1. Projeto de Sistemas de Apoio à Decisão (2024-2025)

Use as ferramentas da aula para construir algo interessante!

- Os projetos devem ser feitos em grupos de 3 a 5 alunos, obrigatoriamente.
- Há 3 entregáveis para o projeto até ao final do trimestre: relatório, vídeo e código R.
- Cada grupo de projeto terá mentoria para ter feedback e respostas a perguntas.
- O projeto será avaliado nos eixos descritos abaixo: definição da tarefa, abordagem, dados e experimentação, análise.
- Relatório final, código e vídeo de apresentação (devido até **15 de junho** de 2025, às 23h59)

2. Diretrizes do Projeto Final

No projeto final, deverão trabalhar em grupos de **três a cinco elementos obrigatoriamente**, para aplicar as técnicas que aprenderam na disciplina a um novo cenário. Observem que, independentemente do tamanho do grupo, todos os grupos devem enviar o trabalho conforme definido e serão classificados nos mesmos critérios.

Além disso, espero que cada equipa envie um projeto concluído. Todos os projetos exigem que os alunos invistam tempo a recolher dados e a configurar a infraestrutura para chegar a um resultado final e, portanto, equipas de 3 a 5 pessoas podem partilhar essas tarefas muito melhor. Isso permite que a equipa se concentre mais nos resultados interessantes que pretende obter e na discussão do projeto. Cada membro da equipa deve contribuir nos **componentes** técnicos e não técnicos do projeto.

Devem construir um sistema para resolver uma tarefa bem definida. Os métodos que usarem devem basear-se nos lecionados na disciplina.

É possível que tenham várias iterações para encontrar o projeto certo, por isso sejam pacientes; essa descoberta é uma parte essencial da pesquisa, e como tal aprendam com ela. Divirtam-se e não esperem até o último minuto!

3. Vídeo final (5 minutos):

O vídeo final deve ser entregue até **15 de junho**. O vídeo final é uma apresentação de vídeo de 5 minutos do vosso projeto. No vídeo, devem descrever brevemente a vossa motivação, definição do problema, desafios, abordagens, resultados e análise. Devem incluir diagramas, figuras e gráficos para ilustrar os destaques do vosso trabalho. Se possível, tentem criar visualizações criativas do vosso projeto. Isso pode incluir diagramas de sistema, exemplos mais detalhados de dados que não cabem no espaço do vosso relatório

ou demonstrações ao vivo para sistemas de ponta a ponta. O objetivo do vídeo é transmitir as ideias importantes de alto nível e dar intuição, em vez de ser uma especificação superdetalhada de tudo o que fizeram. Usem diagramas e exemplos concretos e evitem slides muito densos ou com equações extremamente complexas.

4. Relatório Final (até 10 páginas)

O relatório final deve ser entregue até **15 de junho de 2025**. O vosso relatório final deve ser um relato abrangente do vosso projeto. A estrutura do relatório final do vosso projeto será também alvo de avaliação.

5. Submissão

Enviem os 3 entregáveis do projeto (código, vídeo e relatório final) no eLearning e certifiquem-se de que **todos os membros do grupo** sejam adicionados no envio.

Todos os documentos devem ser entregues até às **23h59**, de dia 15 de junho de 2025.

Para o **Relatório Final do Projeto**, certifiquem-se que incluir um link para o vosso **código** (carregado no Github/Bitbucket/Google Drive/etc.) e **dados** no eLearning. Também devem incluir um arquivo README.md dentro do repo/zip documentando tudo e quais os comandos que executaram.

6. Visão geral do Projeto Final

Neste projeto aplicará várias habilidades e técnicas de Ciência de Dados que aprendeu como parte desta disciplina.

Assumirá o papel de um Cientista de Dados que recentemente se juntou a uma empresa de análise de dados meteorológicos alimentada por IA e será apresentado a um desafio que requer recolha de dados, análise, teste de hipóteses básicas, visualização, modelação e painel de visualização a serem realizados em conjuntos de dados do mundo real.

