



Trabajo Práctico Final

CAMBIO CLIMÁTICO - INDIA

MANUEL HANONO - 62265

DANA NABEL - 62197

NICOLE REIMAN - 62407

SOL WINKEL - 62409



Contenido





Business Case

El cambio climático y sus efectos en diversas regiones del mundo son desafíos cruciales para el siglo XXI.

Asia | Wet bulb hot

India's deadly heatwaves are getting even hotter

The consequences of climate change will be horrific for the Indo-Gangetic



50 °C en India: ya se reportan al menos 3 muertes por la ola de calor



GETTY IMAGES

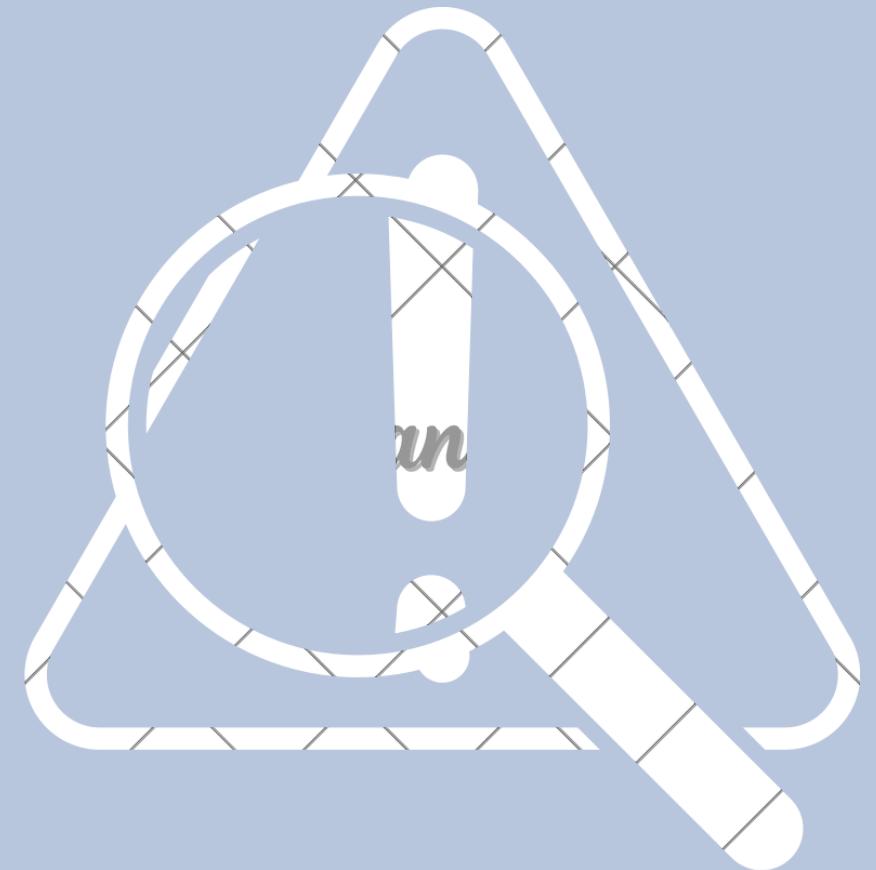
| La oficina meteorológica vaticina más días consecutivos de calor extremo en junio.

Las actuales olas de calor en India 2024
imático alarmante

La capacidad de predecir temperaturas permite a las industrias, gobiernos y comunidades anticiparse a los cambios y planificar de manera efectiva.

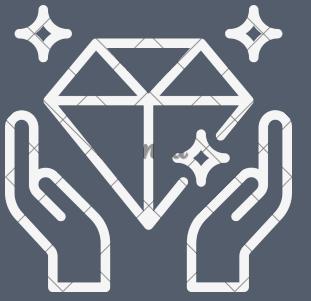
The
Economist





Problema

La necesidad de modelos precisos que predigan temperaturas promedio en base a datos históricos ya que los modelos actuales presentan limitaciones en la captura de patrones estacionales y tendencias no lineales.



Propuesta de Valor

Propuesta de valor

Evaluación de
Riesgos y
Vulnerabilidades

Planificación de
Continuidad del
Negocio

Diversificación
de Suministros

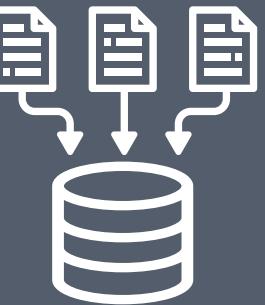
Seguro
Climático

Innovación
Tecnológica



Objetivo

Predecir el clima promedio en un país específico la cantidad de meses a futuro indicada



Base de Datos

Base de datos

- El DataFrame tiene 8.599.212 filas y 10 columnas
- Recopila información a largo plazo sobre las tendencias climáticas. Basado en el Berkeley Earth Surface Temperature Study, combina 1.6 mil millones de registros de temperatura provenientes de 16 archivos preexistentes. Incluye temperaturas de fechas de entre 1743 y 2013 para 159 países
- Variables del dataset:
 - dt (Date): Fecha en la que se observó cada temperatura
 - Average Temperature: temperatura promedio de la ciudad en grados Celsius.
 - AverageTemperatureUncertainty: intervalo de confianza del 95% alrededor de la temperatura promedio.
 - City: Ciudad.
 - Country: País.
 - Latitude: Latitud.
 - Longitude: Longitud

Limpieza y preprocesamiento de datos

DATASET CAMBIO CLIMÁTICO

Unificación formato de fechas

Había fechas en español y en inglés
01/03/2013 vs 2013-11-02

Crear un vector ensamblado

Combinar múltiples columnas numéricas en una sola columna features, que es el formato requerido para la mayoría de los algoritmos de Machine Learning en PySpark.

Limpieza de coordenadas

formato geometry

Acotación de la base

Tenía datos de muchos países y se filtró para contemplar solamente a India.

