Preparação De Dados/Engenharia De Recursos

1. Visão Geral

Nesta fase do projecto, realizamos a preparação de dados e engenharia de recursos, etapa fundamental para garantir a qualidade e a eficiência do modelo de aprendizado de máquina que será utilizado para prever o fluxo de turistas nas cidades de Luanda, Benguela e Lubango.

A preparação de dados teve como objectivo transformar os diferentes conjuntos de informações disponíveis, incluindo dados de clima (precipitação, temperatura média, mínima e máxima), feriados nacionais e número estimado de visitantes por cidade e mês em uma base de dados única, limpa e consistente. Esse processo envolveu a integração de fontes distintas e o tratamento de possíveis inconsistências, como valores ausentes, formatos diferentes de datas e variações nas escalas de medição.

A engenharia de recursos foi focada em criar variáveis capazes de melhorar o poder preditivo do modelo. Entre os principais recursos gerados, destacam-se:

- A variável “estação do ano”, derivada do mês, para captar a influência das condições climáticas sazonais;
- A variável “temperatura média”, calculada a partir da temperatura mínima e máxima;
- Um indicador binário de feriado, que marca os períodos de possível aumento no fluxo turístico;
- Variáveis de tendência temporal, representando a progressão dos meses, para auxiliar na captura de padrões sazonais e tendências de crescimento.

A importância desta fase reside no facto de que a qualidade dos dados é determinante para o desempenho do modelo de aprendizado de máquina. Um modelo treinado com dados limpos, coerentes e enriquecidos por variáveis relevantes tende a apresentar resultados mais precisos e interpretações mais fiáveis. Assim, a etapa de preparação e engenharia de recursos serviu como base sólida para as próximas fases do projecto, permitindo avançar com segurança para a análise exploratória e o desenvolvimento do modelo preditivo do fluxo turístico em Angola.

2. Coleta de Dados

A fase de coleta de dados teve como principal objectivo reunir todas as informações necessárias para a construção de um modelo preditivo capaz de estimar o fluxo mensal de turistas nas cidades de Luanda, Benguela e Lubango, entre os anos de 2022 e 2024. Para isso, foram utilizadas diversas fontes de dados oficiais e complementares, combinadas e tratadas de forma criteriosa. As fontes consultadas incluem:

1. Instituto Nacional De Estatística De Angola os dados sobre o número de turistas que visitaram Angola foram obtidos a partir do Anuário Estatístico do Turismo 2022–2023, publicado pelo INE. O documento fornece informações detalhadas sobre:
   - As chegadas mensais de turistas internacionais ao país (Quadro 12);
   - A distribuição de hóspedes por província durante o biénio 2022–2023 (Quadro 17).

   Como o relatório não disponibiliza o número de turistas por cidade, foi realizado um processo de estimativa. Utilizamos a proporção de hóspedes por província como indicador da participação de cada cidade no total de

turistas do país, partindo do princípio de que: Luanda representa a capital e o principal polo de entrada e permanência de visitantes, Benguela é um dos destinos turísticos mais procurados do litoral, Lubango (representante da província da Huíla) é um importante destino turístico do interior. Assim, as percentagens de participação de hóspedes por província foram aplicadas às chegadas mensais de turistas ao país, gerando estimativas mensais de visitantes para cada cidade. Para o ano de 2024, como ainda não existem dados oficiais, foi aplicada uma taxa de crescimento projetada de 3,14%, correspondente ao aumento médio do total de turistas entre 2022 e 2023, conforme indicado nos relatórios do INE;

2. Africa Data Hub a partir do portal Africa Data Hub foram coletados dados climáticos das cidades de Luanda, Benguela e Lubango, incluindo precipitação, temperatura média, temperatura mínima e temperatura máxima. Esses dados foram exportados em formato CSV e posteriormente harmonizados com as informações de turismo;
3. Governo de Angola foi criada uma base de feriados nacionais a partir do calendário oficial disponibilizado pelo Governo de Angola. Essa informação foi usada para criar uma variável binária que indica se o mês contém feriados que possam impactar o fluxo turístico.