Irá realizar as tarefas de:

- Recolha e compreensão de dados de várias fontes
- Execução de disputa e preparação de dados com expressões regulares e Tidyverse
- Realização de análise exploratória de dados com SQL e visualização utilizando Tidyverse e ggplot2
- Realização de modelação de dados com regressões lineares usando Tidymodels
- Construção de um painel interativo usando R Shiny

O projeto culminará com a apresentação do seu relatório de análise de dados, com um resumo executivo para os vários stakeholders da organização. Será avaliado tanto no seu

trabalho para as várias etapas do processo de análise de dados, bem como no resultado final.

Este projeto é uma ótima oportunidade para mostrar as suas habilidades em Sistemas de Suporte à Decisão e demonstrar a sua proficiência para potenciais empregadores.

6.1. Cenário do projeto

Imagine que acabou de ser contratado por uma empresa de análise de dados meteorológicos alimentada por IA como Cientista de Dados.

O seu primeiro projeto é analisar como o clima afetaria a procura por partilha de bicicletas em áreas urbanas. Para concluir este projeto, precisa primeiro de recolher e processar dados relacionados ao clima e à procura de partilha de bicicletas de várias fontes, realizar análises exploratórias de dados sobre os dados e construir modelos preditivos para prever a procura de partilha de bicicletas. Combinará os seus resultados e ligá-los-á a um painel ao vivo exibindo um mapa interativo e a visualização associada do clima atual e da procura estimada de bicicletas.

A última tarefa é criar uma apresentação de slides perspicaz e informativa e apresentá-la.

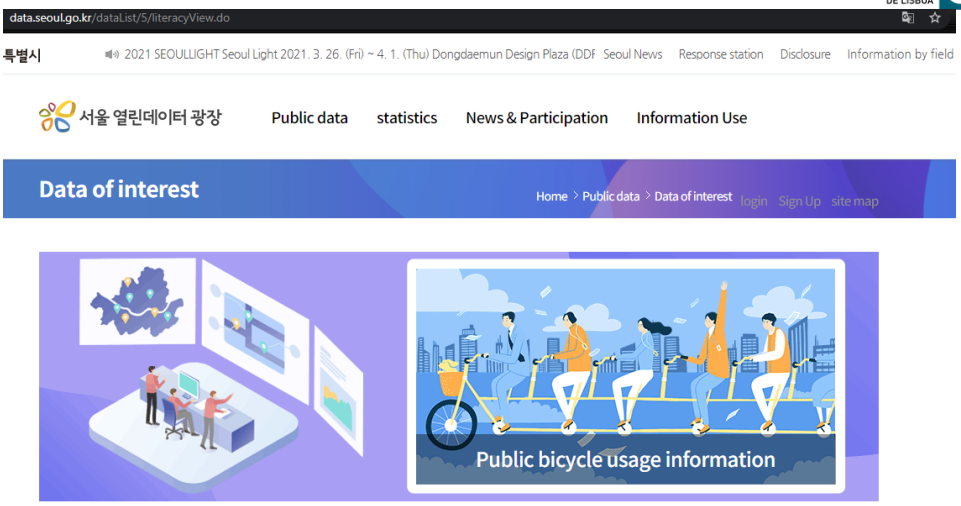
6.2. Compreender os dados de origem

Aqui apresento a sugestão das fontes de dados que utilizará no seu projeto.

As bicicletas de alugues estão disponíveis em muitas cidades ao redor do mundo. É importante para cada uma dessas cidades fornecer um conjunto confiável de bicicletas de aluguer para otimizar a disponibilidade e acessibilidade ao público em todos os momentos.

Também importante é minimizar o custo desses programas, em parte minimizando o número de bicicletas fornecidas para atender à procura. Assim, para ajudar a otimizar o abastecimento, seria útil ser capaz de prever o número de bicicletas necessárias a cada hora do dia, com base nas condições atuais, como o clima.