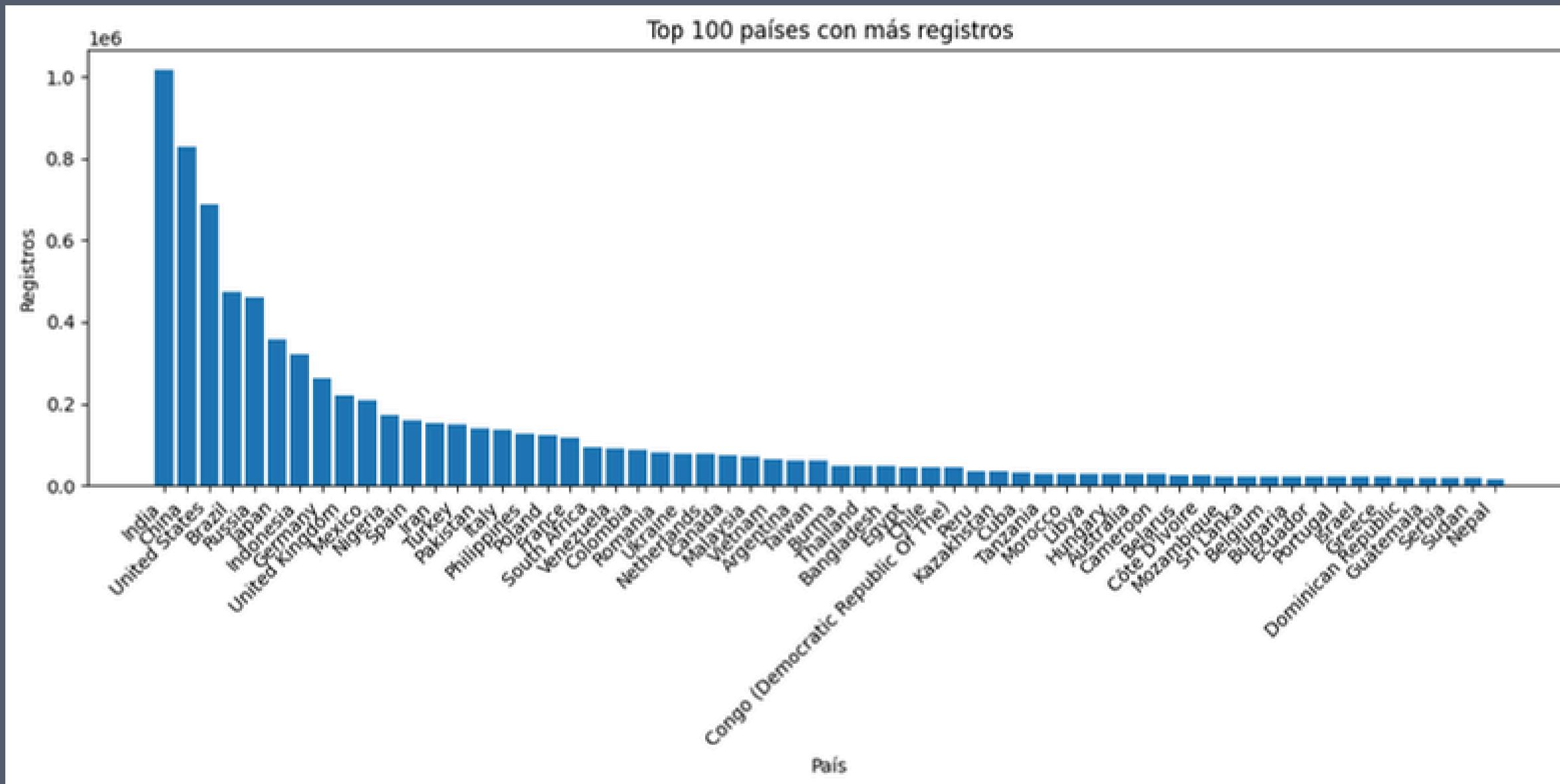
Tratamiento de missings

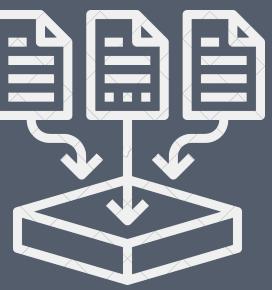
Los filtramos por ser el 5% de los datos.

Indexar la columna categórica 'City'

Utilizamos StringIndexer para asignar un índice numérico único a cada ciudad.

Filtro por país





Serie de tiempo

Serie de tiempo - India

DATASET CAMBIO CLIMÁTICO



1.014.906 registros de india

Los datos comprenden desde el SXVIII hasta el año 2013.