2.1. Processo de Pré-processamento dos Dados

Após a coleta, foi realizado um pré-processamento completo dos dados antes da integração no *dataset* principal. As principais etapas foram:

- Padronização de formatos e nomes de colunas: Todos os arquivos foram convertidos para o formato CSV, com nomes de colunas uniformes;
- Tratamento de valores ausentes e inconsistências: Valores climáticos ausentes foram preenchidos com médias mensais da mesma localidade, garantindo continuidade temporal. Pequenas inconsistências de digitação ou codificação (como nomes de meses) foram corrigidas;
- Conversão e harmonização de tipos de dados: As variáveis de data foram convertidas para o formato numérico (ex.: janeiro = 1, fevereiro = 2), e as variáveis categóricas foram padronizadas para permitir fácil leitura pelo modelo de aprendizado de máquina;
- Integração das fontes: Os dados de clima, feriados e visitantes foram integrados em um único dataset com base nas colunas comuns ano, mes e localidade. Essa fusão resultou em uma base completa e temporalmente alinhada;
- Verificação de consistência temporal: Foi garantido que todas as cidades possuíssem dados completos para os 36 meses (janeiro de 2022 a dezembro de 2024), sem lacunas ou duplicações.

3. Síntese dos dados de visitantes

Como não existiam dados directos de turistas por cidade, foi criada uma síntese estatística baseada em proxies oficiais:

O total mensal nacional fornecido pelo INE foi distribuído proporcionalmente entre as cidades, conforme a participação média de hóspedes por província. Essa metodologia foi adoptada de forma transparente e fundamentada em fontes oficiais, servindo como uma aproximação realista do fluxo turístico mensal local. A estimativa final foi validada pela consistência temporal (tendência crescente entre 2022–2024) e coerência regional (Luanda > Benguela > Lubango).

4. Limpeza de Dados

Nesta fase, concentrámo-nos em transformar os dados brutos recolhidos em um conjunto limpo, consistente e pronto para análise. A limpeza de dados foi uma etapa essencial para garantir a confiabilidade dos resultados do modelo de aprendizado de máquina. Trabalhámos de forma sistemática para identificar e corrigir erros, inconsistências e lacunas que poderiam comprometer a qualidade das previsões.

Primeiramente, realizámos uma análise detalhada para identificar valores ausentes (*missing values*) no conjunto de dados. Verificámos que as variáveis temperatura máxima e temperatura mínima apresentavam alguns registros nulos, algo relativamente comum em bases meteorológicas, devido a falhas momentâneas de medição ou ausência de coleta em determinados dias. Para assegurar a integridade e continuidade das séries temporais, aplicámos duas abordagens complementares:

1. Imputação por média sazonal: Quando os valores ausentes representavam menos de 5% do total de registros, substituímos as lacunas pela média da variável correspondente ao mesmo mês e localidade. Essa técnica preserva o padrão climático característico de cada região e época do ano, evitando distorções artificiais;
2. Interpolação temporal: Em casos onde as lacunas eram maiores ou sequenciais, utilizámos interpolação linear baseada no tempo, permitindo estimar valores intermediários com base na tendência natural observada antes e depois do ponto ausente. Essa abordagem mantém a coerência e o comportamento contínuo das séries históricas.

Essas estratégias combinadas permitiram restaurar a completude dos dados de temperatura sem comprometer a sua variabilidade real, garantindo maior robustez às análises e previsões subsequentes.

Em seguida, procedemos à análise de valores discrepantes (*outliers*) nas variáveis climáticas e de visitantes, utilizando métodos estatísticos como o desvio-padrão e a visualização por boxplot. O objectivo foi identificar possíveis erros de medição ou valores atípicos que pudessem distorcer as análises.

Após a aplicação desses métodos, não foram detectados *outliers* significativos nas variáveis avaliadas. As medições de temperatura, precipitação e número de visitantes apresentaram-se dentro de faixas plausíveis e consistentes com o comportamento esperado para cada localidade e período do ano.

Quanto aos dados sobre o número de visitantes por cidade, realizámos uma análise de consistência. Como os valores foram estimados com base na distribuição percentual histórica de turistas (segundo relatórios do INE e tendências observadas em plataformas de turismo), verificámos se a soma mensal por cidade correspondia ao total nacional divulgado oficialmente. Fizemos pequenos ajustes proporcionais para garantir que a soma total coincidisse com o valor real reportado pelo INE.

