



PROGRAMA DE CAPACITAÇÃO

DATA SCIENCE + DATA INTELLIGENCE



DATA SCIENCE

TRABALHO FASE 03

GRUPO 12

Keper Matheus Aquino Vida
Luciano Ferreira Fernandes
Marcelo Muniz de Alencar



PREVISÃO DA TEMPERATURA COM DADOS HISTÓRICOS

Base 017:
Daily Temperatures of Major Cities

RESUMO EXECUTIVO

oi_masterdados



FACULDADE
norte

RESUMO EXECUTIVO

Somos da G12 Turismo, nossa empresa oferece pacotes de turismo para 20 cidades dos Estados Unidos. Além de disponibilizar todos os serviços normais que as empresas de turismo oferecem, temos um grande diferencial.

Nosso modelo de previsão de temperatura utiliza Séries Temporais para fazer as previsões com base em dados históricos de 26 anos, com esta informação nossos clientes poderão se preparar bem melhor para suas viagens.

DADOS ANALISADOS

DADOS ANALISADOS

Para a realização deste projeto foi analisado dados de temperatura de várias cidades do mundo, valores informados por média diária em Fahrenheit ao longo de todos os meses e dias do ano. São dados de 26 anos (de 01-01-1995 até 13-05-2020) com informações de 321 cidades espalhadas pelo mundo.

Nesta base temos 8 variáveis que são divididas em três grupos de informações: Localização Geográfica (Região, País, Estado e Cidade), Informações de Data (Ano, Mês e Dia) e o valor da temperatura (°F). Os valores das temperaturas informadas são a média diária e estão no formato FAHRENHEIT.

DADOS ANALISADOS

Na tabela abaixo temos o Dicionário de Dados:

DICIONÁRIO - GRUPO 12 - BASE 017 Daily temperatures of Major Cities (city_temperature.csv)

VARIÁVEL	NOME VARIÁVEL	TIPO DE VARIÁVEL	DESCRIÇÃO	VALORES PERMITIDOS	POSSUI VALORES NULOS	ANOTAÇÕES
Região - Continente	Region	CATEGORICA	Região/Continente onde fica o País	Texto	NÃO	
País	Country	CATEGORICA	País onde fica a Cidade	Texto	NÃO	
Estado	State	CATEGORICA	Estado onde fica a Cidade	Texto	SIM	
Cidade	City	CATEGORICA	Nome da Cidade	Texto	NÃO	
Mês	Month	NÚMERICA	Mês em números	1-12	NÃO	
Dia	Day	NÚMERICA	Dia	1-13	SIM	
Ano	Year	NÚMERICA	Ano	0000-9999	NÃO	
Média da Temperatura	AvgTemperature	NÚMERICA	Valor da média da temperatura em °F	-999 a 999	NÃO	Necessário fazer a conversão para °C

DADOS ANALISADOS

Com estes dados verificamos os seguintes problemas de negócios que podemos abordar:

- **Aquecimento global:** Através dos dados de temperatura verificar o comportamento da temperatura ao longo do tempo e propor soluções para a diminuição do efeito estufa;
- **Informação de temperatura para viagens:** Com os dados históricos poder fazer previsões e informar quais as temperaturas máximas e mínimas das cidades no período escolhido;
- **Confecção de Roupas:** Saber a temperatura nos países para saber qual tipo de produto confeccionar. Cidades com temperaturas altas, produtos mais leves e frescos, temperaturas baixas, produtos mais resistentes ao frio.



LIMITAÇÕES DO TRABALHO

LIMITAÇÕES DO TRABALHO

Verificamos que algumas cidades apresentaram grande falta de dados, impossibilitando fazer uma previsão com qualidade.

A base de dados só forneceu dados das temperaturas até 13/05/2020, com isso foi necessário realizar a previsão desta data até 31/12/2023.



PREMISSAS E SUPOSIÇÕES

PREMISSAS E SUPOSIÇÕES

Primeiramente fazer o tratamento da base e eliminar os erros e discrepâncias.

Após a realização da análise exploratória, decidimos escolher 20 cidades nos Estados Unidos para servir de escopo para nossa análise. Escolhemos os Estados Unidos porque os dados estão mais completos, possui menos erros, maior quantidade de cidades e com grande variações de temperatura. Com isso podemos entregar um resultado mais assertivo.

Verificar o comportamento da temperatura ao longo do tempo, informar os valores de máxima, mínima e média para as datas previstas para auxiliar o público em geral quando for fazer o planejamento de viagem.



METODOLOGIA



oi_masterdados



FACULDADE
norte

METODOLOGIA

Verificamos, em um primeiro momento, que embora a base tivesse informações de um longo período de tempo (mais de 20 anos), tínhamos poucos campos (variáveis). Outro problema encontrado foram as falhas nos registros de medições das temperaturas, além da existência de datas inválidas.

Decidimos então criar um procedimento de ETL antes de darmos qualquer outro passo adiante. Imputamos novos campos a partir dos existentes, fizemos as transformações pertinentes e as limpezas necessárias. O objetivo final era criar um modelo multidimensional confiável e com informações de cluster que facilitasse as atividades tanto da área de inteligência como da área de analytics da empresa.

METODOLOGIA

CRIAÇÃO DA ÁREA STAGE

Criação de tabelas no banco DSTG_OiMasterDados tendo como origem das informações o arquivo csv.

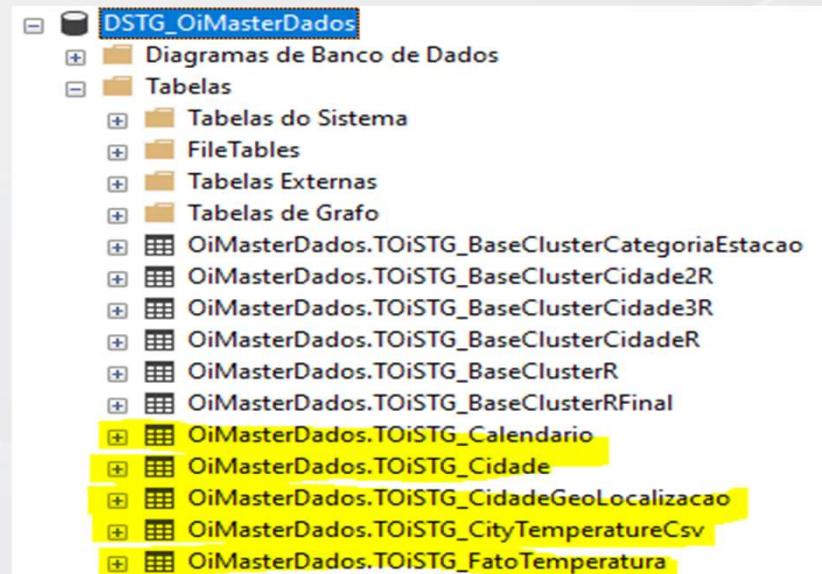
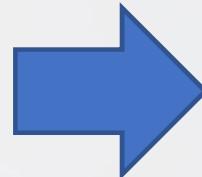
1. TOiSTG_CityTemperatureCsv;
2. TOiSTG_Calendar;
3. TOiSTG_Cidade;
4. TOiSTG_FatoTemperatura;
5. TOiSTG_CidadeGeolocalizacao.

METODOLOGIA

Criação da Área Stage (SQL Server Express)



A	Region,Country,State,City,Month,Day,Year,AvgTemperature
Africa,Algeria,,Algiers,1,1,1995,64.2	
Africa,Algeria,,Algiers,1,2,1995,49.4	
Africa,Algeria,,Algiers,1,3,1995,48.8	
Africa,Algeria,,Algiers,1,4,1995,46.4	
Africa,Algeria,,Algiers,1,5,1995,47.9	
Africa,Algeria,,Algiers,1,6,1995,48.7	
Africa,Algeria,,Algiers,1,7,1995,48.9	
Africa,Algeria,,Algiers,1,8,1995,49.1	
Africa,Algeria,,Algiers,1,9,1995,49.0	
Africa,Algeria,,Algiers,1,10,1995,51.9	
Africa,Algeria,,Algiers,1,11,1995,51.7	



METODOLOGIA

CRIAÇÃO DA ÁREA DIMENSIONAL

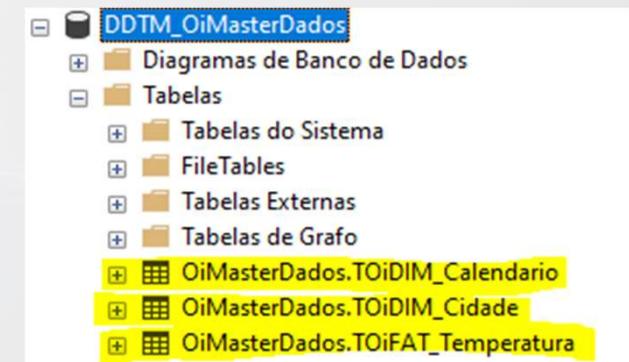
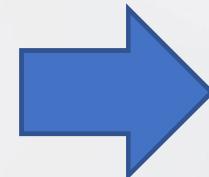
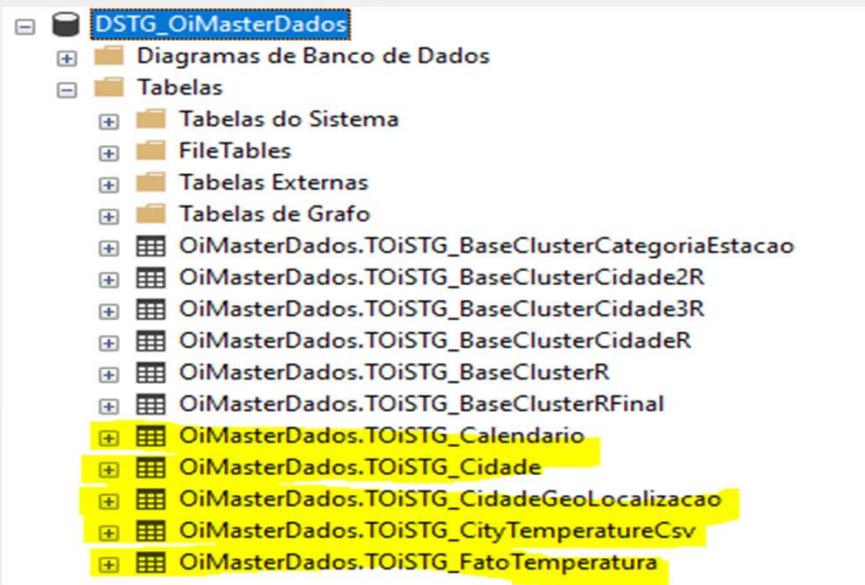
Criação de tabelas no banco DDTM_OiMasterDados tendo como origem informações da Área Stage.

1. TOiDIM_Calendario (dimensão tempo);
2. TOiDIM_Cidade (dimensão localização);
3. TOiFAT_Temperatura (tabela fato com a métrica temperatura).

METODOLOGIA

Criação da Área Dimensional (SQL Server Express)

