Análise do Impacto das Condições Meteorológicas em Serviços Urbanos de Partilha de Bicicletas com Apoio de Modelos Preditivos em R

3ºano LEI - Sistemas de Apoio à Decisão

Hugo Pedro	Guilherme Carvalho	Mário Félix
30010791	30010987	30011520

1. Introdução

Nos dias de hoje, é possível reparar que a utilização de bicicletas tem vindo a aumentar especialmente quando de zonas europeias como os Países Baixos, Alemanha, suécia, entre outros [1]. A verdade é que condições do clima influenciam diretamente a disposição que as têm pessoas de sair à rua e andar de bicicleta, seja por prazer ou simplesmente como meio de transporte [2].

O objetivo deste projeto é analisar como variáveis temporais e meteorológicas podem afetar a procura de serviços de partilha de bicicletas.

Esta análise será feita a partir de métodos de análise de dados e modelação preditiva com suporte de técnicas de Inteligência artificial.

2. Objetivos do Sistema

Uma vez que o sistema esteja funcional, este poderá oferecer informações relevantes que fará possível a redistribuição de bicicletas para zonas com maior probabilidade de utilização, reduzindo o número de bicicletas paradas por pouca procura, minimizando perdas por falta de bicicletas disponíveis em horários de pico e também permitindo avaliar o impacto de mudanças climáticas em negócio a longo prazo [3].

3. Visão Geral

O programa analisa a influência do tempo e clima na procura por partilha de bicicletas em áreas populosas e conhecidas. Faz isto ao recolher dados meteorológicos de vários datasets, e integra-os com outros datasets de utilização de bicicletas, que contêm dados relevantes para uma análise complexa.

Os dados são pré-processados e limpos para garantir a qualidade e consistência para as próximas fases. São realizadas análises visuais que identificam relações e comportamentos (Exemplo: mais procura em dias de sol). Por fim são aplicados modelos preditivos de regressão linear para prever a procura com base em condições meteorológicas, que são integrados com interface gráfica para uma melhor visualização e entendimento do que se procura neste projeto.

Na figura é possível ver este processo sumarizado (Fig.1).

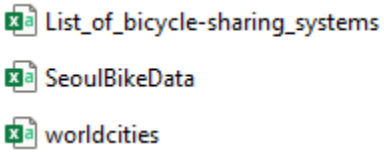
Fig. 1- Resumo Enquadrado dos Processos do Projeto

O projeto foi inteiramente realizado no RStudio [6], um IDE(Integrated Development Environment) dedicado a programação em R.

Em relação à organização de ficheiro e pastas, o projeto encontra-se por inteiro dentro da pasta "Projeto", que possui 2 executáveis de código, um para a recolha, processamento e analise dos dados e outro para a interface gráfica. Aqui também se encontra a base de dados. A suportar estes scripts, existem três pastas principais que organizam os dados: "dados_entrada", que contém os ficheiros de dados originais, "dados_processados" onde são guardadas as tabelas já limpas e transformadas e a pasta modelos, que armazena o ficheiro do modelo de regressão para as previsões futuras.

4. Dados de Entrada

Aqui estão mencionados alguns dos sites onde foram encontrados os datasets iniciais, como também o tipo de dados que incluem.



- **"World_cities" (Wikipedia):** contêm a informação sobre os nomes das cidades, latitudes e longitudes, países associados, população, entre outras [5].

- **"SeoulBikeData" (Kaggle):** têm valores de hora/dia/mês/ano, temperatura, velocidade do vento, visibilidade, radiação solar, chuva e neve, estação do ano, e feriados [4].

- **"List_of_Bicycle-sharing_systems" (Wikipedia):** país e cidade, nome do sistema, operador, data de lançamento, nº de estações e bicicletas totais, nº de utilizadores diarios, e se os sistemas estão em funcionamento [8].

5. Explicação do Código

A partir deste ponto, de maneira descritiva, são explicados diversos pontos importantes do nosso trabalho como foco na programação e como cada um dos elementos do projeto interage num todo.

O código foi maioritariamente realizado num ficheiro, "" que é executado primeiramente possuindo um formato de execução Snake-case.