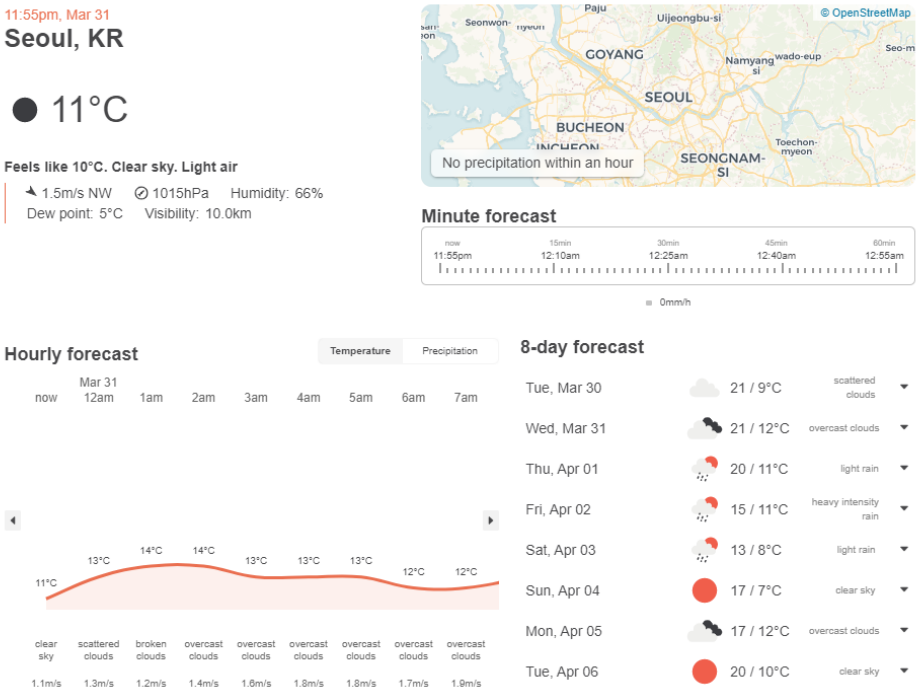
O *Seoul Bike Sharing Demand Data* set foi projetado para essa finalidade. Ele contém informações meteorológicas (temperatura, humidade, velocidade do vento, visibilidade, ponto de orvalho, radiação solar, queda de neve, precipitação), e o número de bicicletas alugadas por hora e data.



Usará esse conjunto de dados para construir um modelo de regressão linear do número de bicicletas alugadas a cada hora, com base no clima.

Informações sobre atributos

- **Data**: ano-mês-dia
- **Contagem de bicicletas alugadas** - Contagem de bicicletas alugadas a cada hora
- **Hora** - Hora do dia Temperatura-Temperatura em Celsius
- **Humidade** - a unidade é %
- **Velocidade do vento** - a unidade é m/s
- **Visibilidade** - unidade 10m
- **Temperatura do ponto de orvalho** - Celsius
- **Radiação solar** - MJ/m2
- **Precipitação** - mm Queda de neve - cm
- **Estações** - inverno, primavera, verão, outono
- **Feriado** - Feriado/Sem feriado
- **Dia Funcional** - NoFunc (Horas Não Funcionais), Diversão (Horas Funcionais)



6.3. Dados abertos da API meteorológica

A *API Open Weather* permite que os utilizadores acedam a dados meteorológicos atuais e previstos para qualquer local, incluindo mais de 200.000 cidades. O OpenWeather recolhe e processa dados meteorológicos de diferentes fontes, como modelos meteorológicos globais e locais, satélites, radares e uma vasta rede de estações meteorológicas. Você pode aceder aos dados necessários para o projeto gratuitamente, registrando-se com uma assinatura gratuita. Os dados aos quais se estará a ligar fornecem a previsão do tempo para cada 3 horas nos próximos 5 dias.

6.4. Conjunto de dados de sistemas globais de partilha de bicicletas

O *Global Bike Sharing Cities Dataset* é uma tabela HTML na página da Wikipédia de sistemas de partilha de bicicletas: https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems.

Ela lista sistemas ativos de partilha de bicicletas em todo o mundo. A maioria dos sistemas listados permite que os utilizadores peguem e deixem bicicletas em qualquer uma das estações automatizadas da rede.