geopy

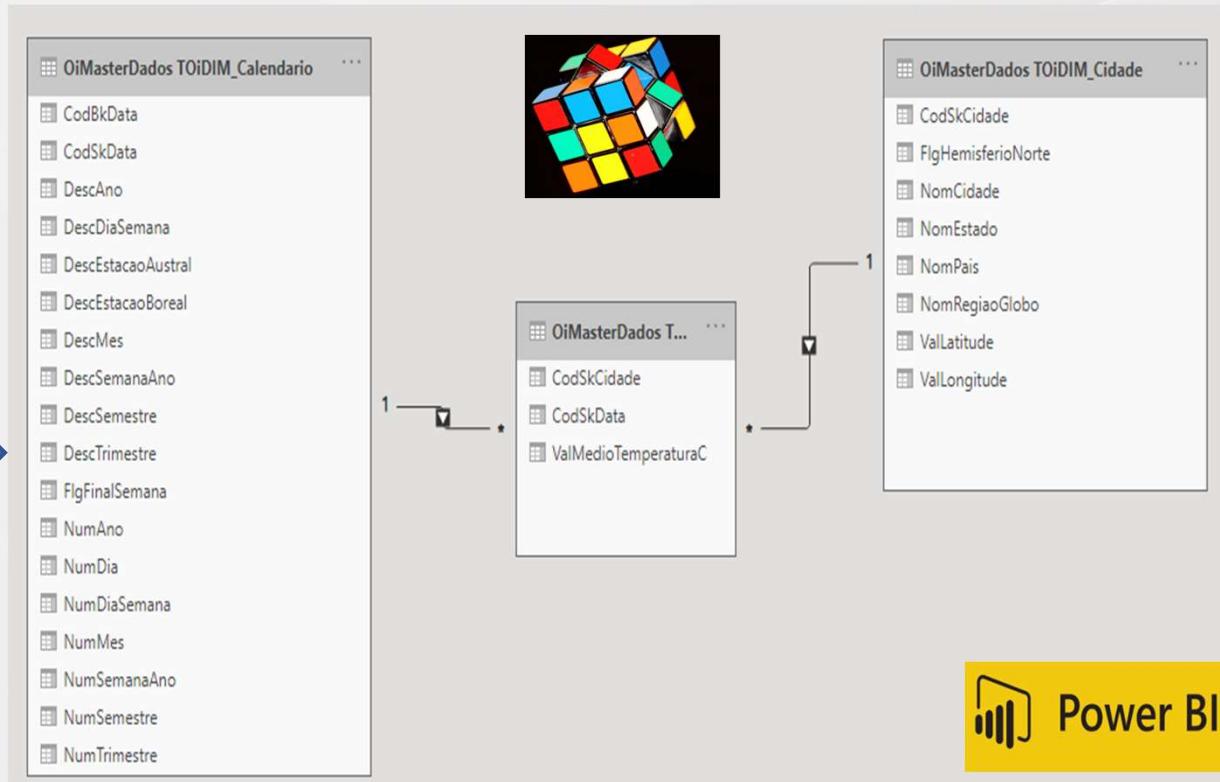


METODOLOGIA

Arquivo CSV Vs. Modelo Multidimensional



A	Region,Country,State,City,Month,Day,Year,AvgTemperature
Africa,Algeria,,Algiers,1,1,1995,64.2	
Africa,Algeria,,Algiers,1,2,1995,49.4	
Africa,Algeria,,Algiers,1,3,1995,48.8	
Africa,Algeria,,Algiers,1,4,1995,46.4	
Africa,Algeria,,Algiers,1,5,1995,47.9	
Africa,Algeria,,Algiers,1,6,1995,48.7	
Africa,Algeria,,Algiers,1,7,1995,48.9	
Africa,Algeria,,Algiers,1,8,1995,49.1	
Africa,Algeria,,Algiers,1,9,1995,49.0	
Africa,Algeria,,Algiers,1,10,1995,51.9	
Africa,Algeria,,Algiers,1,11,1995,51.7	



METODOLOGIA

CLUSTERIZAÇÃO 1

Agrupamento das estações por cidade. As estações variam de 1 a 3, sendo 1 o mais frio e 3 o mais quente. Cada cidade terá 4 registros de estação classificadas de 1 a 3. Ex: Inverno 2 (de 3), Primavera 2 (de 3), etc.

CLUSTERIZAÇÃO 2

A partir da clusterização 1, agrupamos as cidades com características semelhantes em relação às características das estações (mais frias ou mais quentes). No caso serão 5 clusters. Essa base servirá de artefato para criação dos modelos de previsão mais à frente.

METODOLOGIA

Clusterização em R



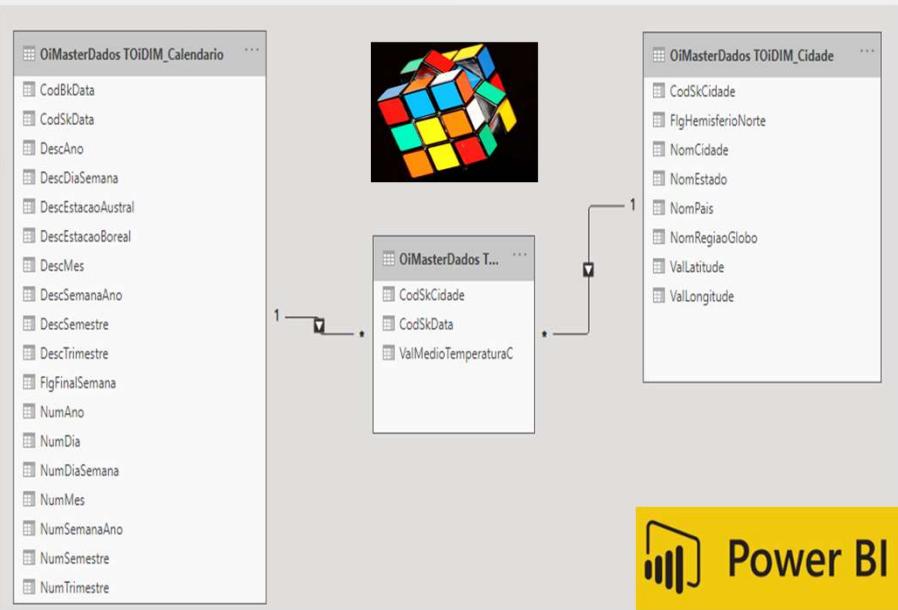
Cluster 1

	CodSkCidade	NomCidade	MinimaCidade	MediaCidade	MaximaCidade	DesvPadCidade	Categoria
1	158	Los Angeles	12,7912359550562	14,1600408157077	15,5545569620253	0,714803405694053	Inverno 2 (de 3)
2	158	Los Angeles	15,8184269662921	17,2652815628192	19,2222471910112	0,943033226475789	Outono 3 (de 3)
3	158	Los Angeles	15,3213978494624	16,5978891991988	18,0543010752688	0,776793883311791	Primavera 2 (de 3)
4	158	Los Angeles	17,8377659574468	20,4223356211393	22,1601063829787	0,948229628975923	Verão 3 (de 3)

Cluster 2

	CodSkCidade	NomCidade	Cluster
1	158	Los Angeles	1
2	1	Algiers	1
3	6	Cairo	1
4	7	Addis Ababa	1
5	13	Nairobi	1
6	14	Antananarivo	1
7	15	Lilongwe	1
8	16	Nouakchott	1
9	17	Rabat	1
10	18	Maputo	1
11	19	Windhoek	1
12	22	Dakar	1
13	24	Capetown	1
14	27	Tunis	1

	CodSkCidade	NomCidade	Cluster
1	2	Cotonou	2
2	3	Bujumbura	2
3	4	Bangui	2
4	5	Brazzaville	2
5	8	Libreville	2
6	9	Banjul	2
7	10	Conakry	2
8	11	Bissau	2
9	12	Abidjan	2
10	20	Lagos	2
11	21	Niamey	2
12	23	Freetown	2
13	25	Dar Es Salaam	2
14	26	Lome	2



oi_masterdados

METODOLOGIA

GERAÇÃO ARQUIVO CSV AJUSTADO

Com as informações do banco modelado dimensionalmente acrescidas das informações da base final de clusters, geramos um arquivo csv bem mais completo tanto para a utilização em ferramentas de visualização quanto para criação de modelos preditivos.

METODOLOGIA

Entrega Final ETL: Arquivo CSV Melhorado



SQL Server Management Studio interface showing two database structures:

- DDTM_OiMasterDados**: Contains Diagramas de Banco de Dados, Tabelas, and several tables including **OiMasterDados.TOIIDIM_Calendario**, **OiMasterDados.TOIIDIM_Cidade**, and **OiMasterDados.TOIIFAT_Temperatura**.
- DSTG_OiMasterDados**: Contains Diagramas de Banco de Dados, Tabelas, and several tables including **OiMasterDados.TOIISTG_BaseClusterCategoriaEstacao**, **OiMasterDados.TOIISTG_BaseClusterCidade2R**, and **OiMasterDados.TOIISTG_BaseClusterCidade3R**.



CSV file content (df.shape) showing data from the ETL process:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
DimCalendario_CodSkData	DimCalendario_DataCalendario	DimCalendario_DimCalendario_NumAno	DimCalendario_DescAno	DimCalendario_NumSemestre	DimCalendario_DescSesmestre	DimCalendario_DimCalendario_NumTrimestre	DimCalendario_DimCalendario													
19950101,1995-01-01,1995,Año 1995,1,1º Semestre,1,Janeiro,1,1ª Semana,1,1,DOM,S,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,17,89,INVERNO 2,11,618556																				
19950102,1995-01-02,1995,Año 1995,1,1º Semestre,1,Janeiro,1,1ª Semana,2,2,SEG,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,9,67,INVERNO 2,11,618556																				
19950103,1995-01-03,1995,Año 1995,1,1º Semestre,1,Janeiro,1,1ª Semana,3,3,TER,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,9,33,INVERNO 2,11,618556																				
19950104,1995-01-04,1995,Año 1995,1,1º Semestre,1,Janeiro,1,1ª Semana,4,4,QUA,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,8,0,INVERNO 2,11,618556																				
19950105,1995-01-05,1995,Año 1995,1,1º Semestre,1,Janeiro,1,1ª Semana,5,5,QUI,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,8,83,INVERNO 2,11,618556																				
19950106,1995-01-06,1995,Año 1995,1,1º Semestre,1,Janeiro,1,1ª Semana,6,6,SEX,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,9,28,INVERNO 2,11,618556																				
19950107,1995-01-07,1995,Año 1995,1,1º Semestre,1,Janeiro,1,1ª Semana,7,7,SAB,S,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,9,39,INVERNO 2,11,618556																				
19950108,1995-01-08,1995,Año 1995,1,1º Semestre,1,Janeiro,2,2ª Semana,8,1,DOM,S,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,9,5,INVERNO 2,11,618556																				
0 19950109,1995-01-09,1995,Año 1995,1,1º Semestre,1,Janeiro,2,2ª Semana,9,2,SEG,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,9,44,INVERNO 2,11,618556																				
1 19950110,1995-01-10,1995,Año 1995,1,1º Semestre,1,Janeiro,2,2ª Semana,10,3,TER,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,11,06,INVERNO 2,11,618556																				
2 19950111,1995-01-11,1995,Año 1995,1,1º Semestre,1,Janeiro,2,2ª Semana,11,4,QUA,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,10,94,INVERNO 2,11,618556																				
3 19950112,1995-01-12,1995,Año 1995,1,1º Semestre,1,Janeiro,2,2ª Semana,12,5,QUI,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,10,72,INVERNO 2,11,618556																				
4 19950113,1995-01-13,1995,Año 1995,1,1º Semestre,1,Janeiro,2,2ª Semana,13,6,SEX,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,8,33,INVERNO 2,11,618556																				
5 19950114,1995-01-14,1995,Año 1995,1,1º Semestre,1,Janeiro,2,2ª Semana,14,7,SAB,N,Inverno,1,Algiers,"** Estado NÃO Informado,Algeria,Africa,36.7753606,3.0601882,5,8,38,INVERNO 2,11,618556																				

df.shape
(2845082, 36)

METODOLOGIA

Selecionamos 20 cidades dos Estados Unidos, fizemos clusterização, e previsão (*Forecast*).

Foi escolhido usar como tema uma agência de turismo e fornecer informações de temperatura para seus clientes auxiliando na decisão de qual época/cidade escolher, será disponibilizado a informação histórica da temperatura e uma previsão baseado no modelo de Séries Temporais.

METODOLOGIA

Lista das cidades escolhidas:

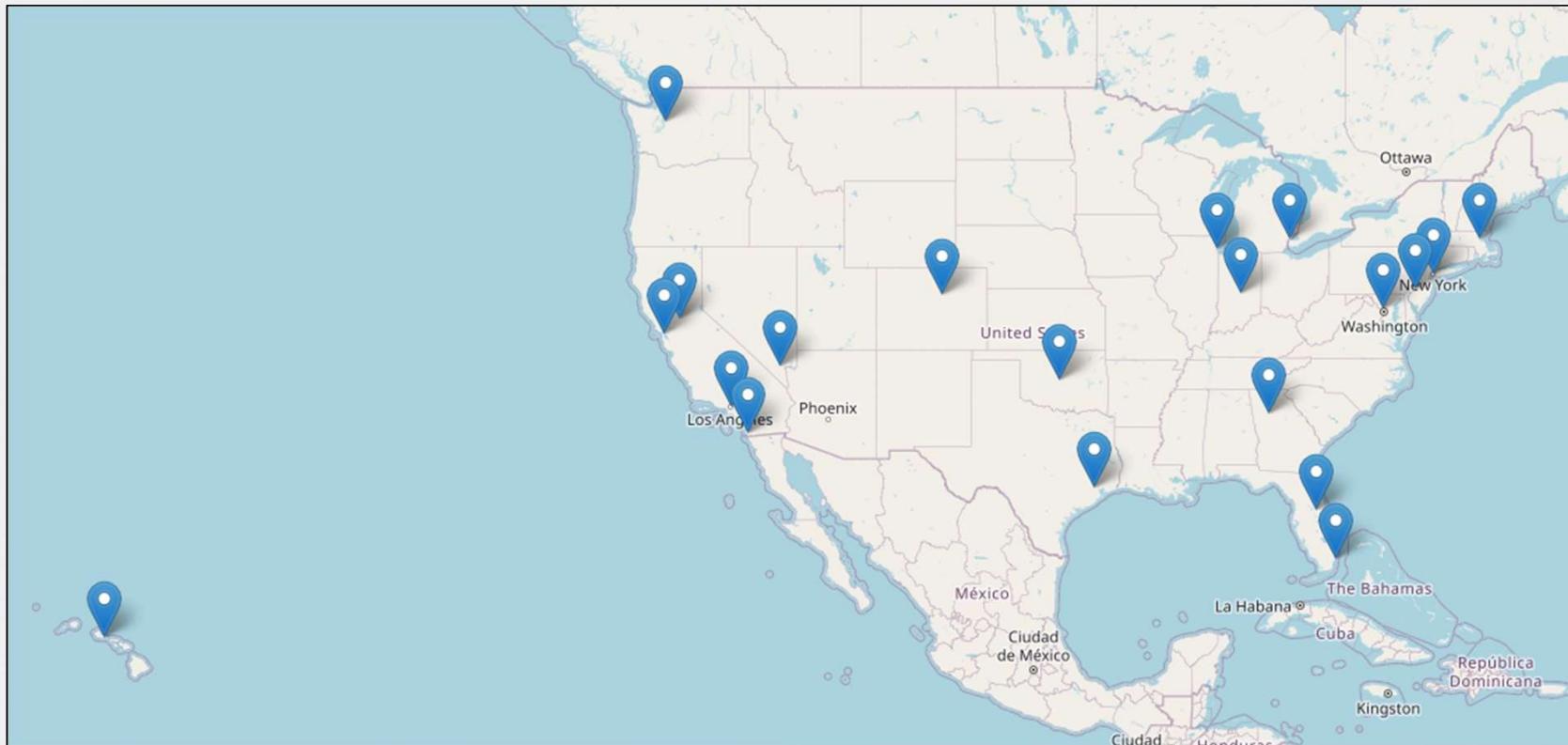
Atlanta, Boston, Chicago, Denver, Detroit, Honolulu, Houston, Indianapolis, Las Vegas, Los Angeles, Miami Beach, New York City, Oklahoma City, Orlando, Philadelphia, Sacramento, San Diego, San Francisco, Seattle e Washington DC.