5.1 Configuração Inicial e Bibliotecas

São utilizadas as bibliotecas que vão ser usadas ao longo do projeto, sendo esta: rvest, httr, jsonlite, dplyr, readr, opurrr, data.table, lubridate, stringr, fastDummies, RSQLite, DBI, ggplot2, caTools, glmnet e o caret. Muitas destas libraries são usadas para simplificar e tornar os processos de manipulação e processamento de dados mais rápidos, sendo muitas OpenSource.

Nesta fase inicial, são também definidos os diretórios onde serão armazenados os dados de entrada, os dados já processados e os modelos criados ao longo do projeto.

5.2 Padronização de Colunas e Dados associados

Para consistência dos dados, criamos funções que diretamente interagem com os dados de entrada sendo estes responsáveis por uniformizar os nomes das colunas de todos os datasets. Estas função transformam os nomes em minúsculas, substitui caracteres especiais por underscores e remove espaços em branco. Com isto temos o trabalho facilitado quando se tratar da manipulação posterior dos dados.

5.3 Processamento dos Sistemas de Bicicletas

Durante a leitura, são resolvidos eventuais problemas de codificação de caracteres e aplicada a padronização dos nomes de colunas. Quando necessário, criamos colunas novas de dados que se enquadram melhor com o conjunto de dados que estamos a mexer atualmente, tomando como por exemplo a coluna “bicycles_numerica”, criada na tabela de Seoul para estar unicamente na forma numérica. Todos os conjuntos de dados resultante é então gravado num ficheiro CSV já processado.

Todas a ações aqui referidas são igualmente aplicadas aos dados futuros.

5.4 Obtenção de Dados Meteorológicos da API OpenWeather

Para recolher previsões meteorológicas, utilizamos a API da OpenWeather. Após definir a chave da API e a lista de cidades requeridas, é feito o pedido à API, sendo a resposta processada em formato JSON e os dados organizados num tibble. Com funções de organização, os dados de todas as cidades são agregados num único conjunto, sendo as datas e horas convertidas para o formato UTC. Este conjunto de dados meteorológicos é gravado em CSV para utilização futura.

5.5 Processamento dos Dados das Cidades do Mundo

O ficheiro "worldcities.csv" é igualmente lido e padronizado. As colunas de latitude, longitude e população são convertidas para valores numéricos, sendo lat e lng renomeadas para latitude e longitude. O resultado é guardado num ficheiro CSV já normalizado.

5.6 Processamento dos Dados de Aluguer de Bicicletas de Seoul

No caso do ficheiro de aluguer de bicicletas de Seoul , usamos à função fread para uma leitura mais eficiente dos dados. Os nomes das colunas são alterados para consistência e depois padronizados com as funções já criadas. A coluna de data é convertida para o formato certo, e variáveis como estação do ano, feriado e dia de funcionamento são transformadas em variáveis dummy, que vão ser uteis para utilização nos modelos de regressão linear. O conjunto de dados completo é então exportado para um ficheiro CSV processado.

5.7 Criação e Carregamento da Base de Dados SQL

Para realizar a análise detalhada de dados, criamos uma base de dados SQL onde os vários datasets processados são carregados como tabelas. Aqui foram criadas funções de consulta SQL para garantir que estas são feitas de forma fácil e que permitem verificar antes se os dados já existem garantindo que qualquer erro por falta de dados é devidamente identificado.

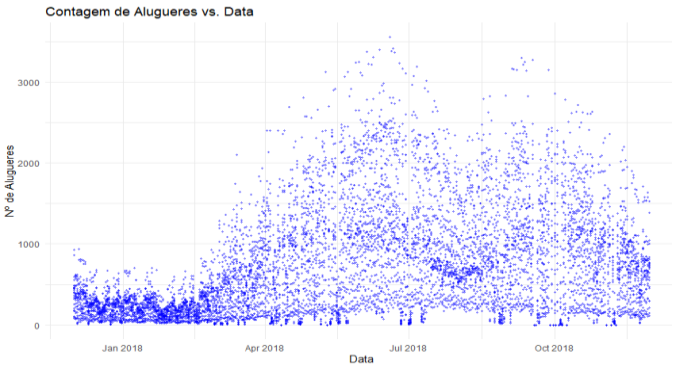
5.8 Execução das Tarefas de Análise com SQL