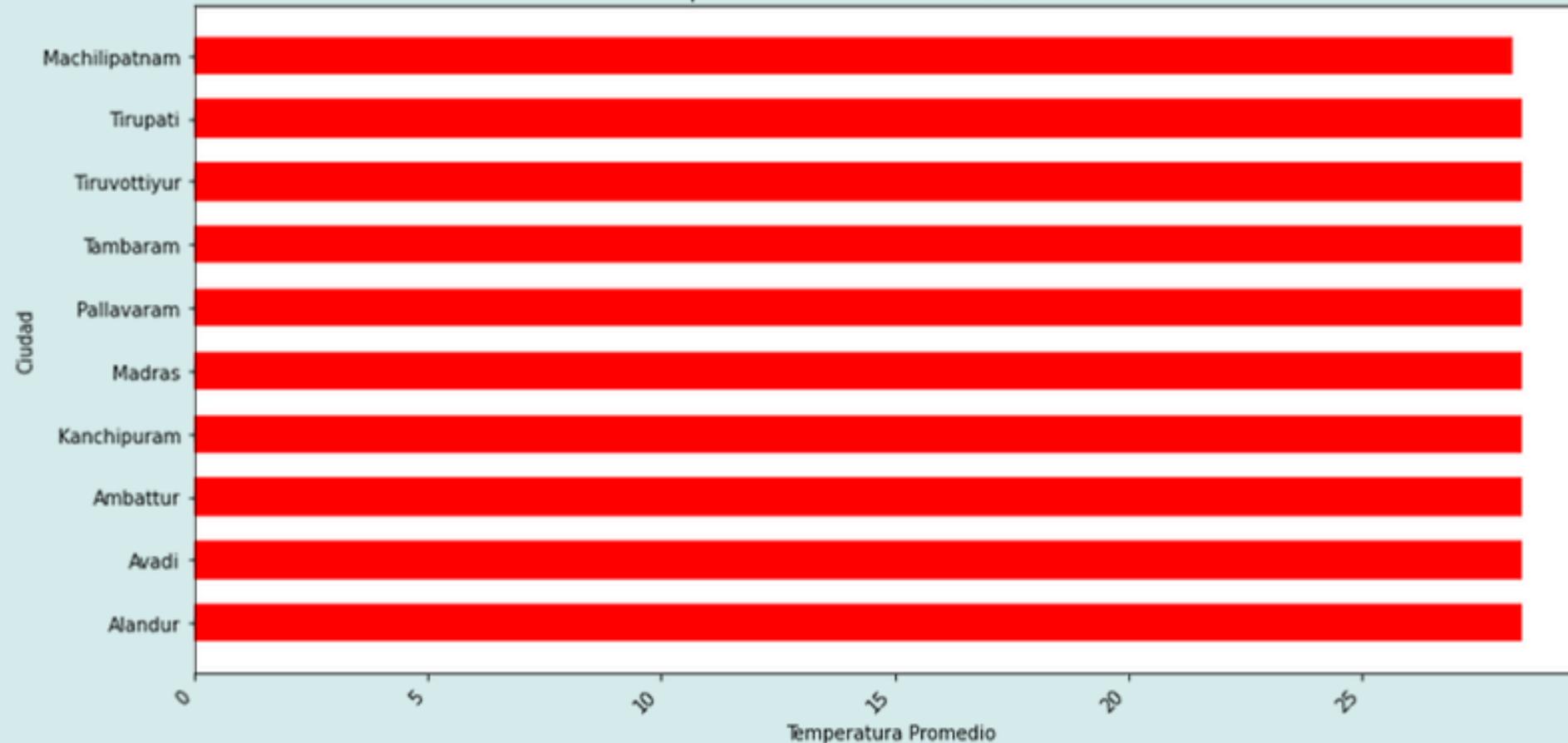
391 ciudades de India

El dataset incluye 391 ciudades de India, incluyendo los puntos cardinales.