Cidades por Clusters:

Cluster 1:	Cluster 2:	Cluster 3:	Cluster 4:
Houston	Honolulu	Atlanta	Boston
Las Vegas		Sacramento	Chicago
Los Angeles		San Francisco	Denver
Miami Beach			Detroit
Orlando			Indianapolis
San Diego			New York City
			Oklahoma City
			Philadelphia
			Seattle
			Washington DC

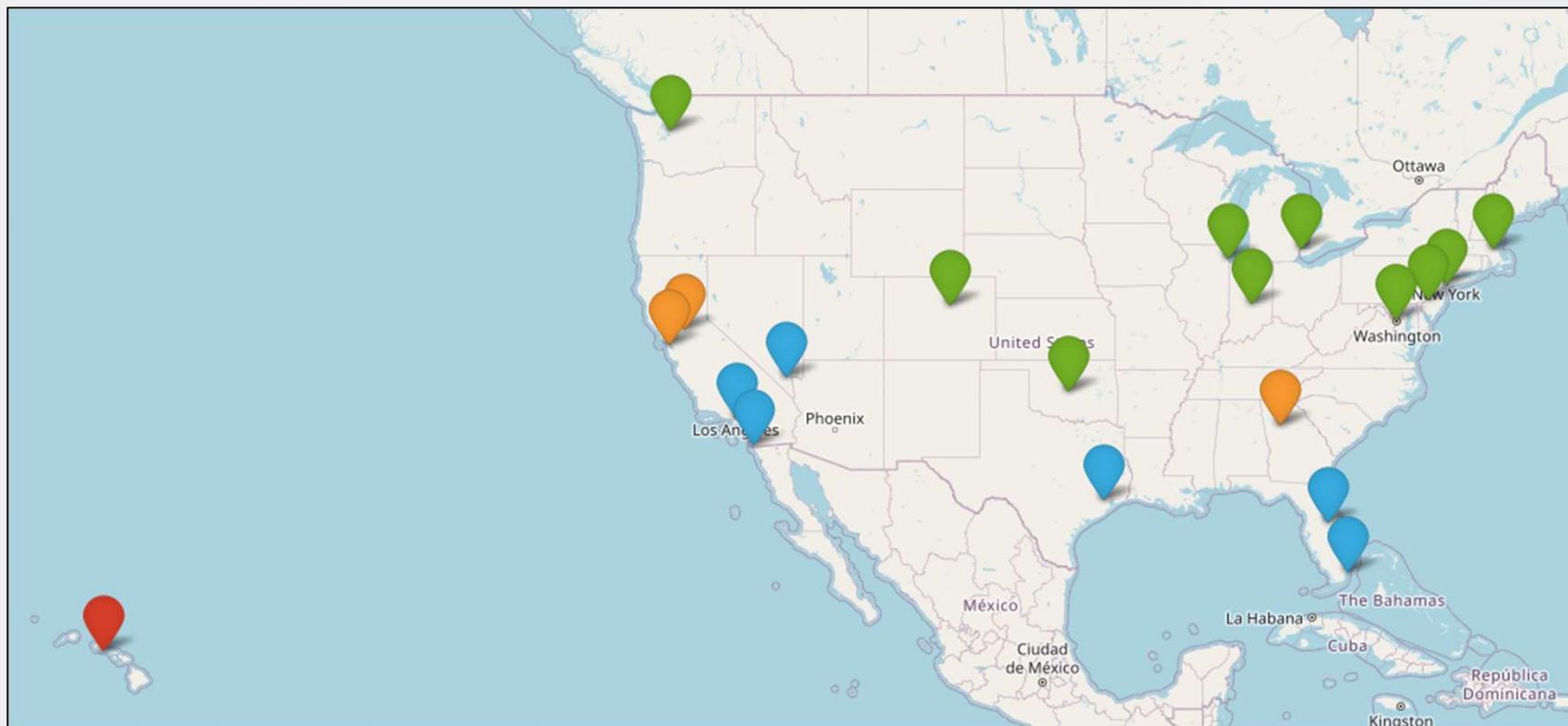
METODOLOGIA

Mapa com a localização das cidades escolhidas:



METODOLOGIA

Mapa com a localização das cidades escolhidas divididas por cluster:



Legenda:

- Cluster 1: Azul**
- Cluster 2: Vermelho**
- Cluster 3: Laranja**
- Cluster 4: Verde**

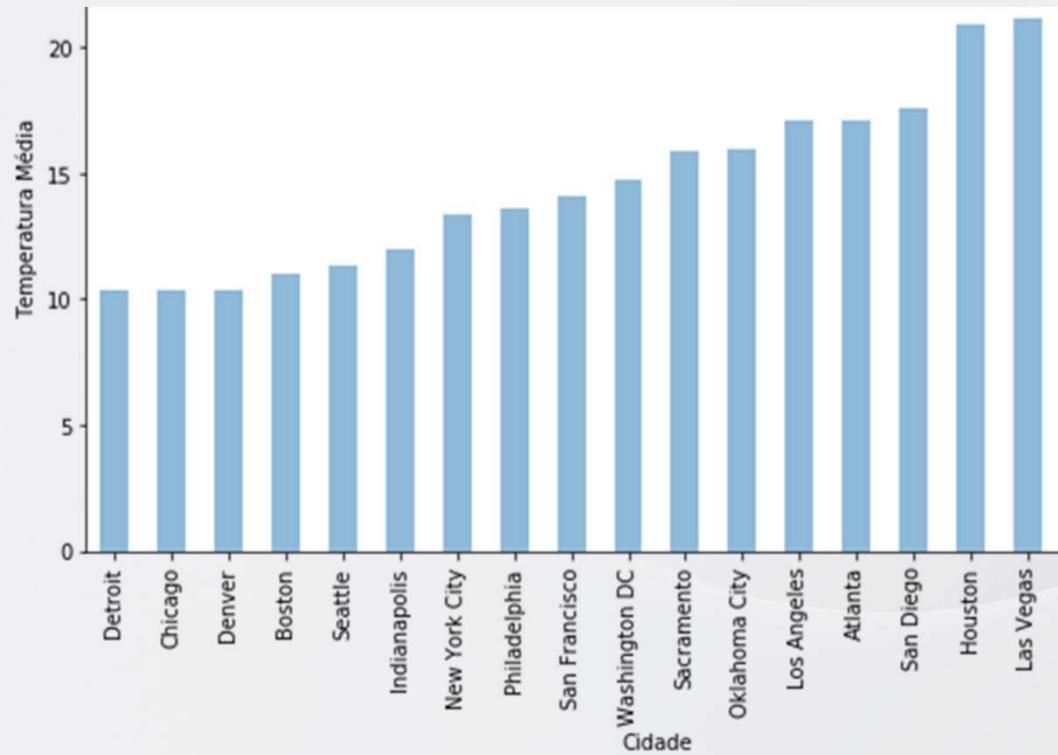
ANÁLISES REALIZADAS

ANÁLISES REALIZADAS

- Após a realização do tratamento da base e escolha das cidades, foi realizado algumas análise dos dados para verificar o comportamento dos dados e tendências.

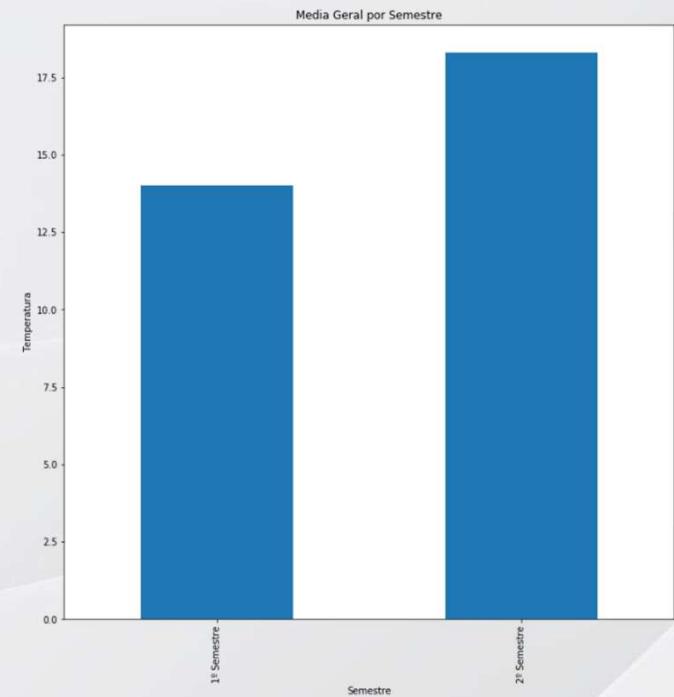
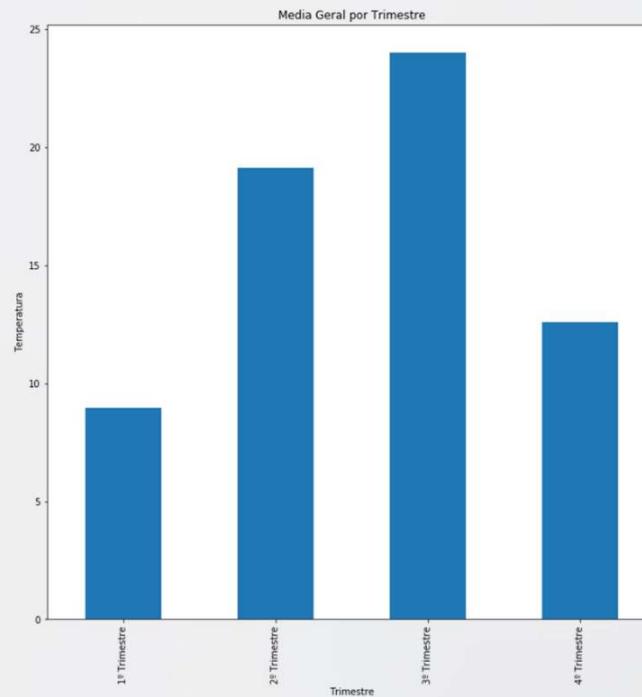
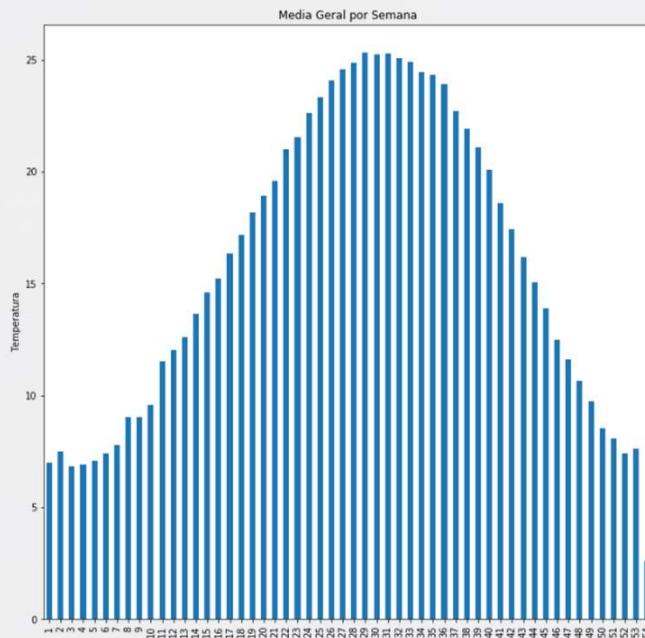
ANÁLISES REALIZADAS

- Gráfico com os valores de temperatura média por cidade de todo o período dos dados históricos.



