

Trabalho Final - Elementos de Programação para Estatística

Prova 5

Antonio Paulo Steffen Neto

Ingrid Giacomeli

Jhullya da Rosa Shalders

Pedro Henrique D'Andrea

2025-09-12

1 Introdução

Este trabalho se propõe a ser a parte de conclusão do Módulo 5 da disciplina CE302 - Elementos de Programação para Estatística. Para tal, ele consiste em uma análise de um conjunto de dados disponibilizado pelos professores responsáveis e subsequente desenvolvimento de um aplicativo em Shiny para visualização dos resultados encontrados. Assim, este trabalho apresenta a parte técnica e teórica do que foi realizado pelo grupo. A parte da apresentação em um aplicativo em Shiny, por outro lado, será feita por meio de um vídeo de apresentação. Nesta introdução apresentamos, então, o conjunto de dados escolhido pelo grupo e, também, como a parte escrita do trabalho está estruturada.

Em relação à escolha do conjuntos de dados disponibilizados, o grupo optou pelo conjunto de dados chamado *Allrecipes*. Neste constam dados retirados do site [Allrecipes](#) sobre receitas, seus países de origem e autores, informações nutricionais, tempo de preparo, avaliações e vezes em que foram reproduzidas. Assim sendo, todas as informações a serem referenciadas no texto têm como fonte a base de dados da própria plataforma “Allrecipes | Recipes, How-Tos, Videos and More” (*S.d.J.*).

Como definido pelas instruções dadas pela proposta de trabalho, a formatação deste segue a ordem: Introdução; Materiais e Métodos; Resultados e Discussão; Conclusão e Referências. Muitas das partes integrantes da ordem originalmente indicada apresentam suas próprias subseções, assim sendo, esta introdução conta com duas subseções: Contextualização e Definição do Problema e Objetivos da Análise. A parte de Materiais e Métodos, por sua vez, tem como subdivisões: Descrição do Conjunto de Dados, Dicionário de Variáveis, Tratamento de Dados e Técnicas Estatísticas e Computacionais (conforme detalhado na Section 2). Na sequência, são apresentados os Resultados e Discussão, divididos de acordo com temas de análise, com suas respectivas análises e interpretações como será, posteriormente, fácil de analisar com o aplicativo Shiny feito em conjunto com esta parte escrita. Por fim, a Conclusão e as Referências Bibliográficas não apresentam quaisquer divisões.

1.1 Contextualização e Definição do Problema

A alimentação é a base de toda a cultura humana e o momento de comumhão em diferentes culturas e religiões. Isso estende-se desde o cômodo onde as pessoas se alimentam até a como grandes festividades por todo o globo costumam ter alguma refeição como parte central das celebrações. Nos últimos tempos, o destaque adquirido por grandes chefes de cozinha e por produtos culturais como reality show e séries sobre culinária são sintomas deste fenômeno.

Mais recentemente, contudo, a relação das pessoas com a comida foi muito modificada e tornada mais complexa. Com o advento de internet, fast food e do fácil acesso a ingredientes de todo o mundo, sites como o *Allrecipes* se tornaram um meio imprescindível de compartilhamento de hábitos alimentares e, para pesquisadores, uma fonte preciosa de

dados para entender como as pessoas se alimentam. Mas sobre o que se trata o site em questão? O *Allrecipes* consiste em um site de compartilhamento de receitas. Nele, usuários de todo o mundo podem postar suas receitas para que outras pessoas possam prová-las, acompanhar as avaliações dos outros usuários e, também, acessar, provar e avaliar receitas de outros usuários. O site *Allrecipes*, portanto, e já olhando para os dados de interesse na presente análise, nos traz dados a respeito da origem das receitas e por quem foram postadas. Além disso, também inclui as informações acerca de tempos de preparo e, também, de questões nutricionais.

A análise de dados não estruturados provenientes da plataforma online *Allrecipes* insere-se na emergente área da Gastronomia Computacional, que tem por objetivo o emprego de técnicas *data-driven* para o estudo e a análise de comidas. Esse campo busca quantificar e modelar fenômenos culinários e de consumo alimentar. Neste contexto, o volume e a variedade dos dados de receitas (incluindo as avaliações, informações nutricionais e origem geográfica) constituem um conjunto de dados de nicho, que seria ideal para a aplicação de métodos de aprendizado de máquina e modelagem estatística, caso os autores deste trabalho estivessem mais adiante no curso de Estatística.

1.2 Objetivos da Análise

Dado o conjunto de dados escolhido, portanto, é possível fazer uma análise que cubra de maneira completa todas as variáveis a disposição, e, deste modo, explorar e documentar as dinâmicas relacionadas à popularidade de receitas publicadas online. Assim, abrem-se diversas direções de investigação, pois é possível entrelaçar as informações nutricionais com as de ingredientes, além das informações de países de origem, avaliações, tempos de preparo, datas de publicação e sucesso de reprodução.

O objetivo principal da análise, é o de entender quais as dinâmicas por trás da popularidade de receitas publicadas online na plataforma *Allrecipes*. Assim sendo, a análise é conduzida em partes e em diferentes frentes que englobam os diferentes tipos de variáveis à disposição nos conjuntos de dados disponíveis. As divisões, como foram feitas, se apresentam da seguinte maneira:

1. Panorama Geral: Caracterizar o perfil nutricional predominante das receitas mais populares. Isto envolverá a limpeza e harmonização das variáveis contínuas (calorias, gordura, proteína, sódio) e o uso de métricas de dispersão para comparar a centralidade desses atributos em relação ao sucesso de reprodução.
2. Países e Culinárias: O objetivo é identificar visualmente (via gráficos de dispersão) se existe uma correlação descritiva entre a complexidade do preparo e as avaliações médias.
3. As avaliações de acordo com os ingredientes: Dividir e listar os ingredientes apresentados e relacioná-los com as avaliações médias e totais distribuídas por cada receita.
4. Análise Nutricional: Partindo das informações acerca de calorias, proteínas, gordura e carboidratos à disposição, serão traçados perfis a serem relacionados, também, com as avaliações das receitas.
5. As avaliações de acordo com tempos de preparo: Seguindo a mesma lógica dos itens anteriores, as diferentes métricas de tempo de preparo das receitas serão analisadas e avaliadas de acordo com a existência, ou não, de relação com as avaliações das receitas.
6. Popularidades das cozinhas e países: Relacionando tudo o que foi mostrado anteriormente, serão utilizadas medidas de dispersão para relacionar os itens anteriores com as informações de origens de cada prato e culinária.

2 Materiais e Métodos

Teremos como principal material de apoio, suplementar aos materiais da disciplina, o livro *R for data science: import, tidy, transform, visualize, and model data*, de autoria de Wickham; Grolemund (2017). Também serão utilizados para fundamentação da análise de observações e para o tratamento de variáveis categóricas os conceitos desenvolvidos por Agresti (2019).

Além de ambos, o presente estudo é informado por trabalhos prévios que investigaram a culinária online. Em particular, serão utilizados os materiais de Trattner; Elsweiler; Howard (2017) e Hussain et al. (2025). O trabalho de Trattner et al. (2017), que analisou a saudabilidade das receitas no [Allrecipes](#), é um referencial metodológico e contextual crucial para a presente análise. Além disso, também será utilizada como base conceitual o trabalho desenvolvidor por Silva Da Costa; Amorim (2021), que aborda a questão da percepção de internautas sobre receitas publicadas online.

