Pontifícia Universidade Católica do Rio de Janeiro – PUC-RJ

Disciplina de Engenharia de Dados

Projeto MVP - Aluna: Caroline Carrilho Correia

Objetivo do MVP

Este trabalho tem como objetivo analisar e investigar, de forma crítica e multidimensional, os padrões e anomalias dos dados meteorológicos de 2024 no Brasil, buscando compreender não apenas dos números, mas também os fatores geográficos que podem estar associados aos fenômenos da natureza. Assim alguns questionamentos devem ser feitos:

1. Sobre a Radiação Global Solar:
    a. Quais foram os 10 dias com maior radiação?
    b. Qual a média de radiação brasileira?
    c. Qual a média de radiação por Estado?

O fator radiação busca gerar insights sobre agricultura, saúde pública e busca sustentável de meios energéticos.

2. Correlação entre umidade e precipitação:

Esses dois fatores devem ser correlacionados pois, ao se aplicar o coeficiente de Pearson, podemos obter várias interpretações dobre a qualidade do nosso banco de dados e sobre pontos de anomalia.

3. Qual a ocorrência de eventos extremos?

Esse fator nos faz pensar em possíveis prevenções futuras em determinadas regiões ao longo de determinados períodos.

4. Quam a média mensal de temperatura e precipitação por Estado?

Esses dois fatores associados aos meses ao longo do ano e divididos por Estados, nos faz ter uma visão Macro dos eventos meteorológicos no nosso país.

Modelagem:

O modelo da tabela para o desenvolvimento deste MVP, os dados foram armazenados em uma única tabela, seguindo o modelo flat file (arquivo plano). A estrutura é composta por:

• 26 colunas: Representando as categorias de dados meteorológicos.

• 2472289 linhas: Onde cada entrada corresponde a um dado individual. Nesse formato, cada linha do arquivo equivale a um registro completo, enquanto as colunas organizam os atributos conforme suas classificações.

Linhagem dos dados:

Os dados utilizados para essa análise foram obtidos na plataforma Kaggle, um repositório público de conjuntos de dados. A fonte original desses registros é do INMET (Instituto Nacional de Metrologia, Qualidade e Tecnologia) tem dados disponíveis desde 2018, garantindo a confiabilidade e relevância das informações para o escopo do projeto. A base de dados pode ser acessada pelo link:

https://www.kaggle.com/datasets/gnomows/dados-metereologicos-2018-2024-inmet

Como a base de cada ano é muito grande, decidi por usar somente o ano de 2024. Mas o ideal seria uma análise de todos os anos disponíveis para a retirada de insight mais robustos.

Catálogo de dados:

Este catálogo descreve a estrutura, os atributos e as descrições dos dados utilizados neste projeto, garantindo transparência e facilitando a interpretação das análises realizadas. As colunas foram organizadas conforme os termos a seguir:

- Nome variável: nome da coluna no banco de dados
- Tipo de variável: categórica ou numérica
- Descrição completo do dado
- Valores permitidos: intervalo de dados ou valores máximos e mínimos

Nome Coluna	Tipo de dado	Descrição	Valores Permitidos
_c0	String	Data	Valores de 01/01/2024 a 31/12/2024
_c1	String	Hora UTC	Valores de 00:00 UTC a 23:00 UTC
_c2	String	PRECIPITAÇÃO TOTAL, HORÁRIO (mm)	0,2 - 9,8 mm
_c3	String	PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA (mB)	980 - 1030 mB
_c4	String	PRESSÃO ATMOSFERICA MAX.NA HORA ANT. (AUT) (mB)	980 - 1030 mB
_c5	String	PRESSÃO ATMOSFERICA MIN. NA HORA ANT. (AUT) (mB)	980 - 1030 mB

_c6	String	RADIACAO GLOBAL (Kj/m²)	0 - 999 Kj/m²
_c7	String	TEMPERATURA DO AR - BULBO SECO, HORARIA (°C)	14 Negativos - 44,8 ºC
_c8	String	TEMPERATURA DO PONTO DE ORVALHO (°C)	14 Negativos - 44,8 ºC
_c9	String	TEMPERATURA MÁXIMA NA HORA ANT. (AUT) (°C)	14 Negativos - 44,8 ºC
_c10	String	TEMPERATURA MÍNIMA NA HORA ANT. (AUT) (°C)	14 Negativos - 44,8 ºC
_c11	String	TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT) (°C)	14 Negativos - 44,8 ºC
_c12	String	TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT) (°C)	14 Negativos - 44,8 ºC
_c13	String	UMIDADE REL. MAX. NA HORA ANT. (AUT) (%)	10 % - 99%
_c14	String	UMIDADE REL. MIN. NA HORA ANT. (AUT) (%)	10 % - 99%
_c15	String	UMIDADE RELATIVA DO AR, HORARIA (%)	10 % - 99%
_c16	String	VENTO, DIREÇÃO HORARIA (gr) (° (gr))	0,1 - 9,9 m/s
_c17	String	VENTO, RAJADA MAXIMA (m/s)	0,1 - 9,9 m/s
_c18	String	VENTO, VELOCIDADE HORARIA (m/s)	0,1 - 9,9 m/s
_c19	String	REGIAO	N = Norte , NO = Nordeste, CO = Centro-oeste, SU = Sudeste e S = Sul
_c20	String	UF	AC, AL, AP, AM, BA, CE, DF, ES, GO, MA, MT, MS, MG, PA, PB, PR, PE, PI, RJ, RN, RS, RO, RR, SC, SP, SE, TO.
_c21	String	Estação	Refere-se as estações meteorológicas, como por exemplo Montes Claros
_c22	String	CODIGO (WMO)	**Código da Organização Meteorológica Mundial:** é um sistema padronizado internacionalmente para representar dados meteorológicos de forma estruturada e universal
_c23	String	LATITUDE	Latitudes referentes as regiões descritas
_c24	String	LONGITUDE	Longitudes referentes as regiões descritas
_c25	String	ALTITUDE	Altitudes referentes as regiões descritas
_c26	String	DATA DE FUNDACAO	Data que a estação meteorológica foi fundada

