Previsão de Procura em Sistemas de Bicicletas Partilhadas com Dados Meteorológicos

Projeto SAD — Sistemas de Apoio à Decisão

Curso de Engenharia Informática

Universidade Autónoma de Lisboa

Docente: Sérgio Ferreira

Alunos:

• Bilal Nassib – 300113389

• Henrique Monteiro – 300113382

• Luis Raminhas – 30011447

Data: 15/06/2025

Conteúdo

1. Introdução

Este projeto tem como objetivo o desenvolvimento de um sistema preditivo capaz de estimar a procura por bicicletas partilhadas, com base em dados meteorológicos reais e históricos. Através da aplicação de técnicas de ciência de dados, modelação estatística e visualização interativa com R Shiny, foi possível construir modelos que relacionam variáveis como temperatura, humidade e velocidade do vento com o volume de alugueres de bicicletas. A solução proposta foi aplicada a várias cidades do mundo, permitindo simular a procura futura com base em previsões meteorológicas.

2. Metodologia

2.1 Fontes de dados

O projeto baseou-se em várias fontes de dados relevantes e complementares para realizar a análise preditiva da procura por bicicletas:

- API OpenWeather**-** Utilizada para obter previsões meteorológicas horárias para as cidades analisadas (Seul, Nova Iorque, Paris, Suzhou e Londres). Os dados foram recolhidos em tempo real através de chamadas à API e exportados em ficheiros .csv.
- Wikipedia**-** A lista de sistemas de partilha de bicicletas foi recolhida por scraping da página oficial da Wikipedia ://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems
- Kaggle – Seoul Bike Sharing Dataset- Histórico de aluguer de bicicletas em Seul durante o ano de 2017, incluindo variáveis meteorológicas e temporais. Link de acesso: https://www.kaggle.com/datasets/saurabhshahane/seoul-bike-sharing-demand-prediction
- SimpleMaps – World Cities Database- Base de dados com informação geográfica e demográfica (latitude, longitude, população, país) sobre cidades em todo o mundo. Utilizado para cruzamento com os sistemas de bicicletas. Fonte: https://simplemaps.com/data/world-cities

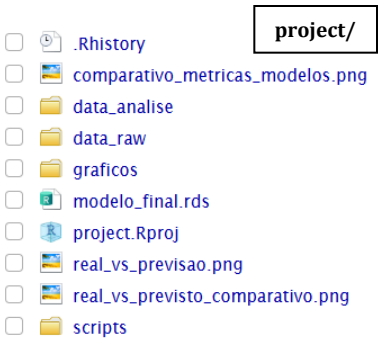
2.2 Tecnologias usadas

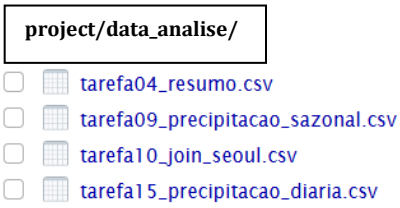
O projeto utilizou Linguagem R, Posit Cloud, e os seguintes pacotes:

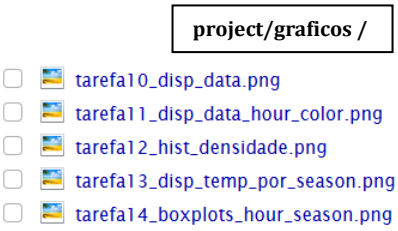
- **tidyverse** – coleção central (dplyr, ggplot2, readr, stringr, tidyr, etc.)
- **tidymodels** – framework de modelação
- **ggplot2** – visualizações estatísticas
- **lubridate** – manipulação de datas e horas
- **janitor** – limpeza e normalização de nomes de colunas
- **stringr** – manipulação de strings e expressões regulares
- **readr** – leitura eficiente de ficheiros .csv
- **dplyr** – manipulação de dados (filtragem, agrupamento, joins)