2.1 Descrição do Conjunto de Dados

Para fazer a descrição dos conjuntos de dados fornecidos, começamos carregando os dados e utilizando a função `str()` para vermos as descrições estruturais de ambos os conjuntos.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr    1.3.1
v purrr    1.1.0
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidyTuesdayR)
```

Warning: pacote 'tidytuesdayR' foi compilado no R versão 4.5.2

```
# Carregamento dos dados originais
all_recipes <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/2023/01/03_recipes.csv')
cuisines <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/2023/01/03_cuisines.csv')

str(all_recipes)
```

```
spc_tbl_ [14,426 x 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
$ name      : chr [1:14426] "Chewy Whole Wheat Peanut Butter Brownies" "Pumpkin Pie Eggnog" "Eggs Poa...
$ url       : chr [1:14426] "https://www.allrecipes.com/recipe/140717/chewy-whole-wheat-peanut-butter...
$ author     : chr [1:14426] "DMOMMY" "Bobbie Susan" "Bren" "Sarah Brekke" ...
$ date_published: Date[1:14426], format: "2020-06-18" "2022-09-26" ...
$ ingredients: chr [1:14426] " cup margarine, softened, cup white sugar, ½ cup packed brown sugar, 2
$ calories   : num [1:14426] 222 477 354 356 366 709 466 782 355 395 ...
$ fat        : num [1:14426] 13 31 18 9 22 47 27 61 15 12 ...
$ carbs      : num [1:14426] 24 43 32 53 23 31 1 19 33 33 ...
$ protein    : num [1:14426] 6 8 20 19 19 37 52 40 23 37 ...
$ avg_rating : num [1:14426] 4.4 5 4.8 4.3 4.7 4.2 4.4 4.6 4.7 4.7 ...
$ total_ratings: num [1:14426] 47 1 4 14 84 5 648 347 129 195 ...
$ reviews    : num [1:14426] 36 1 4 13 67 3 468 259 102 153 ...
```

```

$ prep_time      : num [1:14426] 20 10 10 20 30 45 15 10 30 20 ...
$ cook_time      : num [1:14426] 35 5 75 40 95 80 45 190 480 55 ...
$ total_time     : num [1:14426] 55 495 85 60 125 155 60 200 510 75 ...
$ servings       : num [1:14426] 16 8 4 8 8 12 6 8 12 6 ...
- attr(*, "spec")=
.. cols(
..   name = col_character(),
..   url = col_character(),
..   author = col_character(),
..   date_published = col_date(format = ""),
..   ingredients = col_character(),
..   calories = col_double(),
..   fat = col_double(),
..   carbs = col_double(),
..   protein = col_double(),
..   avg_rating = col_double(),
..   total_ratings = col_double(),
..   reviews = col_double(),
..   prep_time = col_double(),
..   cook_time = col_double(),
..   total_time = col_double(),
..   servings = col_double()
.. )
.. )
- attr(*, "problems")=<externalptr>

```

```
str(cuisines)
```

```

spc_tbl_ [2,218 x 17] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
$ name          : chr [1:2218] "Saganaki (Flaming Greek Cheese)" "Coney Island Knishes" "Diana's Hawaiian ...
$ country       : chr [1:2218] "Greek" "Jewish" "Australian and New Zealander" "Chilean" ...
$ url           : chr [1:2218] "https://www.allrecipes.com/recipe/263750/flaming-greek-cheese-saganaki/"
$ author         : chr [1:2218] "John Mitzewich" "John Mitzewich" "CHIPPENDALE" "Heidi" ...
$ date_published: Date[1:2218], format: "2024-02-07" "2024-11-26" ...
$ ingredients   : chr [1:2218] "1 (4 ounce) package kasseri cheese, 1 tablespoon water, or as needed, ¼ c ...
$ calories       : num [1:2218] 391 301 64 106 449 958 378 90 157 322 ...
$ fat            : num [1:2218] 25 17 3 9 23 24 10 5 6 16 ...
$ carbs           : num [1:2218] 15 31 9 7 58 144 59 10 25 39 ...
$ protein         : num [1:2218] 16 7 1 1 7 46 14 1 2 7 ...
$ avg_rating     : num [1:2218] 4.8 4.6 4.3 5 3.8 4.4 4.3 NA 4.6 5 ...
$ total_ratings  : num [1:2218] 25 10 126 1 13 40 3 NA 65 2 ...
$ reviews         : num [1:2218] 22 9 104 1 11 32 3 NA 55 2 ...
$ prep_time       : num [1:2218] 10 30 20 10 30 30 30 40 0 5 ...
$ cook_time       : num [1:2218] 5 75 15 0 15 165 75 30 0 5 ...
$ total_time      : num [1:2218] 15 180 180 10 45 675 585 155 0 10 ...
$ servings        : num [1:2218] 2 16 12 6 15 6 6 84 24 1 ...
- attr(*, "spec")=
.. cols(
..   name = col_character(),
..   country = col_character(),
..   url = col_character(),
.. 
```