Extração, Carga e Transformação dos Dados:

Para a ingestão e preparação dos dados no ambiente Databricks, foram executadas as seguintes etapas:

Carga:

- Configuração do DBFS (Databricks File System): É um sistema de arquivos distribuído utilizado para interação com o armazenamento em nuvem.
- Estruturação do Diretório: criação de um notebook específico no catálogo para armazenamento do arquivo.
- Upload do CSV: carregamento do arquivo para o ambiente Databricks.
- Leitura dos Dados: uso do comando df = spark.read.csv para importação do dataset no notebook.

Transformação:

- Verificação dos nomes das colunas e tipos de dados com o comando DESCRIBE
- Criação de nova tabela para renomear os nomes das colunas com AS
- Verificação de todos os dados então dentro do período proposto no projeto
- Verificação de existência de dados nulos
- Transformação dos valores data (string) em YYYY/MM/DD
- Cálculo de médias, máximo e mínimo e correlações

Análise:

Qualidade dos dados: Apesar de os dados terem sido obtidos de uma fonte confiável como o INMET, o alto índice de valores nulos, a correlação entre umidade e precipitação ter dado negativa e o intervalo de dados tão ter sido provado entro de 2024, acende-se um ponto de alerta quanto a qualidade dos dados em si.

Não foi feito um tratamento para valores nulos, uma vez que média ou máximos e mínimos já os desconsideram, sendo assim:

- Nomes de colunas, tipos de dados e formatos todos padronizados, como temperatura, radiação, ventos, data e precipitação.
- As unidades de medidas todas estavam dentro das normas.
- Todas as latitudes e longitudes eram válidas.
- Dados limpos apesar verificar valores nulos, mas não críticos, e pela ausência de valores extremos (outliers)

Devido a boa qualidade dos dados, não se fez necessária a aplicação das etapas tradicionais de ingestão (bronze) e limpeza/enriquecimento (silver), permitindo um foco maior na análise exploratória e geração de insights. Assim, a economia de tempo e a redução da complexidade de pipelines e custos de processamento pela eliminação de etapas de ETL, foi visível e os recursos foram todos direcionados para a modelagem estatística, visualizações e tomada de decisão. Mesmo sem camadas intermediárias, foram adotadas práticas para garantir qualidade

- Validação pontual: Checagem de integridade via queries (ex: DISTINCT, IS NOT NULL).
- Versionamento: não foi necessário pois somente eu trabalhei no projeto e não se fez necessário o histórico de inserções, atualizações e exclusões.
- Metadados: Documentação das colunas e fontes no próprio Databricks (comentários em tabelas)."

Os dados trabalhados nesse projeto nos revelam uma consistência de eventos meteorológicos no Brasil. Assim, pude concluir:

- De um modo geral deve-se procurar uma igualdade energética no Brasil. Trabalhar no combate à desertificação com tecnologias de reuso de água no Nordeste e infraestrutura para lidar com chuvas intensas e variações sazonais no Sul são pontos de adaptação climática. A) Desigualdade energética: Estados com alta radiação podem tornar-se exportadores de energia limpa, reduzindo a dependência de hidrelétricas no Sudeste. Manter a floresta Amazônica é crucial para regular a radiação e o clima regional. B) A radiação solar é um recurso estratégico, e seu mapeamento ajuda a direcionar políticas públicas, investimentos e adaptações setoriais no Brasil.
- A correlação de umidade e precipitação deu negativa, o que é não é normal, pois já é de conhecimento comum que as duas métricas são diretamente proporcionais. Com isso, podemos tirar alguns insights: possivelmente erro na coleta/processamento dos dados, a amostra que escolhemos pode ser pequena ou também foram escolhidos valores extremos (como um dia com umidade baixa e chuva intensa devido a uma tempestade isolada) que podem distorcer a correlação.
- De acordo com o INMET em 2025, para ser caracterizado um evento climático extremo, deve-se ter as seguintes características:

- Ventos > m/s

- Chuvas > 50mm/h

No período analisado, pode verificar que não houve nenhum evento extremo, mas houve sim precipitações e ventos pontuais com essas características.

- Podemos verificar que a Maior temperatura média foi em Roraima e a menor foi no Rio Grade do Sul. Também foi possível relacionar as altas temperaturas com baixas precipitações, ou seja períodos de grande seca, geram altas temperaturas.

Autoavaliação:

Quando comecei o MVP, acreditava que não tinha conhecimento suficiente para o fazer, mas fui estudando muito e acredito que as amizades colaborativas que fiz com meus colegas de turma me ajudou a me encorajar e a melhorar cada vez mais. Hoje me sinto muito disposta a aprender. Quanto ao desenvolvimento em si do projeto, hoje após finalizado, eu já tenho condições de fazer de uma forma muito mais evoluída, o que me deixa muito feliz, com o a pós em si. Optei por me apegar ao escopo do MVP e não sair disso, pois acredito que no básico bem-feito. Quero agradecer aos professores e amigos que m apoiaram e sempre me deram um norte na conclusão desse projeto. Estou muito feliz, pois sei que o grupo do Discord e WhatsApp foram essenciais na conclusão. Obrigada.