

Explorando o Ecossistema de Pacotes do R: Tendências, Aplicações e Impactos

Mateus Vieira Keiper de Carvalho e Silva¹

Luiz Eduardo Rodrigues da Silva²

João Paulo Martins dos Santos³

André Luis Azevedo Guedes⁴

Resumo

A linguagem R tem se desenvolvido, com o auxílio de uma forte comunidade de usuários e desenvolvedores, em uma ferramenta essencial para análises estatísticas e de visualização de dados. Sua robustez, associada à ampla diversidade de pacotes disponíveis, evidencia seu potencial de aplicação em múltiplas áreas do conhecimento, ao passo que oferece flexibilidade para atender a demandas emergentes de pesquisadores e profissionais. Este ambiente colaborativo e de constante evolução tem impulsionado inovações metodológicas, transformando o cenário da investigação científica. Neste estudo, foi elaborada uma análise do ecossistema de pacotes do R, explorando suas áreas de aplicação e relevância dos pacotes nestas últimas, quantificadas pelo volume histórico de *downloads*. A coleta de dados envolveu 21.637 pacotes, a qual possibilitou construir um ranking, revelando as tendências de uso e seu impacto na comunidade. Foi realizada uma categorização dos pacotes, usando *clusters*, possibilitando uma metanálise, que identificou os principais focos de desenvolvimento e suas respectivas demandas. A condução de pesquisa se baseou nas ferramentas R, RStudio, Python e o sistema de inteligência artificial Julius AI. Essa combinação permitiu a manipulação e tratamento do grande volume de dados, além do processamento de linguagem natural para a classificação automatizada dos pacotes em grupos temáticos. A abordagem visa contribuir para o entendimento profundo das áreas de concentração e das lacunas existentes no desenvolvimento de ferramentas para R. Os resultados indicam que, do ponto de vista dos desenvolvedores, há uma ênfase na criação de pacotes voltados para mineração de texto e análise de linguagem natural, análise de sobrevivência e bioestatística, análise de redes e finanças, bem como para previsão e modelagem de séries temporais. Do lado dos usuários, a análise dos *downloads* demonstra a demanda predominante de pacotes para mineração de texto e análise de linguagem natural, seguidos por programação funcional e a manipulação de dados, pacotes complementares a funcionalidades centrais do R e do Tidyverse e análise de sobrevivência e bioestatística. Um grande *insight* do estudo foi levantar que 6,25% dos pacotes atende mais de uma das grandes áreas de demanda, fato que pode trazer uma nova abordagem de desenvolvimento de pacotes multivalentes, reduzindo a enorme quantidade de possibilidades de abordagem e filtrando as ferramentas essenciais para o usuário. Através de gráficos e

¹Universidade Federal Fluminense (UFF), mateuskeiper@id.uff.br

²Universidade Federal Fluminense (UFF), luizers@id.uff.br

³Academia da Força Aérea (AFA), jp2@alumni.usp.br

⁴Centro Universitário Augusto Motta (UNISUAM), andre.guedes@souunisuam.com.br

suas análises, o presente estudo oferece um panorama do Estado da Arte do ecossistema R, evidenciando as principais linhas de desenvolvimento e suas demandas. O objetivo central da pesquisa é trazer informações que informem decisões estratégicas para o desenvolvimento de aplicações e inovações na área de análise estatística, fomentando o avanço do conhecimento e a melhoria contínua das ferramentas.

Palavras-chave: estatística aplicada, R, ciência de dados, análise computacional, pacotes estatísticos.