```

.. author = col_character(),
.. date_published = col_date(format = ""),
.. ingredients = col_character(),
.. calories = col_double(),
.. fat = col_double(),
.. carbs = col_double(),
.. protein = col_double(),
.. avg_rating = col_double(),
.. total_ratings = col_double(),
.. reviews = col_double(),
.. prep_time = col_double(),
.. cook_time = col_double(),
.. total_time = col_double(),
.. servings = col_double()
.. )
- attr(*, "problems")=<externalptr>

```

Isto feito, podemos dizer que o primeiro conjunto, o maior, chamado `all_recipes`, é constituído de 14.426 observações e 16 variáveis; o segundo, `cuisines`, 2.218 e 17 variáveis. O segundo conjunto tem, portanto, uma variável a mais, esta sendo a variável `country`, uma variável de caracteres que descreve o país de origem da receita. Nas outras variáveis, presentes em ambos os conjuntos de dados, há informações que descrevem as receitas postadas no site, trazendo informações dos autores, url's das receitas, ingredientes e informações nutricionais, tempos de preparo, cozimento e total, avaliações e vezes reproduzidas. Grande parte das variáveis é do tipo numérico, três do tipo caractere e uma do tipo data.

As observações vêm em uma ordem padrão de índice (definida pelo fornecedor), mas podem ser facilmente classificadas (reordenadas) de acordo com as variáveis disponíveis, como, por exemplo, ordem crescente ou decrescente de calorias, classificação média, ou tempo de preparo, etc.

2.2 Dicionário de Variáveis

Abaixo, temos um dicionário das variáveis existentes em ambos os conjuntos de dados:

1. `all_recipes`:

Variável	Tipo de Dado	Descrição Completa
<code>author</code>	Caractere/Fator	Nome do usuário que postou a receita
<code>recipe_name</code>	Caractere	Nome da receita
<code>url</code>	Caractere	Link da receita
<code>ingredients</code>	Caractere	Base para contagem de ingredientes
<code>prep_time</code>	Numérico	Tempo de preparo (minutos)
<code>cook_time</code>	Numérico	Tempo de cozimento (minutos)
<code>total_time</code>	Numérico	Tempo total (preparo + cozimento). Variável chave para Complexidade
<code>servings</code>	Numérico	Número de porções que a receita produz
<code>calories</code>	Numérico	Conteúdo energético (Kcal). Variável chave Nutricional
<code>fat</code>	Numérico	Quantidade de gordura (g)

Variável	Tipo de Dado	Descrição Completa
protein	Numérico	Quantidade de proteína (g)
carbs	Numérico	Quantidade de carboidratos (g)
avg_rating	Numérico	Avaliação média (1-5). Métrica de popularidade
total_rating	Numérico	Número de avaliações. Métrica de popularidade
reviews	Numérico/Inteiro	Número de avaliações recebidas
date_published	Data	Data em que a receita foi publicada no site

2. cuisines:

Variável	Tipo de Dado	Descrição Completa
author	Caractere/Fator	Nome do usuário
name	Caractere	Nome da receita
url	Caractere	Link da receita
ingredients	Caractere	Usado para derivar a contagem de ingredientes
prep_time	Numérico	Tempo de preparo (minutos)
cook_time	Numérico	Tempo de cozimento (minutos)
total_time	Numérico	Tempo total (preparo + cozimento)
servings	Numérico	Número de porções
calories	Numérico	Conteúdo energético (Kcal)
fat	Numérico	Quantidade de gordura (g)
protein	Numérico	Quantidade de proteína (g)
carbs	Numérico	Quantidade de carboidratos (g)
avg_rating	Numérico	Avaliação média (1-5). Métrica de popularidade
total_rating	Numérico	Número de avaliações. Métrica de popularidade
reviews	Numérico/Inteiro	Número de avaliações recebidas
date_published	Data	Data de publicação
country	Caractere/Fator	País de origem. Variável chave para análise geocultural

2.3 Tratamento dos Dados

1. Fusão de Conjuntos de Dados: Inicialmente, o dataset principal (`all_recipes`) foi unido ao dataset auxiliar (`cuisines`) utilizando-se a variável de chave primária (`recipe_id` ou equivalente), resultando em um único data frame analítico.
2. Tratamento de Valores Faltantes (NA): Verificou-se a presença de valores NA (Not Available) nas variáveis nutricionais. Foi adotada a estratégia de imputação por mediana para os NAs em variáveis contínuas, como calorias, onde a ausência de informação era considerada Missing At Random (MAR), para maximizar o número de observações úteis. Para a variável `country`, os NAs foram categorizados como “Não Especificado”.
3. Padronização e Normalização: Variáveis com alta variabilidade, como reproduções, foram consideradas para transformação logarítmica ($\log(X + 1)$) para aproximar suas distribuições de uma normalidade e mitigar a influência de outliers extremos no ajuste dos modelos de regressão.

- Codificação de Variáveis Categóricas: A variável `country` foi convertida para o tipo fator (`factor`) para uso em modelos estatísticos, sendo implementado o esquema de codificação dummy (variáveis binárias) para as análises de regressão. Esta abordagem é fundamental para o tratamento rigoroso de variáveis nominais em contextos de modelagem, conforme diretrizes da análise de dados categóricos Agresti (2019).

2.4 Técnicas Estatísticas e Computacionais

Detalhando os métodos utilizados para o processamento, análise e visualização dos dados, podemos começar apontando o fato de que o estudo foi conduzido utilizando a linguagem R e o ecossistema de pacotes `tidyverse` (incluindo `dplyr` e `ggplot2`). As etapas iniciais de preparação de dados incluíram algumas tarefas distintas. Dentre elas pode-se apontar duas direções principais: a de padronização geográfica e a de derivação das variáveis. Ambas serão definidas nos próximos parágrafos.

Primeiramente, a padronização geográfica. Esta consistiu na padronização de países e regiões para categorias consistentes (por exemplo, o agrupamento de sub-culinárias regionais dos EUA sob o rótulo “US American”). Posteriormente, as culinárias foram agrupadas em seis categorias continentais (Ásia, Europa, América do Norte, América do Sul, África e Oceania) para análises macro.

Em segundo lugar, e menos para a simples categorização de variáveis e observações, foram criadas duas variáveis-chave para medir a complexidade e o esforço. A primeira a Quantidade de Ingredientes (\hat{I}): Calculada pela contagem de vírgulas nos *strings* da lista de ingredientes, adicionada de um, como proxy da complexidade. A segunda, o Tempo Total de Preparo (\hat{T}): Variável `total_time` convertida para formato numérico (minutos), representando o esforço temporal.

2.5 Métodos Estatísticos e Modelagem

Olhando para as análises estatísticas, elas se concentraram em quatro vertentes principais. A primeira é dedicada à estatística descritiva e ao ranqueamento de observações. Foram reaisizados cálculo de média de avaliação (\bar{R}_g), desvio padrão (s_R) e contagem de receitas (N) por país/região. Após isto, foi aplicado um filtro mínimo de $N \geq 10$ receitas para garantir a estabilidade das médias de avaliação. A segunda vertente diz respeito à análise das quantidades de ingredientes e dos tempos de preparo, foi empregada a análise de correlação, utilizando o Coeficiente de Correlação de Pearson (r) para quantificar a relação linear entre as variáveis de complexidade/esforço e a qualidade percebida (média de avaliação).

Além delas, a análise bivariada foi empregada para demonstrar a relação entre Popularidade (Média de Avaliações por Receita) e Qualidade (Média de Nota). Um **Gráfico de Bolhas** foi escolhido para visualizar simultaneamente estas duas métricas e o número de receitas (Tamanho da Bolha) por culinária.

2.6 Visualização dos Dados

A visualização de dados é parte central para a interpretação dos resultados, sendo o pacote `ggplot2` utilizado para a criação de todos os gráficos estatísticos, como o ranking de culinárias, os gráficos de dispersão e o Gráfico de Bolhas. Os pacotes `gridExtra` e `grid` foram cruciais para a renderização e o arranjo das tabelas comparativas e seus títulos (`tableGrob`, `grid.arrange`, `textGrob`), garantindo a apresentação organizada dos sumários de dados.

3 Resultados e discussão

3.1 Relação entre Países, Avaliações e Porções Servidas

Os resultados atingidos em cada parte do trabalho seguem em conformidade com as diferentes direções tomadas em cada análise. Inicialmente, a análise das médias de avaliação revelou um ranking onde culinárias com alta qualidade percebida (alta \bar{R}_g) frequentemente exibiam um baixo desvio padrão (s_R), sugerindo consistência na excelência. O agrupamento continental permitiu observar tendências regionais na satisfação do usuário. A Figura 1 apresenta o ranking das top 15 culinárias, destacando a média de avaliação com barras de erro representando o desvio padrão. Esta análise da variação na qualidade entre continentes sugere que a tradição e a identidade gastronômica específica de uma região exercem maior influência na satisfação do que fatores universais de preparação.