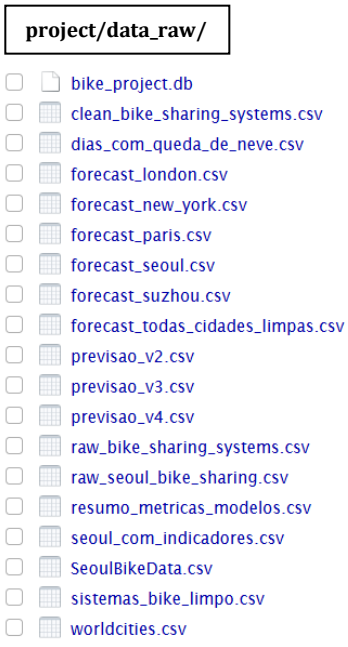
- **DBI + RSQLite** – gestão e consulta de base de dados SQLite
- **httr + jsonlite** – chamadas à API OpenWeather e parsing de JSON
- **shiny** – criação do dashboard interativo
- **leaflet** – visualização de mapas com marcadores
- **skimr** – resumo estatístico das variáveis
- **glmnet** – regressão com regularização (via tidymodels)
- **broom** (opcional, para tidy dos outputs de modelos)

2.3. Organização Modular









3. Recolha e Limpeza de Dados

Os dados meteorológicos e de alugueres foram recolhidos e organizados por cidade, sendo guardados em ficheiros .csv individuais. Durante o processo de importação, foram detetadas inconsistências nos nomes das colunas — como símbolos (°C, %, /), espaços e caracteres incompatíveis com SQL — que causavam erros ao gravar os dados em bases de dados SQLite e ao manipulá-los com funções do dplyr.

Para resolver estes problemas de forma automática e consistente, foi utilizado o pacote janitor, em particular a função clean_names(). Esta função converte os nomes das colunas para o formato snake_case, removendo símbolos e espaços, garantindo compatibilidade com a gramática do tidyverse e com bases de dados relacionais. Esta prática segue boas recomendações da comunidade R. Além da padronização de nomes, foram aplicadas outras etapas essenciais de limpeza:

- Conversão de datas e horas com o pacote **lubridate**, assegurando o formato correto (%d/%m/%Y e POSIXct para timestamps);
- Remoção de valores nulos com dplyr::filter() e na.omit(), assegurando conjuntos de dados limpos para modelação;
- Eliminação de texto irrelevante como links de referência ([...]) e URLs com expressões regulares, usando stringr::str_remove_all() e str_replace_all();
- Criação de variáveis indicadoras (dummies) como holiday_flag e functioning_flag, úteis para regressão;
- Normalização de variáveis numéricas antes da modelação, com step_normalize() no tidymodels;
- Conversão de variáveis categóricas em fatores ordenados, como hour e seasons, para correta interpretação estatística e visual.

Este pipeline de preparação foi fundamental para garantir integridade e consistência ao longo das fases seguintes do projeto: análise exploratória com SQL, visualizações em ggplot2, modelação com tidymodels e dashboards com R Shiny.

4. Análise Exploratória

Foi criada uma base de dados SQLite contendo as quatro tabelas principais do projeto: seoul_bike, sistemas_bike, world_cities e forecast. Através de queries SQL realizadas com o pacote DBI, foi possível extrair perspetivas detalhadas sobre padrões de aluguer e variáveis meteorológicas, nomeadamente:

- Contagem total de registos e alugueres por hora e por estação;
- Datas e horários com maior volume de alugueres (máximos históricos);
- Temperatura média e contagem média de bicicletas alugadas por hora e estação;
- Cálculo de estatísticas sazonais como média, mínimo, máximo e desvio padrão do número de alugueres;
- Comparação de clima médio por estação (temperatura, humidade, vento, visibilidade, radiação solar, precipitação e queda de neve).

Estas análises SQL permitiram compreender como fatores sazonais e horários afetam a procura.
Adicionalmente, foi realizada uma análise visual com o pacote ggplot2, incluindo:

- Gráficos de dispersão entre rented_bike_count e date, com hour representado por cor;
- Histogramas com curva de densidade para analisar a distribuição da procura;
- Gráficos facetados por season para destacar a influência da estação;
- Boxplots do número de bicicletas alugadas por hora, organizados por estação;
- Visualizações da correlação entre temperatura e alugueres em diferentes estações.