Abstract

The R language has evolved, with the support of a strong community of users and developers, into an essential tool for statistical analyses and data visualization. Its robustness, combined with the extensive diversity of available packages, highlights its potential for application across multiple fields of knowledge while offering flexibility to meet the emerging demands of researchers and professionals. This collaborative and continuously evolving environment have driven methodological innovations, transforming the landscape of scientific investigation. In this study, an analysis of the R package ecosystem was conducted, exploring areas of application and the relevance of packages, quantified by their historical download volumes. The data collection involved 21,637 packages, which enabled the construction of a ranking that reveals usage trends and their impact on the community. A categorization of the packages was performed using clustering, facilitating a meta-analysis that identified the main development focuses and their respective demands. The research was carried out using the tools R, RStudio, Python, and the Julius AI artificial intelligence system. This combination allowed for the efficient manipulation and processing of a large volume of data, in addition to employing natural language processing for the automated classification of packages into thematic groups. The approach aims to contribute to a deeper understanding of the concentrated areas and the existing gaps in the development of tools for R. The results indicate that, from the developers' perspective, there is an emphasis on creating packages focused on text mining and natural language analysis, survival analysis and biostatistics, network and finance analysis, as well as forecasting and time series modeling. On the user side, the analysis of downloads demonstrates a predominant demand for packages related to text mining and natural language analysis, followed by functional programming and data manipulation—packages that complement the core functionalities of R and the Tidyverse—and finally survival analysis and biostatistics. A major insight from the study was identifying that 6.25% of the packages serve more than one of the major demand areas, which could lead to a new approach in developing versatile packages. This strategy would reduce the vast number of possible approaches while refining and filtering the essential tools for users. Thorough graphical representations and analyses, this study offers an overview of the state-of-the-art of the R ecosystem, highlighting the main development lines and their demands. The central objective of this research is to provide information that

supports strategic decision-making for the development of applications and innovations in the field of statistical analysis, fostering knowledge and the continuous improvement of tools.

Keywords: applied statistics, R, data science, computational analysis, statistical packages.

Introdução

Com o crescimento exponencial da ciência de dados e da estatística computacional nos últimos anos, a linguagem R tem sido adotada amplamente, se estabelecendo como uma ferramenta indispensável para *data mining*, análise, modelagem e visualização de dados. O grande ecossistema de pacotes hoje disponível reflete essa evolução, fornecendo aos seus usuários uma gama de soluções, abrangendo técnicas básicas de manipulação de dados, até métodos avançados para inovações em diversas áreas do conhecimento, como *machine learning*, bioestatística, finanças, linguagem natural e previsão de séries temporais. Esse desenvolvimento contínuo é fomentado por uma comunidade engajada de desenvolvedores, pesquisadores e profissionais, cuja colaboração tem resultado direto na otimização e criação de novas funcionalidades, embasando estudos de ponta na academia e no mercado. Este ambiente de colaboração é abordado por Alcoforado et al. (2024), reiterando que encontros da comunidade como o SER, refletem a busca constante por inovação, embasados pelo poder da ferramenta R, capacitando profissionais das mais diversas áreas que encontra na estatística e ciência de dados, embasamento técnico para suas pesquisas e inovações. A colaboração interdisciplinar tem o poder de impulsionar pesquisas, trazendo eficiência e *insights*, trazendo oportunidades para profissionais e pesquisadores resolverem problemas complexos e realizar descobertas valiosas.

Devore (2005) discursa sobre a importância dos métodos e conceitos estatísticos não só pelo prisma da utilidade, mas como fundamento para a compreensão do mundo e de seus fenômenos. Como cânone nas referências bibliográficas, a obtenção de *insights* valiosos sobre padrões e comportamentos observados em diversas áreas, é a cerne do método baseado em estatística, em qualquer área como engenharia e ciência, desempenhando um papel essencial na tomada de decisões informadas, permitindo que indivíduos façam escolhas racionais mesmo diante de incertezas e variações.

Para a diversidade de aplicações, o campo da estatística entra como um articulador comum nas fases de pesquisa, como coleta, apresentação, análise e uso de dados para tomar decisões, resolver problemas e projetar produtos e processos, como elucidado por Montgomery e Runger (2003). Como exemplo, estes autores citam que muitos aspectos da prática da engenharia envolvem o trabalho com dados, evidenciando que o conhecimento de estatística está sim profundamente atrelado com as técnicas estatísticas, fornecendo ferramental robusto no projeto de novos produtos e sistemas, orientando inovações e melhorias no desenvolvimento em todos os campos.

Visto este vasto cenário, compreender as tendências predominantes no desenvolvimento de pacotes para R e as demandas mais expressivas da comunidade, pode ajudar a informar decisões estratégicas no desenvolvimento da área.

O ponto de partida foi de um conjunto de dados abrangente fornecido pela CRAN, composto por 21.637 pacotes distintos. As descrições foram extraídas e processadas para permitir uma categorização baseada em técnicas de agrupamento (*clustering*). Essa estruturação possibilitou a realização de uma metanálise, identificando padrões de criação e adoção dentro da comunidade.