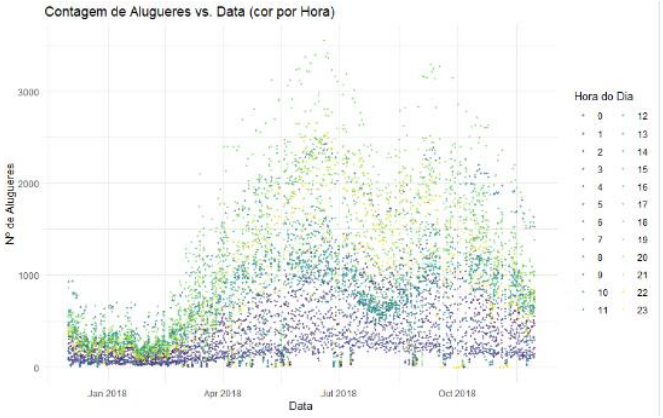
As várias tarefas analíticas previstas no projeto são então executadas através de consultas SQL, que permitem realizar buscas e filtragens sobre os dados carregados. Por motivos de cache, memoria e uma quantidade elevanda de variáveis, os resultados de cada uma destas consultas são visualizados diretamente na consola.

5.9 Análise de Dados (EDA) com Visualização

Esta fase refere-se à criação de gráficos requeridos inicialmente pelo documento do projeto e que servem para entender visualmente os dados.

Aqui utilizamos a library ggplot2 para gerar os vários gráficos:





Gráficos 1 e 2 - Gráficos de pontos para ver a tendência dos alugueres ao longo do tempo.

Os gráficos 1 e 2 mostram a contagem de alugueres de bicicletas ao longo de um ano inteiro. Analisando melhor é possível reparar que no início e no fim do ano, que correspondem ao inverno em Seul, a quantidade de alugueres é mais baixa. Na primavera e no verão, a contagem de alugueres aumenta constantemente, com um pico nos meses mais quentes. É esperado, pois as condições do clima são melhores e isso é um dos principais fatores que influenciam a decisão de andar de bicicleta.

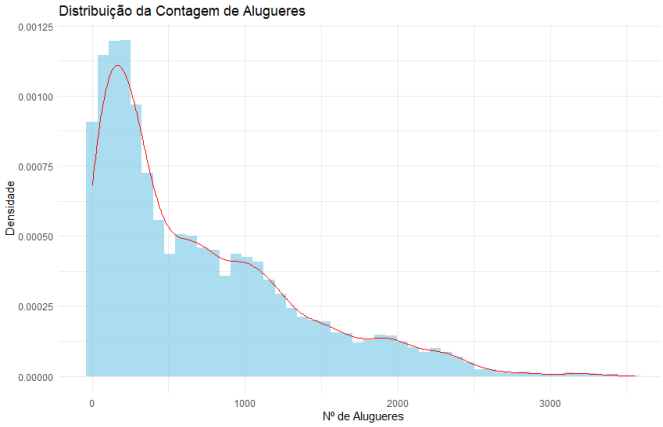
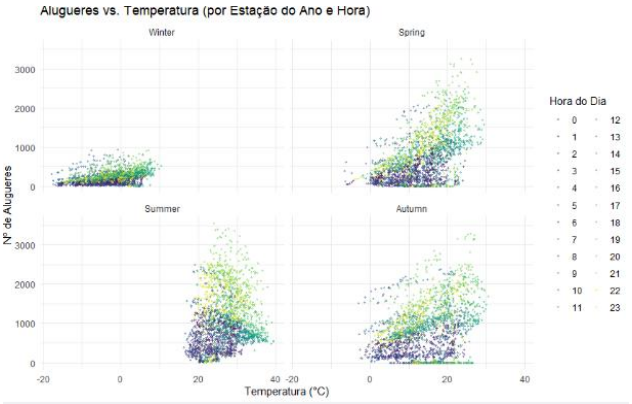


