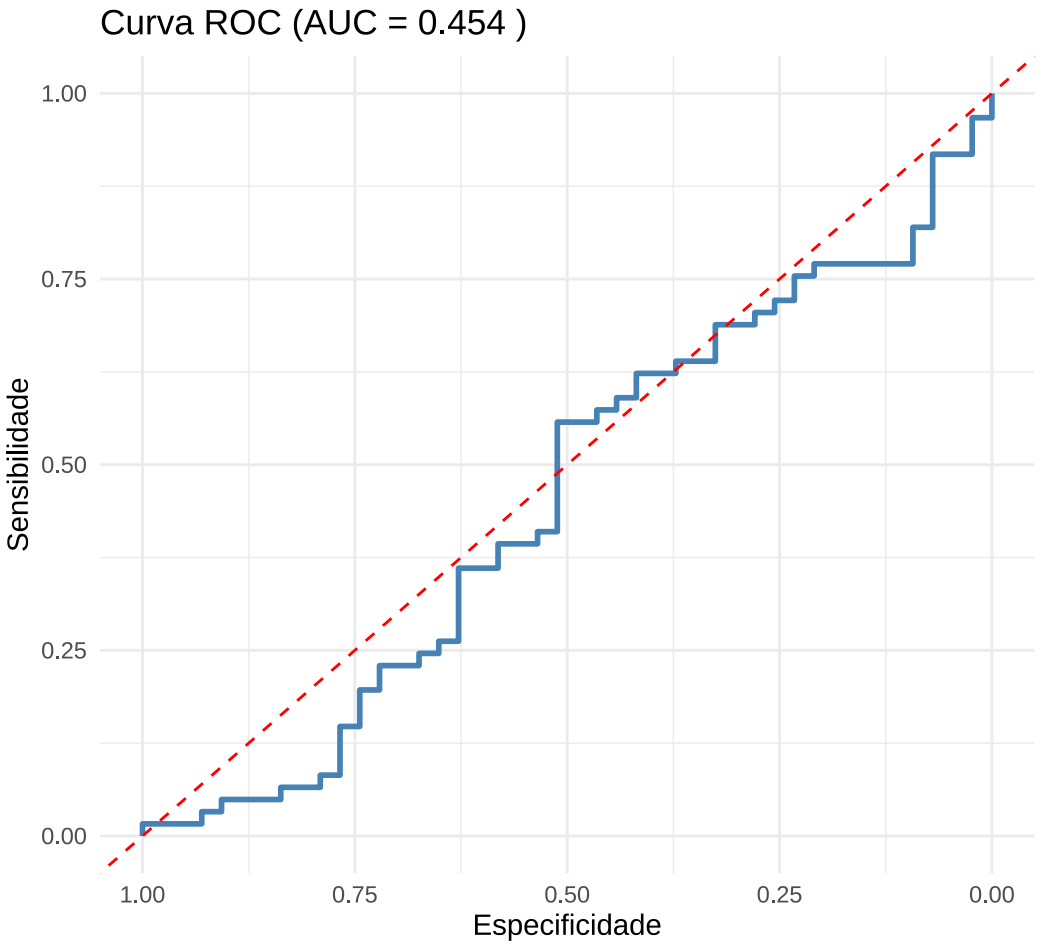

Figura 4: Curva ROC para o modelo de teste, indicando a relação entre sensibilidade e especificidade.

Idem e)

O modelo KNN com $K = 1$ apresentou um desempenho insatisfatório, com uma acurácia de $50,96\%$, valor consideravelmente inferior à taxa de não informação ($58,65\%$) (Tabela 12). Isso indica que o ajuste tem um resultado pior do que uma estratégia simples de apostar na subida do mercado em todas as semanas. Diferente da regressão logística, o KNN com $K = 1$ distribui as predições de forma mais equilibrada entre as classes, com sensibilidade de $52,46\%$ e especificidade de $48,84\%$.

Tabela 12: Matriz de Confusão para o modelo KNN (K=1) no período de teste (2009-2010).

Predito / Real	Down	Up
Down	21	29
Up	22	32

Estatísticas: Acurácia: 0.5096 | Sensibilidade: 0.5246 | Especificidade: 0.4884 | NIR: 0.5865

Item f)

O método que fornece os melhores resultados é a Regressão Logística. Enquanto o modelo logístico alcançou uma acurácia de 62,5%, superando a taxa de acerto base (NIR de 58,65%), o modelo KNN

apresentou um desempenho de apenas 50,96%, sendo inferior a uma estratégia de aposta cega na tendência de alta.

EXERCÍCIO 3

Item a)

O código da Tabela 13 cria a variável `mpg1`, em que o valor 0 indica os veículos que fazem menos milhas medianas (mediana igual a 22,75) por galão, enquanto que o valor 1 indica os veículos que fazem mais milhas medianas (mediana igual a 22,75) por galão, e renomeia a variável *origin* conforme descrito no `R` ao rodar `?Auto`.

Tabela 13: Amostra das primeiras linhas do dataset Auto após a criação da variável mpg1 e recodificação de origem.

MPG (Original)	**mpg1 (Alvo)**	displacement	horsepower	weight	Origem
18	Abaixo Mediana	307	130	3504	Americano
15	Abaixo Mediana	350	165	3693	Americano
18	Abaixo Mediana	318	150	3436	Americano
16	Abaixo Mediana	304	150	3433	Americano
17	Abaixo Mediana	302	140	3449	Americano

Item b)

A eficiência de combustível apresenta uma forte ligação com as características físicas e a origem dos veículos (Figura 5). Carros com menor autonomia possuem tipicamente mais cilindros e são predominantemente de origem americana, enquanto os modelos mais econômicos mostram uma distribuição mais equilibrada entre fabricantes americanos, europeus e japoneses. Nota-se um padrão onde veículos menos eficientes possuem motores maiores, maior potência e peso elevado. Além disso, esses modelos tendem a ser mais antigos e, apesar do peso, conseguem acelerar em menos tempo (são mais rápidos) do que os carros mais econômicos.