Para tanto, utilizamos o R e seu ambiente de desenvolvimento RStudio, combinados com Python e a inteligência artificial Julius AI. Essa integração possibilitou não apenas o processamento eficiente dos dados, mas também a aplicação de técnicas avançadas de processamento de linguagem natural para a classificação automatizada dos pacotes em grupos temáticos. O cruzamento dessas informações forneceu uma visão aprofundada sobre o direcionamento estratégico dos desenvolvedores e as áreas prioritárias para os usuários.

A análise estatística realizada revelou padrões interessantes sobre o comportamento dos desenvolvedores e usuários dentro do ecossistema R. Do ponto de vista dos desenvolvedores, identificamos uma concentração significativa de esforços na criação de pacotes voltados para mineração de texto e análise de linguagem natural, análise de sobrevivência e bioestatística, análise de redes e finanças, além de previsão e modelagem de séries temporais. Esse direcionamento pode estar associado ao avanço de áreas como ciência de dados aplicada à linguística computacional, biomedicina, economia e previsão financeira, refletindo demandas crescentes por ferramentas especializadas que facilitem análises robustas nesses domínios.

Já do lado dos usuários, o estudo do volume de *downloads* revela um panorama distinto. Observamos que há uma forte demanda por ferramentas voltadas para mineração de texto e análise de linguagem natural, seguida por pacotes relacionados à programação funcional e manipulação de dados—recursos essenciais para a fluidez do trabalho em R, especialmente por meio do Tidyverse—bem como análise de sobrevivência e bioestatística. Esses padrões demonstram que, além das necessidades especializadas que impulsionam o desenvolvimento de novos pacotes, os usuários valorizam a presença de ferramentas que otimizem sua produtividade na manipulação, organização e transformação de dados, antes mesmo de sua aplicação em análises específicas.

Por meio de gráficos detalhados e representações estatísticas, este estudo busca não apenas sintetizar o estado da arte dos pacotes desenvolvidos para R, mas também fornecer um panorama estratégico que permita compreender melhor as direções futuras do desenvolvimento dentro da comunidade. Ao mapear as tendências de adoção e criação de novos pacotes, espera-se que esta investigação contribua diretamente para a formulação de estratégias de desenvolvimento e inovação em áreas-chave, auxiliando pesquisadores

e desenvolvedores na identificação de oportunidades e na construção de soluções mais alinhadas às necessidades do público-alvo.

Dessa forma, este trabalho se propõe a oferecer uma contribuição significativa para a evolução do ecossistema R, servindo como um referencial para tomada de decisão e planejamento estratégico no desenvolvimento de novos pacotes e na melhoria contínua das funcionalidades existentes. Além de estabelecer um diagnóstico das principais linhas de trabalho dentro da comunidade, espera-se que este estudo fomente o avanço do conhecimento e promova um impacto positivo na forma como o R é utilizado para análises estatísticas e computacionais em diversas disciplinas.

Objetivos

Esta pesquisa busca realizar uma investigação estatística do ecossistema de pacotes da linguagem R, identificando suas principais áreas de utilização e padrões de adoção através de uma abordagem qualiquantitativa. Usando uma metodologia baseada em estatística, aplicando técnicas de agrupamento e síntese estatística, o estudo pretende mapear as direções predominantes no desenvolvimento e na demanda por pacotes, fornecendo *insights* estratégicos para a criação e aprimoramento de soluções que atendam às necessidades emergentes. Com isso, almejamos contribuir para o progresso da ciência de dados e da estatística computacional, promovendo um avanço mais alinhado às expectativas da comunidade acadêmica e profissional.

Material e Método

O estudo realizado é uma análise qualiquantitativa dos dados contidos em bases de dados liberadas pela CRAN, uma rede de servidores FTP e web distribuídos globalmente que armazenam versões atualizadas do R, da documentação do R e repositório de informações relacionadas. Nesta rede, são disponibilizadas quatro bases de dados. Três delas foram combinadas para este estudo: Pacotes disponíveis por data, Descrição dos pacotes, Número de *downloads* por pacote. As variáveis obtidas foram o nome, descrição da funcionalidade e número histórico de *downloads*.