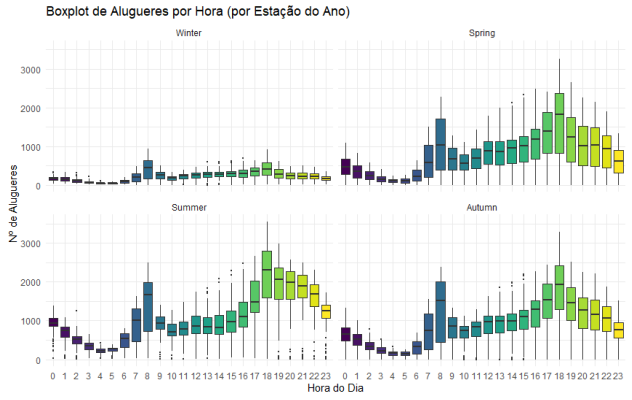
Gráfico 3- Histograma da distribuição da contagem de alugueres

O gráfico 3 demonstra que em certas horas o aluguer de bicicletas tendo a certo perto de nulo. Faz sentido tendo em conta horas mortas como a noite e a madrugada.



Gráficos 4 – Gráficos de dispersão para analisar a relação entre alugueres e temperatura, divididos por estação do ano

O conjunto de gráficos (Gráficos 4) demonstra um comportamento previsível em estações como o Inverno, o Outono e a Primavera. A mantem-se baixa no Inverno e tende a aumentar nas outras estações. Um dado interessante é no Verão onde apesar de haver procura alta, quando as temperaturas ficam mais elevadas, existe uma discrepância na procura.



Graficos 5 - Boxplots para mostrar a distribuição dos alugueres por hora do dia e estação.

No gráfico 5 é possível ver melhor o quanto as horas dos dias influenciam na procura. Dá para ver picos às 8 horas, provavelmente devido a movimentações para o trabalho, e outro às 18h, remetendo ao oposto. O tema das estações já se confirmou com os gráficos anteriores (Gráficos 4).

5.10 Modelação com Regressão Linear

Com os dados devidamente tratados, avançamos para a construção de modelos de regressão linear [7]. Após garantir que não existem valores em falta, o conjunto de dados de Seul é dividido em dois subconjuntos: um para treino (70% dos dados), e outro para teste (30%). São então desenvolvidos vários modelos: um baseado apenas em variáveis meteorológicas, outro em variáveis temporais, um terceiro com termos quadráticos, e finalmente um modelo com interações entre variáveis. Cada modelo é avaliado através do erro quadrático médio (RMSE) [9], que permite medir a precisão das previsões.

A fórmula do RMSE pode ser a seguir:

$$RSME = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}}$$

Onde:

- yi é o valor real da i-ésima observação.
- ŷi é o valor previsto da i-ésima observação.
- N é o número total de observações.
- P é o número de parâmetros estimados, incluindo a constante.

Quanto menor este valor, melhor é a previsão.

Dá para notar na tabela 1 os vários modelos e os resultados de RMSE que obtiveram (existem outras estatísticas, mas o foco está nesta pelo simples facto de estar diretamente ligada ao desempenho dos modelos. Após vários testes

realizados pelo grupo concluímos que o modelo com melhor desempenho foi o que incluiu interações, o qual foi guardado num ficheiro RData, ficando assim pronto para ser utilizado pela aplicação Shiny na realização de previsões em tempo real.

Modelo/Algoritmo	RMSE
Mateológico	538.31
Temporal	510.94
Polinomial	489.83
Iterações	470.48

Tabela 1 – Tabela dos modelos treinados com os dados de Seul e respetivos resultados RMSE.

6. Aplicativo R Shiny Dashboard

A análise seguinte foca-se no script da aplicação (app.R), que é executado após o script principal (projeto_codigo.R), pois necessita do modelo preditivo que foi gerado e guardado por este. Tudo o que executado neste ficheiro refere-se à parte gráfica do projeto.

Usando um painel interativo R Shiny, podemos prever a procura de bicicletas em tempo real com base nas condições meteorológicas atuais, utilizando o modelo previamente treinado.