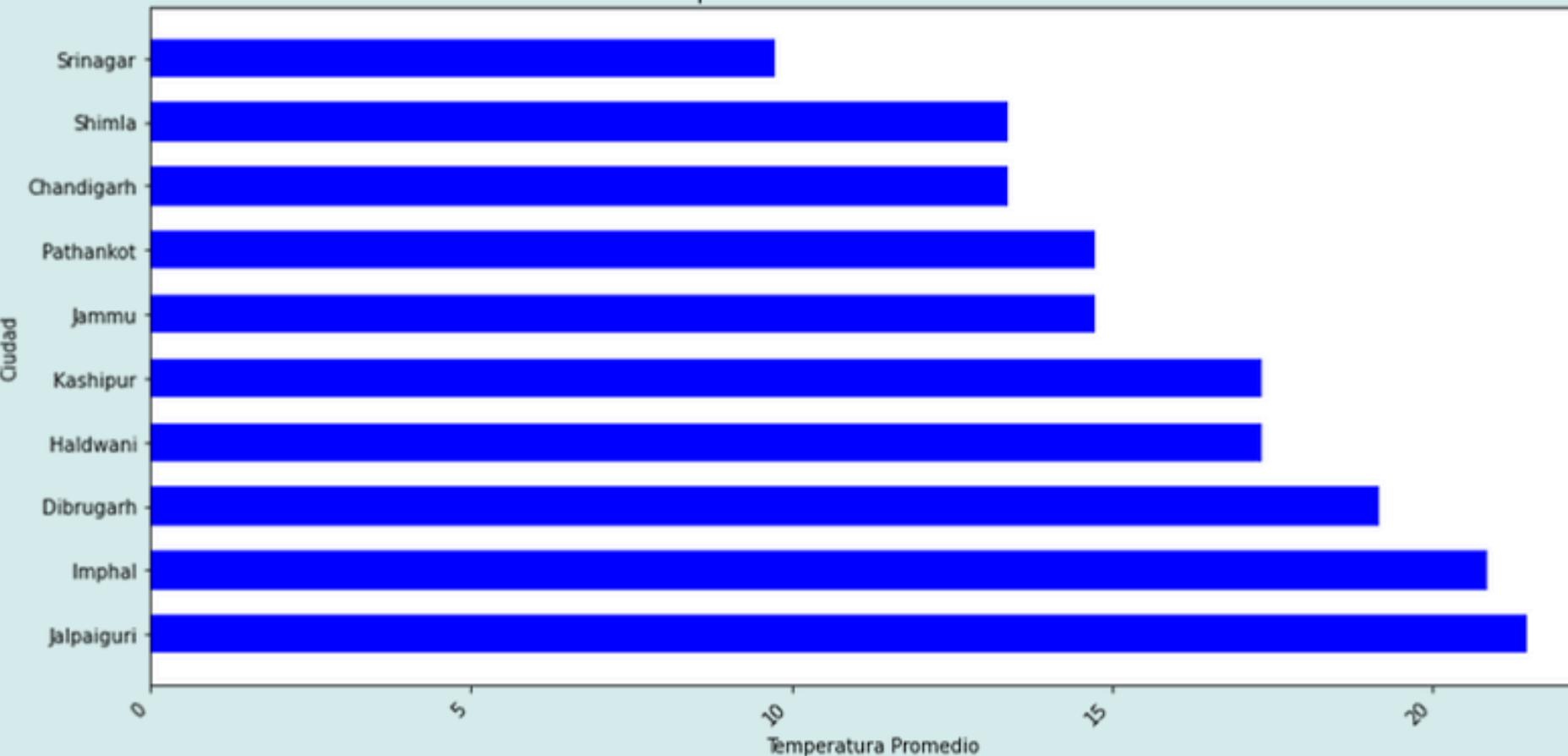


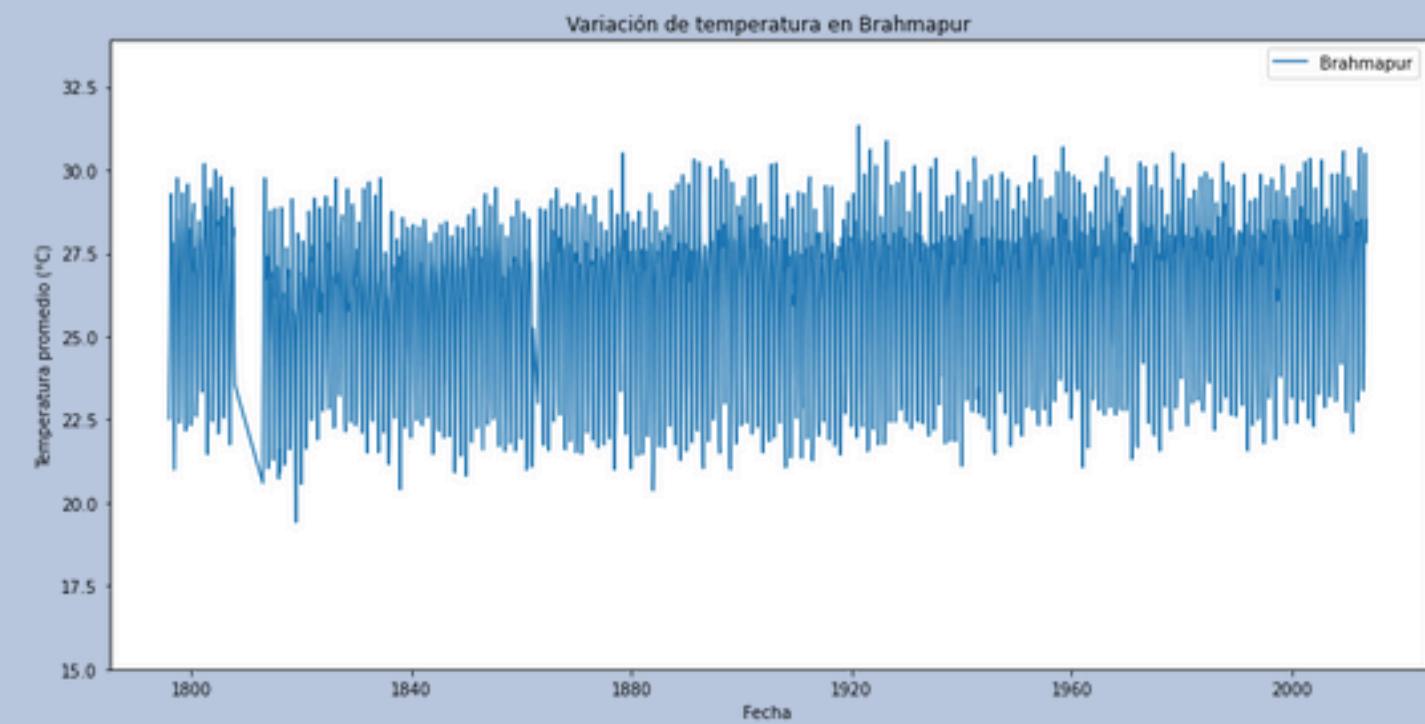