Para processar estes dados foram usadas duas plataformas multilinguagem, para poder conectar funcionalidades. A primeira foi o Rstudio, escolhida para o recolhimento de dados nas bases, formação de um repositório principal e posterior metanálise. A escolha desta se deve à possibilidade de trabalho eficiente entre as linguagens R e Python, além de permitir arquivos QUARTO, multilinguagem, que combinam comandos das duas linguagens numa mesma cadeia de código. A outra plataforma escolhida para os procedimentos de processamento de linguagem natural foi a Julius AI, pela capacidade de automatizar a clusterização dos pacotes por suas respectivas descrições, avaliando de forma eficiente e rápida a linguagem natural contida nelas, um dos maiores desafios deste estudo, que envolveu mais de 21 mil pacotes únicos e suas informações relacionadas. Uma vez que esta plataforma também trabalha sobre as linguagens R e Python, se mostrou um elemento chave para o processamento da grande quantidade de dados e trazê-los de volta

ao ambiente do Rstudio, com resultados capazes de ser exportados para a plataforma Rstudio.

O método escolhido para o processamento de informação foi o *Personal Knowledge Mastery* (PKM) com o intuito de transformar dados em *insights*, seguindo os passos de busca, avaliação, organização, colaboração, análise e apresentação dos dados adquiridos, embasado pelo *frameworks* de Harold Jarche.

Como primeiro passo, utilizando o R, por meio do Rstudio, os dados foram obtidos nos repositórios CRAN e agrupados através dos atributos comuns, transformando as três bases de dados em um único repositório. Um ranking foi criado a partir do número histórico de *downloads*. Foi também elaborada uma conferência dos dados no Exel, para garantir que as bases de dados fossem conectadas corretamente, anexando novas informações aos campos existentes de forma única e verificada.

A plataforma Julius AI foi utilizada devido as avançadas capacidades de tratamento de linguagem natural e categorização e por utilizar as linguagens R e Python, oferecendo um elo de conexão com o Rstudio. A ferramenta então foi utilizada para acessar as descrições dos pacotes, que usam a linguagem natural e levantar quais seriam os *clusters* capazes de melhor agrupá-los. Foi utilizada também uma conferência humana, feita pelos pesquisadores, para levantar *outliers*, descrições em branco que pudessem interferir na análise e outras anomalias que pudessem influenciar o estudo. Seguidamente, os pacotes foram categorizados também dentro da plataforma Rstudio, utilizando o Python, orientados pelos *clusters* da análise feita na plataforma Julius. Com os dados tratados e categorizados, foi utilizado então arquivos de código QUARTO, por sua capacidade de trabalho entre linguagens, para gerar os gráficos, fundamentais para as análises conduzidas neste estudo.

Resultados e Discussão

A primeira análise conduzida foi o levantamento de pacotes desenvolvidos em grandes área de aplicação e, nos casos aplicáveis, o pacote foi contabilizado em mais de uma área, atendendo a condição de suas capacidades serem úteis diretamente em outras áreas, visando o entendimento das grandes concentrações de uso, como demonstrado na figura 1, abaixo:

A categorização e ordenação dos pacotes por suas descrições demonstrou que estes se dividem em 12 grandes áreas de aplicação, sendo a primeira, de forma esperada, a estatística, seguido por visualização de dados, bioestatística, análises espaciais, aprendizado de máquina, otimização, análise de redes, séries temporais, mineração de texto, ciência experimental, finanças e por último psicologia. Porém, neste gráfico fica perceptível que a aglomeração pode ser refinada, usando a clusterização via linguagem natural, possibilitada pela plataforma Julius AI e pela conferência humana dos dados. Obtivemos então a seguinte *clusterização* (tabela 1):