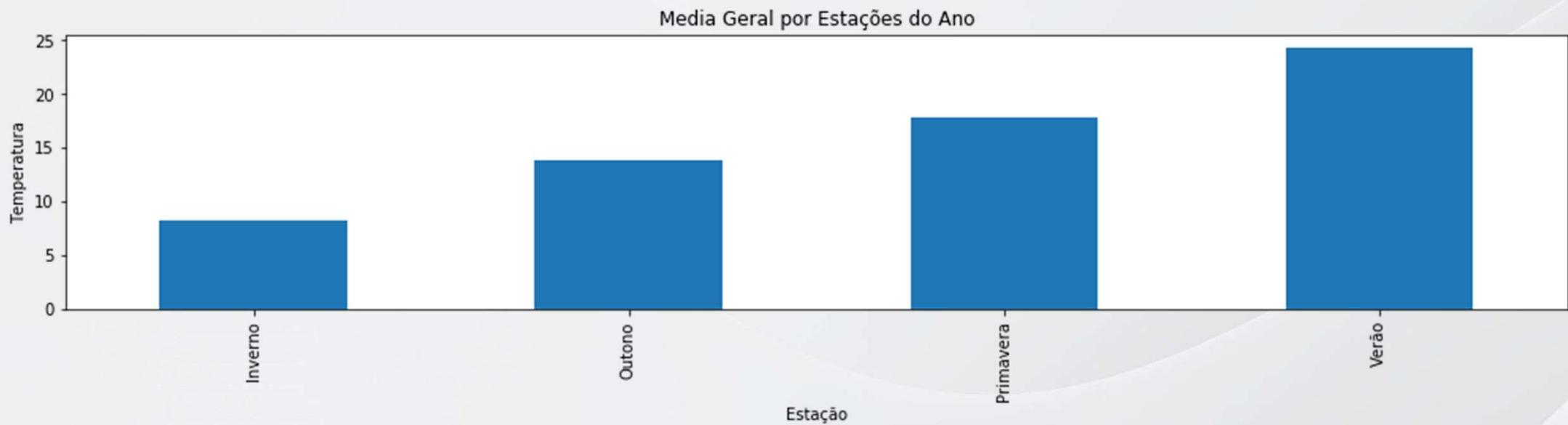
ANÁLISES REALIZADAS

- Gráfico com os valores de temperatura média de todas as cidades por Semana, Trimestre e Semestre dos dados históricos.



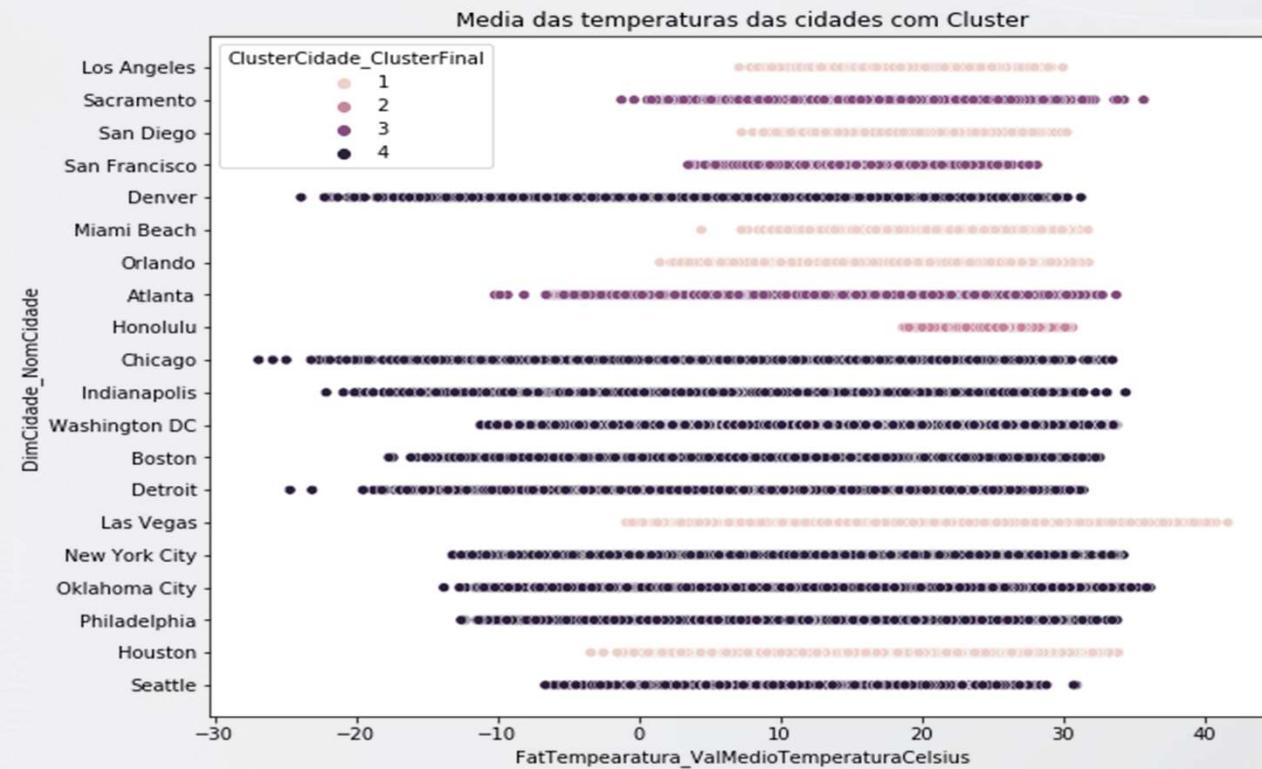
ANÁLISES REALIZADAS

- Gráfico com os valores de temperatura média de cada estação da base geral dos dados históricos.



ANÁLISES REALIZADAS

- Gráfico de Scatterplot com a média de temperatura por cidade com a informação dos Clusters.



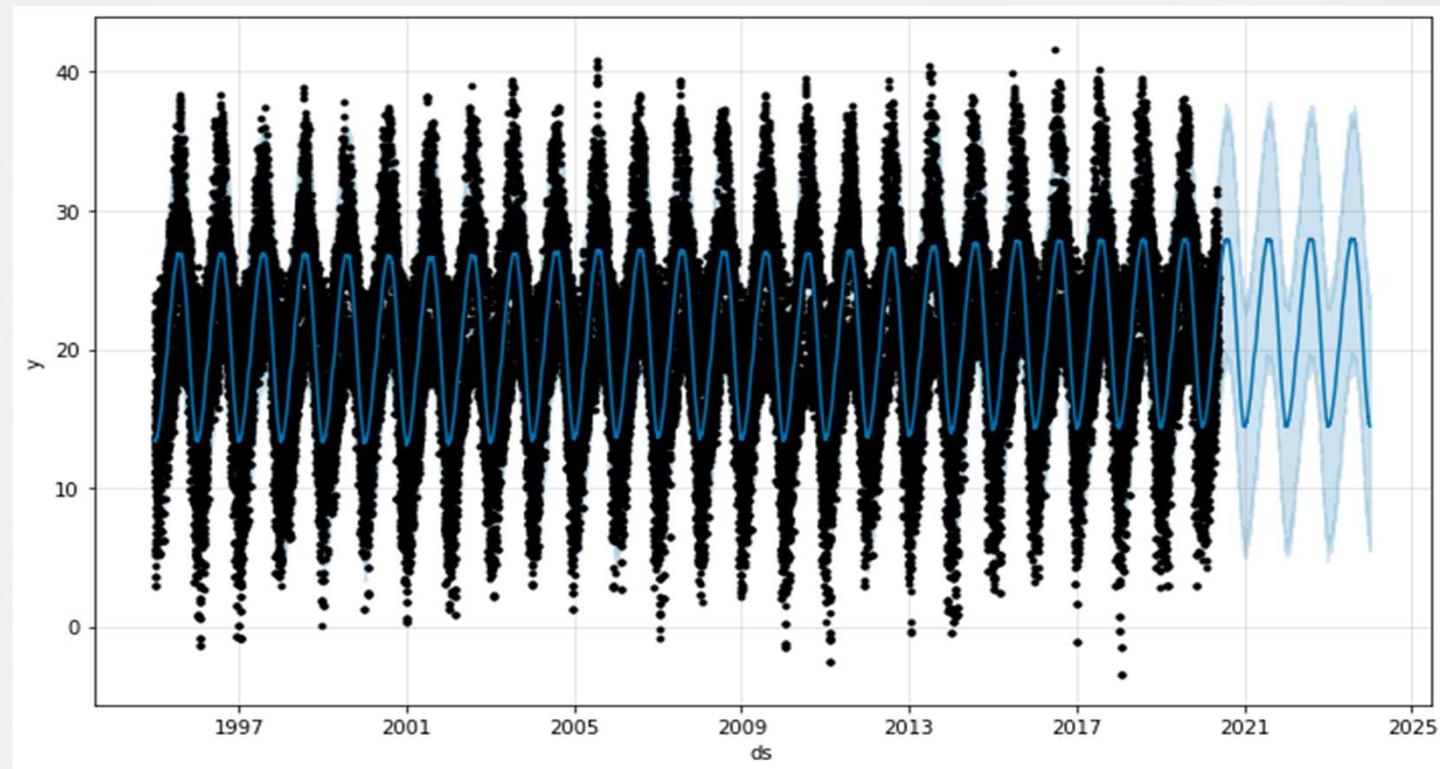
ANÁLISES REALIZADAS

- Foi utilizado o modelo **Prophet** para a realização da predição do modelo de Séries Temporais. O modelo foi rodado uma vez para cada cluster gerando assim um resultado mais assertivo.

O modelo **Phophet** foi criado pelo Facebook e segue o princípio de uma série temporal poder ser decomposta em 4 componentes: tendência, sazonalidade, uma componente que comporte as variações devido a feriados e como quarta componente, as informações de erro.

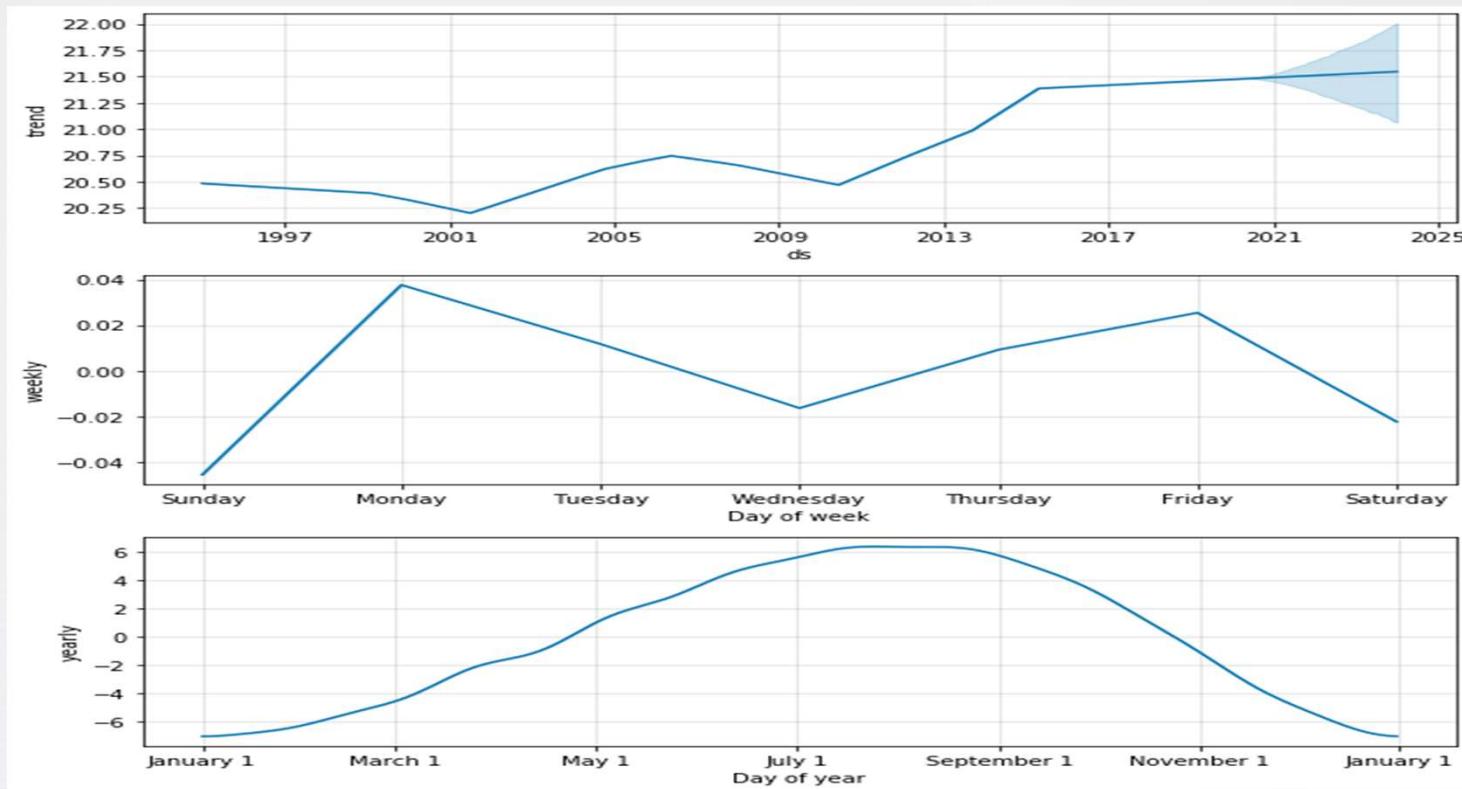
ANÁLISES REALIZADAS

- Gráfico com a previsão do cluster 1.