```
# limpeza e organização dos dados

cuisines_limpo <- cuisines %>%
  mutate(
    country_padrao = case_when(
      # Padroniza "Jewish"
      country %in% c("Jewish") ~ "Israeli/Jewish",
      # Padroniza culinárias regionais dos EUA
      country %in% c("Cajun and Creole", "Southern Recipes", "Tex-Mex", "Southwestern Recipes", "Amish and
      # Mantém o nome do país para todos os outros
      .default = country
    )
  )

# Exibe a contagem das categorias padronizadas (opcional)
table(cuisines_limpo$country_padrao)
```

Argentinian	Australian and New Zealander	
30		65
Austrian	Bangladeshi	
22		12
Belgian	Brazilian	
6		67
Canadian	Chilean	
67		22
Chinese	Colombian	
65		11
Cuban	Danish	
65		33
Dutch	Filipino	
22		66
Finnish	French	
18		65
German	Greek	
62		62
Indian	Indonesian	
65		24

Israeli		Israeli/Jewish	
23		61	
Italian		Jamaican	
64		43	
Japanese		Korean	
63		56	
Lebanese		Malaysian	
51		24	
Norwegian		Pakistani	
26		25	
Persian		Peruvian	
45		38	
Polish		Portuguese	
61		53	
Puerto Rican		Russian	
60		65	
Scandinavian		South African	
38		19	
Spanish		Swedish	
61		31	
Swiss		Thai	
10		62	
Turkish		US American	
36		292	
Vietnamese			
62			

```

analise_por_pais <- cuisines_limpo %>%
# Remove linhas onde a avaliação é NA
filter(!is.na(avg_rating)) %>%
group_by(country_padrao) %>%
summarise(
  contagem_receitas = n(),
  media_rating = mean(avg_rating, na.rm = TRUE),
  desvio_padrao_rating = sd(avg_rating, na.rm = TRUE),
  media_porcoes = mean(servings, na.rm = TRUE),
  .groups = 'drop'
) %>%
# Filtra países com poucas receitas
filter(contagem_receitas >= 10) %>%
arrange(desc(media_rating))

# Exibe o resultado da análise
print(analise_por_pais)

```

```

# A tibble: 42 x 5
  country_padrao contagem_receitas media_rating desvio_padrao_rating
  <chr>                  <int>        <dbl>                <dbl>
1 French                   62          4.69                0.238
2 Greek                    57          4.68                0.221

```

3 Italian	57	4.65	0.373
4 Puerto Rican	59	4.61	0.458
5 Korean	55	4.60	0.294
6 Persian	38	4.60	0.384
7 Israeli/Jewish	60	4.59	0.266
8 Chinese	64	4.58	0.272
9 Russian	65	4.57	0.270
10 US American	282	4.55	0.354
# i 32 more rows			
# i 1 more variable: media_porcoes <dbl>			

```
# A tibble: 6 x 4
  continente      contagem_total media_rating_continente desvio_padrao_continente
  <chr>              <int>                  <dbl>                      <dbl>
1 América do Norte       515                   4.54                      0.362
2 Europa                  658                   4.52                      0.356
3 África                   19                    4.49                      0.371
```

```

4 Ásia           697          4.49          0.383
5 Oceania        62           4.36          0.478
6 América do Sul 146           4.34          0.515
# i abbreviated name: 1: desvio_padrao_continente

grafico_colorido <- analise_por_continente %>%
  # Limita aos Top 15 países/regiões para visualização
  slice_max(media_rating, n = 15) %>%

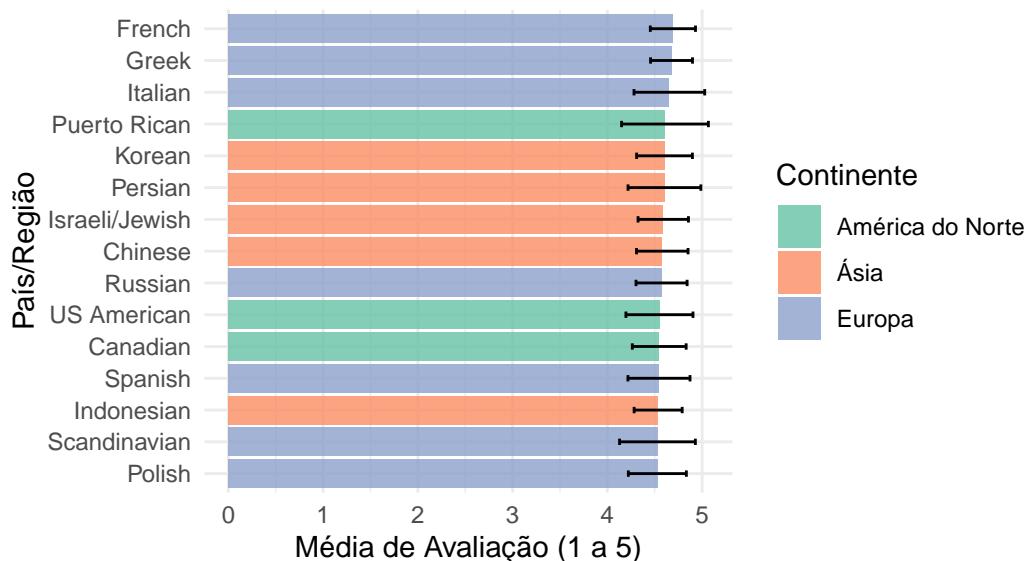
  # Estética principal, usando 'fill' para o continente
  ggplot(aes(x = reorder(country_padrao, media_rating),
             y = media_rating,
             fill = continente)) +
    # Adiciona as barras coloridas
    geom_col(alpha = 0.8) +
    # Adiciona as Barras de Erro (Desvio Padrão)
    geom_errorbar(aes(ymin = media_rating - desvio_padrao_rating,
                       ymax = media_rating + desvio_padrao_rating),
                  width = 0.2, color = "black") + # Cor das barras de erro
    # Escala de cores (Set2 é uma paleta amigável)
    scale_fill_brewer(palette = "Set2") +
    # Inverte os eixos para facilitar a leitura
    coord_flip() +
    # Títulos e Rótulos
    labs(
      title = "Top 15 Culinárias por Média de Avaliação (Coloridas por Continente)",
      subtitle = "Barras de erro representam o Desvio Padrão do Rating.",
      x = "País/Região",
      y = "Média de Avaliação (1 a 5)",
      fill = "Continente" # Legenda da cor
    ) +
    # Tema
    theme_minimal() +
    theme(plot.title = element_text(face = "bold"))

print(grafico_colorido)

```

Top 15 Culinárias por Média de Avaliação (Coloridas)

Barras de erro representam o Desvio Padrão do Rating.



3.2 Relação entre Ingredientes e Avaliações

A análise dos dados revela uma correlação positiva, porém extremamente fraca, entre a quantidade de ingredientes e a avaliação final das receitas em ambos os bancos de dados. A Figura 2 confirma que, com coeficientes de 0,1 para “cuisines” e 0,08 para “all_recipes”, conclui-se que a complexidade do prato (medida pelo número de itens) praticamente não influencia a nota atribuída pelos usuários; ou seja, embora exista uma tendência matemática muito sutil de aumento na avaliação conforme o número de ingredientes cresce, na prática essa relação é desprezível, indicando que receitas simples têm tanta chance de serem bem avaliadas quanto as mais complexas.