6.5. Dados das Cidades do Mundo

Os Dados das Cidades do Mundo contêm informações como nome, latitude e longitude, sobre as principais cidades ao redor do mundo.

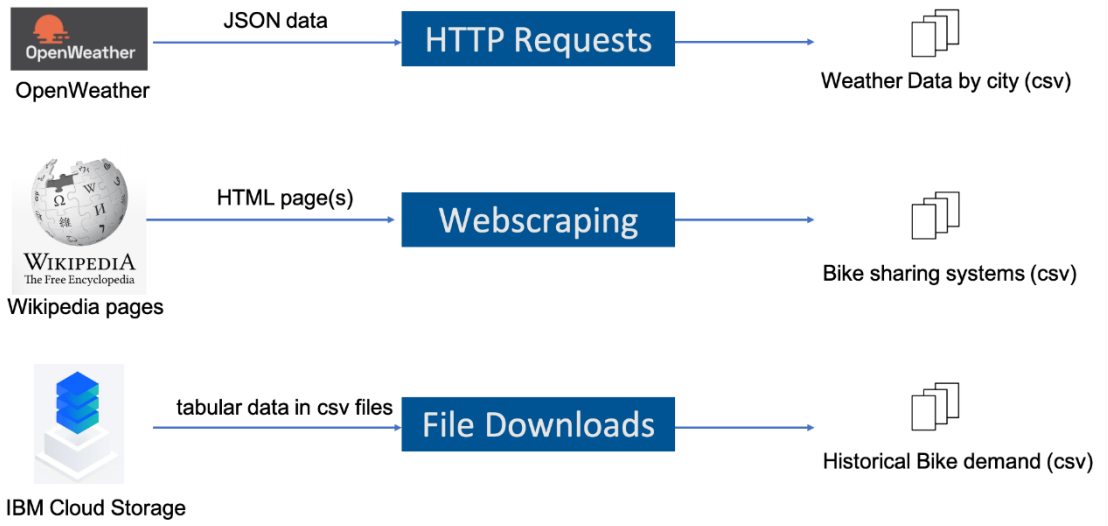
6.6. R-Studio na Posit Cloud

Para este projeto, pode utilizar o Posit Cloud (anterior R-Studio Cloud) como o seu ambiente de desenvolvimento principal. Este ambiente de laboratório não requer nenhuma instalação e é conveniente de usar.

Tenha em atenção que o serviço baseado na nuvem terá limitações no número de horas de computação ou no número de dias durante os quais terá de concluir o seu trabalho. Portanto, gira-o com base nas suas necessidades e recursos.

7. Recolha de dados meteorológicos e de procura de partilha de bicicletas

Para ser capaz de analisar e prever a procura de partilha de bicicletas para uma cidade, precisará de recolher dados relevantes de várias fontes, incluindo sistemas globais de partilha de bicicletas em páginas da web públicas, dados meteorológicos por meio de APIs OpenWeather e dados tabulares agregados do armazenamento em nuvem, etc., como mostrado abaixo:



Este é o processo de recolha de dados sugerido:

- Web scrape a Global Bike-Sharing Systems Wiki Page
    - **TAREFA:** Extrair dados do sistema de partilhamento de bicicletas de uma página Wiki e converter os dados num quadro de dados
- Chamadas de APIs OpenWeather
    - **Leitura**: Configuração da API OpenWeather
    - **Prática de codificação**: Obtenha os dados meteorológicos atuais de uma cidade usando a API OpenWeather

- **TAREFA:** Obter previsão do tempo para 5 dias para uma lista de cidades usando a API OpenWeather
- **TAREFA**: Baixar conjuntos de dados em arquivos csv do armazenamento em nuvem

Depois de o processo de recolha de dados ser concluído, deve ser capaz de obter os seguintes dados brutos, todos armazenados no formato de arquivo csv:

- **raw_bike_sharing_systems.csv**: Uma lista de sistemas ativos de partilha de bicicletas em todo o mundo
- **raw_cities_weather_forecast.csv**: Previsão do tempo para 5 dias para uma lista de cidades de OpenWeather API
- **raw_worldcities.csv**: Uma lista de informações das principais cidades (como nome, latitude e longitude) em todo o mundo
- **raw_seoul_bike_sharing.csv**: Contém informações meteorológicas (temperatura, humidade, velocidade do vento, visibilidade, ponto de orvalho, radiação solar, queda de neve, precipitação), o número de bicicletas alugadas por hora e informações de dados dos sistemas de partilha de bicicletas de Seul.