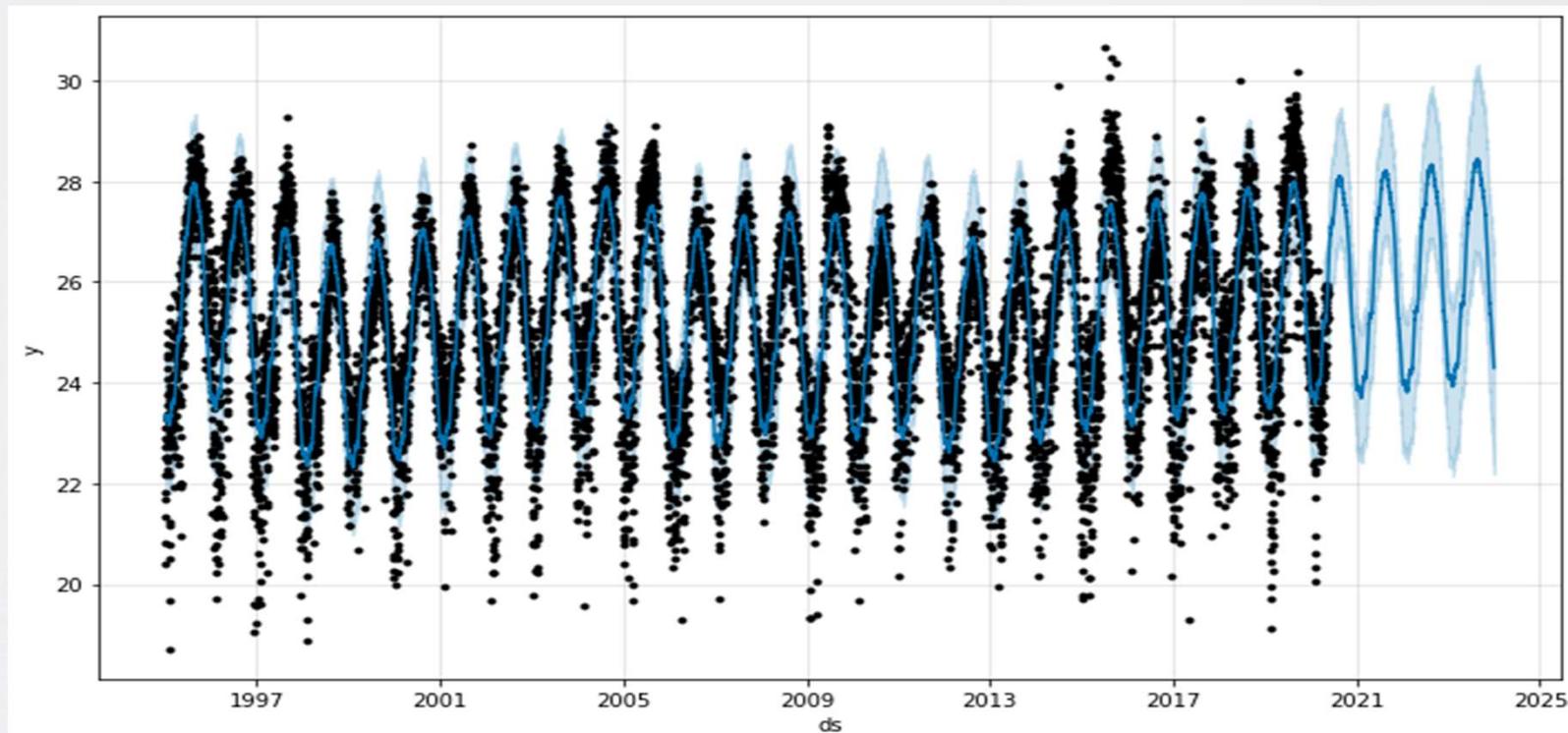
ANÁLISES REALIZADAS

- Gráficos de tendência e sazonalidade (Semanal e Anual) do cluster 1.



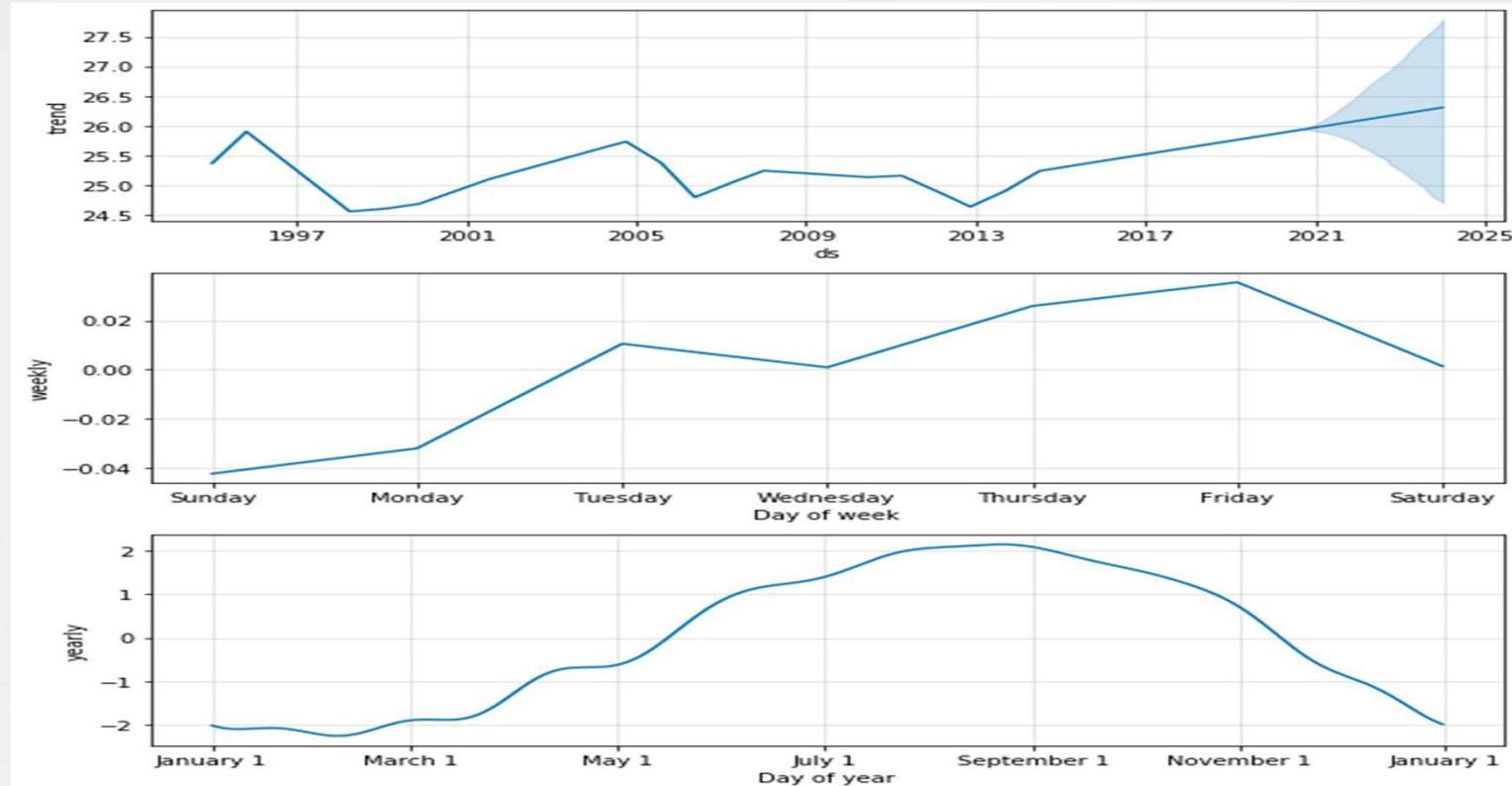
ANÁLISES REALIZADAS

- Gráfico com a previsão do cluster 2.



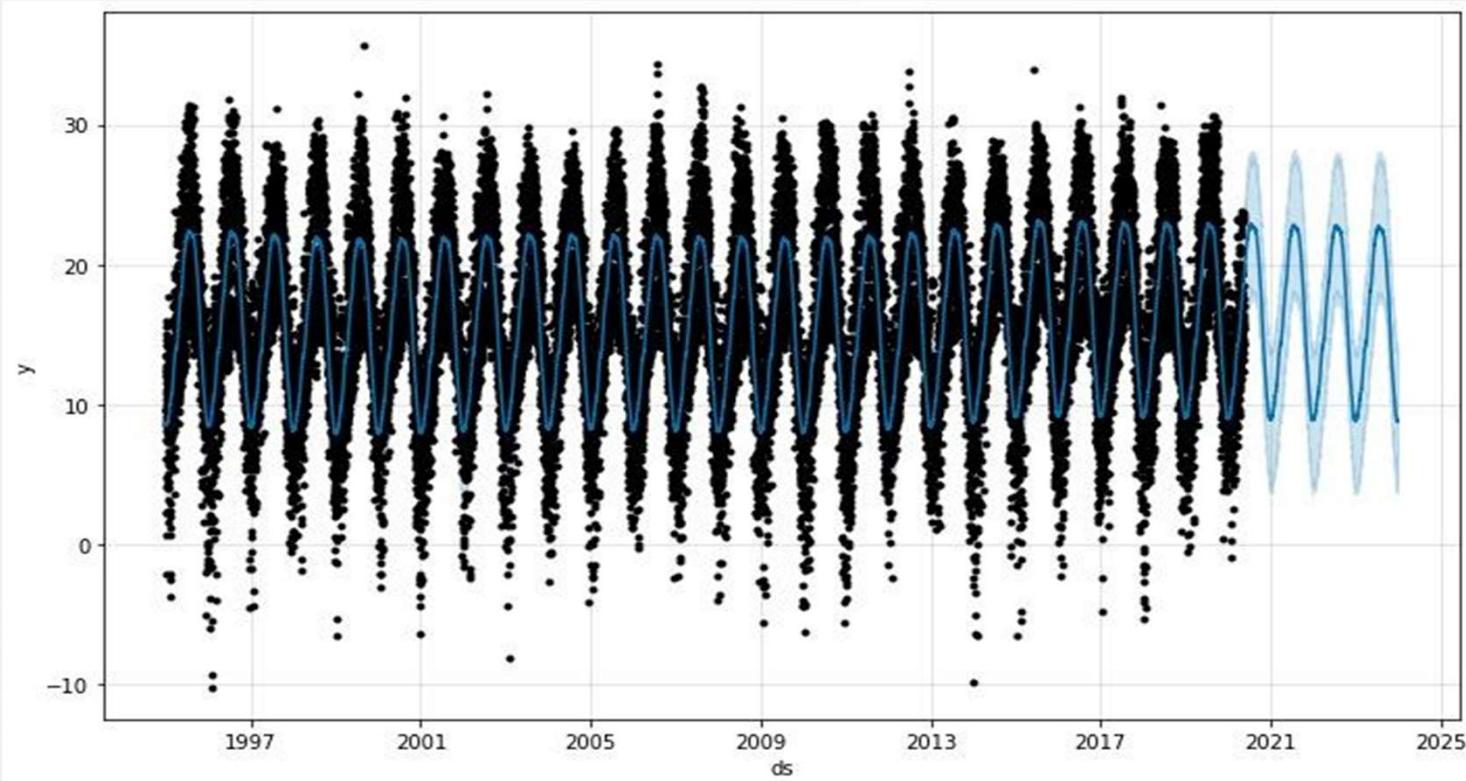
ANÁLISES REALIZADAS

- Gráficos de tendência e sazonalidade (Semanal e Anual) do cluster 2.