```
#ingredientes x nota
ingredientes_nota = function(dados, nome_do_banco) {

  #coluna de quantidade de ingredientes
  dados_processados = dados %>%
    filter(!is.na(avg_rating), !is.na(ingredients)) %>%
    mutate(
      qtd_ingredientes = str_count(ingredients, ",") + 1,
      avg_rating = as.numeric(avg_rating)
    )

  # Correlação
  correlacao = cor(dados_processados$qtd_ingredientes,
                    dados_processados$avg_rating,
                    use = "complete.obs")

  cat(paste0("\nAnálise para: ", nome_do_banco, "\n"))
  cat(paste("Correlação:", round(correlacao, 3), "\n"))
}
```

```

# Gráfico
grafico = ggplot(dados_processados, aes(x = qtd_ingredientes, y = avg_rating)) +
  geom_jitter(alpha = 0.2, color = "steelblue") +
  geom_smooth(method = "lm", color = "darkred") +
  labs(
    title = paste("Ingredientes vs Nota:", nome_do_banco),
    subtitle = paste("Correlação de Pearson:", round(correlacao, 3)),
    x = "Quantidade de Ingredientes",
    y = "Nota Média"
  ) +
  theme_minimal()

print(grafico)
}

ingredientes_nota(cuisines, "Banco Cuisines")

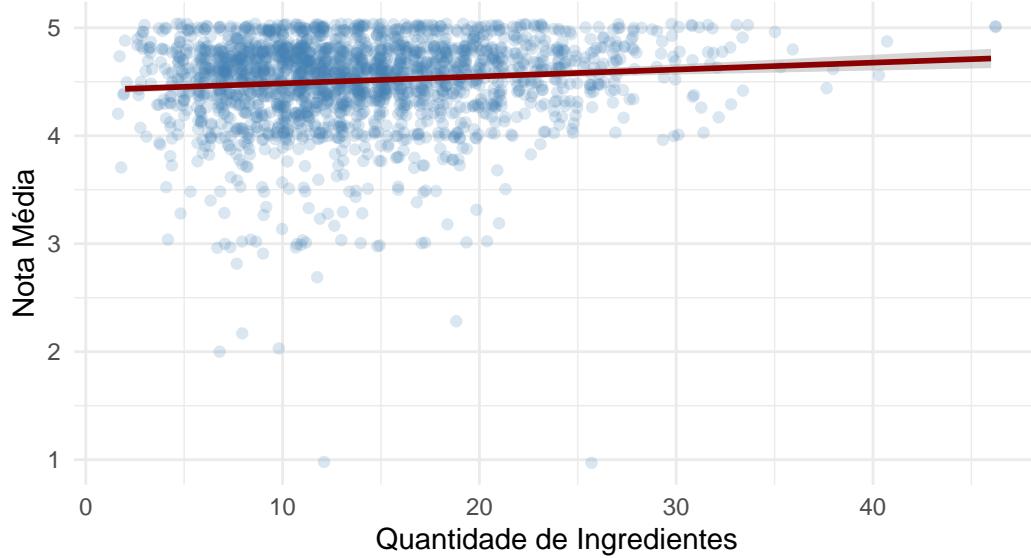
```

Análise para: Banco Cuisines
 Correlação: 0.1

`geom_smooth()` using formula = 'y ~ x'

Ingredientes vs Nota: Banco Cuisines

Correlação de Pearson: 0.1



```
ingredientes_nota(all_recipes, "Banco All Recipes")
```

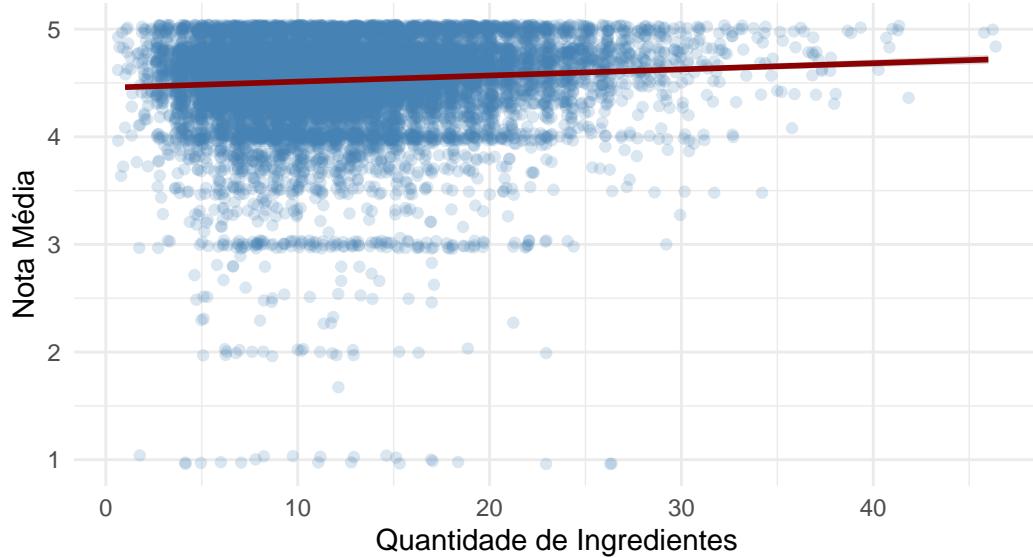
Análise para: Banco All Recipes

Correlação: 0.08

```
`geom_smooth()` using formula = 'y ~ x'
```

Ingredientes vs Nota: Banco All Recipes

Correlação de Pearson: 0.08



3.3 Relação entre Tempo de Preparo e Avaliações

Ao analisar a influência do tempo total de preparo (total_time) sobre a avaliação das receitas (avg_rating), com uma visualização (Figura 3) abrangendo até 480 minutos (8 horas), observou-se uma ausência de correlação significativa em ambos os bancos de dados. Com coeficientes irrelevantes de 0,047 para o banco ‘cuisines’ e 0,008 para ‘all_recipes’, os dados demonstram estatisticamente que a duração da receita não é um fator determinante para a sua nota final. Isso sugere que a satisfação dos usuários independe do tempo investido na cozinha, com pratos rápidos recebendo avaliações tão altas quanto preparos de longa duração.

```
tempo_nota <- function(dados, nome_do_banco) {  
  # 1. Limpeza e Tratamento  
  dados_processados <- dados %>%  
    mutate(  
      avg_rating = as.numeric(avg_rating),  
      total_time = as.numeric(total_time)  
    ) %>%  
    # Removemos linhas vazias ou com tempo zerado/negativo  
    filter(!is.na(avg_rating), !is.na(total_time), total_time > 0)  
  
  # 2. Cálculo da Correlação (considerando todos os dados)  
  correlacao <- cor(dados_processados$total_time,
```

```

dados_processados$avg_rating,
use = "complete.obs")

cat(paste0("\n--- Análise: ", nome_do_banco, " ---\n"))
cat(paste("Correlação geral:", round(correlacao, 3), "\n"))

# 3. Gráfico de Dispersão
# OBS: Fiz um filtro NO GRÁFICO para mostrar apenas receitas até 480 min (8h)
# Se não fizer isso, uma única receita de 24h esmaga o gráfico todo.
grafico <- ggplot(dados_processados %>% filter(total_time <= 480),
                   aes(x = total_time, y = avg_rating)) +
  geom_jitter(alpha = 0.2, color = "forestgreen") + # Pontos verdes
  geom_smooth(method = "lm", color = "black") +      # Linha de tendência preta
  labs(
    title = paste("Tempo de Preparo vs Nota:", nome_do_banco),
    subtitle = paste("Correlação de Pearson:", round(correlacao, 3)),
    x = "Tempo de Preparo (minutos)",
    y = "Nota (0-5)"
  ) +
  theme_minimal()
print(grafico)
}