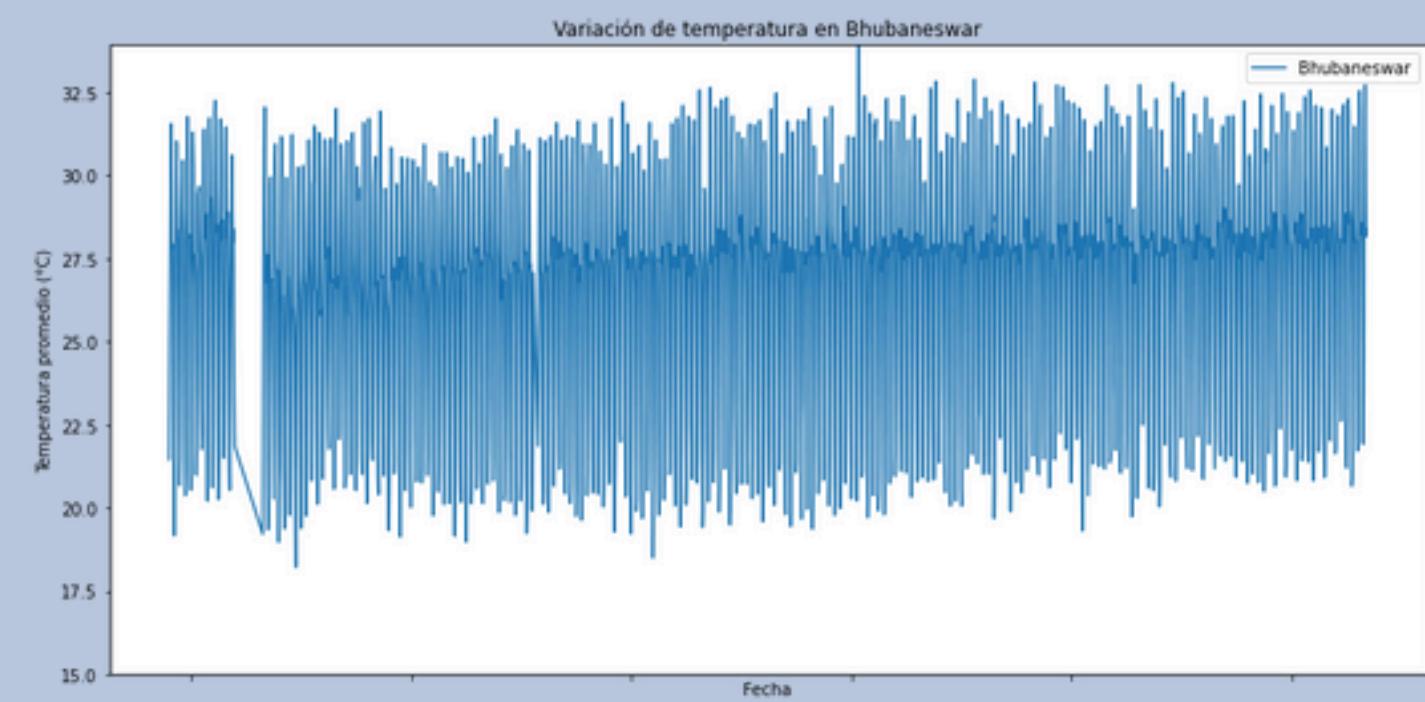
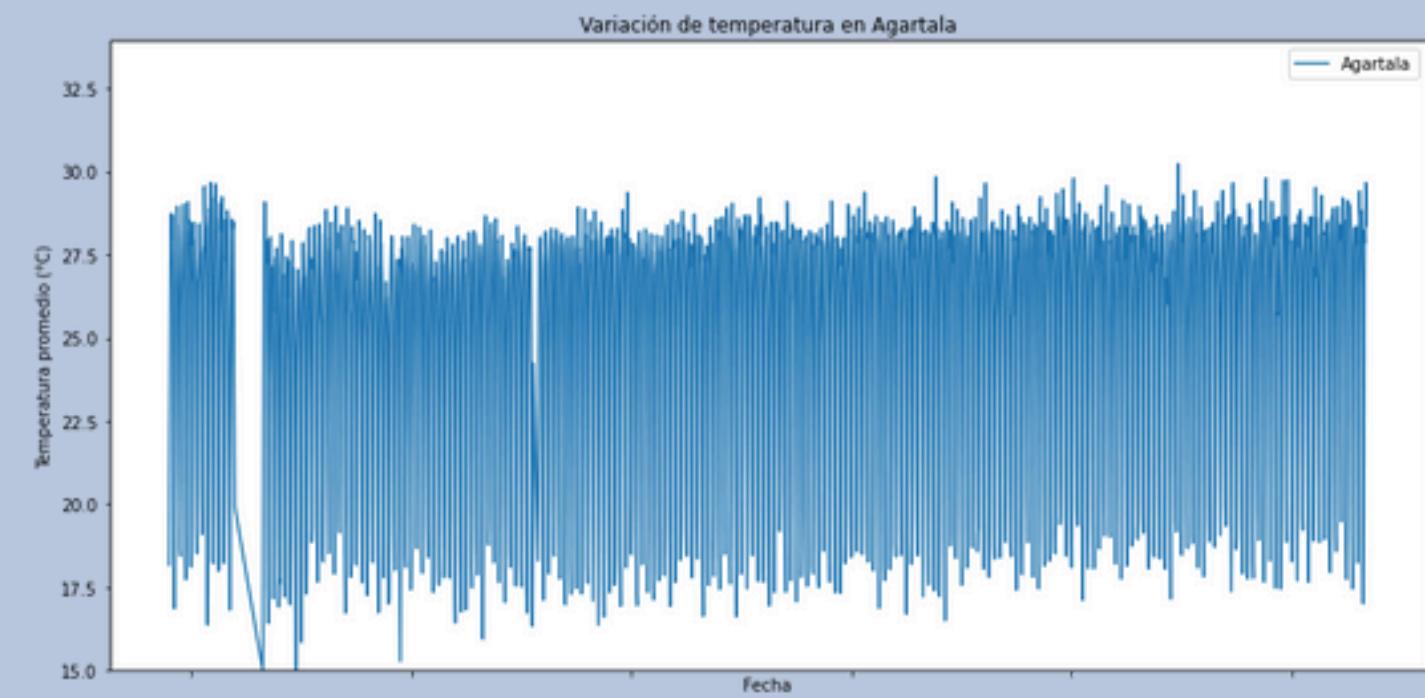
Análisis exploratorio de datos