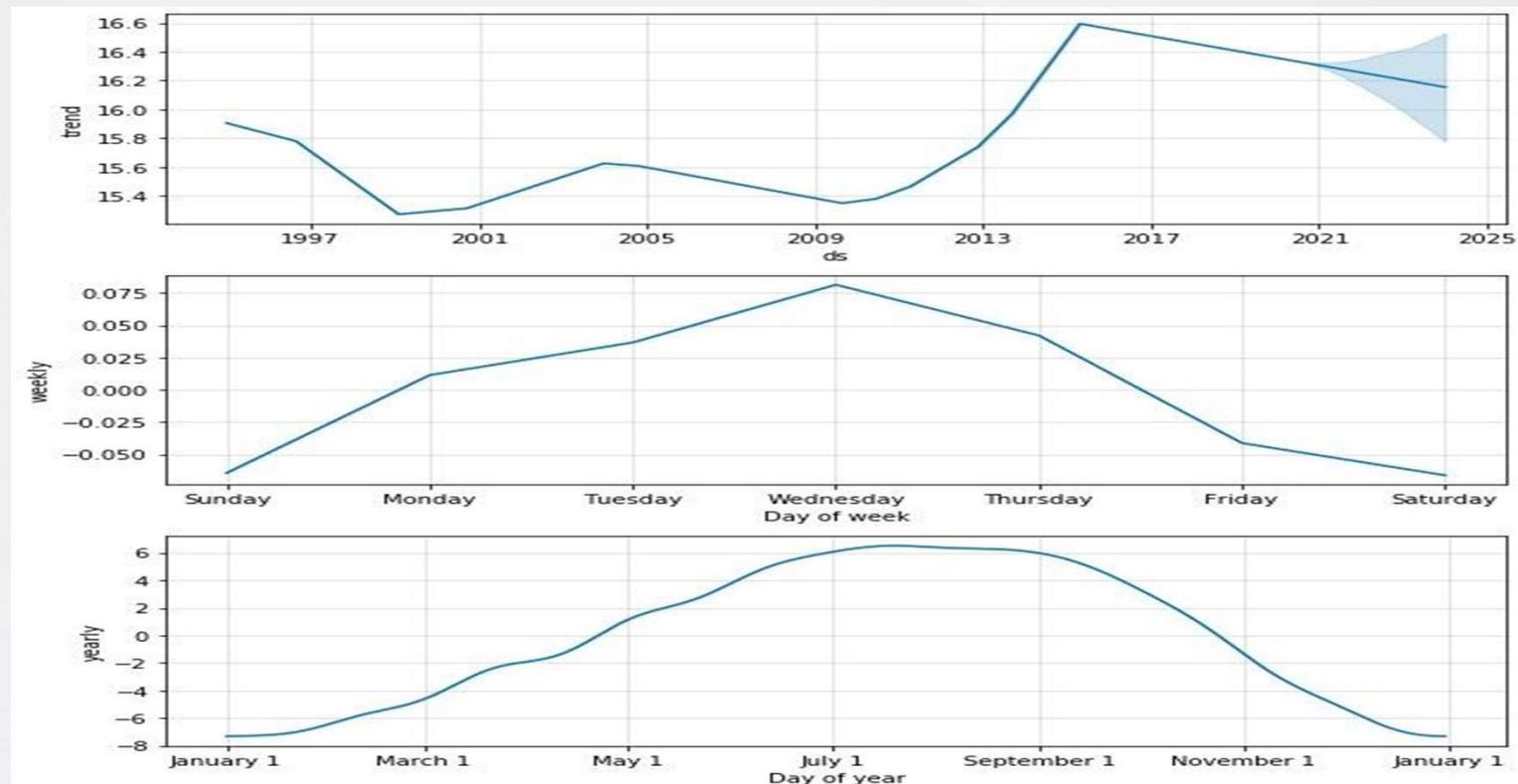
ANÁLISES REALIZADAS

- Gráfico com a previsão do cluster 3.



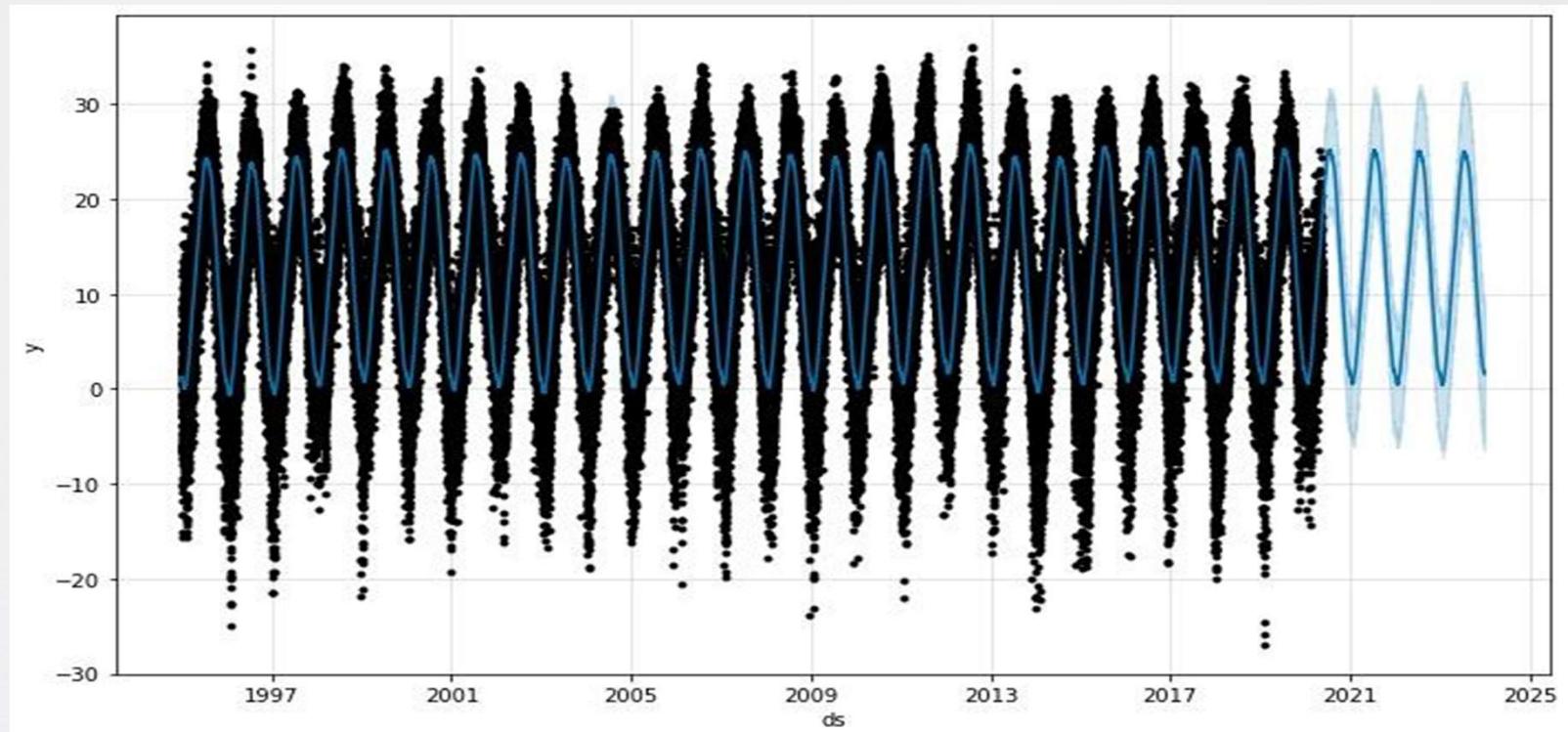
ANÁLISES REALIZADAS

- Gráficos de tendência e sazonalidade (Semanal e Anual) do cluster 3.



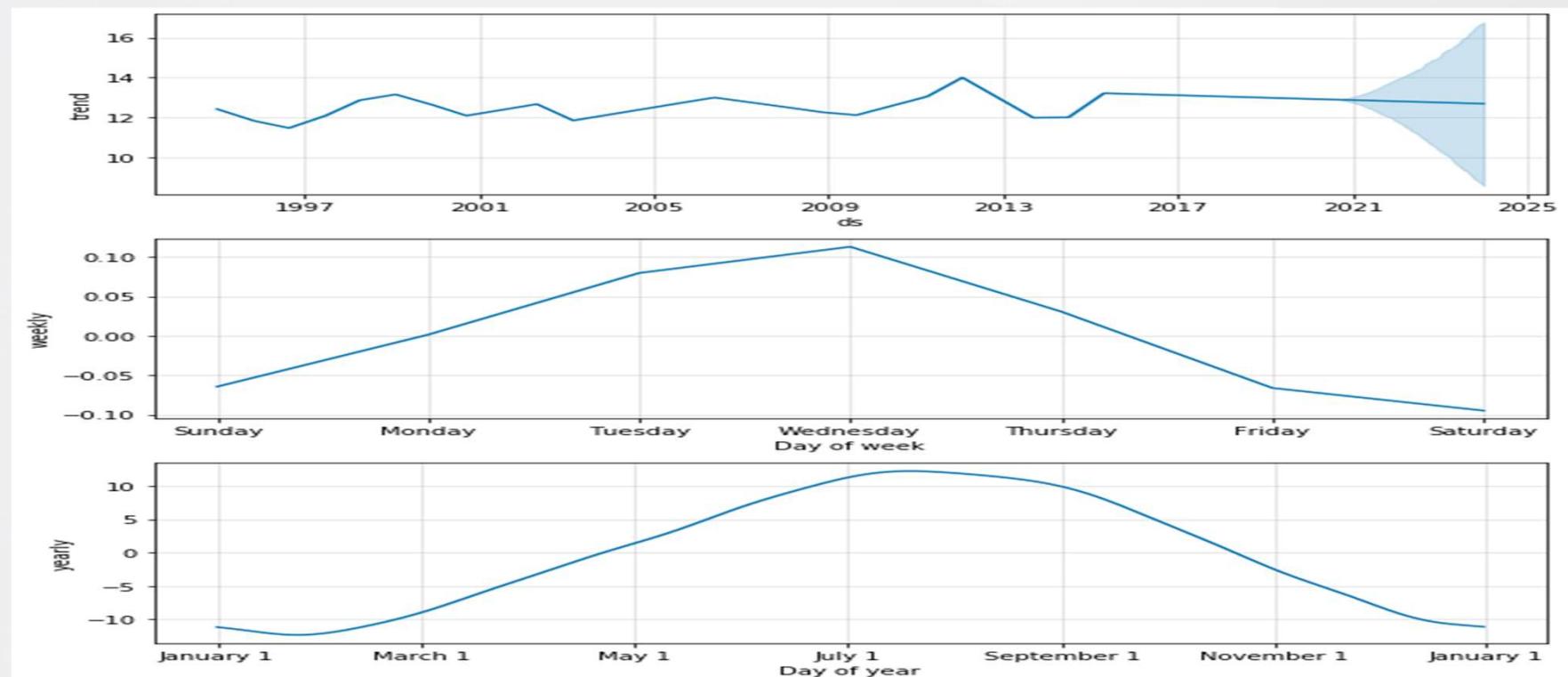
ANÁLISES REALIZADAS

- Gráfico com a previsão do cluster 4.