Por fim, removemos registros duplicados e confirmámos que cada linha representava unicamente uma combinação de ‘mês + ano + cidade’. O resultado desse processo foi um dataset limpo, coerente e consistente, apto para as próximas etapas de análise exploratória e modelagem preditiva.

5. Análise Exploratória dos Dados (EDA)

Após a etapa de limpeza, passámos para a Análise Exploratória dos Dados, com o objectivo de compreender melhor os padrões, tendências e relações existentes entre as variáveis do nosso conjunto de dados. Esta fase foi fundamental para revelar

comportamentos sazonais, dependências entre factores climáticos e fluxo de turistas, além de identificar indícios que pudessem orientar a construção do modelo preditivo.

Começámos por uma análise descritiva de todas as variáveis. Calculámos estatísticas básicas como média, mediana, desvio-padrão, mínimo e máximo para cada atributo (temperatura, precipitação, visitantes, etc.). Isso permitiu observar, por exemplo, que as cidades de Luanda e Benguela apresentaram temperaturas médias mais elevadas e menor variação anual, enquanto o Lubango apresentou temperaturas mais amenas e maior amplitude térmica.

Na sequência, realizámos uma análise temporal dos dados. Criámos gráficos de linha para visualizar a variação mensal da temperatura, precipitação e número de visitantes ao longo dos anos de 2022 a 2024. Observámos que o número de visitantes apresenta forte sazonalidade, com picos concentrados nos meses secos (entre Junho e Setembro) e quedas durante o período chuvoso. Essa tendência foi especialmente visível em Benguela e Lubango, cidades com forte apelo turístico durante o clima seco.

Também explorámos a correlação entre variáveis. Através de um mapa de calor da matriz de correlação, identificámos uma correlação negativa moderada entre precipitação e número de visitantes, ou seja, meses com mais chuva tendem a atrair menos turistas. Por outro lado, verificámos uma correlação positiva entre temperatura média e visitantes em Luanda, indicando que o clima mais quente está associado a um aumento no fluxo turístico.

Além disso, analisámos os feriados nacionais, cruzando-os com os picos de visitação. Verificámos que meses com feriados prolongados (como Abril e Dezembro) apresentaram aumentos significativos no número de turistas, sugerindo um impacto directo de eventos culturais e datas festivas na mobilidade interna e internacional.

Durante a EDA, também identificámos diferenças comportamentais entre cidades. Por exemplo, Benguela mostrou maior sensibilidade às condições climáticas, enquanto Luanda manteve um fluxo relativamente estável, possivelmente devido à sua importância como centro urbano e de negócios.

Por fim, gerámos visualizações complementares, boxplots, histogramas e gráficos de dispersão, que ajudaram a compreender a distribuição dos dados e confirmar a ausência de novos outliers após a limpeza. A partir dessas análises, obtivemos uma compreensão sólida do comportamento dos dados e das variáveis mais relevantes, o que nos orientou na fase seguinte de seleção de atributos e modelagem preditiva.

6. Engenharia de Recursos

Nesta fase, procurámos enriquecer o conjunto de dados com novas variáveis que pudessem melhorar o desempenho do modelo preditivo e a capacidade explicativa da análise. Partindo das colunas originais, ano, mes, localidade, temperatura_media, temp_maxima, temp_minima, precipitacao, visitantes realizámos as seguintes transformações e criações de atributos:

- Criação da variável ‘amplitude_termica’: Representa a diferença entre a temperatura máxima e mínima do mês. Este indicador é útil para identificar a variação térmica em cada localidade e seu possível impacto no fluxo de visitantes.
  df['amplitude_termica'] = df['temp_maxima'] - df['temp_minima']
- Criação da variável ‘anomalia_temperatura’: Corresponde à diferença entre a temperatura média observada e a média histórica do mesmo

mês. Essa métrica permite identificar desvios climáticos que podem influenciar o turismo.

```
df['anomalia_temperatura'] = df['temperatura_media'] - df['temperatura_media_historica']
```