Top 10 Ciudades Más Calurosas en India



Top 10 Ciudades Más Frías en India

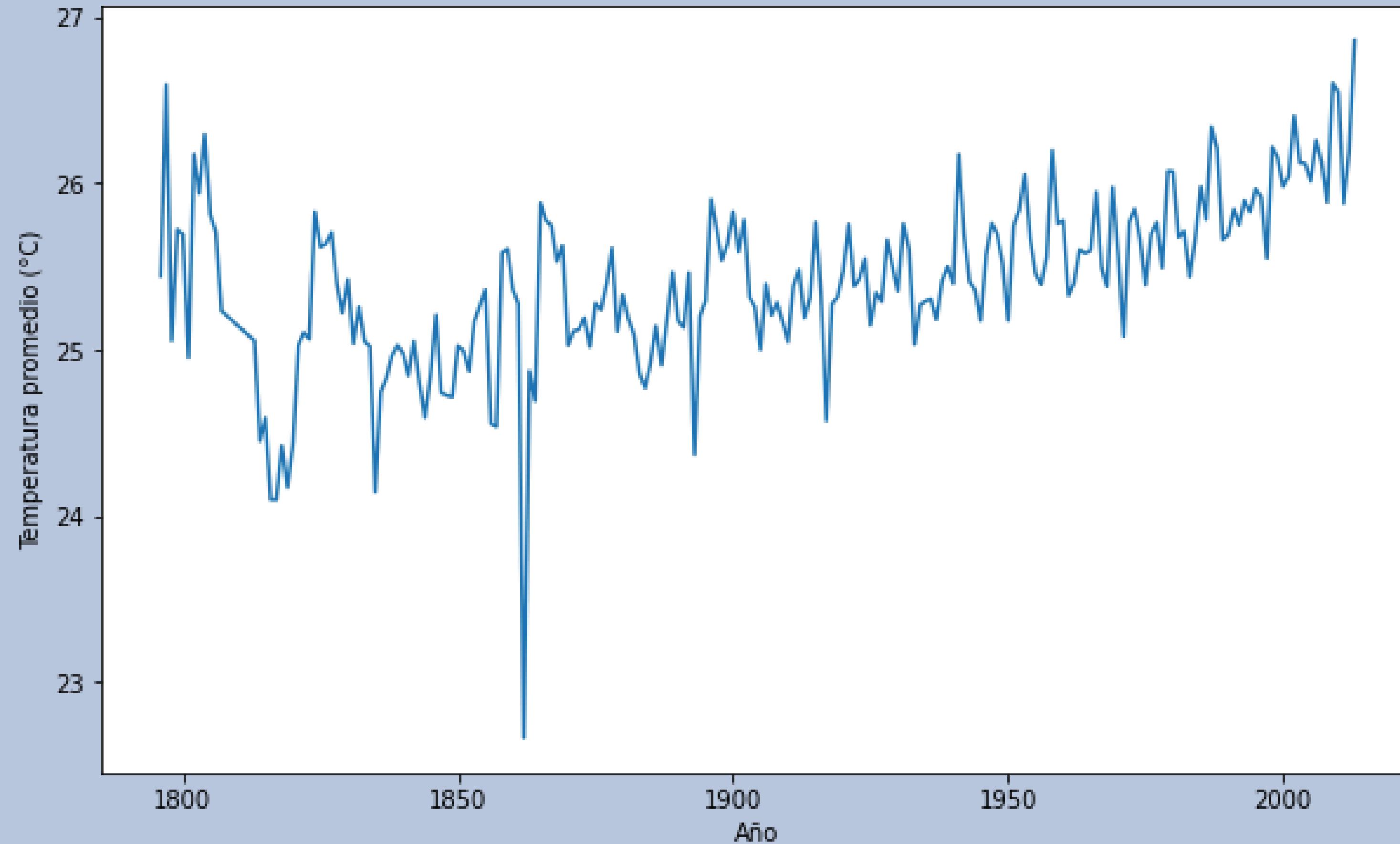




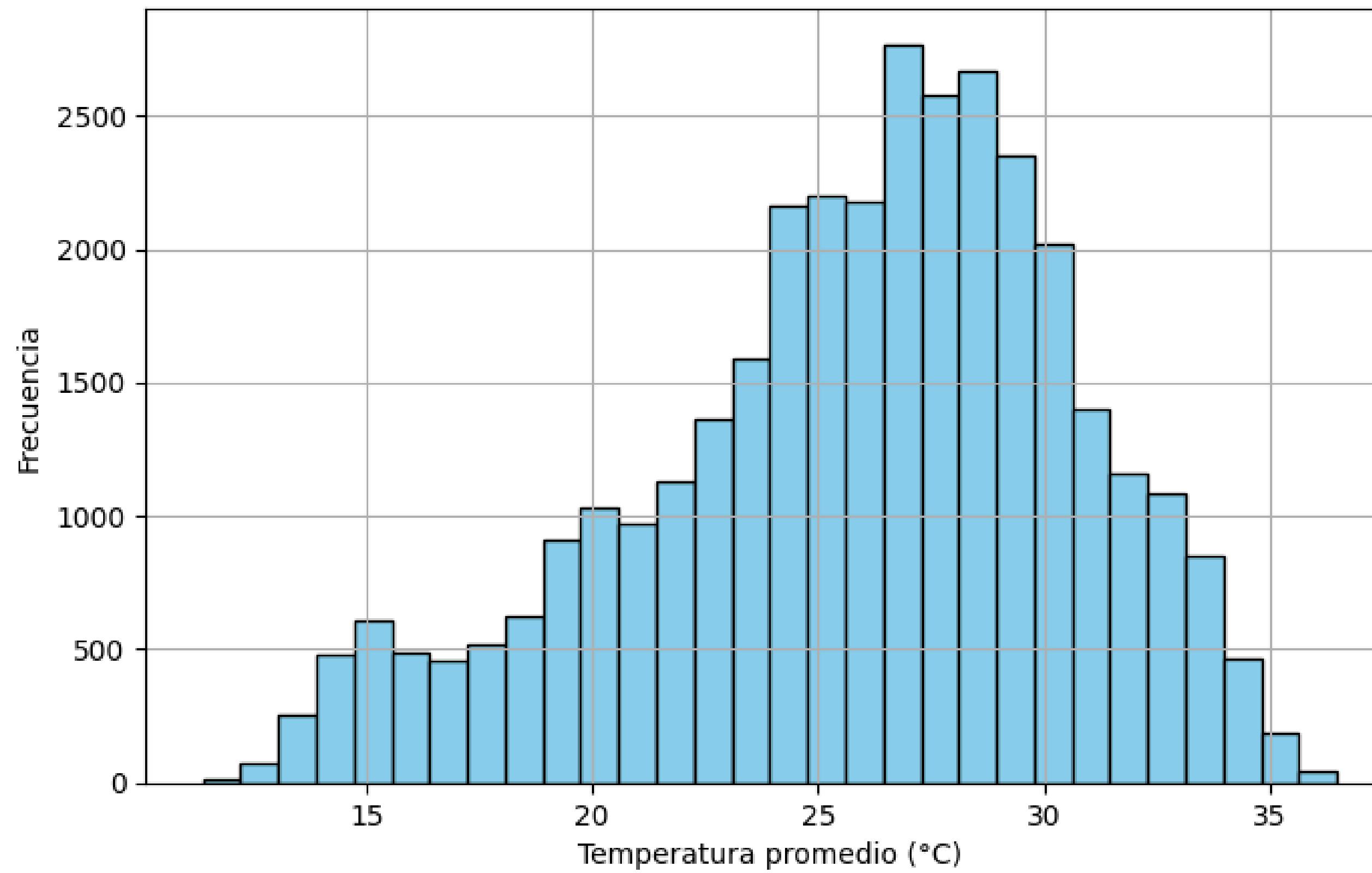
¿Como cambia el clima a
través del tiempo?