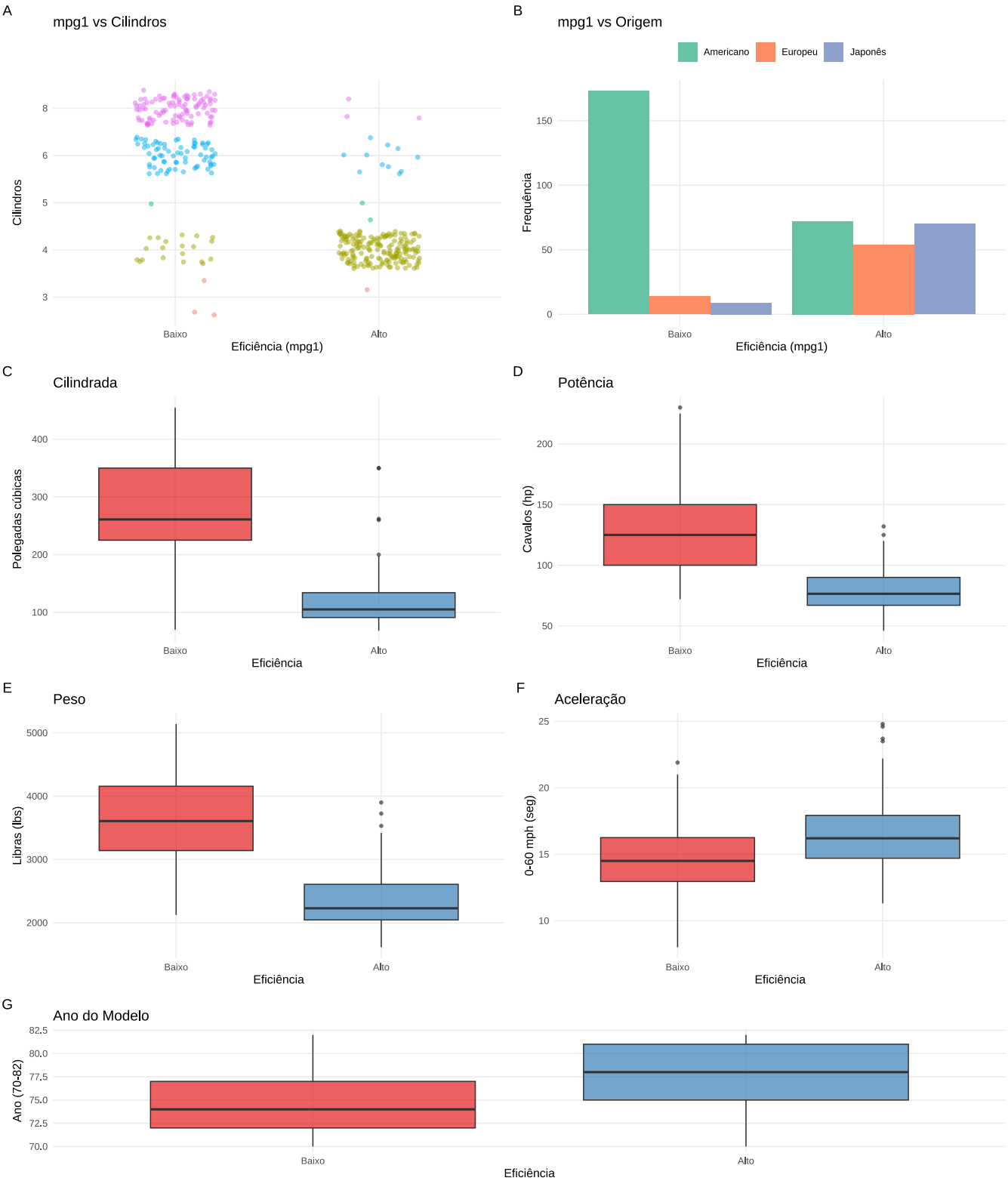

Figura 5: Distribuição das variáveis do dataset Auto segmentadas pela classe de eficiência (mpg1).

A interdependência entre as variáveis também é evidente (Figura 6), especialmente a forte correlação entre o tamanho do motor, a potência e o peso, que costumam variar de forma conjunta. Os dados foram divididos de tal forma que cerca de 80% das observações fossem destinadas ao conjunto de treino e cerca de 20% ao conjunto de teste.