ANÁLISES REALIZADAS

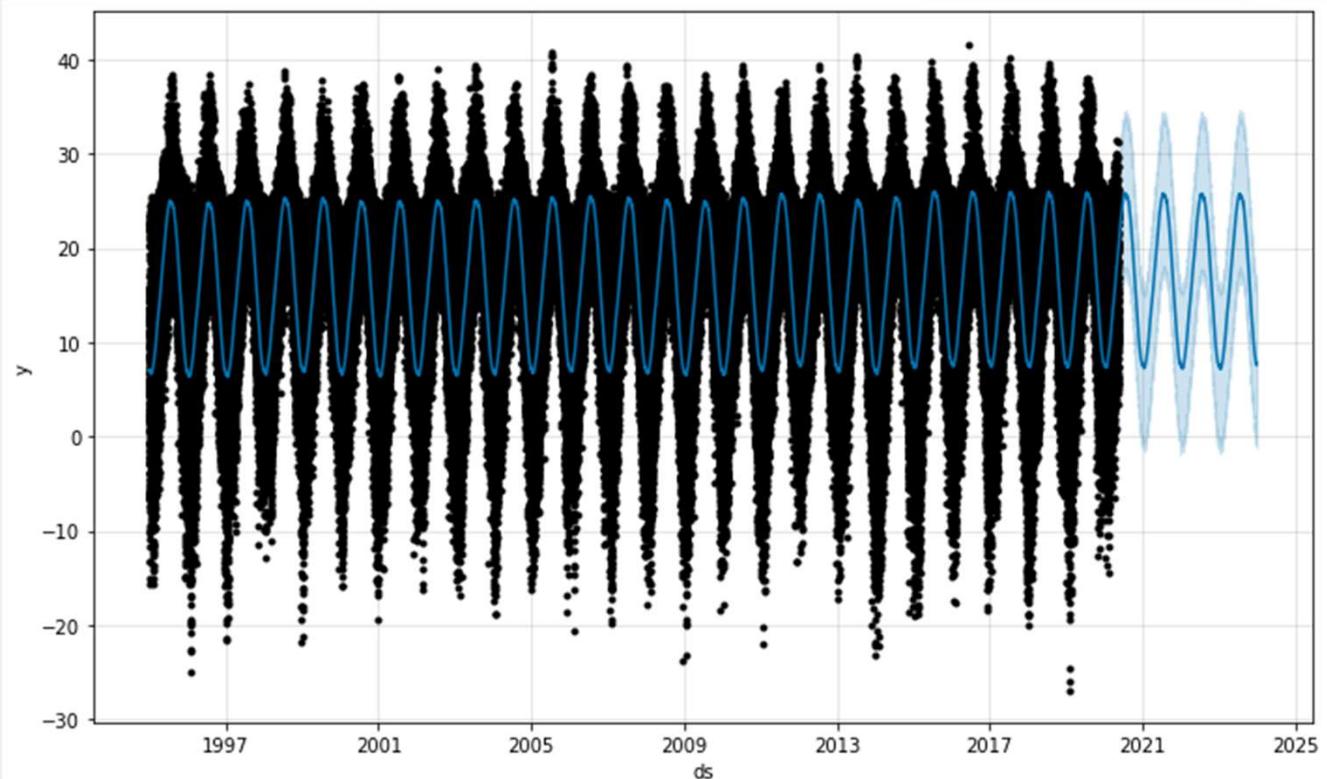
- Gráficos de tendência e sazonalidade (Semanal e Anual) do cluster 4.



ANÁLISES REALIZADAS

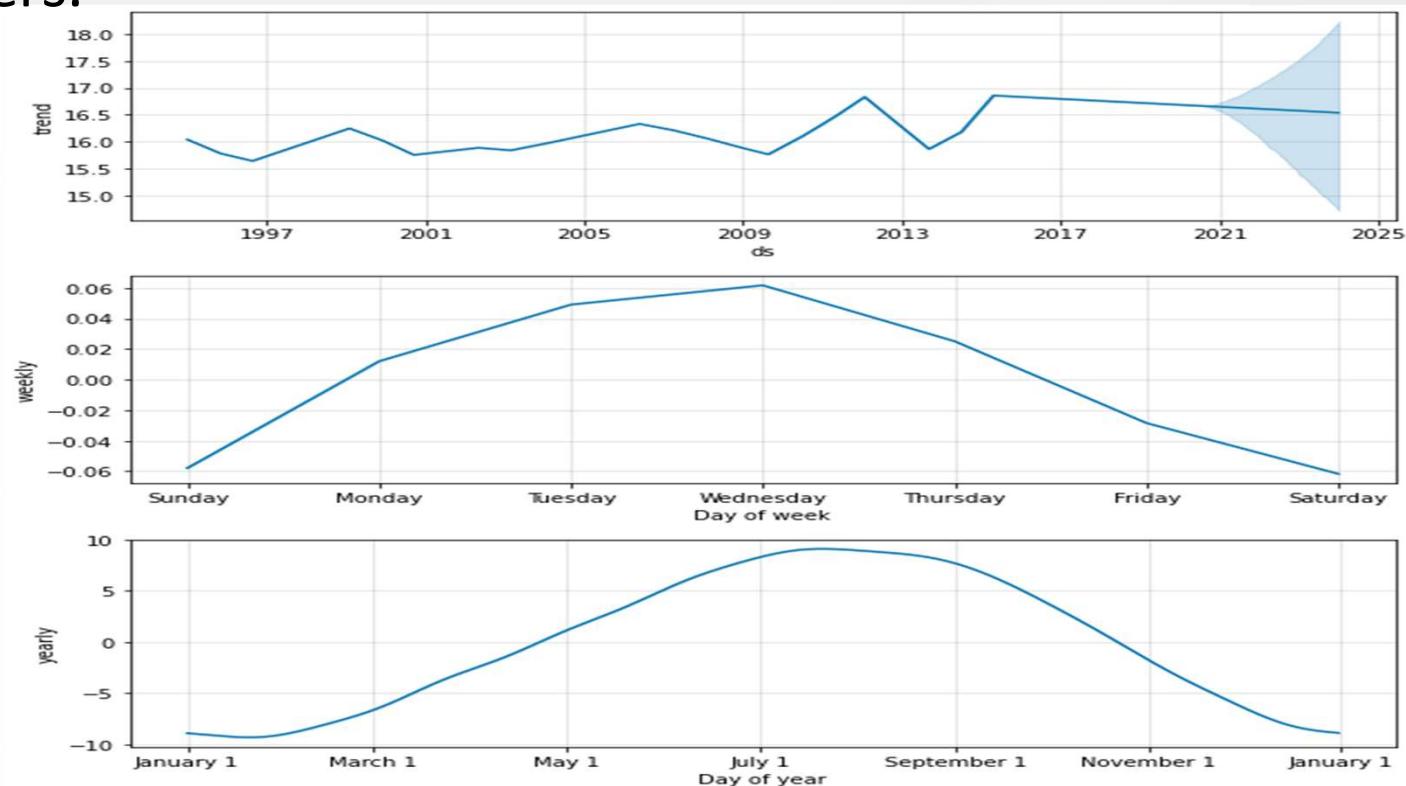
Realizamos também uma previsão com o modelo sem a separação dos clusters, com os dados de todas as cidades.

- Gráfico com a previsão da base toda sem a separação dos clusters.



ANÁLISES REALIZADAS

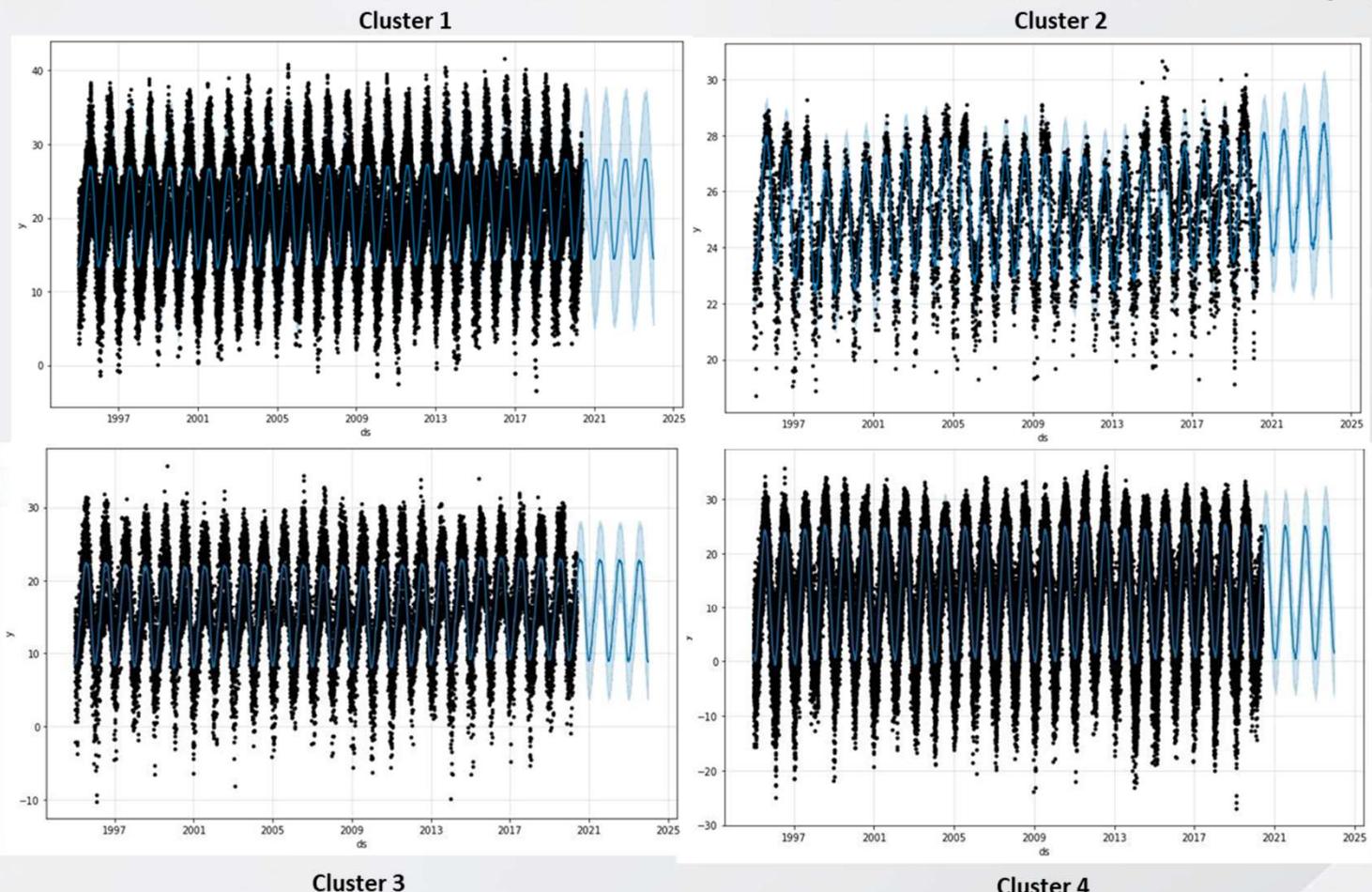
- Gráficos de tendência e sazonalidade (Semanal e Anual) sem a separação dos clusters.



ANÁLISES REALIZADAS

Ao lado podemos notar as diferenças do comportamento de cada cluster e o resultado da predição.

Cluster 1:	Cluster 2:	Cluster 3:	Cluster 4:
Houston	Honolulu	Atlanta	Boston
Las Vegas		Sacramento	Chicago
Los Angeles		San Francisco	Denver
Miami Beach			Detroit
Orlando			Indianapolis
San Diego			New York City
			Oklahoma City
			Philadelphia
			Seattle
			Washington DC



ANÁLISES REALIZADAS

• Não tiveram muitas complicações relacionadas à análise, o único fator que o algoritmo Prophet em si, solicita, é que haja colunas como o nome ‘ds’ e ‘y’, que são respectivamente as datas e o alvo, que será feita a previsão.



CONCLUSÕES RECOMENDAÇÕES ESTRATÉGIAS

CONCLUSÕES, RECOMENDAÇÕES e ESTRATÉGIAS

Com este estudo foi possível verificar a possibilidade de fazer previsões com base em dados históricos usando Séries Temporais. Observamos que com a clusterização é possível ter resultados bastante confiáveis e assertivos.

A seguir apresentaremos a interface gráfica criada para visualização do resultado obtido no estudo. De acordo com os dados de entrada, o sistema lhe mostra um resultado.

CONCLUSÕES, RECOMENDAÇÕES e ESTRATÉGIAS



Opção 1: Inserir a cidade e a temperatura desejada como entradas;

Resultado: Será informado qual a faixa de data aquela cidade terá a temperatura desejada.