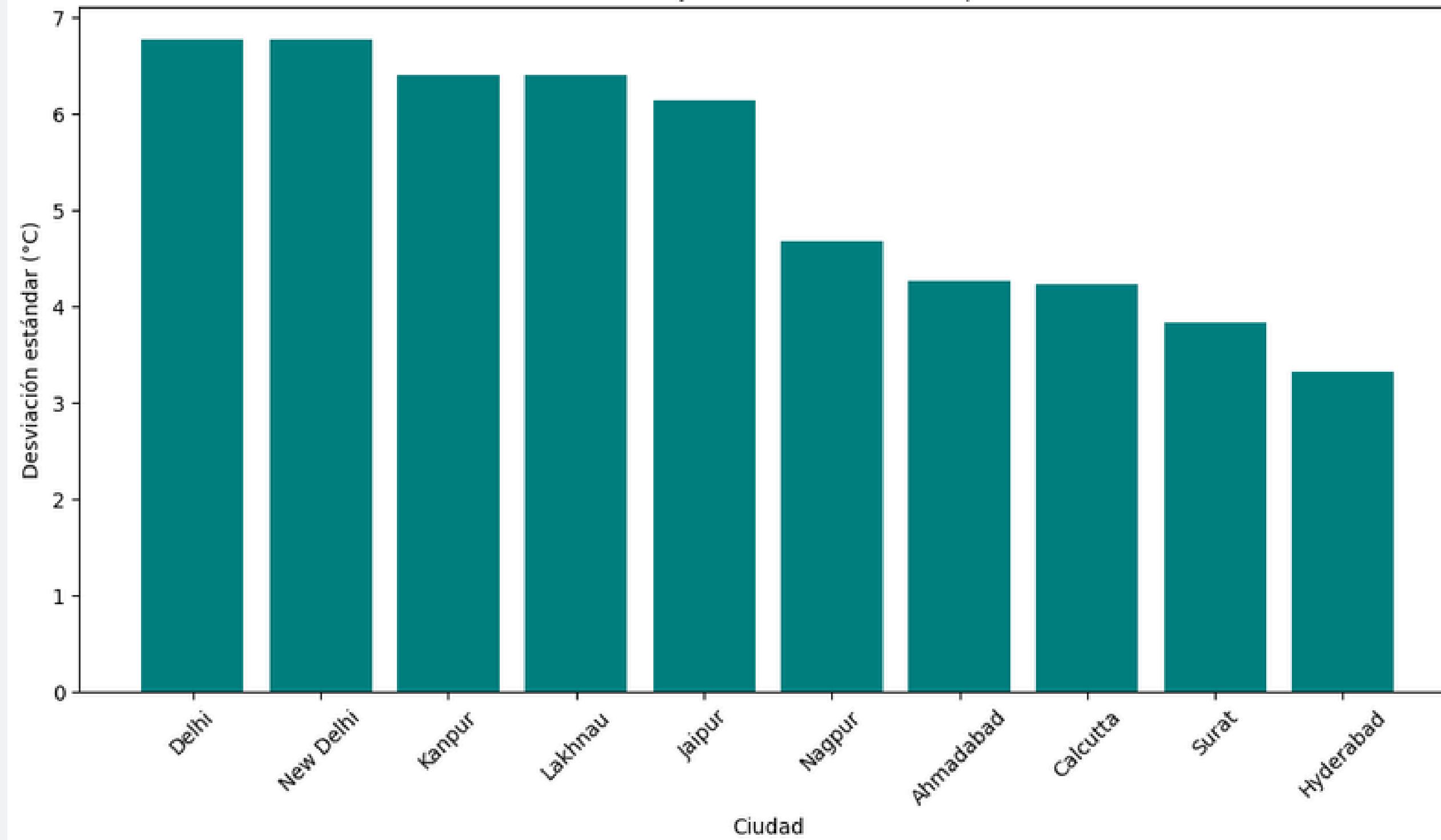
Temperatura promedio anual



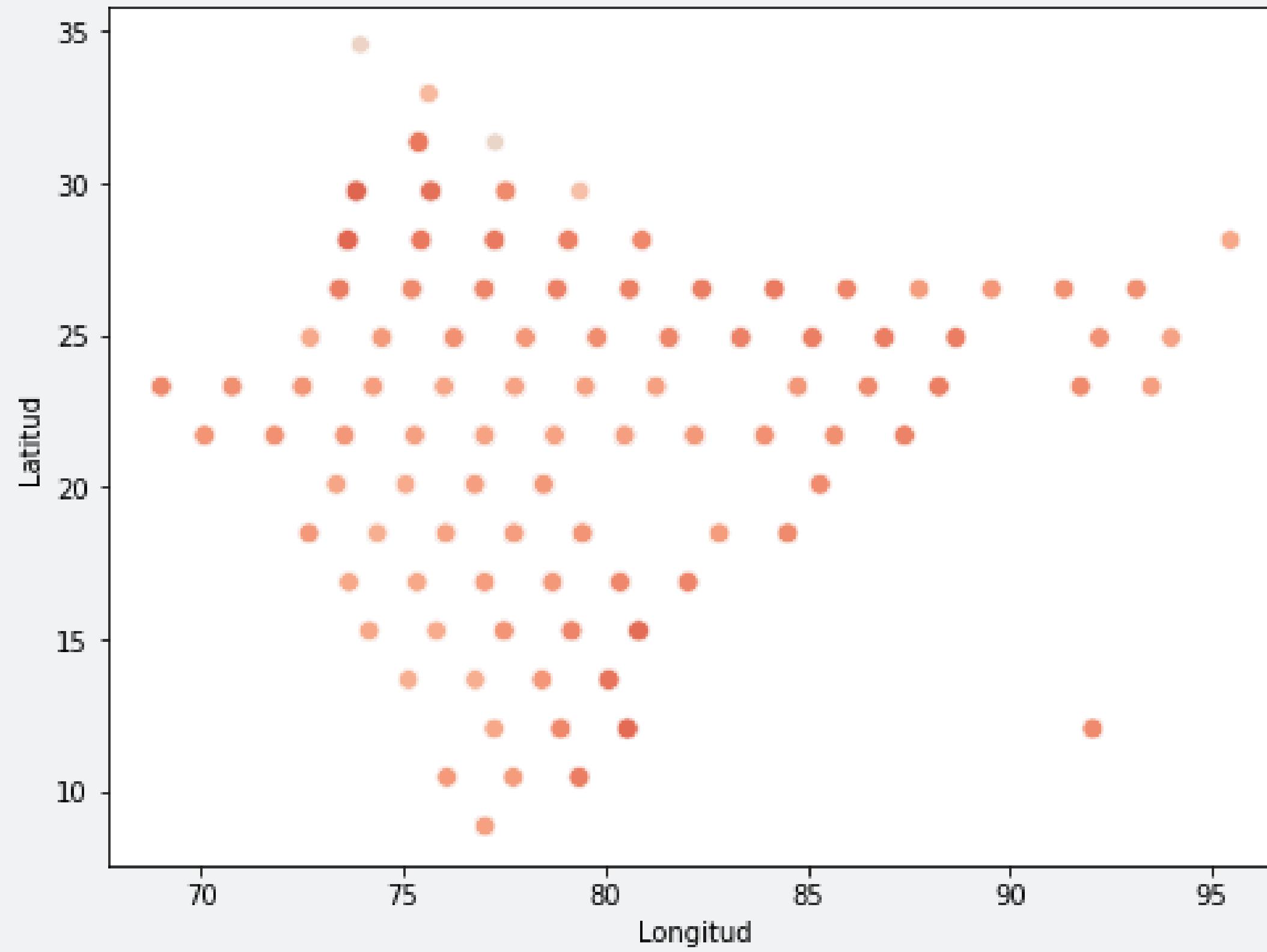
Distribución de temperaturas promedio



Ciudades con mayor variabilidad de temperatura



Relación entre ubicación y temperatura promedio

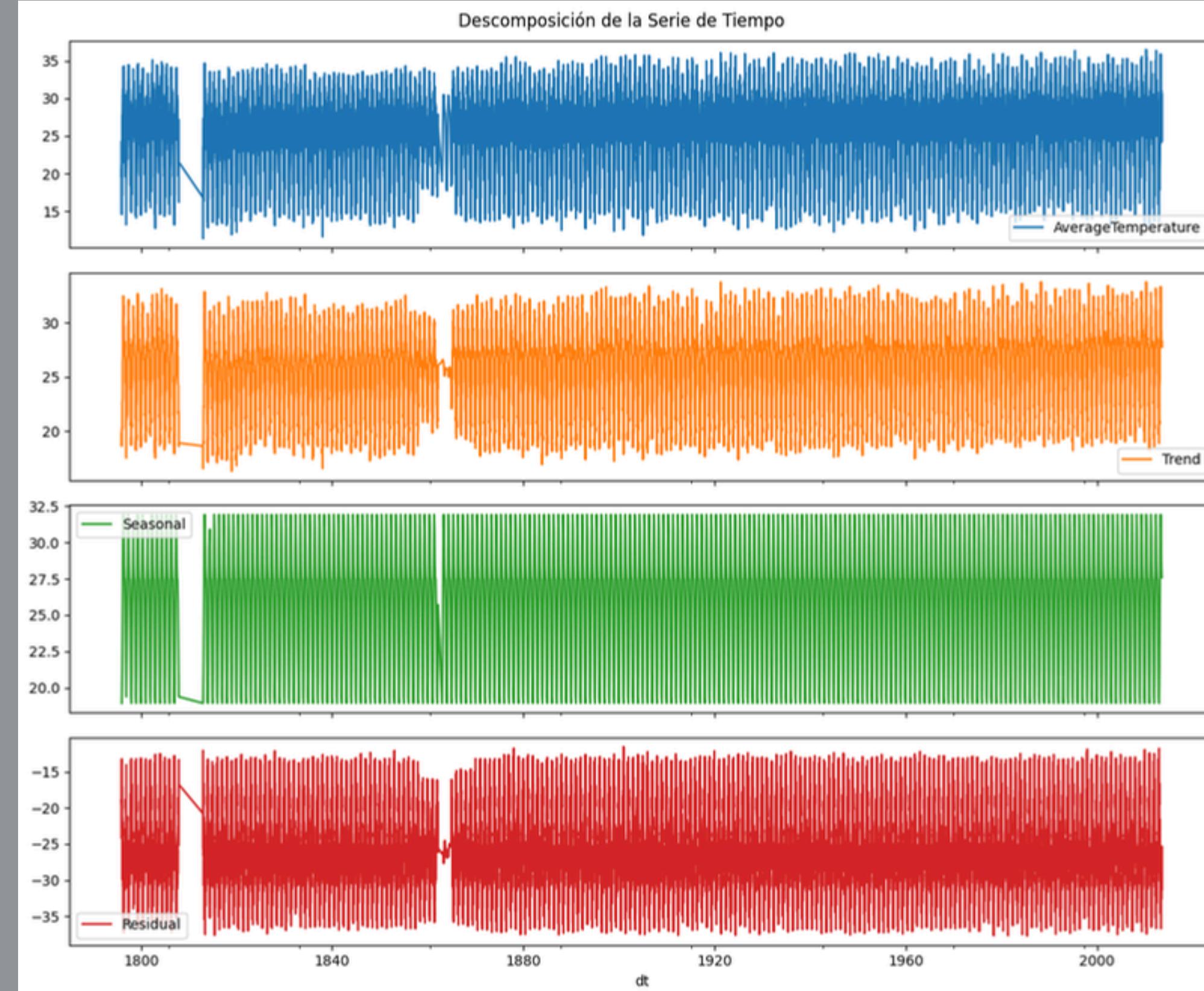


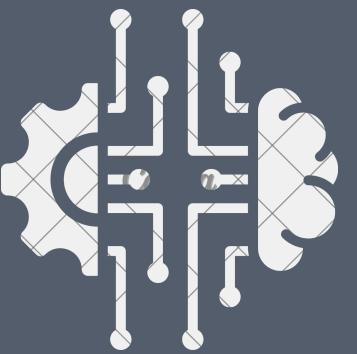
Temperatura
promedio

Tendencia

Estacionalidad

Componente
aleatoria





Modelos predictivos

Modelos implementados

Random Forest

- Utiliza múltiples árboles de decisión que combinan sus predicciones para mayor estabilidad y precisión.
- Captura relaciones complejas y no lineales entre las variables.
- Es robusto frente a datos ruidosos y no necesita que la serie sea estacionaria o siga un patrón específico.