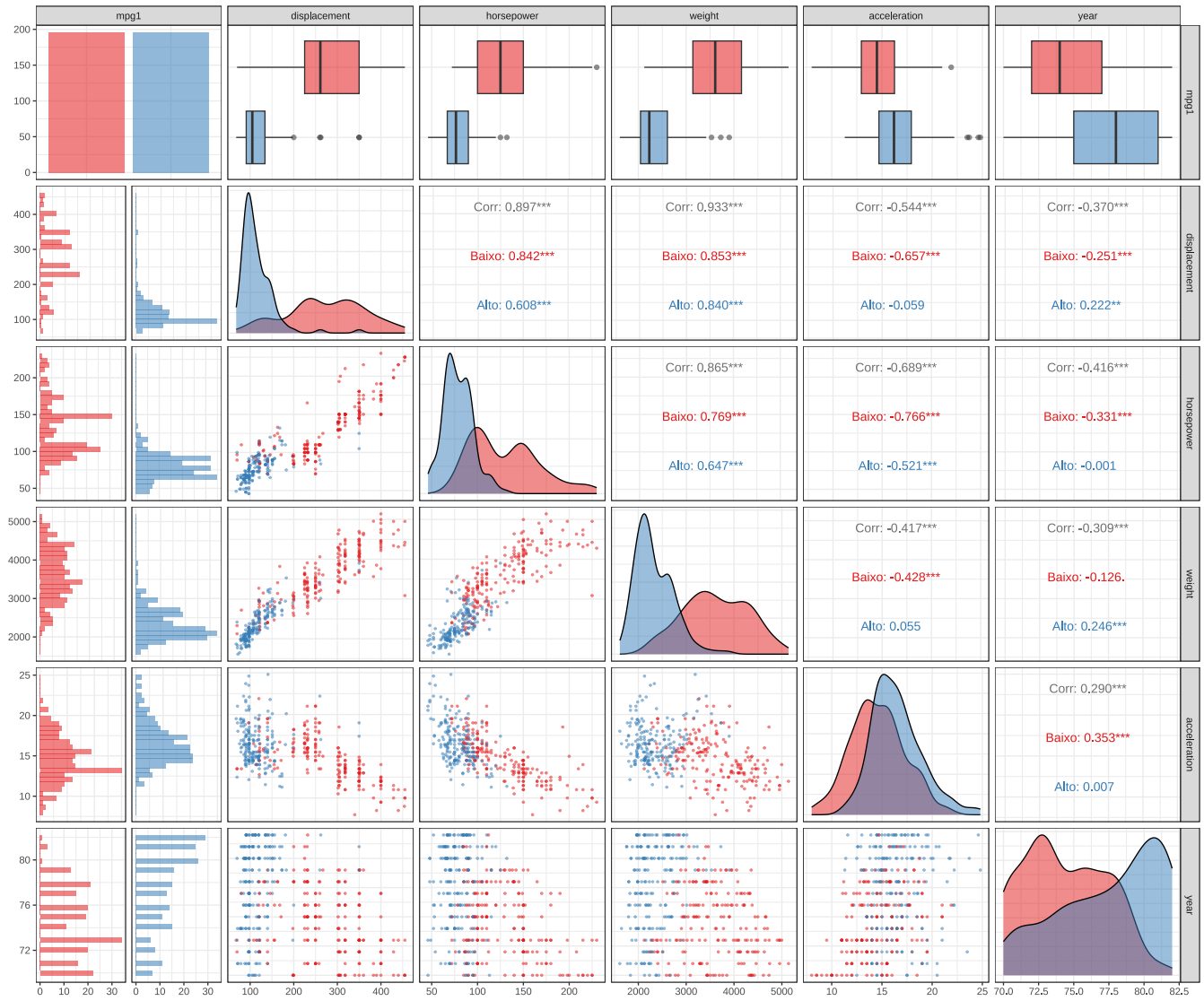

Figura 6: Matriz de correlação e dispersão estratificada por eficiência (mpg1).

Item c)

A Análise Discriminante Linear (LDA) demonstrou alta eficácia na separação dos veículos entre as classes de eficiência. De acordo com a Tabela 14, o modelo que utiliza todas as características técnicas apresentou uma taxa de erro de 10,13% no conjunto de teste. Já o modelo simplificado, focado apenas no número de cilindros e no ano, obteve um desempenho ainda superior, com uma taxa de erro de apenas 8,86% e uma acurácia de 91,14%.

Tabela 14: Desempenho comparativo dos modelos LDA no conjunto de teste.

Modelo	Acurácia	Taxa de Erro	Sensibilidade	Especificidade
Completo (Todos)	89.87%	10.13%	90.00%	89.66%
Simples (Cyl + Year)	91.14%	8.86%	92.00%	89.66%

A Análise Discriminante Linear (LDA) demonstrou eficácia na separação dos veículos. O modelo busca maximizar a diferença entre os grupos em relação à variação interna de cada um:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Os centros de cada grupo e os pesos de cada característica são apresentados na Tabela 15. Com base nesses valores, o cálculo para classificar cada veículo segue a equação:

$$\begin{aligned}\text{Escore} = {} & (3,7248 \cdot \text{Cyl4}) + (2,2689 \cdot \text{Cyl5}) + (1,2951 \cdot \text{Cyl6}) + (1,9370 \cdot \text{Cyl8}) \\ & + (-0,0016 \cdot \text{Displacement}) + (-0,0001 \cdot \text{Horsepower}) + (-0,0008 \cdot \text{Weight}) \\ & + (-0,0289 \cdot \text{Acceleration}) + (0,1193 \cdot \text{Year})\end{aligned}$$

O ponto de decisão é definido pelo valor de corte (4,31), calculado pela média dos centros dos dois grupos. Assim, veículos com resultado abaixo de 4,31 são classificados como de baixa eficiência (mpg1 = 0), enquanto valores acima indicam alta eficiência (mpg1 = 1). A análise dos pesos confirma que ter menos cilindros (Cyl4) e ser um modelo mais novo (Year) são os fatores que mais elevam a pontuação para a categoria de alta eficiência. Por outro lado, o aumento do tamanho do motor, da potência e do peso atua diminuindo o escore, o que classifica o veículo como menos econômico.

Tabela 15: Diagnóstico do Modelo LDA: Coeficientes Discriminantes e Médias dos Grupos.

Variável	**Coeficiente (LD1)**	**Média (Baixo)**	**Média (Alto)**
cylinders4	**3.7248**	0.1018	0.9178
cylinders5	**2.2689**	0.0060	0.0068
cylinders6	**1.2951**	0.3772	0.0616
cylinders8	**1.9370**	0.4970	0.0137
displacement	**-0.0016**	271.8383	114.1884
horsepower	**-0.0001**	129.2156	78.3151
weight	**-0.0008**	3,588.9641	2,314.6233
acceleration	**-0.0289**	14.5461	16.4925
year	**0.1193**	74.2934	77.4110

Nota: Os coeficientes (LD1) indicam a direção e força da separação entre os grupos.

O aumento da potência, do peso e do tamanho do motor são os fatores que mais penalizam a eficiência, deslocando o veículo para a categoria de baixo desempenho. Embora exista uma mínima sobreposição entre as classes, a distinção entre os grupos é clara, como ilustrado na Figura 7.