# Executando para os dois bancos
tempo_nota(cuisines, "Banco Cuisines")

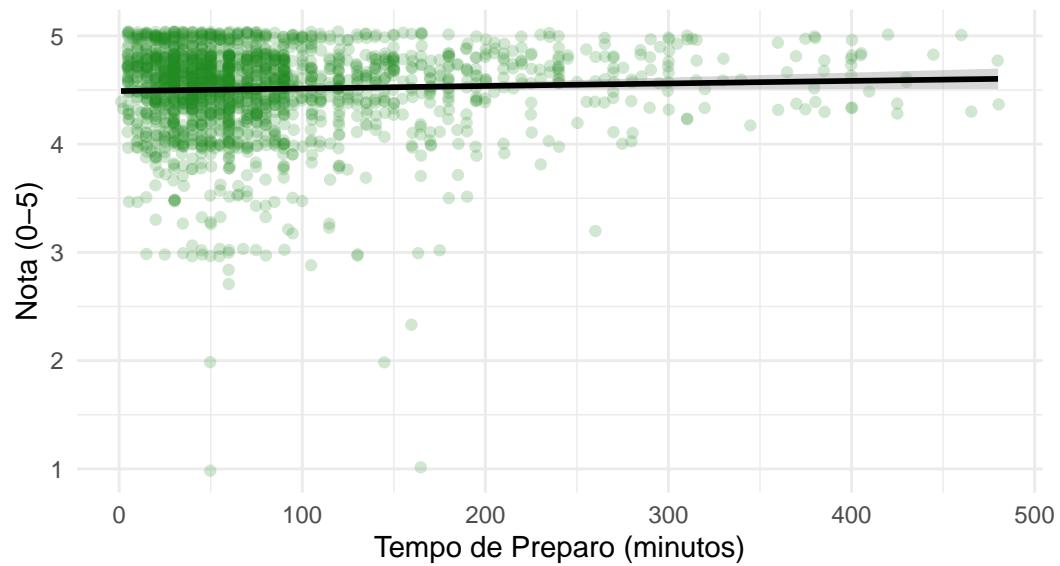
```

--- Análise: Banco Cuisines ---
 Correlação geral: 0.047

`geom_smooth()` using formula = 'y ~ x'

Tempo de Preparo vs Nota: Banco Cuisines

Correlação de Pearson: 0.047



```
tempo_nota(all_recipes, "Banco All Recipes")
```

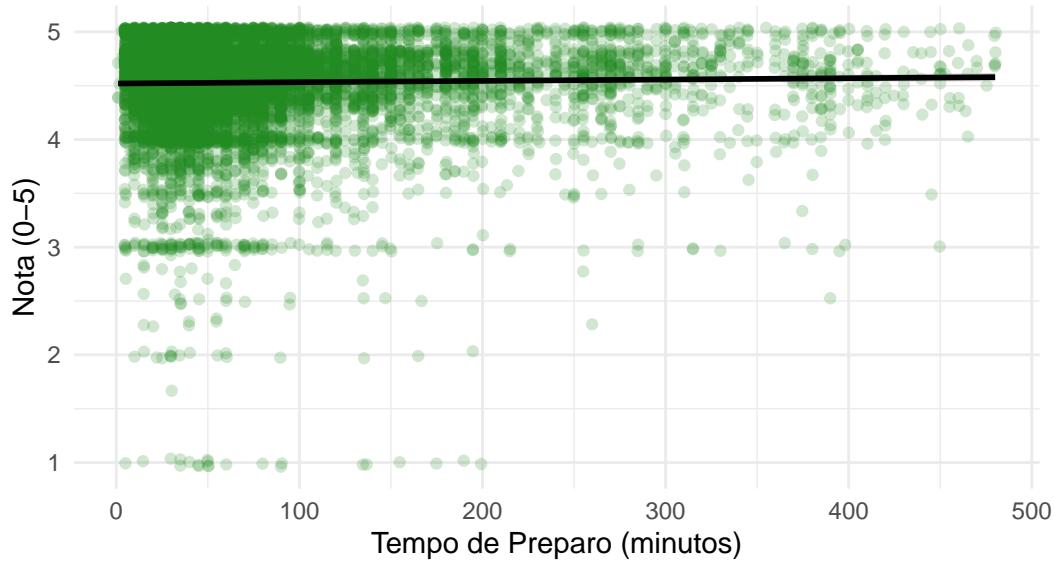
--- Análise: Banco All Recipes ---

Correlação geral: 0.008

```
`geom_smooth()` using formula = 'y ~ x'
```

Tempo de Preparo vs Nota: Banco All Recipes

Correlação de Pearson: 0.008



3.4 Relação entre Culinárias e Qualidade de acordo com uma análise bivariada

Ao olharmos para a análise bivariada, por fim, estabeleceu-se uma clara diferenciação entre as culinárias mais populares e as de maior qualidade. O Gráfico de Bolhas (Figura 4) ilustra que a alta Popularidade, medida pela média de avaliações por receita, nem sempre se traduz em alta Qualidade, medida pela média de nota. Esta divergência indica que a popularidade pode ser impulsionada por fatores de acessibilidade ou tendência, desvinculando-se da excelência na avaliação. A análise de tendência temporal complementou este achado ao mostrar que o Total de Avaliações no Ano é dinâmico, refletindo o ciclo de vida e a ascensão do interesse em certas culinárias ao longo do tempo.

```
# 1) Instalar e carregar pacotes
packages <- c("dplyr", "ggplot2", "readr", "DT")

# Verifica e instala pacotes faltantes (opcional em R Markdown)
installed <- packages %in% rownames(installed.packages())
if (any(!installed)) {
  install.packages(packages[!installed])
}

# Carrega os pacotes
lapply(packages, library, character.only = TRUE)
```

Warning: pacote 'DT' foi compilado no R versão 4.5.2

```
[[1]]
[1] "tidytuesdayR"   "lubridate"      "forcats"        "stringr"        "dplyr"
[6] "purrr"          "readr"          "tidyverse"      "tibble"         "ggplot2"
[11] "tidyverse"       "stats"          "graphics"      "grDevices"     "utils"
```

```

[16] "datasets"      "methods"       "base"

[[2]]
[1] "tidytuesdayR" "lubridate"     "forcats"      "stringr"      "dplyr"
[6] "purrr"        "readr"         "tidyR"        "tibble"       "ggplot2"
[11] "tidyverse"    "stats"         "graphics"    "grDevices"   "utils"
[16] "datasets"      "methods"       "base"

[[3]]
[1] "tidytuesdayR" "lubridate"     "forcats"      "stringr"      "dplyr"
[6] "purrr"        "readr"         "tidyR"        "tibble"       "ggplot2"
[11] "tidyverse"    "stats"         "graphics"    "grDevices"   "utils"
[16] "datasets"      "methods"       "base"

[[4]]
[1] "DT"           "tidytuesdayR" "lubridate"     "forcats"      "stringr"
[6] "dplyr"        "purrr"       "readr"        "tidyR"        "tibble"
[11] "ggplot2"      "tidyverse"    "stats"        "graphics"    "grDevices"
[16] "utils"        "datasets"    "methods"      "base"

```

```

# 2) Carregar e preparar os dados
cuisines <- readr::read_csv(
  "https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/2025/2025-09-16/cuisines.csv"
)

```

```

Rows: 2218 Columns: 17
-- Column specification -----
Delimiter: ","
chr  (5): name, country, url, author, ingredients
dbl  (11): calories, fat, carbs, protein, avg_rating, total_ratings, reviews...
date (1): date_published