8. Visão geral da disputa de dados

Até agora, já recolheu os dados necessários de partilha de bicicletas para outras tarefas de análise exploratória, visual e preditiva. No entanto, alguns dados podem conter ruídos ausentes, mal formatados e/ou inesperados. Estas fontes de ruído podem reduzir significativamente o desempenho da análise. Assim, precisa realizar a disputa de dados antes de analisar melhor os dados.

A disputa de dados visa remover o ruído dos dados e converter o formato de dados indesejado para um formato que provavelmente será melhor para a análise.

Disputa de dados com stringr e expressões regulares:

- TAREFA: Padronizar nomes de colunas para todos os conjuntos de dados recolhidos
- TAREFA: Remover links de referência indesejados usando expressões regulares
- TAREFA: Extrair valores numéricos usando expressões regulares

Disputa de dados com dplyr (Tempo estimado: 60 minutos)

- TAREFA: Detetar e manipular valores ausentes
- TAREFA: Criar variáveis indicadoras (fictícias) para variáveis categóricas
- TAREFA: Normalizar dados

https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems

9. Realizar Análise Exploratória de Dados com SQL, Tidyverse & ggplot2

Criar as tabelas e carregá-las para a base de dados, 4 tabelas para os 4 conjuntos de dados.

Tarefa 1 - Contagem de Registos

Determine quantos registos estão no conjunto de dados seoul_bike_sharing.

Tarefa 2 - Horário de Funcionamento

Determinar quantas horas tiveram uma contagem de bicicletas alugadas diferente de zero

Tarefa 3 - Perspetivas meteorológicas

Consulte a previsão do tempo para Seul nas próximas 3 horas. Lembre-se de que os regitros no conjunto de dados CITIES_WEATHER_FORECAST têm 3 horas de intervalo, portanto, precisamos apenas do primeiro registo da consulta.

Tarefa 4 - Estações

Descubra que estações estão incluídas no conjunto de dados de partilha de bicicletas Seul.

Tarefa 5 - Intervalo de datas

Encontre a primeira e a última data no conjunto de dados Seoul Bike Sharing.

Tarefa 6 - Subconsulta - 'máximo histórico'

Determinar qual data e hora que teve mais alugueres de bicicletas.

Tarefa 7 - Popularidade horária e temperatura por estação

Determine a temperatura média horária e o número médio de alugueres de bicicletas por hora em cada estação. Liste os dez melhores resultados por contagem média de bicicletas.

Tarefa 8 - Sazonalidade do aluguer

Encontre a contagem horária média de bicicletas durante cada estação. Inclua também o mínimo, o máximo e o desvio padrão da contagem horária da bicicleta para cada estação.

Tarefa 9 - Sazonalidade Meteorológica

Considere o clima ao longo de cada estação. Em média, quais foram as TEMPERATURAS, HUMIDADE, WIND_SPEED, VISIBILIDADE, DEW_POINT_TEMPERATURE, SOLAR_RADIATION, PRECIPITAÇÃO e QUEDA de NEVE por estação? Inclua também a

contagem média de bicicletas e classifique os resultados por contagem média de bicicletas para que possa ver se está correlacionado com o clima.

Tarefa 10 - Contagem total de bicicletas e informações sobre a cidade de Seul

Use uma junção implícita nas tabelas de WORLD_CITIES e BIKE_SHARING_SYSTEMS para determinar o número total de bicicletas disponíveis em Seul, além das seguintes informações da cidade sobre Seul: CIDADE, PAÍS, LAT, LON, POPULAÇÃO, numa única visualização.