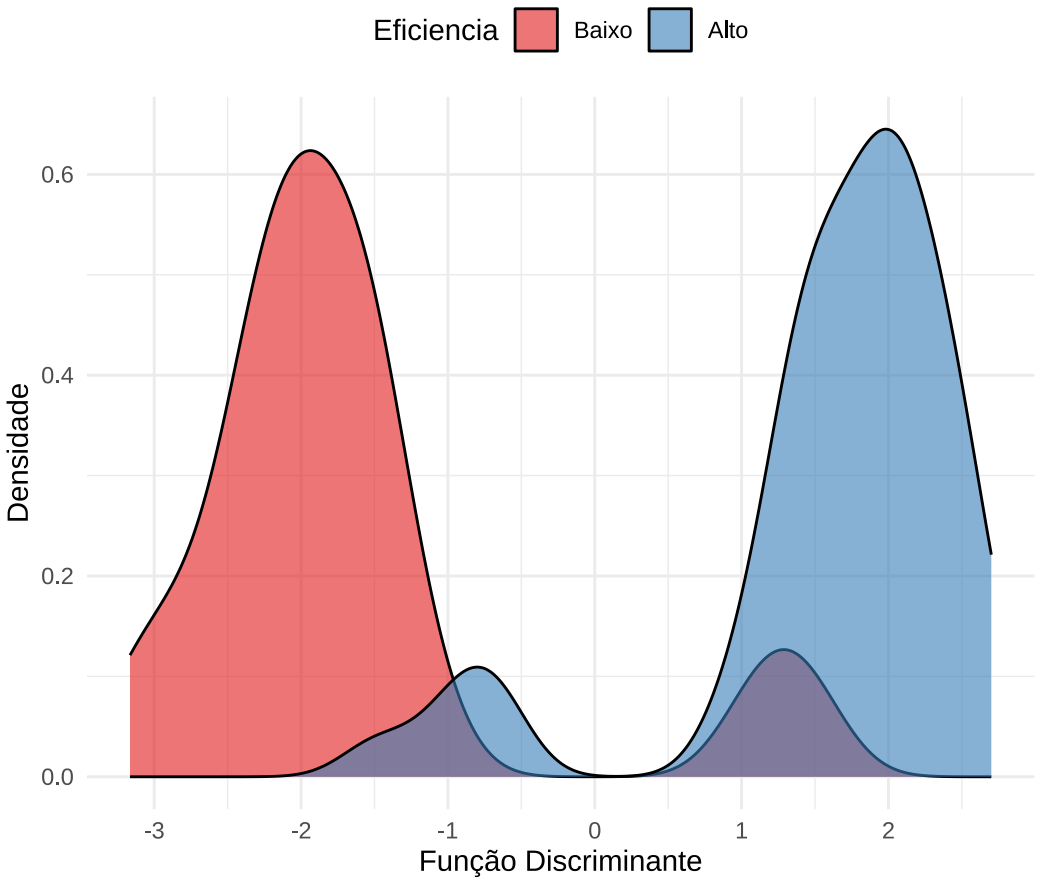

Figura 7: Separação das classes de eficiência através dos escores discriminantes (LD1).

> **!** Importante
>
> Vale ressaltar que os resultados obtidos são específicos aos conjuntos de treinamento e teste usados. Pequenas diferenças podem surgir a depender da separação dos conjuntos realizada.

Item d)

Os melhores resultados ocorreram com $K = 9$ e $K = 10$, com acerto de $93,67\%$ e erro de $6,33\%$ (Tabela 16). Nessa configuração, o modelo identificou corretamente 96% dos veículos econômicos. A opção por $K = 9$ é preferível por ser mais estável. O desempenho confirma que veículos com características parecidas têm consumo similar, garantindo previsões corretas em mais de 93% dos casos.

Tabela 16: Desempenho do modelo KNN para diferentes valores de K (1 a 10).

K	Acurácia	Taxa de Erro	Sensibilidade	Especificidade
1	92.41%	7.59%	92.00%	93.10%
2	92.41%	7.59%	92.00%	93.10%
3	92.41%	7.59%	92.00%	93.10%
4	92.41%	7.59%	92.00%	93.10%
5	92.41%	7.59%	92.00%	93.10%
6	92.41%	7.59%	92.00%	93.10%
7	92.41%	7.59%	92.00%	93.10%
8	92.41%	7.59%	92.00%	93.10%
9	**93.67%**	**6.33%**	**96.00%**	**89.66%**
10	**93.67%**	**6.33%**	**96.00%**	**89.66%**

A Tabela 17 demonstra que o modelo com $K = 9$ alcançou um desempenho equilibrado, com acurácia de $93,67\%$, valor superior à taxa de não informação ($63,29\%$). O modelo é eficaz em identificar veículos econômicos, com acerto de 96% (sensibilidade), mantendo também uma alta precisão para os modelos menos eficientes, com $89,66\%$ de acerto (especificidade).

Tabela 17: Matriz de Confusão para o modelo KNN com K=9.

Predito / Real	0	1
0	26	2
1	3	48

Estatísticas: Acurácia: 0.9367 | Sensibilidade: 0.96 | Especificidade: 0.8966 | NIR: 0.6329

> **!** Importante
>
> Vale ressaltar que os resultados obtidos são específicos aos conjuntos de treinamento e teste usados. Pequenas diferenças podem surgir a depender da separação dos conjuntos realizada.

Item e)

O KNN com $K = 9$ é o melhor classificador para este cenário, apresentando a menor taxa de erro (6,33%) em comparação à LDA (8,86%). A Análise Discriminante de Fisher fundamenta-se em suposições rígidas, como a necessidade de os dados seguirem uma distribuição normal e possuírem variações iguais entre os grupos (ver MORETTIN e SINGER; 2023; FISHER, 1936 e EFRON, 1975)[1]. Quando essas condições não são plenamente atendidas, o desempenho da LDA é prejudicado. Já o KNN é um método livre de distribuição, o que significa que ele não faz suposições sobre o formato dos dados. Essa flexibilidade permite ao KNN capturar padrões complexos e curvas de separação que a rigidez linear da LDA não consegue descrever. Portanto, para fins de precisão na previsão, o KNN é a escolha mais robusta e eficiente (ver Cover e Hart, 1967).

[1]EFRON (1975) apresenta uma boa discussão sobre LDA vs Regressão Logistica, recomendo a leitura.

Referências

MORETTIN, P. A. e SINGER, J. M. **Estatística e Ciências de Dados**. Rio de Janeiro:LTC, 2023. (Já existe segunda edição do recente do livro, caso haja interressado pode comprar pelo link

NHANCOLOLO, A. M.; SCALON, J. D.; ALENCAR, A. P. Capítulo 1. In: _____. Introdução à Estatística Espacial: Teoria e Aplicações com R. São Paulo: Instituto de Matemática, Estatística e Ciência da Computação, Universidade de São Paulo, 2026. Disponível em: https://introducao-a-estatistica-espacial.netlify.app/.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21-27, jan. 1967.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v. 7, n. 2, p. 179-188, set. 1936.

EFRON, Bradley. The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. **Journal of the American Statistical Association**, [S. l.], v. 70, n. 352, p. 892-898, Dec. 1975.

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, [S. l.], v. 58, n. 1, p. 267-288, 1996.

HOERL, Arthur E.; KENNARD, Robert W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics**, [S. l.], v. 12, n. 1, p. 55-67, fev. 1970.

ZOU, Hui; HASTIE, Trevor. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, [S. l.], v. 67, n. 2, p. 301-320, 2005.

Apêndice

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load(xtable, gt, caret, tidyverse, kableExtra, ISLR, astsa, stargazer, knitr,
               dplyr,  e1071, class, ggpubr, GGally, MASS, broom, pROC, flextable,
               finalfit, texreg, equatiomatic, dlookr, gridExtra, ggplot2, dplyr)
```

EXERCÍCIO 1

Item a)

```
set.seed(2023)
X = rnorm(100)
erros = rnorm(100)
dataf <- data.frame(
   X, erros
 )
dataf |>
  head()|>
  gt()
```

Item b)

```
beta_0=30;beta_1= 7.3;beta_2= -3.6;beta_3=0.65
Y = beta_0 + beta_1 * X + beta_2 * X^2 + beta_3 * X^3 + erros
Y[1:10]
```

Item c)

```
pacman::p_load(gtsummary)

OLS %>%
  tbl_regression(
    intercept = TRUE,
    estimate_fun = ~ style_number(.x, digits = 3)
  ) %>%
  add_global_p() %>%
  bold_p(t = 0.05) %>%
  add_glance_source_note(
    label = list(r.squared ~ "$R^2$", AIC ~ "AIC")
  )

p_load(ggfortify, performance, see)
check_model(OLS)
#ou
#autoplot(OLS, which = 1:4, ncol = 2, label.size = 3) + theme_minimal()
```

Item d)