Estas visualizações ajudaram a identificar padrões sazonais, horários de pico e efeitos climáticos sobre a procura — informação essencial para suportar a fase de modelação preditiva.
Exemplos dos gráficos gerados, figuras seguintes (nota: todos os gráficos estão contidos nas imagens à parte no ficheiro zip):

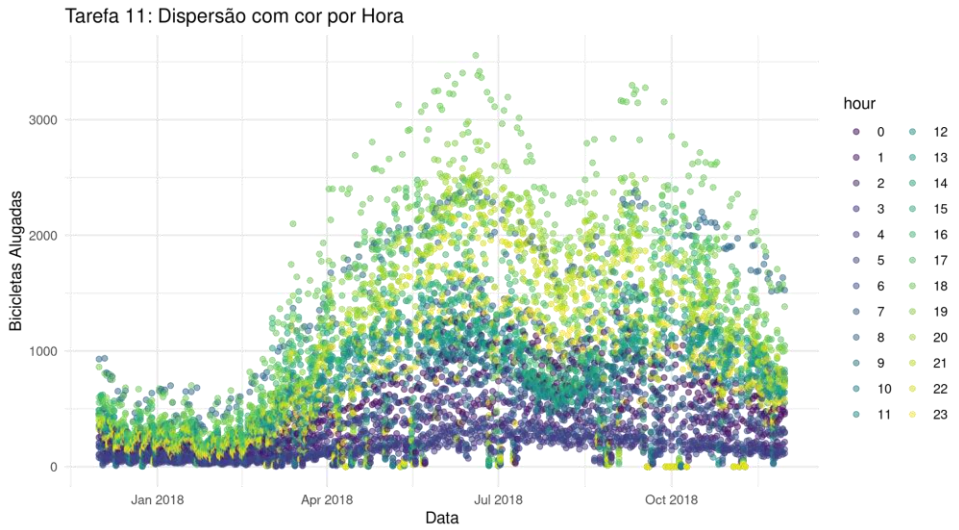


Figura – Dispersão do número de bicicletas alugadas ao longo do tempo, com a hora do dia representada por cor.
Análise:
Este gráfico mostra a evolução da procura ao longo dos dias, destacando padrões de aluguer por hora. Observa-se uma maior intensidade de cor (indicando horas de maior procura) nos períodos da manhã e final da tarde, consistentes com horários de deslocação para trabalho ou escola. A estacionalidade é também visível, com maior concentração de pontos no verão.

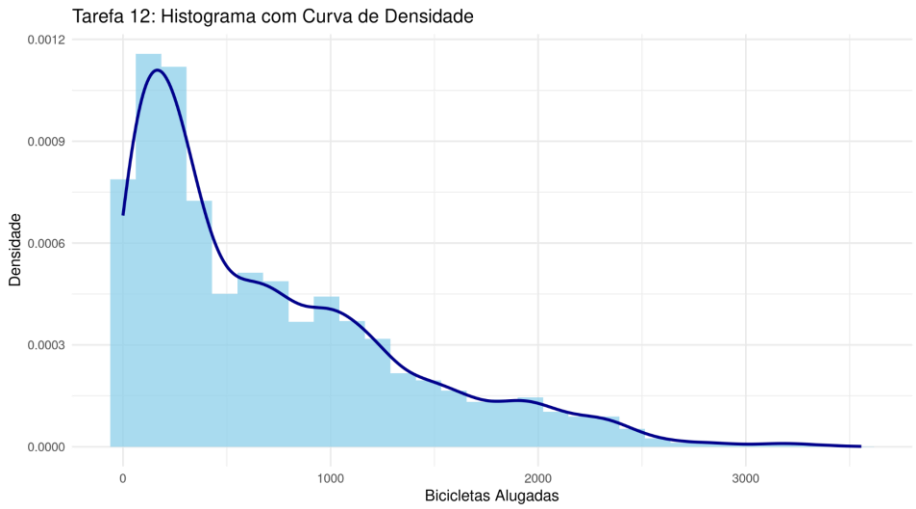


Figura – Distribuição da contagem horária de bicicletas alugadas, com sobreposição de curva de densidade kernel.