Tarefa 11 - Encontrar todos os nomes de cidades e coordenadas com escala de bicicleta comparável ao sistema de partilha de bicicletas de Seul

Encontre todas as cidades com contagens totais de bicicletas entre 15000 e 20000. Retorne os nomes da cidade e do país, além das coordenadas (LAT, LNG), população e número de bicicletas para cada cidade.

10. EDA com Vsualização

Tarefa 1 - Carregar o conjunto de dados

Carregue os dados seoul_bike_sharing num dataframe, isso será o CSV limpo antes de fazer a normalização e converter algumas colunas em colunas numéricas.

Tarefa 2 - Reformular DATE como data

Use o formato dos dados, ou seja, "%d/%m/%Y"

Tarefa 3 - Transmitir HORAS como uma variável categórica

Além disso, torne os seus níveis a serem uma sequência ordenada. Isso garantirá que as suas visualizações utilizem corretamente HOURS como uma variável discreta com o pedido esperado.

Tarefa 4 - Resumo do conjunto de dados

Tarefa 5 - Com base nas estatísticas acima, calcule quantos feriados existem

Tarefa 6 - Calcular a percentagem de registos que caem num feriado.

Tarefa 7 - Dado que há exatamente um ano inteiro de dados, determine quantos registos esperamos ter

Tarefa 8 - Dadas as observações para o 'FUNCTIONING_DAY', quantos registos devem existir?

Tarefa 9 - Carregue o pacote dplyr, agrupe os dados por SEASONS e use a função summarize() para calcular a precipitação total sazonal e a queda de neve

Tarefa 10 - Criar um gráfico de dispersão de RENTED_BIKE_COUNT vs DATE

Tarefa 11 - Crie o mesmo gráfico da RENTED_BIKE_COUNT série temporal, mas agora adicione HORAS como cor.

Tarefa 12 - Criar um histograma sobreposto com uma curva de densidade do kernel

Normalize o histograma para que o eixo y represente 'densidade'.

Tarefa 13 - Use um gráfico de dispersão para visualizar a correlação entre RENTED_BIKE_COUNT e TEMPERATURA por ESTAÇÕES, Use HORA como cor.

Tarefa 14 - Crie uma exibição de quatro boxplots de RENTED_BIKE_COUNT vs. HORA agrupados por SEASONS.

Tarefa 15 - Agrupe os dados por DATA e use a função summarize() para calcular a precipitação total diária e a queda de neve

Tarefa 16 - Determinar quantos dias tiveram queda de neve

11. Prever a procura de partilha de bicicletas usando modelos de regressão

Preveja a contagem horária de bicicletas alugadas usando modelos básicos de regressão linear:

- TAREFA: Dividir dados em conjuntos de dados de treino e teste
- TAREFA: Construir um modelo de regressão linear usando apenas as variáveis meteorológicas
- TAREFA: Construir um modelo de regressão linear usando variáveis de tempo e data/hora
- TAREFA: Avaliar os modelos e identificar variáveis importantes

Refine os modelos de regressão da linha de base:

- TAREFA: Adicionar termos de ordem mais alta
- TAREFA: Adicionar termos de interação
- TAREFA: Adicionar regularização
- TAREFA: Experimentar para encontrar o modelo com melhor desempenho

12. Criação de um aplicativo R Shiny Dashboard

Agora, a sua equipa deve construir um painel interativo R Shiny para poder visualizar dados de previsão do tempo e procura de partilha de bicicletas por hora prevista para as seguintes cidades: Nova York, EUA, Paris, França, Suzhou, China e Londres, Reino Unido (eles têm tamanhos de frota de aluguer de bicicletas semelhantes aos de Seul, Coreia do Sul).

Neste projeto, construirá um mapa interativo que mostra a procura máxima prevista de partilha de bicicletas nos próximos 5 dias. As previsões serão baseadas no resultado de um modelo de regressão, que por sua vez utilizará dados de previsão do tempo da API OpenWeather como entrada.