```r
gerar_resumo_regularizacao <- function(dados, target = "Y", n_folds = 10) {
  pacman::p_load(tidyverse, caret, gt, glmnet)

  covariáveis <- setdiff(names(dados), target)
  formula_modelo <- as.formula(paste(target, "~ ."))

  lambda_grid <- 10^seq(-3, 3, length = 1000)
  ctrl <- trainControl(method = "cv", number = n_folds)

  ajustar <- function(a) {
    train(formula_modelo, data = dados, method = "glmnet",
          trControl = ctrl, tuneGrid = expand.grid(alpha = a, lambda = lambda_grid))
  }

  m_ridge <- ajustar(0)
  m_lasso <- ajustar(1)

  extrair_tudo <- function(model) {
    co <- as.matrix(coef(model$finalModel, s = model$bestTune$lambda))
    p <- predict(model, dados)
    metricas <- c(
      alpha  = model$bestTune$alpha,
      lambda = model$bestTune$lambda,
      RMSE   = RMSE(p, dados[[target]]),
      R2     = R2(p, dados[[target]])
    )
    return(c(co[,1], metricas))
  }

  res_ridge <- extrair_tudo(m_ridge)
  res_lasso <- extrair_tudo(m_lasso)

  df_final <- data.frame(
    Parametro = names(res_ridge),
    Ridge = res_ridge,
    Lasso = res_lasso,
    row.names = NULL
  ) %>%
    mutate(Parametro = case_when(
      Parametro == "(Intercept)" ~ "Intercepto ($\\beta_0$)",
      Parametro == "alpha"       ~ "Alpha ($\\alpha$)",
      Parametro == "lambda"      ~ "Lambda ($\\lambda$)",
      Parametro == "R2"          ~ "$R^2$",
      TRUE ~ str_replace_all(Parametro, "([a-zA-Z])([0-9])", "$\\1^\\2$")
    ))

  n_coefs <- length(covariáveis) + 1
```

```
  df_final %>%
    gt() %>%
    fmt_markdown(columns = Parametro) %>%
    cols_label(
      Parametro = "Parâmetro",
      Ridge = md("Ridge (*L*~2~)"),
      Lasso = md("Lasso (*L*~1~)")
    ) %>%
    fmt_number(
      columns = c(Ridge, Lasso),
      rows = c(1:n_coefs, (nrow(df_final)-1):nrow(df_final)),
      decimals = 4
    ) %>%
    fmt_scientific(
      columns = c(Ridge, Lasso),
      rows = (n_coefs + 2),
      decimals = 2
    ) %>%
    tab_style(
      style = cell_text(weight = "bold"),
      locations = cells_body(rows = (nrow(df_final)-1):nrow(df_final))
    ) %>%
    tab_options(
      table.width = pct(90),
      column_labels.background.color = "#f9f9f9",
      data_row.padding = px(4)
    )
}

dados_reg <- tibble(Y = Y, X = X, X2 = X^2, X3 = X^3)
gerar_resumo_regularizacao(dados_reg)
```

EXERCÍCIO 2

Item a)

```
Weekly <- data(Weekly)

  Tab_3 <- ISLR::Weekly |>
    dlookr::describe() |>
    slice(2:8) %>%
  dplyr::select(described_variables, mean, sd, IQR, skewness, kurtosis, p25, p50, p75)

Tab_3 |>
  gt() |>
  cols_label(
    described_variables = "Variável",
    mean = md("Média ($\\bar{x}$)"),
    sd = md("Desvios-padrão ($\\sigma$)"),
```

```
    IQR = "IQR",
    skewness = "Assimetria",
    kurtosis = "Curtose",
    p25 = md("$Q_1$"),
    p50 = md("Mediana ($Q_2$)"),
    p75 = md("$Q_3$")
  ) %>%
  fmt_number(
    columns = -described_variables,
    decimals = 2
  ) %>%
  tab_options(
    table.width = pct(100),
    column_labels.background.color = "#f9f9f9",
    table.font.size = px(14),
    data_row.padding = px(4)
  ) %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels()
  )
pacman::p_load(patchwork)
data(Weekly)

df_plot <- Weekly %>%
  dplyr::mutate(Year_short = factor(str_sub(Year, 3, 4)))

vars_box <- c("Volume", "Today", paste0("Lag", 1:5))

lista_boxplots <- vars_box %>%
  map(~{
    ggplot(df_plot, aes(x = Year_short, y = .data[[.x]])) +
      geom_boxplot(fill = "gray95", outlier.size = 0.8, color = "black") +
      labs(title = paste(.x, "vs Ano"), x = "Ano", y = .x) +
      theme_minimal(base_size = 9) +
      theme(panel.grid.minor = element_blank())
  })

plot_bar <- ggplot(df_plot, aes(x = Year_short, fill = Direction)) +
  geom_bar(position = "dodge", alpha = 0.8) +
  scale_fill_brewer(palette = "Set1") +
  labs(title = "Distribuição de Direction por Ano", x = "Ano", y = "Frequência") +
  theme_minimal(base_size = 9) +
  theme(legend.position = "top", panel.grid.minor = element_blank())

wrap_plots(lista_boxplots, ncol = 2) + plot_bar +
  plot_layout(ncol = 2) +
  plot_annotation(tag_levels = 'A')
data(Weekly)
```

```r
dependent <- "Direction"
explanatory <- c("Lag1", "Lag2", "Lag3", "Lag4", "Lag5", "Volume", "Today")