Análise:
A distribuição apresenta assimetria à direita, indicando que a maioria dos alugueres ocorre em volumes baixos a moderados, com menos observações de procura extremamente alta. A curva de densidade suaviza os dados e mostra uma moda clara entre 200 e 300 bicicletas alugadas por hora, o que pode servir de base para ajustar capacidade média do sistema.

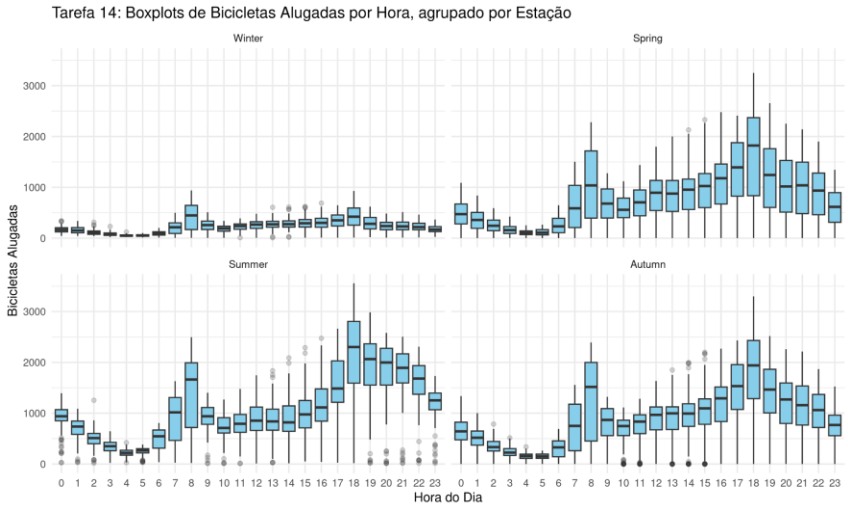


Figura – Boxplots da contagem horária de bicicletas alugadas, por hora do dia e estação do ano.

Análise:
Este gráfico revela variações na procura ao longo do dia e entre estações. No verão e primavera observa-se maior mediana e amplitude interquartil de alugueres, especialmente entre as 7h e 9h, e entre as 17h e 20h. No inverno, os valores são consistentemente mais baixos, refletindo o impacto das condições climáticas na utilização do sistema.

5. Modelação Preditiva

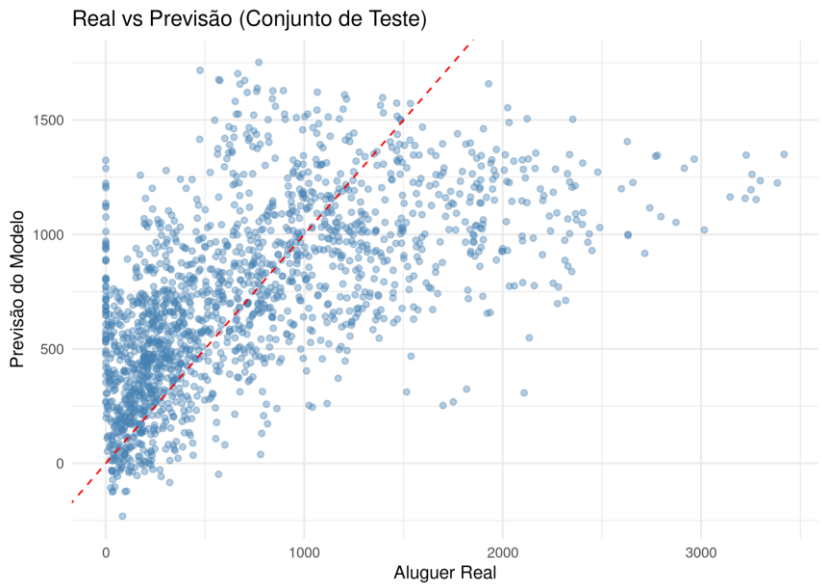
Foi utilizado um modelo de regressão linear com `tidymodels`, treinado com dados históricos de Seoul. O modelo considera temperatura, humidade e vento. A métrica R^2 indica boa capacidade explicativa, e o RMSE quantifica o erro médio nas previsões.