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```

```

# Ajustar data e extrair ano
cuisines <- cuisines |>
  mutate(
    date_published = as.Date(date_published),
    year = as.integer(format(date_published, "%Y"))
  )

year_min <- min(cuisines$year, na.rm = TRUE)
year_max <- max(cuisines$year, na.rm = TRUE)

# Resumo por país/cozinha (usado para ranking e popularidade x qualidade)
cuisine_country_summary <- cuisines |>
  group_by(country) |>
  summarise(
    n_receitas      = n(),
    ...
  )

```

```

media_rating      = mean(avg_rating, na.rm = TRUE),
total_avaliacoes = sum(total_ratings, na.rm = TRUE),
media_avaliacoes = mean(total_ratings, na.rm = TRUE),
total_reviews    = sum(reviews, na.rm = TRUE),
media_reviews    = mean(reviews, na.rm = TRUE),
.groups = "drop"
)

# Dados por país e ano (para tendência)
cuisine_year <- cuisines |>
  filter(!is.na(year)) |>
  group_by(country, year) |>
  summarise(
    n_receitas      = n(),
    total_avaliacoes = sum(total_ratings, na.rm = TRUE),
    .groups = "drop"
  )

# --- Definições para o R Markdown (Valores de Filtro Fixo) ---
# Como não temos os inputs do Shiny, definimos valores fixos para simular os filtros:
MIN_RECEITAS_FIXO <- 10 # Simula input$min_receitas
METRICA_RANK_FIXO <- "total_avaliacoes" # Simula input$metrica_rank
TOP_N_FIXO        <- 10 # Simula input$top_n
ANO_MIN_FIXO      <- year_min # Simula input$ano_range[1]
ANO_MAX_FIXO      <- year_max # Simula input$ano_range[2]
PAIS_TENDENCIA_FIXO <- "top" # Simula input$pais_tendencia

# Aplica filtro fixo de mínimo de receitas
resumo_filtrado <- cuisine_country_summary |>
  filter(n_receitas >= MIN_RECEITAS_FIXO)

# Cria o ranking fixo
col_ord <- switch(
  METRICA_RANK_FIXO,
  "total_avaliacoes" = resumo_filtrado$total_avaliacoes,
  "n_receitas"        = resumo_filtrado$n_receitas,
  "media_rating"      = resumo_filtrado$media_rating
)

ranking_paises <- resumo_filtrado |>
  arrange(desc(col_ord)) |>
  slice_head(n = TOP_N_FIXO)

```

```

# Gráfico de ranking por país
df_ranking <- ranking_paises

y_lab <- switch(
  METRICA_RANK_FIXO,
  "total_avaliacoes" = "Total de avaliações",
  "n_receitas"        = "Número de receitas",

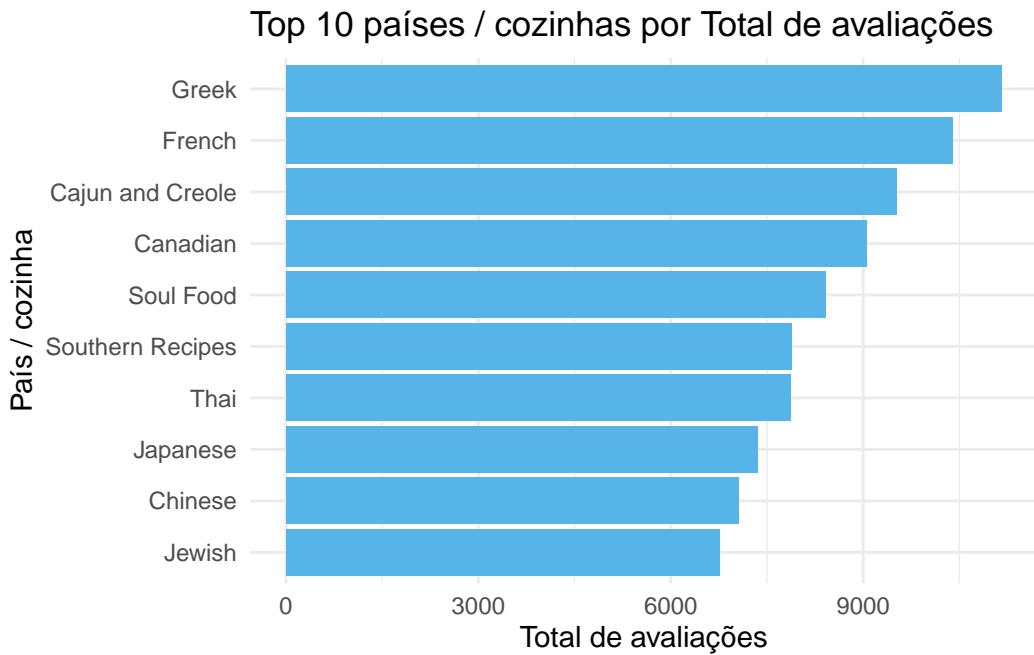
```

```

    "media_rating"      = "Média de nota"
)

ggplot(df_ranking, aes(x = reorder(country,
                                    if (METRICA_RANK_FIXO == "media_rating") media_rating else get(METRICA_
                                    y = if (METRICA_RANK_FIXO == "media_rating") media_rating else get(METRICA_RANK_FIXO
                                    geom_col(fill = "#56B4E9") +
                                    coord_flip() +
                                    labs(
                                        title = paste("Top", TOP_N_FIXO, "países / cozinhas por", y_lab),
                                        x = "País / cozinha",
                                        y = y_lab
                                    ) +
                                    theme_minimal()

```



```

# Gráfico: popularidade x qualidade por culinária
df_pop_rating <- resumo_filtrado

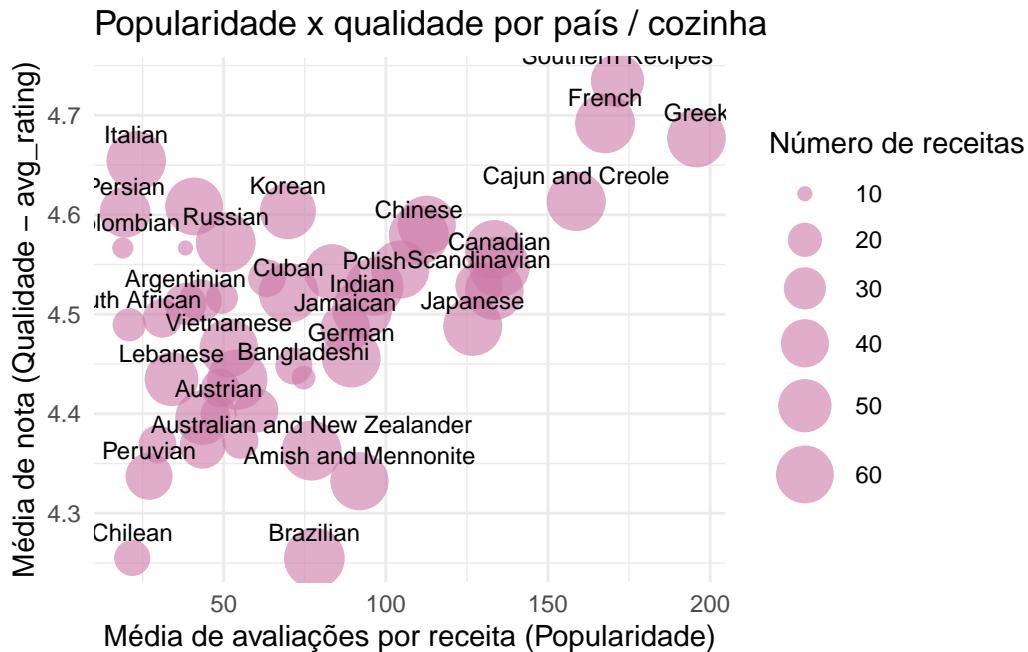
ggplot(df_pop_rating, aes(
    x = media_avaliacoes,
    y = media_rating,
    size = n_receitas,
    label = country
)) +
  geom_point(alpha = 0.6, color = "#CC79A7") +
  geom_text(check_overlap = TRUE, vjust = -1, size = 3) +
  scale_size_continuous(range = c(2, 10)) + # Ajusta o tamanho das bolhas
  labs(