Weekly %>%
  dplyr::select(all_of(c(dependent, explanatory))) %>%
  tbl_summary(
    by = Direction,
    statistic = list(all_continuous() ~ "{mean} ({sd})"),
    digits = list(all_continuous() ~ 3),
    label = list(
      Volume ~ "Volume",
      Today ~ "Retorno Atual",
      Lag1 ~ "Lag 1", Lag2 ~ "Lag 2", Lag3 ~ "Lag 3",
      Lag4 ~ "Lag 4", Lag5 ~ "Lag 5"
    )
  ) %>%
  modify_header(
    label = "**Variável**",
    stat_1 = "**Down** (N = {n})",
    stat_2 = "**Up** (N = {n})"
  ) %>%
  add_overall(last = FALSE, col_label = "**Geral** (N = {N})") %>%
  as_gt() %>%
  tab_options(
    table.width = pct(95),
    column_labels.background.color = "#f9f9f9",
    table.font.size = px(14)
  ) %>%
  opt_footnote_marks(marks = "standard") %>%
  tab_source_note(source_note = "Nota: Valores expressos como Média (Desvio-Padrão).")
pacman::p_load(GGally, ggplot2, ISLR)

df_biv <- Weekly %>% dplyr::select(-Year)

painel_prof <- ggpairs(
  df_biv,
  mapping = aes(color = Direction, alpha = 0.4),
  upper = list(continuous = wrap("cor", size = 3.5, color = "black")),
  lower = list(continuous = wrap("points", alpha = 0.3, size = 0.4)),
  diag = list(continuous = wrap("densityDiag", alpha = 0.5))
) +
  scale_fill_brewer(palette = "Set1") +
  scale_color_brewer(palette = "Set1") +
  theme_minimal(base_size = 10) +
  theme(axis.text = element_text(size = 7), panel.grid.major = element_blank(),
    strip.background = element_rect(fill = "gray95", color = NA),
    strip.text = element_text(face = "bold")) ; painel_prof
```

Item b)

```
conf_table <- as.data.frame(conf_mat$table) %>%
  rename(Predito = Prediction, Real = Reference, Freq = Freq) %>%
  pivot_wider(names_from = Real, values_from = Freq)

conf_table %>%
  gt() %>%
  tab_header(title = "") %>%
  cols_label(Predito = md("**Predito / Real**")) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "Down", rows = Predito == "Down")
  ) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "Up", rows = Predito == "Up")
  ) %>%
  tab_source_note(
    source_note = md(paste0(
      "**Estatísticas:** ",
      "Acurácia: ", round(conf_mat$overall["Accuracy"], 4), " | ",
      "Sensibilidade: ", round(conf_mat$byClass["Sensitivity"], 4), " | ",
      "Especificidade: ", round(conf_mat$byClass["Specificity"], 4), " | ",
      "NIR: ", round(conf_mat$overall["AccuracyNull"], 4)
    ))
  ) %>%
  tab_options(
    table.width = pct(70),
    table.font.size = px(14),
    column_labels.background.color = "#f9f9f9"
  )

pacman::p_load(ISLR, gtsummary, gt, equatiomatic, caret)
data(Weekly)

mod_log1 <- glm(Direction ~ Lag1, data = Weekly, family = binomial(link = "logit"))

equatiomatic::extract_eq(mod_log1, use_coefs = TRUE)

mod_log1 %>%
  tbl_regression(
    intercept = TRUE,
    label = list(Lag1 ~ "Retorno da Semana Anterior (Lag1)"),
    estimate_fun = ~ style_number(.x, digits = 4)
  ) %>%
  add_glance_source_note(
    label = list(logLik ~ "Log-Likelihood", AIC ~ "AIC")
  ) %>%
  as_gt() %>%
```

```
  tab_options(
    table.width = pct(90),
    column_labels.background.color = "#f9f9f9",
    table.font.size = px(14)
  )

probs <- predict(mod_log1, type = "response")
preds <- factor(ifelse(probs >= 0.5, "Up", "Down"), levels = c("Down", "Up"))
conf_mat <- confusionMatrix(preds, Weekly$Direction, positive = "Up")
```

Item c)

```
pacman::p_load(ISLR, gtsummary, gt, equatiomatic, caret, tidyverse)
data(Weekly)

mod_log2 <- glm(Direction ~ Lag1 + Lag2, data = Weekly, family = binomial(link = "logit"))

equatiomatic::extract_eq(mod_log2, use_coefs = TRUE)

mod_log2 %>%
  tbl_regression(
    intercept = TRUE,
    label = list(
      Lag1 ~ "Retorno da Semana Anterior (Lag1)",
      Lag2 ~ "Retorno de Duas Semanas Atrás (Lag2)"
    ),
    estimate_fun = ~ style_number(.x, digits = 4)
  ) %>%
  add_glance_source_note(
    label = list(logLik ~ "Log-Likelihood", AIC ~ "AIC")
  ) %>%
  as_gt() %>%
  tab_options(
    table.width = pct(90),
    column_labels.background.color = "#f9f9f9",
    table.font.size = px(14)
  )

probs2 <- predict(mod_log2, type = "response")
preds2 <- factor(ifelse(probs2 >= 0.5, "Up", "Down"), levels = c("Down", "Up"))
conf_mat2 <- confusionMatrix(preds2, Weekly$Direction, positive = "Up")
conf_table2 <- as.data.frame(conf_mat2$table) %>%
  rename(Predito = Prediction, Real = Reference, Freq = Freq) %>%
  pivot_wider(names_from = Real, values_from = Freq)

conf_table2 %>%
  gt() %>%
  tab_header(title = "") %>%
  cols_label(Predito = md("**Predito / Real**")) %>%
```

```
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "Down", rows = Predito == "Down")
  ) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "Up", rows = Predito == "Up")
  ) %>%
  tab_source_note(
    source_note = md(paste0(
      "**Estatísticas:** ",
      "Acurácia: ", round(conf_mat2$overall["Accuracy"], 4), " | ",
      "Sensibilidade: ", round(conf_mat2$byClass["Sensitivity"], 4), " | ",
      "Especificidade: ", round(conf_mat2$byClass["Specificity"], 4), " | ",
      "NIR: ", round(conf_mat2$overall["AccuracyNull"], 4)
    ))
  ) %>%
  tab_options(
    table.width = pct(70),
    table.font.size = px(14),
    column_labels.background.color = "#f9f9f9"
  )
```

Item d)

```
roc_obj <- roc(teste$Direction, probs_teste, levels = c("Down", "Up"), quiet = TRUE)

ggroc(roc_obj, color = "steelblue", size = 1) +
  geom_abline(slope = 1, intercept = 1, linetype = "dashed", color = "red") +
  labs(title = paste("Curva ROC (AUC =", round(auc(roc_obj), 3), ")"),
       x = "Especificidade", y = "Sensibilidade") +
  theme_minimal()
pacman::p_load(ISLR, gtsummary, gt, equatiomatic, caret, pROC, tidyverse)
data(Weekly)

treino <- Weekly %>% filter(Year <= 2008)
teste  <- Weekly %>% filter(Year > 2008)

mod_treino <- glm(Direction ~ Lag2, data = treino, family = binomial)

equatiomatic::extract_eq(mod_treino, use_coefs = TRUE)

mod_treino %>%
  tbl_regression(
    intercept = TRUE,
    label = list(Lag2 ~ "Retorno de Duas Semanas Atrás (Lag2)")
  ) %>%
  add_glance_source_note(
    label = list(logLik ~ "Log-Likelihood", AIC ~ "AIC")
```