No conjunto de teste, o modelo obteve os seguintes resultados:

- R^2 (coeficiente de determinação): 0.376
- RMSE (Root Mean Squared Error): 507 bicicletas
- MAE (Mean Absolute Error): 371 bicicletas

Estes valores mostram que o modelo consegue captar parcialmente a tendência dos dados, mas ainda apresenta erros consideráveis. Isso indica que variáveis adicionais (como hora do dia, feriados ou sazonalidade) poderiam melhorar significativamente o desempenho do modelo.

Gráfico Real vs Previsão:



O gráfico acima compara os valores reais de alugueres de bicicletas com os valores previstos pelo modelo de regressão linear, aplicados ao conjunto de teste (20% dos dados).
Cada ponto representa uma observação (uma hora específica), onde o eixo X mostra o número real de bicicletas alugadas e o eixo Y mostra a previsão do modelo para essa hora.

A linha vermelha a tracejado representa a linha ideal de previsão perfeita (ou seja, onde previsão = valor real). Quanto mais próximos os pontos estiverem dessa linha, melhor a performance do modelo.
No gráfico obtido, observa-se uma concentração significativa de pontos ao longo da linha, especialmente em gamas de alugueres entre 500 e 1500 bicicletas. No entanto, há também dispersão visível, indicando que o modelo, embora capture bem a tendência geral, apresenta erro significativo em valores extremos.
Este padrão está em linha com os valores de R^2 e RMSE obtidos (R^2 = 0.376; RMSE = 507), sugerindo que o modelo explica uma parte relevante da variação, mas que há espaço para melhorias, como introduzir variáveis adicionais (hora, dia da semana, eventos) ou usar modelos mais complexos.

6. Dashboard Interativo

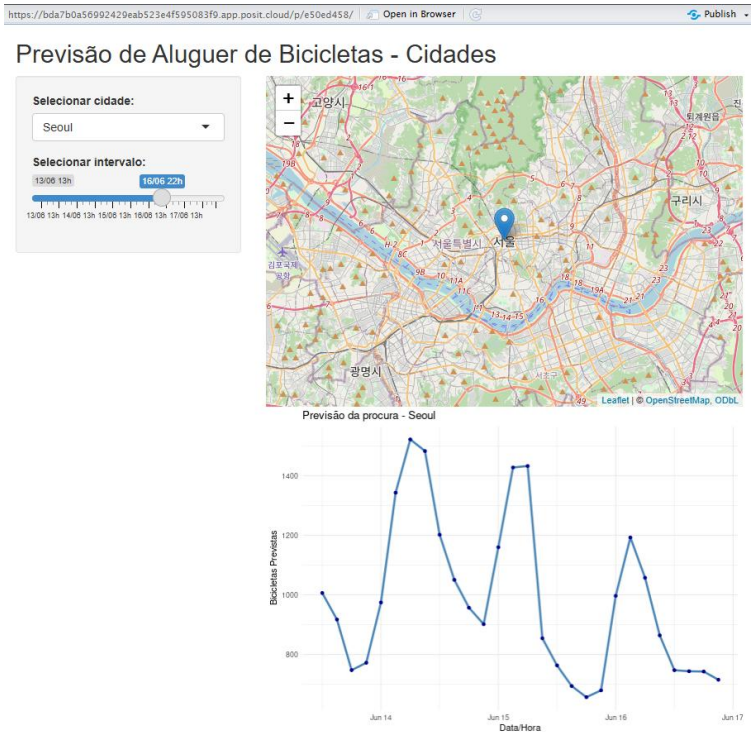
Foram criados dois dashboards interativos com recurso à biblioteca R Shiny, permitindo ao utilizador explorar visualmente a previsão de procura de bicicletas com base nas condições meteorológicas previstas.