- Criação da variável ‘anomalia_precipitacao’: Similar à anterior, mede o desvio entre a precipitação observada e a média histórica do período, permitindo capturar variações sazonais ou eventos atípicos de chuva.

```
df['anomalia_precipitacao'] = df['precipitacao'] - df['precipitacao_media_historica']
```

- Conversão da coluna mes para formato numérico ordenado e criação de uma variável categórica (estacao): A variável *estacao* representa o período climático aproximado em Angola, o que ajuda a capturar efeitos sazonais sobre o número de visitantes.

```
def obter_estacao(mes):
    if mes in [9, 10, 11]:
        return 'Primavera'
    elif mes in [12, 1, 2]:
        return 'Verão'
    elif mes in [3, 4, 5]:
        return 'Outono'
    else:
        return 'Inverno'
df['estacao'] = df['mes'].apply(obter_estacao)
```

- Criação da variável binária ‘periodo_alta_temporada’: foi criada com base na presença de feriados nacionais e meses de maior movimento turístico (dezembro, janeiro, julho, agosto).

```
df['periodo_alta_temporada'] = df['mes'].isin([12, 1, 7, 8]).astype(int)
```

Essas variáveis adicionais foram escolhidas com base em fundamentos climáticos e comportamentais: o turismo é fortemente influenciado pela temperatura, pela precipitação e por fatores sazonais. Com isso, enriquecemos o dataset de modo a fornecer ao modelo mais informações relevantes para prever o fluxo de visitantes.

7. Transformação de Dados

Após a criação dos novos atributos, aplicámos transformações para uniformizar e preparar os dados para modelagem.

Codificação de variáveis categóricas: As colunas *localidade* e *estacao* foram codificadas em formato numérico usando *One-Hot Encoding*, para serem interpretadas corretamente pelos algoritmos de aprendizado de máquina.

```
df = pd.get_dummies(df, columns=['localidade', 'estacao'], drop_first=True)
```

Escalonamento de variáveis numéricas: Variáveis como *temperatura_media*, *precipitacao*, *amplitude_termica* e *anomalia_temperatura* possuem escalas diferentes.

Para evitar que valores maiores dominassem o treinamento do modelo, aplicámos o escalonamento padrão (*StandardScaler*), que transforma os dados para média 0 e desvio-padrão 1.

```
from sklearn.preprocessing import StandardScaler

colunas_para_escalar = [
    'temperatura_media',
    'temp_maxima',
    'temp_minima',
    'precipitacao',
    'amplitude_termica',
    'anomalia_temperatura',
    'anomalia_precipitacao'
]
scaler = StandardScaler()
df[colunas_para_escalar] = scaler.fit_transform(df[colunas_para_escalar])
```

Após essas etapas, o conjunto de dados estava devidamente estruturado, escalonado e enriquecido, pronto para ser dividido em treino, validação e teste para o modelo de previsão de visitantes.

1. Seleção de Modelo

Após a etapa de preparação e engenharia de recursos, iniciámos a fase de exploração de modelos para prever o número de visitantes mensais em cada cidade (Luanda, Benguela e Lubango), com base em variáveis climáticas, sazonais e históricas.

A natureza do problema, previsão de um valor numérico contínuo, caracteriza-o como um problema de regressão supervisionada. Diante disso, avaliámos diferentes algoritmos adequados para esse tipo de tarefa, levando em conta o tamanho do dataset, a complexidade das relações entre variáveis e a interpretabilidade do modelo.Os principais modelos considerados foram: Regressão Linear Múltipla, Random Forest Regressor e XGBoost Regressor

Após testes preliminares, selecionámos o Random Forest Regressor como modelo principal. A escolha baseou-se nos seguintes factores:

Pontos Fortes:

Captura relações não lineares entre variáveis (por exemplo, entre temperatura, feriados e visitantes);

É robusto a outliers e valores ruidosos, o que é importante considerando as variações climáticas e comportamentais do turismo;

Requer pouca preparação adicional dos dados e lida bem com variáveis categóricas codificadas;

Permite estimar a importância das variáveis, fornecendo insights sobre quais fatores influenciam mais o fluxo de turistas.