```
  ) %>%
  as_gt() %>%
  tab_options(
    table.width = pct(90),
    column_labels.background.color = "#f9f9f9",
    table.font.size = px(14)
  )
probs_teste <- predict(mod_treino, newdata = teste, type = "response")
preds_teste <- factor(ifelse(probs_teste >= 0.5, "Up", "Down"), levels = c("Down", "Up"))
conf_teste  <- confusionMatrix(preds_teste, teste$Direction, positive = "Up")

df_conf_teste <- as.data.frame(conf_teste$table) %>%
  rename(Predito = Prediction, Real = Reference, Freq = Freq) %>%
  pivot_wider(names_from = Real, values_from = Freq)

df_conf_teste %>%
  gt() %>%
  tab_header(title = "") %>%
  cols_label(Predito = md("**Predito / Real**")) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "Down", rows = Predito == "Down")
  ) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "Up", rows = Predito == "Up")
  ) %>%
  tab_source_note(
    source_note = md(paste0(
      "**Estatísticas:** ",
      "Acurácia: ", round(conf_teste$overall["Accuracy"], 4), " | ",
      "Sensibilidade: ", round(conf_teste$byClass["Sensitivity"], 4), " | ",
      "Especificidade: ", round(conf_teste$byClass["Specificity"], 4), " | ",
      "NIR: ", round(conf_teste$overall["AccuracyNull"], 4)
    ))
  ) %>%
  tab_options(
    table.width = pct(70),
    table.font.size = px(14),
    column_labels.background.color = "#f9f9f9"
  )
```

Item e)

```
pacman::p_load(class, ISLR, caret, tidyverse, gt)
data(Weekly)

treino <- Weekly %>% filter(Year <= 2008)
teste  <- Weekly %>% filter(Year > 2008)
```

```r
train_X <- as.matrix(treino$Lag2)
test_X  <- as.matrix(teste$Lag2)
train_Direction <- treino$Direction

set.seed(2023)
pred_knn1 <- knn(train = train_X, test = test_X, cl = train_Direction, k = 1)

conf_knn1 <- confusionMatrix(pred_knn1, teste$Direction, positive = "Up")

as.data.frame(conf_knn1$table) %>%
  rename(Predito = Prediction, Real = Reference, Freq = Freq) %>%
  pivot_wider(names_from = Real, values_from = Freq) %>%
  gt() %>%
  tab_header(title = "") %>%
  cols_label(Predito = md("**Predito / Real**")) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "Down", rows = Predito == "Down")
  ) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "Up", rows = Predito == "Up")
  ) %>%
  tab_source_note(
    source_note = md(paste0(
      "**Estatísticas:** ",
      "Acurácia: ", round(conf_knn1$overall["Accuracy"], 4), " | ",
      "Sensibilidade: ", round(conf_knn1$byClass["Sensitivity"], 4), " | ",
      "Especificidade: ", round(conf_knn1$byClass["Specificity"], 4), " | ",
      "NIR: ", round(conf_knn1$overall["AccuracyNull"], 4)
    ))
  ) %>%
  tab_options(
    table.width = pct(70),
    table.font.size = px(14),
    column_labels.background.color = "#f9f9f9"
  )
```

EXERCÍCIO 3

Item a)

```r
pacman::p_load(ISLR, tidyverse, gt)
data(Auto)

Auto_mod <- Auto %>%
  as_tibble() %>%
  mutate(
    mpg1 = factor(if_else(mpg > median(mpg), 1, 0),
```

```
                  levels = c(0, 1),
                  labels = c("Abaixo Mediana", "Acima Mediana")),
    origin = factor(origin,
                    levels = 1:3,
                    labels = c("Americano", "Europeu", "Japonês"))
  )

Auto_mod %>%
  dplyr::select(mpg, mpg1, displacement, horsepower, weight, origin) %>%
  head(5) %>%
  gt() %>%
  tab_header(title = "") %>%
  cols_label(
    mpg = "MPG (Original)",
    mpg1 = md("**mpg1 (Alvo)**"),
    origin = "Origem"
  ) %>%
  tab_options(table.width = pct(90), table.font.size = px(14))
```

Item b)

```
pacman::p_load(tidyverse, patchwork, GGally, ISLR)

data(Auto)
df_auto <- Auto %>%
  mutate(
    mpg1 = factor(if_else(mpg > median(mpg), 1, 0), labels = c("Baixo", "Alto")),
    origin = factor(origin, labels = c("Americano", "Europeu", "Japonês")),
    cylinders = factor(cylinders)
  )

theme_eda <- theme_minimal(base_size = 10) +
  theme(legend.position = "none", panel.grid.minor = element_blank())

p1 <- ggplot(df_auto, aes(x = mpg1, y = cylinders, color = cylinders)) +
  geom_jitter(alpha = 0.5, width = 0.2) +
  labs(title = "mpg1 vs Cilindros", x = "Eficiência (mpg1)", y = "Cilindros") +
  theme_eda

p2 <- ggplot(df_auto, aes(x = mpg1, fill = origin)) +
  geom_bar(position = "dodge") +
  scale_fill_brewer(palette = "Set2") +
  labs(title = "mpg1 vs Origem", x = "Eficiência (mpg1)", y = "Frequência") +
  theme_eda + theme(legend.position = "top", legend.title = element_blank())

create_boxplot <- function(var, title, ylab) {
  ggplot(df_auto, aes(x = mpg1, y = .data[[var]], fill = mpg1)) +
    geom_boxplot(alpha = 0.7, outlier.size = 1) +
    scale_fill_manual(values = c("#E41A1C", "#377EB8")) +
```

```
    labs(title = title, x = "Eficiência", y = ylab) +
    theme_eda
}

p3 <- create_boxplot("displacement", "Cilindrada", "Polegadas cúbicas")
p4 <- create_boxplot("horsepower", "Potência", "Cavalos (hp)")
p5 <- create_boxplot("weight", "Peso", "Libras (lbs)")
p6 <- create_boxplot("acceleration", "Aceleração", "0-60 mph (seg)")
p7 <- create_boxplot("year", "Ano do Modelo", "Ano (70-82)")