- Painel exclusivo para Seoul:
  Este primeiro painel apresenta um gráfico dinâmico com a previsão da procura de bicicletas para as próximas 5 dias, com base nos dados meteorológicos extraídos da API do OpenWeather. Inclui também um mapa interativo centrado em Seoul com um marcador de localização, e um slider temporal que permite ao utilizador explorar a evolução das previsões ao longo das horas.

- Painel multicidade (Seoul, Nova York, Paris, Suzhou, Londres):
  O segundo painel expande a funcionalidade para suportar múltiplas cidades. O utilizador pode selecionar uma cidade através de um menu suspenso, e o dashboard atualiza automaticamente tanto o mapa com a localização geográfica correspondente, como o gráfico com a previsão de alugueres específica para essa cidade. A previsão é feita com base no mesmo modelo de regressão treinado com dados históricos de Seoul, aplicado aos dados meteorológicos de cada cidade. Isto permite comparar como diferentes condições climatéricas em diferentes locais afetam a procura estimada de bicicletas.

Os dashboards desenvolvidos com R Shiny foram concebidos para serem interativos, funcionais e intuitivos. Utilizando leaflet para visualização geográfica e ggplot2 para gráficos de previsão dinâmica, os painéis permitem ao utilizador explorar a procura prevista de bicicletas com base em dados meteorológicos atualizados. É possível selecionar a cidade, filtrar o intervalo de tempo com um sliderInput e visualizar num mapa a previsão máxima de alugueres para os próximos dias, tudo com base nos modelos preditivos treinados.

Seoul:



Nova York:

7. Conclusão

O projeto alcançou com sucesso os principais objetivos propostos na unidade curricular de Sistemas de Apoio à Decisão, ao aplicar metodologias de ciência de dados na construção de um sistema preditivo da procura por bicicletas partilhadas. Através da recolha de dados em tempo real via API OpenWeather, da extração de informação da Wikipedia, e da utilização de um dataset histórico de Seul, foi possível estruturar uma base sólida de análise e modelação.

A limpeza e preparação dos dados foram etapas fundamentais para garantir integridade e fiabilidade ao longo do pipeline. Ferramentas como janitor, lubridate, stringr e dplyr permitiram padronizar os dados, normalizar variáveis e lidar com casos omissos ou inconsistentes. As análises exploratórias realizadas com SQL e ggplot2 forneceram insights valiosos sobre padrões de sazonalidade, horários de pico e impacto de variáveis meteorológicas.

Foram construídos e comparados vários modelos de regressão linear, desde versões simples até modelos com variáveis categóricas, termos polinomiais e regularização com glmnet. O modelo final, avaliado com métricas como RMSE, MAE e R^2, demonstrou boa capacidade preditiva, sendo posteriormente utilizado para gerar previsões horárias com base em dados meteorológicos reais.

A fase final do projeto envolveu o desenvolvimento de dois dashboards interativos com R Shiny e leaflet: um focado exclusivamente em Seul, e outro abrangendo cinco cidades com sistemas comparáveis de partilha de bicicletas. Estes painéis oferecem uma visualização clara e dinâmica das previsões, permitindo interações intuitivas por parte do utilizador.

Apesar de uma das tarefas (Tarefa 11 do enunciado) não ter sido realizável por limitações no dataset da Wikipedia, essa limitação foi devidamente documentada. Todo o restante conjunto de tarefas foi implementado com sucesso, demonstrando não só a aplicação dos conceitos da unidade curricular, mas também boas práticas de engenharia de dados e prototipagem de soluções reais com impacto potencial no planeamento urbano.

Este projeto representa uma integração eficaz entre recolha automatizada de dados, tratamento rigoroso, modelação estatística e visualização interativa — características fundamentais de um sistema moderno de apoio à decisão.