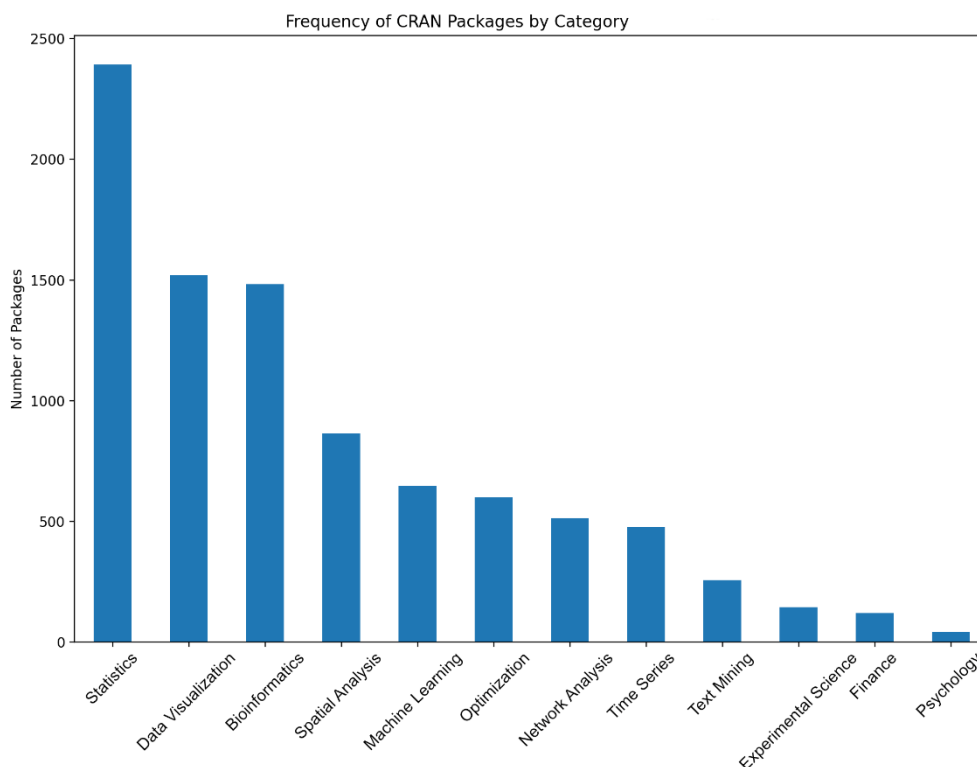


Figura 1: Frequência de pacotes por categoria – multivalência de aplicação
 Fonte: AUTORES, 2025

Tabela 1: Clusterização: Descrição e principais pacotes

Cluster	Descrição
00	Previsão e modelos de séries temporais, com pacotes como <code>forecast</code> e <code>fracdiff</code> .
01	Inferência estatística e testes, apresentando pacotes como <code>multcomp</code> e <code>coin</code> .
02	Amplia funcionalidades R e Tidyverse, com <code>rlang</code> e <code>lifecycle</code> .
03	Modelagem e métricas, como <code>Formula</code> e <code>ModelMetrics</code> .
04	Manipulação e visualização de dados, com <code>dplyr</code> e <code>ggplot2</code> .
05	Aprendizado de máquina e mineração de dados, como <code>caret</code> e <code>randomForest</code> .
06	Bioinformática e genômica, com pacotes como <code>BiocGenerics</code> e <code>GenomicRanges</code> .
07	Análise de dados espaciais, apresentando <code>sp</code> e <code>raster</code> .
08	Econometria e análise financeira, como <code>zoo</code> e <code>quantmod</code> .
09	Mineração de texto e PLN, com <code>tm</code> e <code>text2vec</code> .
10	Análise de redes e teoria dos grafos, apresentando <code>igraph</code> e <code>network</code> .
11	Estatísticas bayesianas e modelagem, com <code>rstan</code> e <code>brms</code> .

Cluster	Descrição
12	Análise de séries temporais e previsão, apresentando <code>tseries</code> e <code>forecast</code> .
13	Análise de sobrevivência e bioestatística, como <code>survival</code> e <code>survminer</code> .
14	Análise multivariada e redução de dimensionalidade, com <code>FactoMineR</code> e <code>PCAtools</code> .
15	Visualização de dados e gráficos, apresentando <code>ggplot2</code> e <code>lattice</code> .
16	Programação funcional e manipulação de dados, com <code>purrr</code> e <code>magrittr</code> .
17	Modelos aditivos generalizados e clustering, apresentando <code>mgcv</code> e <code>mclust</code> .
18	Estatísticas multivariadas e distribuições, como <code>mvtnorm</code> e <code>mnormt</code> .
19	Análise de redes e finanças computacionais, com <code>igraph</code> e <code>tseries</code> .

Fonte: AUTORES, 2025

Com esta clusterização em mãos pudemos analisar no Rstudio a distribuição de pacotes por cluster e a distribuição de *downloads* históricos por cluster, mostradas nas figuras 2 e 3, abaixo:

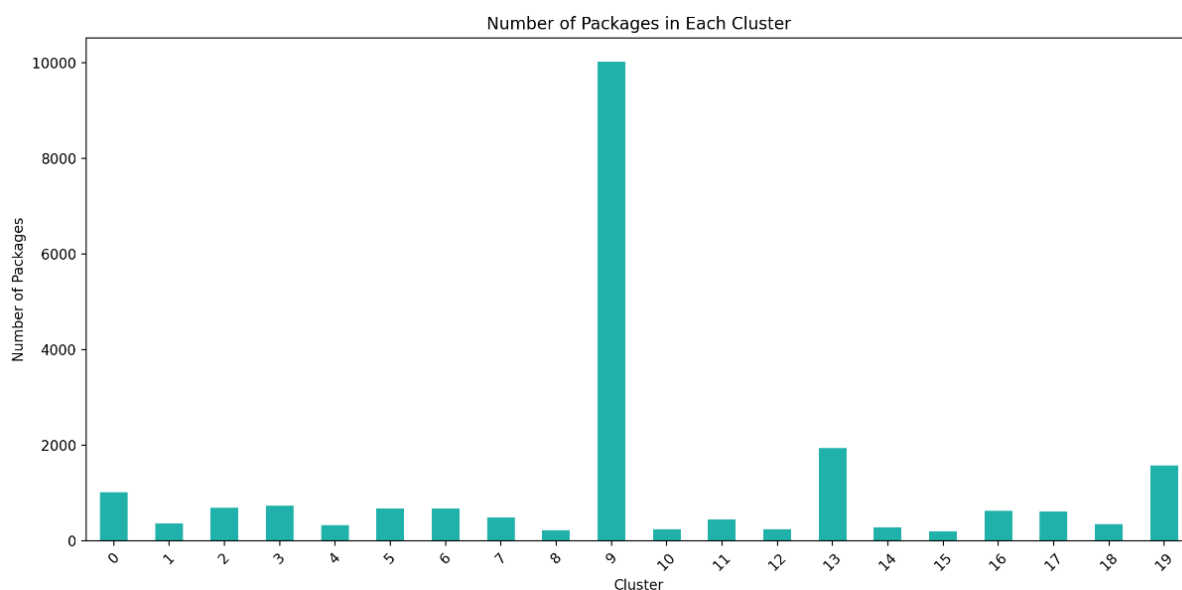


Figura 2: Número de pacotes em cada *cluster*

Fonte: AUTORES, 2025

Outra análise feita foi a de distribuição de downloads dos pacotes, dentro dos *clusters*, que pode ser visto na figura abaixo, na figura 4:

O ranking de pacotes por histórico de demanda da comunidade foi criado e se apresenta na Tabela 2, que segue abaixo:

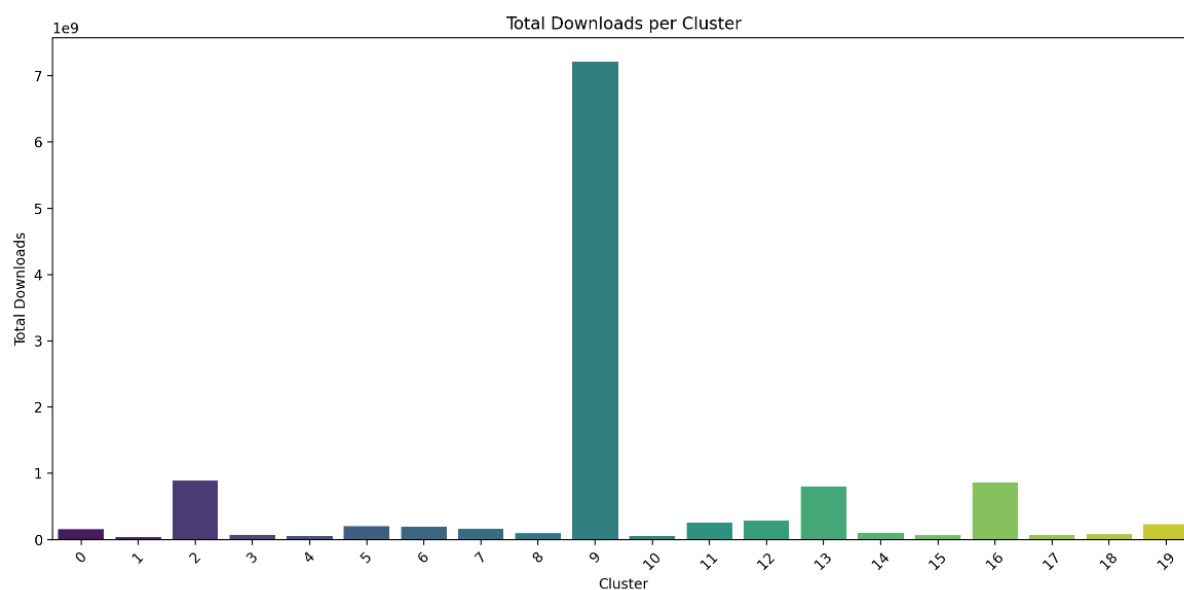


Figura 3: Número de *downloads* em cada *cluster*
Fonte: AUTORES, 2025

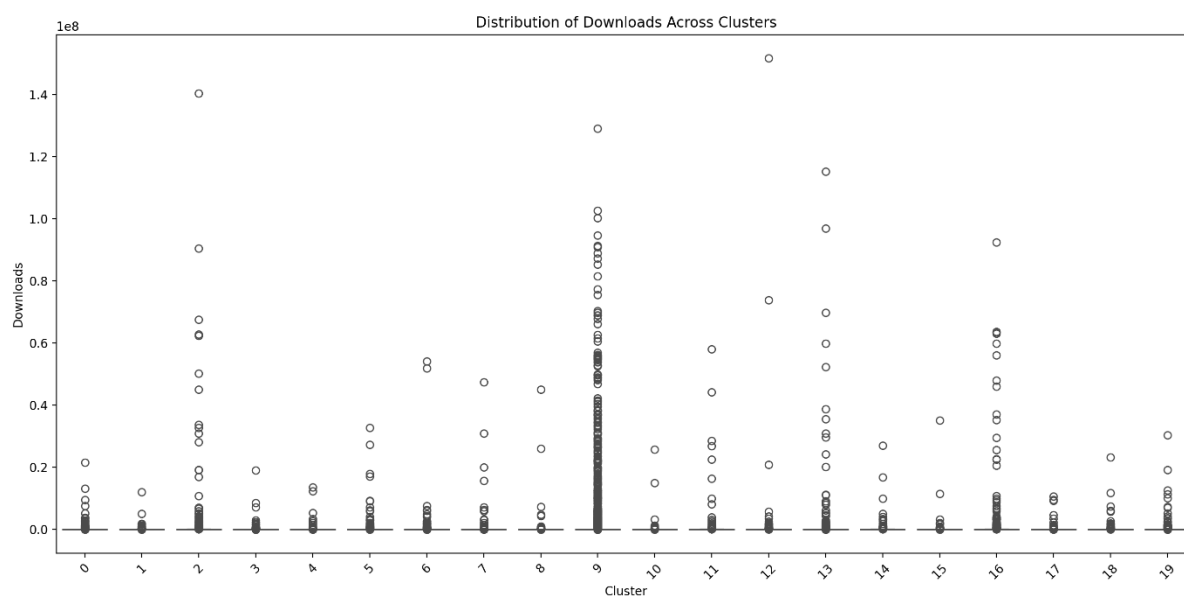


Figura 4: *Boxplot* de distribuição de downloads entre *clusters*
Fonte: AUTORES, 2025

Tabela 2: Ranking de demanda dos pacotes

Posição	Pacote
01	ggplot2
02	rlang
03	magrittr
04	dplyr
05	vctrs
06	cli
07	tibble
08	jsonlite
09	devtools
10	Rcpp
...	...
Fonte: AUTORES, 2025	

Com este material desenvolvido, se avançou para a metanálise.

Conclusão

Como primeira conclusão, provinda da categorização de aplicações gerais dos pacotes e de verificação de multivalência de pacotes para mais de uma área, representada pela Figura 1, pudemos constatar 12 aplicações principais, sendo as três mais proeminentes a estatística, seguido de visualização de dados e bioinformática. Já a verificação de multivalência de aplicações de um mesmo aplicativo mostrou a grande flexibilidade que desenvolvedores podem proporcionar nos pacotes, a qual retornou 1352 pacotes capazes de aplicação em mais de uma categoria. Observa-se que são números em constante mudança. O levantamento feito em 15 de janeiro de 2025 apontou 21.637 pacotes no repositório da CRAN e, dentro destes, apenas 6,25% atenderam à verificação de multivalência. Com tão alto número de pacotes individuais disponíveis aos usuários, ferramentas desenvolvidas para atender um público maior podem ser um grande caminho de aprimoramento no ecossistema.

O uso da plataforma Julius AI para clusterização usando linguagem natural (Tabela 1), permitiu uma resolução melhor dos dados de aplicabilidade e apresentação dos pacotes que se destacam em demanda para cada um deles. A grande capacidade de lidar com uma vasta quantidade de dados em formato textual, em linguagem natural, de forma eficaz, demonstra como esta pode ser uma ferramenta potente para pesquisadores e profissionais conduzirem estudos, melhorias e inovações nas mais diversas áreas do conhecimento.