(p1 + p2) / (p3 + p4) / (p5 + p6) / p7 +
  plot_layout(heights = c(1, 1, 1, 0.5)) +
  plot_annotation(tag_levels = 'A')
ggpairs(
  df_auto %>% select(mpg1, displacement, horsepower, weight, acceleration, year),
  mapping = aes(color = mpg1, alpha = 0.5),
  lower = list(continuous = wrap("points", size = 0.5)),
  diag = list(continuous = wrap("densityDiag", alpha = 0.5))
) +
  theme_bw(base_size = 9) +
  scale_color_manual(values = c("#E41A1C", "#377EB8")) +
  scale_fill_manual(values = c("#E41A1C", "#377EB8"))

set.seed(12345)
indices <- sample(seq_len(nrow(df_auto)), size = 0.8 * nrow(df_auto))
train_set <- df_auto[indices, ]
test_set  <- df_auto[-indices, ]
```

Item c)

```
scores_data <- tibble(
  Escore = predict(mod_lda_full, newdata = test_set)$x[,1],
  Eficiencia = test_set$mpg1
)

ggplot(scores_data, aes(x = Escore, fill = Eficiencia)) +
  geom_density(alpha = 0.6) +
  scale_fill_manual(values = c("#E41A1C", "#377EB8")) +
  labs(title = "", x = "Função Discriminante", y = "Densidade") +
  theme_minimal() +
  theme(legend.position = "top")
lda_scaling <- as.data.frame(mod_lda_full$scaling) %>%
  rownames_to_column("Variável") %>%
  rename(LD1 = LD1)

lda_means <- as.data.frame(t(mod_lda_full$means)) %>%
  rownames_to_column("Variável")

lda_full_report <- lda_scaling %>%
```

```r
  left_join(lda_means, by = "Variável")

lda_full_report %>%
  gt() %>%
  cols_label(
    Variável = md("**Variável**"),
    LD1 = md("**Coeficiente (LD1)**"),
    Baixo = md("**Média (Baixo)**"),
    Alto = md("**Média (Alto)**")
  ) %>%
  fmt_number(columns = where(is.numeric), decimals = 4) %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_body(columns = LD1)
  ) %>%
  tab_source_note(
    source_note = "Nota: Os coeficientes (LD1) indicam a
    direção e força da separação entre os grupos."
  ) %>%
  tab_options(
    table.width = pct(100),
    column_labels.background.color = "#f9f9f9"
  )
pacman::p_load(MASS, caret, tidyverse, gt)

mod_lda_full <- lda(mpg1 ~ cylinders + displacement + horsepower +
                      weight + acceleration + year,
                    data = train_set, prior = c(0.5, 0.5))

mod_lda_simples <- lda(mpg1 ~ cylinders + year,
                       data = train_set, prior = c(0.5, 0.5))

calc_metrics <- function(model, data) {
  pred <- predict(model, newdata = data)
  conf <- confusionMatrix(pred$class, data$mpg1, positive = "Alto")
  return(conf)
}

conf1 <- calc_metrics(mod_lda_full, test_set)
conf2 <- calc_metrics(mod_lda_simples, test_set)

performance_lda <- tibble(
  Modelo = c("Completo (Todos)", "Simples (Cyl + Year)"),
  Acurácia = c(as.numeric(conf1$overall["Accuracy"]),
               as.numeric(conf2$overall["Accuracy"])),
  `Taxa de Erro` = 1 - Acurácia,
  Sensibilidade = c(conf1$byClass["Sensitivity"], conf2$byClass["Sensitivity"]),
  Especificidade = c(conf1$byClass["Specificity"], conf2$byClass["Specificity"])
)
```

```
performance_lda %>%
  gt() %>%
  fmt_percent(columns = -Modelo, decimals = 2) %>%
  tab_header(title = "") %>%
  tab_options(table.width = pct(95), table.font.size = px(14))
```

Item d)

```
pacman::p_load(class, caret, tidyverse, gt)

data(Auto)
df_knn <- Auto %>%
  mutate(mpg1 = factor(if_else(mpg > median(mpg), 1, 0), levels = c(0, 1),
                       labels = c("0", "1")))

set.seed(12345)
indices <- sample(seq_len(nrow(df_knn)), size = 0.8 * nrow(df_knn))
Ex03_Treino <- df_knn[indices, ]
Ex03_Teste  <- df_knn[-indices, ]

train_X <- Ex03_Treino %>%
  mutate(cylinders = as.numeric(as.character(cylinders))) %>%
  select(cylinders, year) %>%
  as.matrix()

test_X  <- Ex03_Teste %>%
  mutate(cylinders = as.numeric(as.character(cylinders))) %>%
  select(cylinders, year) %>%
  as.matrix()

train_Y <- Ex03_Treino$mpg1

knn_metrics <- map_dfr(1:10, function(k) {
  set.seed(12345)
  pred <- knn(train = train_X, test = test_X, cl = train_Y, k = k)
  conf <- confusionMatrix(pred, Ex03_Teste$mpg1, positive = "1")

  tibble(
    K = k,
    Acurácia = conf$overall["Accuracy"],
    `Taxa de Erro` = 1 - Acurácia,
    Sensibilidade = conf$byClass["Sensitivity"],
    Especificidade = conf$byClass["Specificity"]
  )
})

knn_metrics %>%
  gt() %>%
  fmt_percent(columns = -K, decimals = 2) %>%
```

```
  tab_style(
    style = cell_text(weight = "bold", color = "#2e7d32"),
    locations = cells_body(rows = `Taxa de Erro` == min(`Taxa de Erro`))
  ) %>%
  tab_options(table.width = pct(95), table.font.size = px(14))
set.seed(12345)
best_knn_pred <- knn(train = train_X, test = test_X, cl = train_Y, k = 9)
conf_best <- confusionMatrix(best_knn_pred, Ex03_Teste$mpg1, positive = "1")

as.data.frame(conf_best$table) %>%
  rename(Predito = Prediction, Real = Reference, Freq = Freq) %>%
  pivot_wider(names_from = Real, values_from = Freq) %>%
  gt() %>%
  cols_label(Predito = md("**Predito / Real**")) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "0", rows = Predito == "0")
  ) %>%
  tab_style(
    style = cell_fill(color = "gray95"),
    locations = cells_body(columns = "1", rows = Predito == "1")
  ) %>%
  tab_source_note(
    source_note = md(paste0(
      "**Estatísticas:** ",
      "Acurácia: ", round(conf_best$overall["Accuracy"], 4), " | ",
      "Sensibilidade: ", round(conf_best$byClass["Sensitivity"], 4), " | ",
      "Especificidade: ", round(conf_best$byClass["Specificity"], 4), " | ",
      "NIR: ", round(conf_best$overall["AccuracyNull"], 4)
    ))
  ) %>%
  tab_options(
    table.width = pct(70),
    table.font.size = px(14),
    column_labels.background.color = "#f9f9f9"
  )
```