```

```

        title = "Popularidade x qualidade por país / cozinha",
        x = "Média de avaliações por receita (Popularidade)",
        y = "Média de nota (Qualidade - avg_rating)",
        size = "Número de receitas"
    ) +
    theme_minimal()

```



```

# Gráfico: tendência por ano
df_year <- cuisine_year |>
  filter(
    year >= ANO_MIN_FIXO,
    year <= ANO_MAX_FIXO
  )

# Seleciona os países para a tendência
if (PAIS_TENDENCIA_FIXO == "top") {
  paises_sel <- ranking_paises$country # Usa o ranking fixo
  df_tendencia <- df_year |>
    filter(country %in% paises_sel)
} else {
  df_tendencia <- df_year |>
    filter(country == PAIS_TENDENCIA_FIXO)
}

ggplot(df_tendencia, aes(x = year, y = total_avaliacoes, color = country, group = country)) +
  geom_line(linewidth = 1) +
  geom_point() +
  labs(
    title = "Tendência da avaliação média por ano",
    subtitle = "Comparando diferentes países",
    x_label = "Ano",
    y_label = "Média de avaliações"
  )

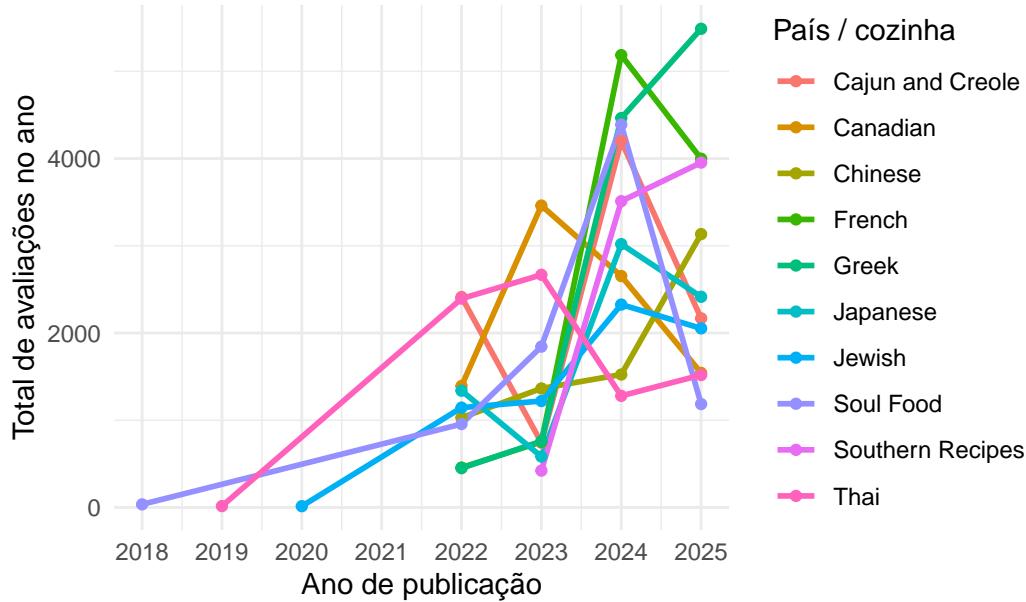
```

```

title = paste("Tendência de popularidade (Total de avaliações)",
             ifelse(PAIS_TENDENCIA_FIXO == "top", "para os Top países", paste("para",
x = "Ano de publicação",
y = "Total de avaliações no ano",
color = "País / cozinha"
) +
theme_minimal() +
scale_x_continuous(breaks = scales::pretty_breaks(n = 10))

```

Tendência de popularidade (Total de avaliações) para os Top



3.5 Análises e Interpretações

As interpretações centrais do estudo convergem na primazia da qualidade intrínseca sobre o esforço extrínseco. A correlação nula entre o tempo e os ingredientes versus a nota média implica que os usuários valorizam o resultado final da receita sobre o custo de preparo, sugerindo que o foco na otimização do sabor e na clareza da instrução é mais determinante do que a complexidade.

Em contraste, a diferença significativa nas médias de avaliação entre culinárias continentais sugere que as tradições gastronômicas consolidadas (origem) são um forte preditor da qualidade percebida.

A distinção entre popularidade e qualidade, por sua vez, reforça a necessidade de se utilizar métricas ponderadas ao ranquear culinárias. Um alto volume de interesse (popularidade) pode não ser um substituto direto para a excelência percebida (alta média de nota), e a dinâmica temporal do conteúdo é um fator chave para o engajamento na plataforma.

4 Conclusão

A aplicação dos métodos e o conjunto de resultados aqui apresentados serviram de base para a construção do aplicativo Shiny, que oferece uma ferramenta de visualização interativa para explorar as dinâmicas de popularidade e qualidade,

tornando os achados acessíveis ao público. Este trabalho conclui, assim, o que foi proposto como Avaliação 5 da disciplina de Elementos de Programação para Estatística.

O presente estudo demonstrou, através de análises estatísticas, que o esforço e a complexidade de uma receita não são fatores determinantes para a sua avaliação final pelos usuários, com o coeficiente de Pearson próximo de zero estabelecendo a independência entre estas variáveis e a satisfação. Em contrapartida, a origem geográfica da culinária influencia a avaliação média, sugerindo a alta relevância da tradição. O estudo também enfatizou a distinção analítica entre popularidade e qualidade, alertando contra a utilização de métricas de volume como substitutas para a percepção de excelência. Em síntese, a satisfação do usuário é primariamente guiada por fatores intrínsecos de sabor e qualidade de execução, e não por custos de tempo ou esforço.

Referências Bibliográficas

- AGRESTI, Alan. **An introduction to categorical data analysis**. Third edition ed. Hoboken, NJ: Wiley, 2019.
- Allrecipes | Recipes, How-Tos, Videos and More. Allrecipes, [S.d.J]. Disponível em: <<https://www.allrecipes.com/>>. Acesso em: 13 nov. 2025
- HUSSAIN, Zahid *et al.* (ORG.). **Innovative Trends Shaping Food Marketing and Consumption:** [S.l.]: IGI Global, 2025.
- SILVA DA COSTA, Ana Carolina; AMORIM, Maria Marta Amancio. [A percepção de internautas sobre as receitas mais acessadas em mídia digital](#). **Research, Society and Development**, v. 10, n. 12, p. e455101220461, set. 2021.
- TRATTNER, Christoph; ELSWEILER, David; HOWARD, Simon. [Estimating the Healthiness of Internet Recipes: A Cross-sectional Study](#). **Frontiers in Public Health**, v. 5, fev. 2017.
- WICKHAM, Hadley; GROLEMUND, Garrett. **R for data science: import, tidy, transform, visualize, and model data**. Sebastopol, CA: O'Reilly Media, Inc, 2017.