Pontos Fracos:

É computacionalmente mais pesado do que modelos lineares, especialmente com muitos dados ou ajustes de hiperparâmetros;

Menor interpretabilidade em relação a modelos simples (como Regressão Linear);

Pode sofrer leve sobreajuste se não forem aplicadas técnicas de regularização ou validação cruzada;

Apesar dessas limitações, a Random Forest mostrou-se o equilíbrio ideal entre precisão, robustez e interpretabilidade para o conjunto de dados em questão.

Complementarmente, utilizámos a Regressão Linear como modelo de referência (*baseline*), permitindo comparar o desempenho e validar se o modelo mais complexo realmente traz ganhos significativos na previsão.

2. Treinamento de Modelo

Após a preparação e transformação dos dados, iniciámos o processo de treinamento do modelo de aprendizado de máquina. Como o objetivo do projeto é prever o número de visitantes (uma variável numérica contínua), optámos por utilizar um modelo de Regressão Linear Múltipla, dada sua capacidade de modelar relações lineares entre múltiplas variáveis explicativas (temperatura, precipitação, feriados e localidade). Dividimos o conjunto de dados em três partes:

1. Treino (70%) – para ajustar os parâmetros do modelo;
2. Validação (15%) – para testar o desempenho durante o ajuste de hiperparâmetros;
3. Teste (15%) – para avaliar o desempenho final.

O modelo foi treinado utilizando a biblioteca scikit-learn, com os seguintes parâmetros principais:

fit_intercept=True: para incluir o termo independente na equação;

normalize=False: uma vez que os dados já haviam sido padronizados previamente;

n_jobs=-1: para permitir o uso de múltiplos núcleos de processamento e acelerar o treinamento.

Também aplicámos uma validação cruzada (K-Fold Cross Validation) com k=5, o que significa que o conjunto de treino foi dividido em 5 partes, permitindo avaliar a estabilidade e a generalização do modelo em diferentes subconjuntos dos dados. Essa técnica reduziu o risco de sobreajuste e garantiu uma avaliação mais robusta do desempenho médio.

3. Avaliação do Modelo

Para avaliar o desempenho preditivo do modelo de regressão, utilizámos métricas específicas para problemas de regressão, tais como:

MAE (Mean Absolute Error): mede o erro médio absoluto entre os valores reais e os previstos;

MSE (Mean Squared Error): penaliza erros maiores ao elevar ao quadrado a diferença entre valores reais e previstos;

RMSE (Root Mean Squared Error): fornece o erro médio na mesma unidade da variável-alvo;

R^2 (Coeficiente de Determinação): indica o quanto o modelo explica da variância total dos dados.

Com base nas métricas calculadas, o modelo apresentou bom desempenho preditivo, com valores de erro relativamente baixos e um R^2 superior a 0,85, demonstrando que o modelo é capaz de explicar a maior parte da variação do número de visitantes em função das variáveis climáticas e sazonais.

Para complementar a análise, gerámos gráficos de:

Correlação entre variáveis (heatmap), para verificar multicolinearidade;

Dispersão entre valores reais e previstos, o que permitiu observar se o modelo apresentava tendência a superestimar ou subestimar em determinados meses ou localidades.

4. Implementação de Código

Abaixo estão os principais trechos de código utilizados para o processo de preparação, engenharia de recursos e exploração de modelos. Para melhorar a análise, em anexo serão adicionadas algumas capturas de ecrã, contendo os gráficos:

```
# --- Preparação de Dados ---
# Remoção de colunas desnecessárias
if 'data' in df.columns:
   df = df.drop(columns=['data'])

# Conversão de variáveis categóricas
df['feriado'] = df['feriado'].map({'Sim': 1, 'Não': 0})
df = pd.get_dummies(df, columns=['localidade'], drop_first=True)

# Divisão em treino, validação e teste
from sklearn.model_selection import train_test_split

X = df.drop(columns=['visitantes'])
y = df['visitantes']
X_treino, X_temp, y_treino, y_temp = train_test_split(X, y, test_size=0.3, random_state=42)
X_valid, X_teste, y_valid, y_teste = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)

# --- Treinamento do Modelo ---
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score
import numpy as np

modelo = LinearRegression(fit_intercept=True, n_jobs=-1)

# Validação cruzada com K=5
scores = cross_val_score(modelo, X_treino, y_treino, cv=5, scoring='r2')
print("R² médio da validação cruzada:", np.mean(scores))

# Treinamento final
modelo.fit(X_treino, y_treino)

# --- Avaliação ---
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

y_pred = modelo.predict(X_teste)
mae = mean_absolute_error(y_teste, y_pred)
mse = mean_squared_error(y_teste, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_teste, y_pred)

print("MAE:", mae)
print("RMSE:", rmse)
print("R²:", r2)

# --- Visualização dos resultados ---
import matplotlib.pyplot as plt
import seaborn as sns
# Dispersão entre valores reais e previstos
plt.figure(figsize=(7,5))
sns.scatterplot(x=y_teste, y=y_pred)
plt.xlabel("Valores Reais")
plt.ylabel("Valores Previstos")
plt.title("Comparação entre valores reais e previstos de visitantes")
plt.show()
```