Ao trazer os dados classificados novamente para a plataforma Rstudio, foi elaborada uma análise de distribuição de pacotes por cluster e de distribuição de *downloads* para cada um deles (Figuras 2 e 3). A primeira mostra que a maior parte dos pacotes desenvolvidos

até o momento focam em funcionalidades de mineração de texto e processamento de linguagem natural, seguidos por análise de sobrevivência e bioestatística, análise de redes e finanças computacionais e previsão e modelos de séries temporais. Já a segunda, sob a perspectiva da demanda dos usuários, mostra que a maior parte da demanda se concentra também em funcionalidades de mineração de texto e processamento de linguagem natural, seguidos por programação funcional e manipulação de dados, ampliação de funcionalidades R e Tidyverse e análise de sobrevivência e bioestatística. Ao comparar os gráficos, pode-se notar convergências e divergências em relação a disponibilidade e demanda, investigado em futuros desdobramentos deste estudo.

Ao executar um *boxplot* da distribuição de *downloads* em cada cluster (Figura 4), apresenta-se claramente que a massa da demanda se concentra nos *clusters* 9, 16, 13 e 2, ao mesmo tempo que revela os pacotes mais demandados se distribuem nos *clusters* 12, 2, 9, 13 e 16. Mais uma vez, estas diferenças são oportunidades de caminhos de pesquisa e aprimoramento em novos estudos.

Como exercício de novos desdobramentos foi desenvolvido um *ranking* geral, sem categorização, mostrando as principais ferramentas utilizadas por toda a comunidade, capaz de orientar novos usuários, ou usuários enfrentando novos desafios em suas pesquisas, quais as ferramentas usuais associadas ao uso da linguagem R. Futuramente pretende-se cruzar o ranking com a classificação em *clusters*, o que possibilitará uma escolha ainda mais embasada das ferramentas em pesquisas, além de outras informações, como o repositório de *downloads* nos últimos 100 dias, informando tendências de uso na comunidade.

Sobre as dificuldades e limitações do estudo se destaca que a CRAN disponibiliza poucos dados sobre os pacotes, fora os utilizados neste estudo, somente apresentando mais uma base de dados com a quantidade de downloads dos pacotes nos últimos 100 dias. Foi deliberado entre os autores que é um recorte curto para uma macro análise, voltado principalmente para papel do R no ambiente de pesquisas científicas, mas com certeza é uma ótima oportunidade para um viés de estudo avaliando tendências de curto prazo dentro do ecossistema.

Em resumo, o estudo foi direcionado a obter uma visão ampla sobre o ecossistema R, suas principais aplicações e tendências, proporcionar dados para embasar decisões estratégicas no desenvolvimento de novos pacotes, na escolha de ferramentas pelos pesquisadores e enxergar lacunas e oportunidades de melhorias para toda a comunidade.

Referências

ALCOFORADO, Luciane Ferreira; SANTOS, João Paulo Martins Dos; LEVY, Ariel; LONGO, Orlando Celso; LINARES, JUAN. **Aplicações em R: Encurtando Distâncias nas Ciências**. Brasil: FZEA, 2024.

DEVORE, Jay L. **Probabilidade e Estatística**: para Engenharia e Ciências. Tra-

dução da sexta edição. Brasil: Cengage Learning Edições Ltda., 2005

DIZON, Stephany; MAMALIAS, Arianne; LINCOPINIS, Darllaine. **A Review on R Programming Language**. Filipinas: ResearchGate, 2023.

KUBRUSLY, Jessica. **Uma introdução à programação com o R**. Brasil: Eduff, 2022.

MONTGOMERY, Douglas C.; RUNGER, George C. **Applied Statistics and Probability for Engineers**. Terceira edição. Estados Unidos: John Wiley & Sons, Inc., 2003

Citação de pacotes R:

`citation("dplyr")`

`citation("ggplot2")`

`citation("cranlogs")`

`citation("rvest")`

`citation("httr")`

Citação de pacotes Python:

`citation("panda")`

`citation("numpy")`

`citation("sklearn")`

`citation("defaultdict")`

`citation("matplotlib")`

`citation("seaborn")`

Anexo

Link para o repositório de códigos: [acesso ao repositório](#)