Entrada:

G12 TURISMO - PREDIÇÃO DE TEMPERATURA

G12 TURISMO

Insira os dados desejados para sua viagem

Cidade:

Data Inicial:

Data Final (Max 2023-12-31):

Temperatura Mínima:

Temperatura Máxima:

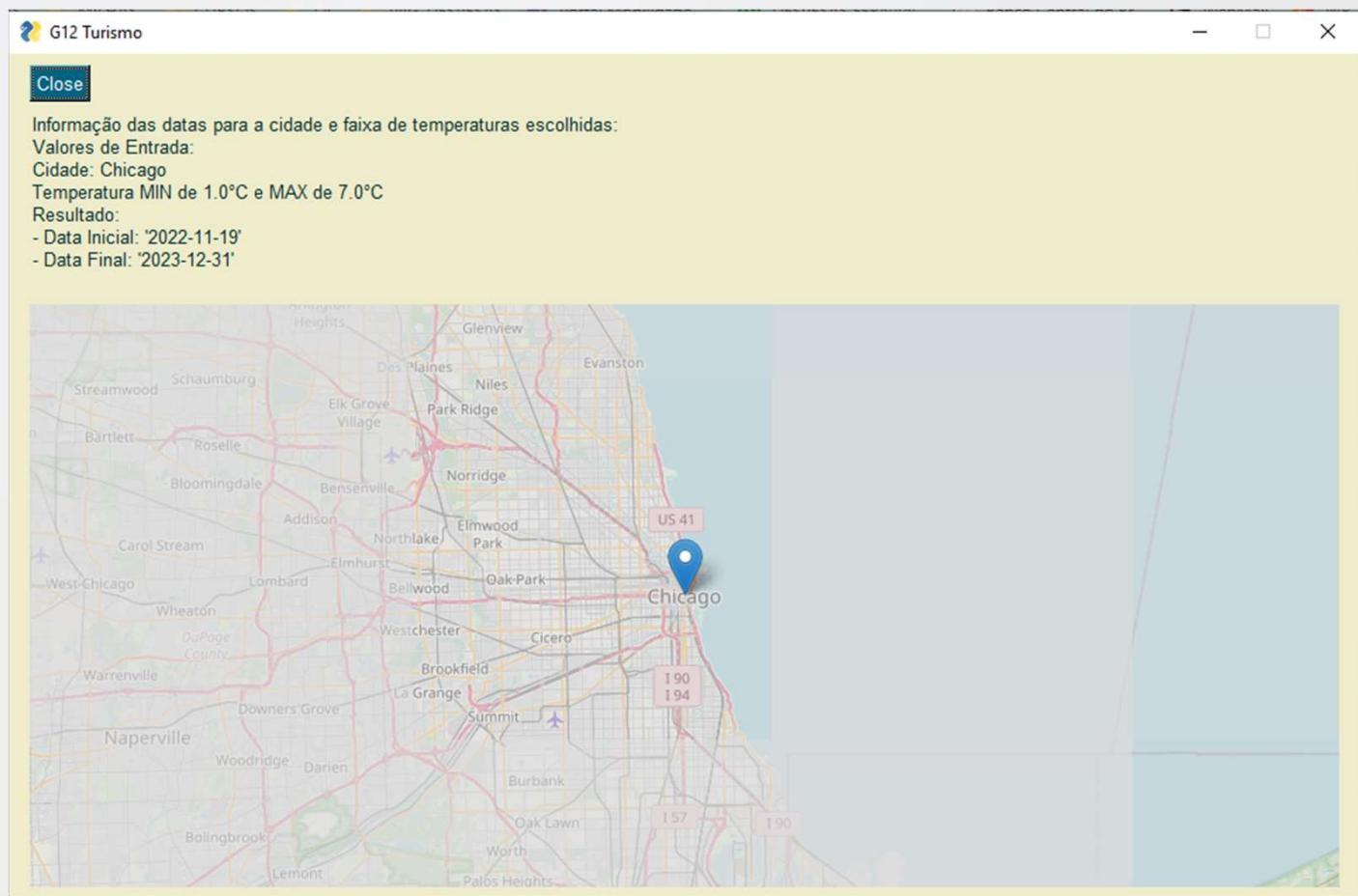
OK CANCEL

CONCLUSÕES, RECOMENDAÇÕES e ESTRATÉGIAS

Opção 1: Saída (Resultado)

- Mapa com a localização da cidade e o resultado.
 - Data Inicial e Data Final.

Saída:



CONCLUSÕES, RECOMENDAÇÕES e ESTRATÉGIAS



Opção 2: Inserir a cidade e a data inicial e final desejada como entradas;

Resultado: Será informado qual a faixa de temperatura aquela cidade terá a nas datas desejadas.

Entrada:

G12 TURISMO - PREDIÇÃO DE TEMPERATURA

G12 TURISMO

Insira os dados desejados para sua viagem

Cidade: San Diego

Data Inicial: 2023-07-12

Data Final (Max 2023-12-31): 2023-07-25

Temperatura Mínima:

Temperatura Máxima:

OK CANCEL

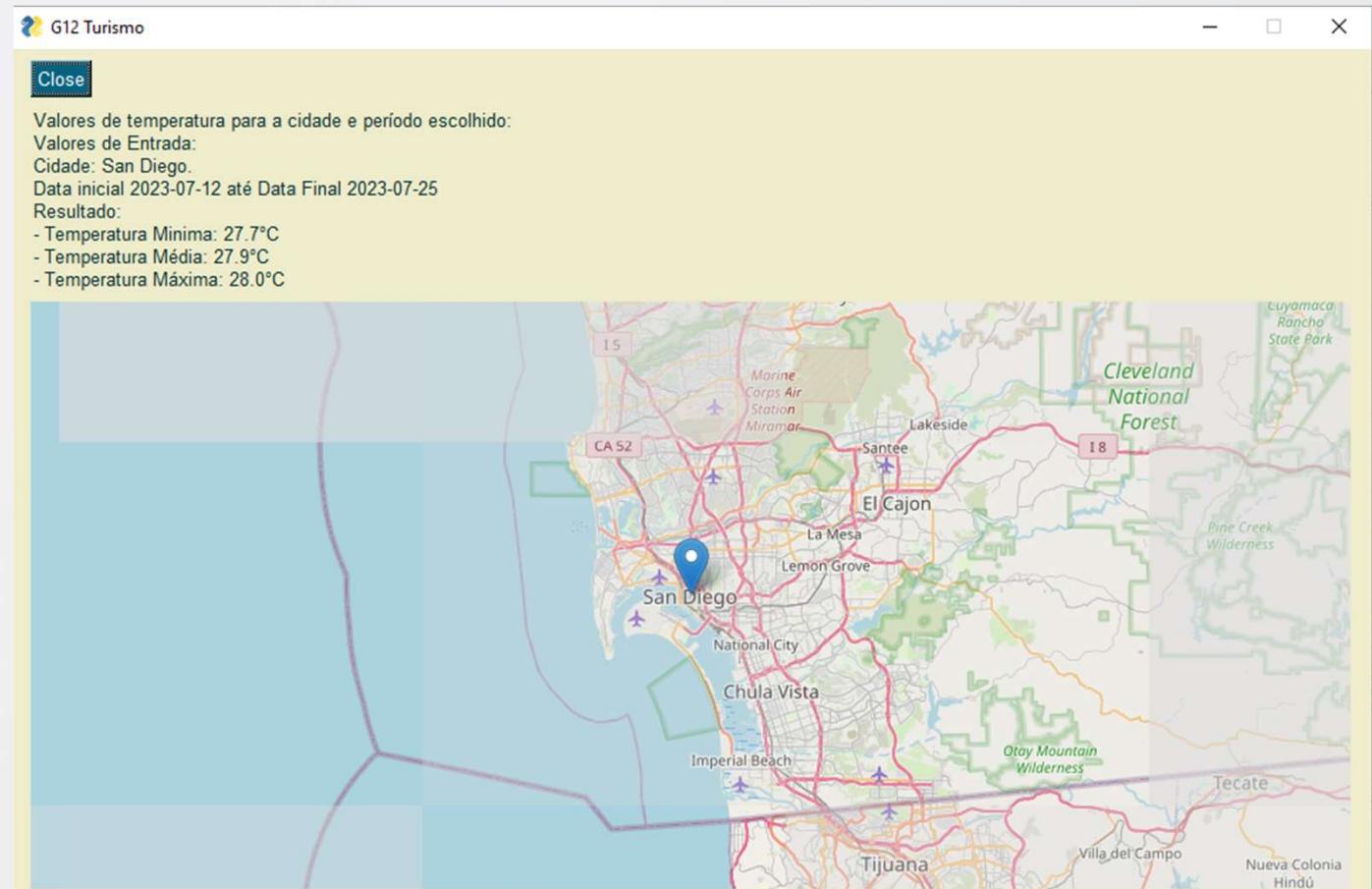
CONCLUSÕES, RECOMENDAÇÕES e ESTRATÉGIAS



Opção 2: Saída (Resultado)

- Mapa com a localização da cidade e o resultado da faixa de temperatura:
- Min, Média e Max.

Saída:



CONCLUSÕES, RECOMENDAÇÕES e ESTRATÉGIAS



Opção 3: Inserir a data inicial e final desejada e também os valores de temperatura Min e Máx como entradas;

Resultado: Será informado uma lista de cidades que terá a faixa de temperatura definida nas datas desejadas.

Entrada:

G12 TURISMO - PREDIÇÃO DE TEMPERATURA

G12 TURISMO

Insira os dados desejados para sua viagem

Cidade:

Data Inicial: 2023-12-20

Data Final (Max 2023-12-31): 2023-12-28

Temperatura Minima: 1

Temperatura Máxima: 7

OK CANCEL

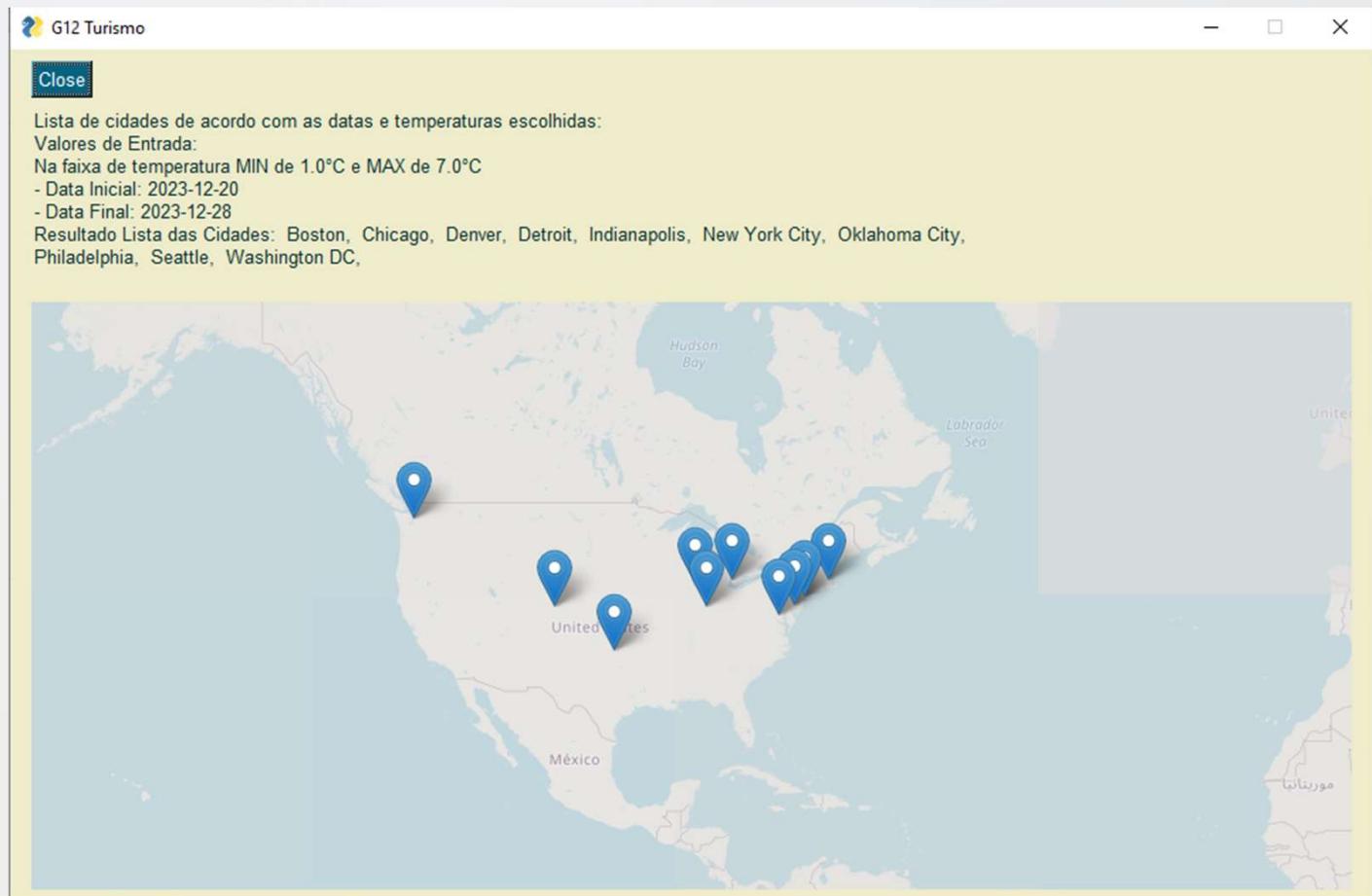
CONCLUSÕES, RECOMENDAÇÕES e ESTRATÉGIAS



Opção 3: Saída (Resultado)

- Mapa com a localização das cidades que terão a temperatura desejada nas datas informadas.
- Lista das Cidades

Saída:



CONCLUSÕES, RECOMENDAÇÕES e ESTRATÉGIAS

Escolhemos a abordagem relacionada ao turismo porque gostaríamos de apresentar uma aplicação que fosse mais presente no nosso dia a dia e que qualquer pessoa pudesse usufruir desta solução.



PRÓXIMOS PASSOS

oi_masterdados



FACULDADE
norte

PRÓXIMOS PASSOS

Primeiramente, melhorar a interface gráfica de saída do resultado, inserindo mais detalhes. Será necessário colocar alguns filtros para facilitar a entrada dos dados pelo usuário.

Inserir o data frame final em um banco de dados e fazer uma aplicação web para a interface gráfica.

Aumentar gradualmente o número de cidades disponíveis no estudo e a atualização da base das temperaturas com valores mais atuais.