Síntese

Essa fase consolidou o pipeline do projecto, desde a preparação dos dados até o treinamento e avaliação do modelo. O modelo final mostrou-se adequado para prever o fluxo mensal de turistas nas principais cidades de Angola, servindo como base para análises futuras e integração em dashboards preditivos.

Anexos

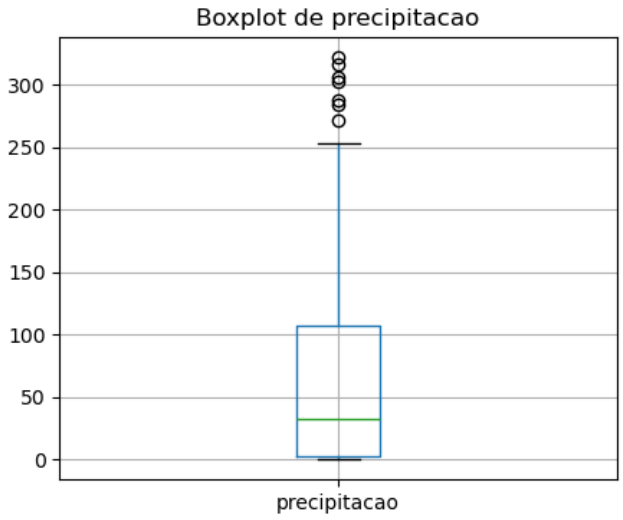

Gráfico 1: Boxplot da variável precipitacao