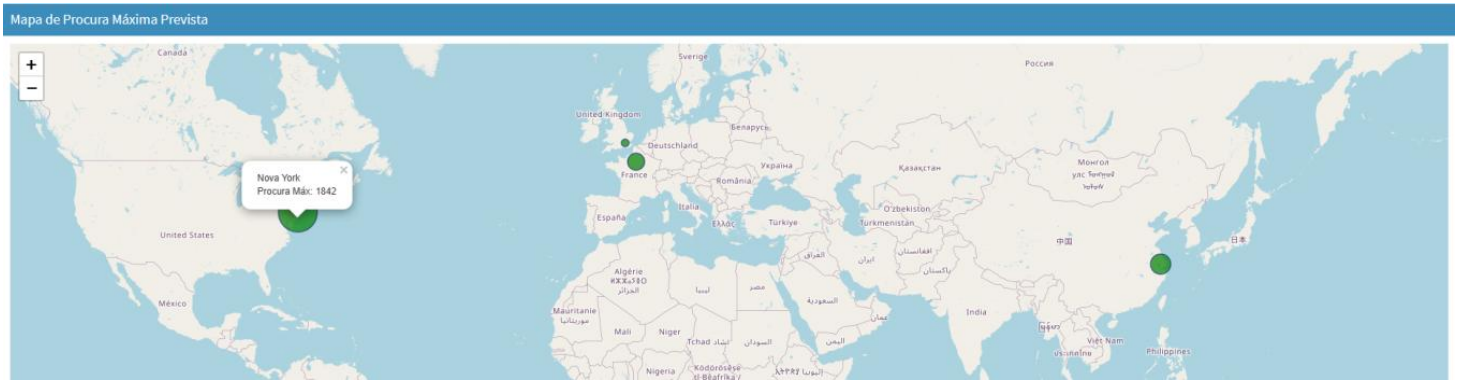


Figura 2 – Mapa identificado com as cidades requeridas e a procura de bicicletas nos próximos 5 dias.

Foram usadas diversas libraries para garantir que a interface gráfica se enquadrava com o necessário para o projeto. Algumas destas já foram utilizadas no ficheiro de execução principal("projeto_codigo"), destacando algumas novas como o próprio shiny em conjunto com o leaflet que serviu para a criação iterativa. Para garantir a atualidade dos dados, tendo em conta que o utilizador pode executar primeiramente o ficheiro principal sem executar a interface gráfica durante muito tempo, definimos outra vez o acesso à API da OpenWeather, e buscamos de novo os dados atuais das cidades de interesse. A partir daqui carregamos o modelo de regressão previamente treinado, que vai prever a partilha de bicicletas para os países de interesse tendo em conta os dados meteorológicos adquiridos da API.

Entrando pelo tópico da interface em si, esta foi feita para ser maioritariamente funcional. O utilizador pode selecionar a localização pretendida e configurar alguns parâmetros necessários. A aplicação apresenta em tempo real várias visualizações interativas, incluindo mapas (Fig.2), gráficos(Fig.3) e tabelas, permitindo assim uma análise da procura prevista [10]. Cada vez que o utilizador seleciona uma nova cidade as previsões são continuamente atualizadas com base nas condições meteorológicas atuais, isto é, cada vez que acontece uma mudança, volta a haver uma

request à API para garantir que os dados estão sempre atualizados.

Analisando melhor o gráfico da figura 3 é possível notar a semelhança que tem em relação aos gráficos vistos anterior na fase de analise de dados(EDA). A procura de bicicletas é menor a horas de madrugada, e a procura maior é quando os fatores de temperatura maior e as 18 horas juntam-se.