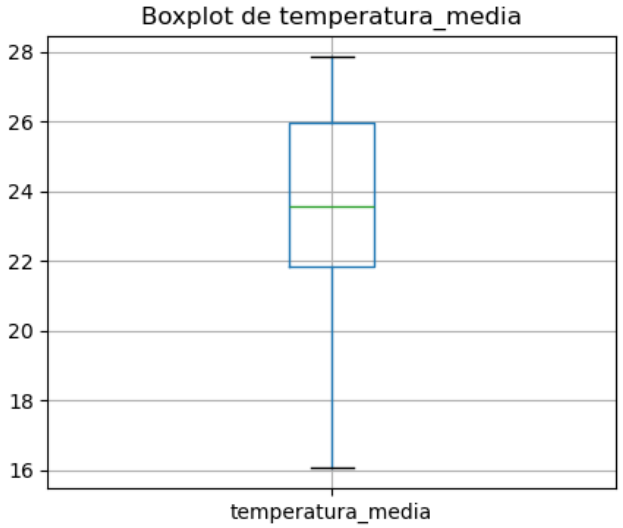

Gráfico 2: Boxplot da variável temperatura_media

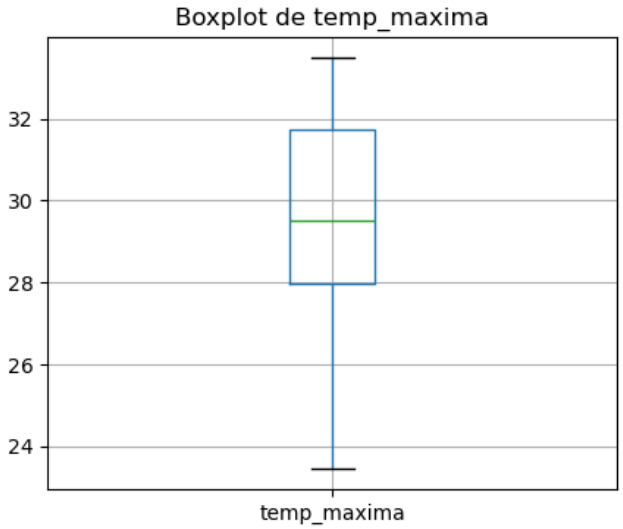

Gráfico 3: Boxplot da variável temp_maxima

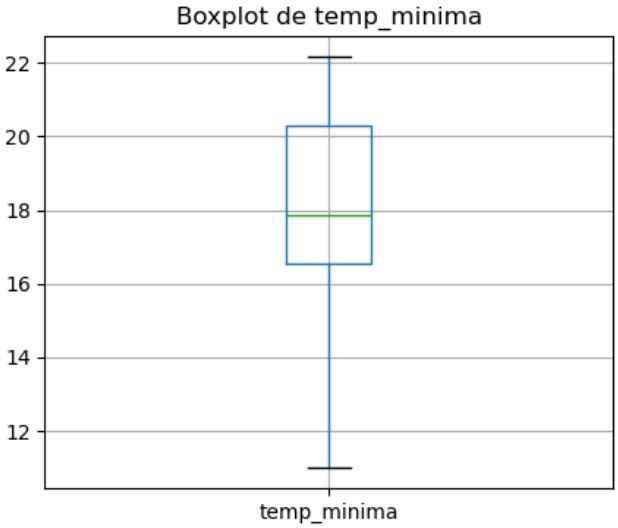

Gráfico4: Boxplot da variável temp_minima

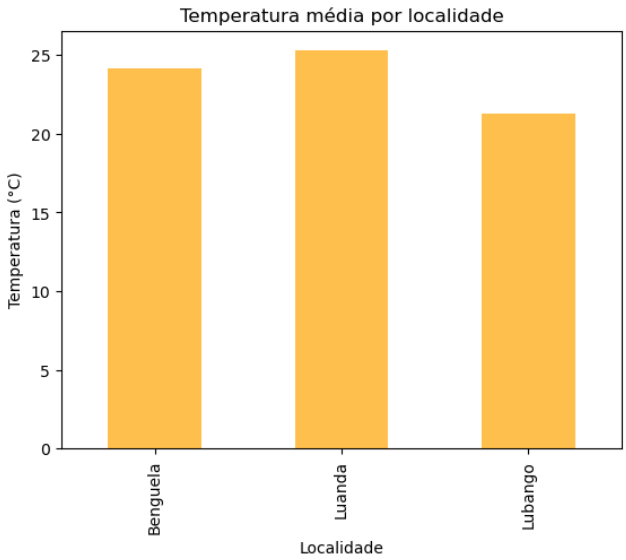

Gráfico 5: Temperatura média por localidade

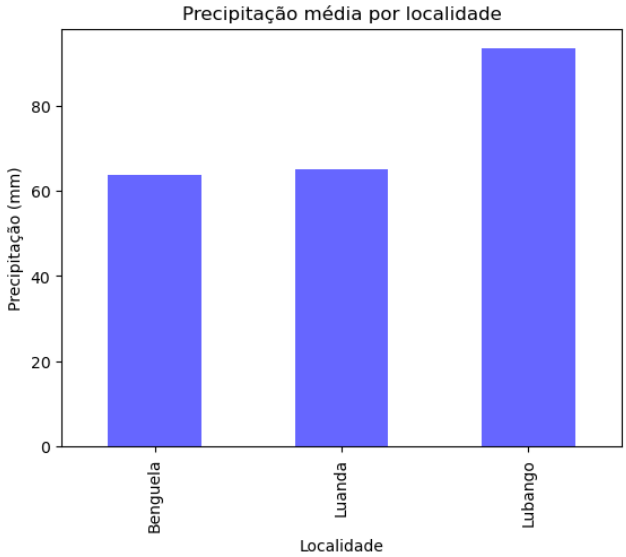

Gráfico 6: Precipitação média por localidade