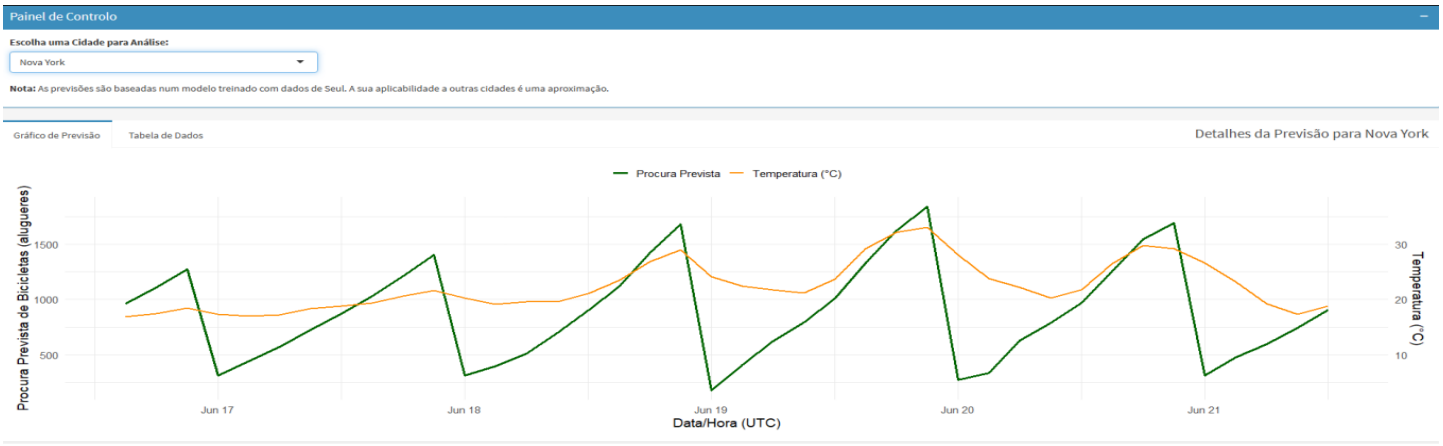


Figura 3– Interface com escolha da cidade e gráfico associado da procura de bicicletas nos próximos 5 dias.

7. Trabalho Futuro

Para aperfeiçoar o projeto, pode-se realizar diversas ações. O código pode ser otimizado para ocupar menos cache aumentando o desempenho. Os modelos podem ser aperfeiçoados a ponto de a margem de erro ser mínima, talvez com mais variáveis ou até com outros tipos de modelos para além dos lineares. Para além destas melhorias, também é possível adicionar mais cidades ao conjunto de previsão e melhorar a interface para acompanhar estas alterações.

8. Código

O código utilizado neste projeto está disponível no repositório do GitHub:

https://github.com/git2026/sistemas_apoio_decisao

9. Referencias

- [1] Euronews. (2024, Março 1). *From Amsterdam to Vilnius: The European cities with the most and least frequent cyclists*. https://www.euronews.com/travel/2024/03/01/from-amsterdam-to-vilnius-the-european-cities-with-the-most-and-least-frequent-cyclists
- [2] Liu, L., Ji, Y., Liu, Z., Cheng, L., & Wu, J. (2021, Outubro). Exploring the impact of built environment on bike-sharing usage and trip duration: A case study of Ningbo, China. Transportation Research Part D: Transport and Environment. https://doi.org/10.1016/j.trd.2021.103033
- [3] ITF Corporate Partnership Board. (2021). *Decarbonising Urban Mobility with Data and New Business Models*. International Transport Forum. https://www.itf-oecd.org/decarbonising-urban-mobility-data-new-business-models

- [4] Beach, J. (2020, Dezembro). *Seoul Bike Sharing Demand*. Kaggle. https://www.kaggle.com/datasets/joebeachcapital/seoul-bike-sharing
- [5] Wikipedia contributors. (2024, Junho 14). *List of largest cities*. Wikipedia. https://en.wikipedia.org/wiki/List_of_largest_cities
- [6] Posit team. (2024). *Posit - The Open Source Data Science Company*. Posit. https://posit.co/
- [7] T, S. (2018, Novembro 13). *Linear Regression-Detailed View*. Medium. https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86
- [8] Wikipedia contributors. (2019, Dezembro 2). *List of bicycle-sharing systems*. Wikipedia. Arquivado de https://web.archive.org/web/20191202194031/https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems
- [9] Frost, J. (2024). *Root Mean Square Error (RMSE)*. *Statistics By Jim*. https://statisticsbyjim.com/regression/root-mean-square-error-rmse/
- [10] Zafar, R. (2020, Janeiro 19). *Interactive maps in R using Leaflet package for beginners*. *Medium*. Interactive maps in R using Leaflet package for beginners