Prophet

- Divide la serie en componentes: tendencia (lineal o logística), estacionalidades y eventos especiales.
- Ajusta automáticamente los patrones estacionales y la tendencia general.
- Maneja datos irregulares, con valores faltantes o ciclos complejos

Arima: Autoregressive Integrated Moving Average

- Se compone de tres componentes principales:
- Autoregresivo (AR): relación lineal entre las observaciones y sus valores pasados
- Media móvil (MA): tiene en cuenta el error residual de las observaciones pasadas
- Integrado (I): la diferenciación de la serie temporal para hacerla estacionaria.

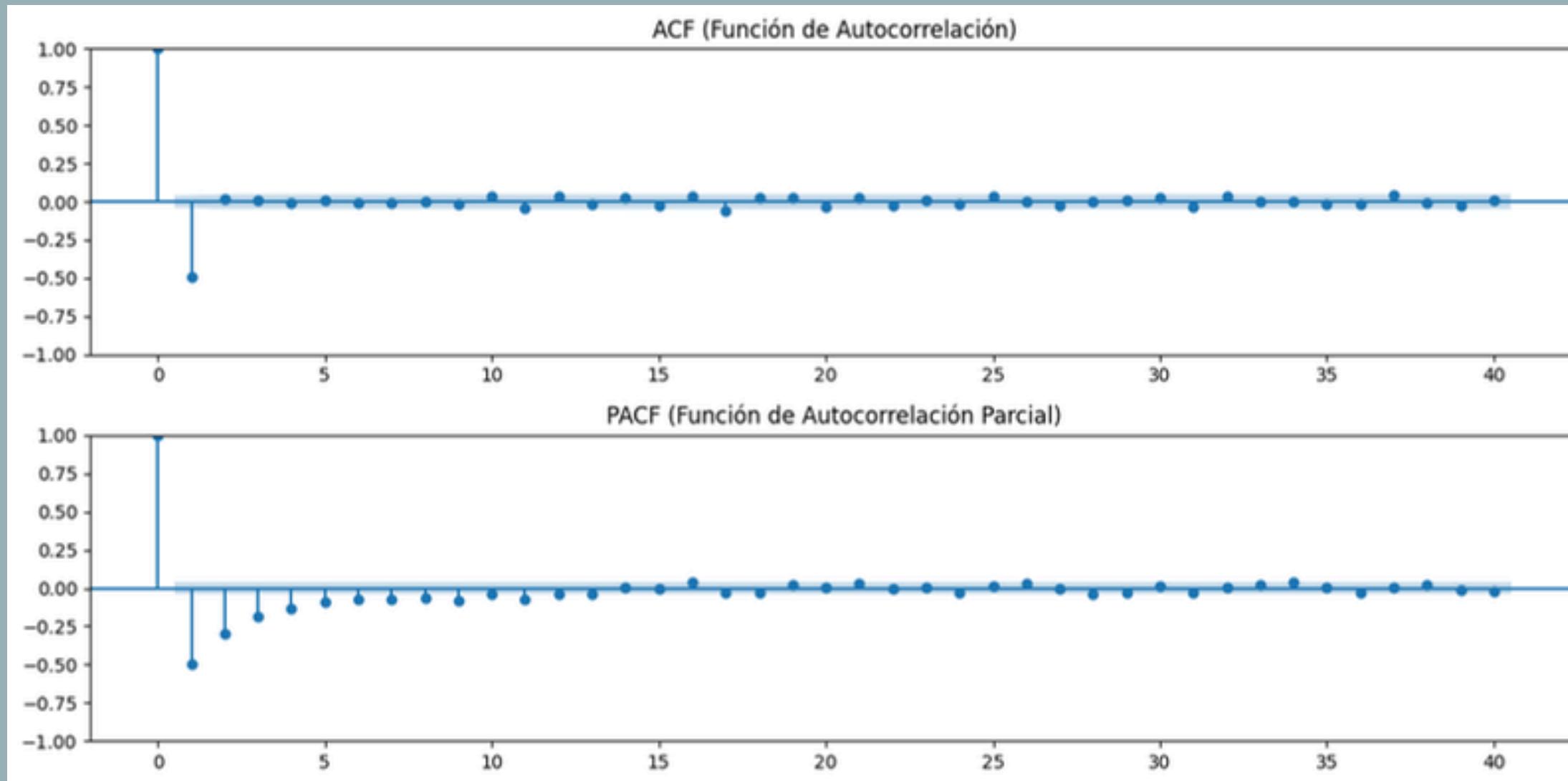
ARIMA

¡Es estacionaria!

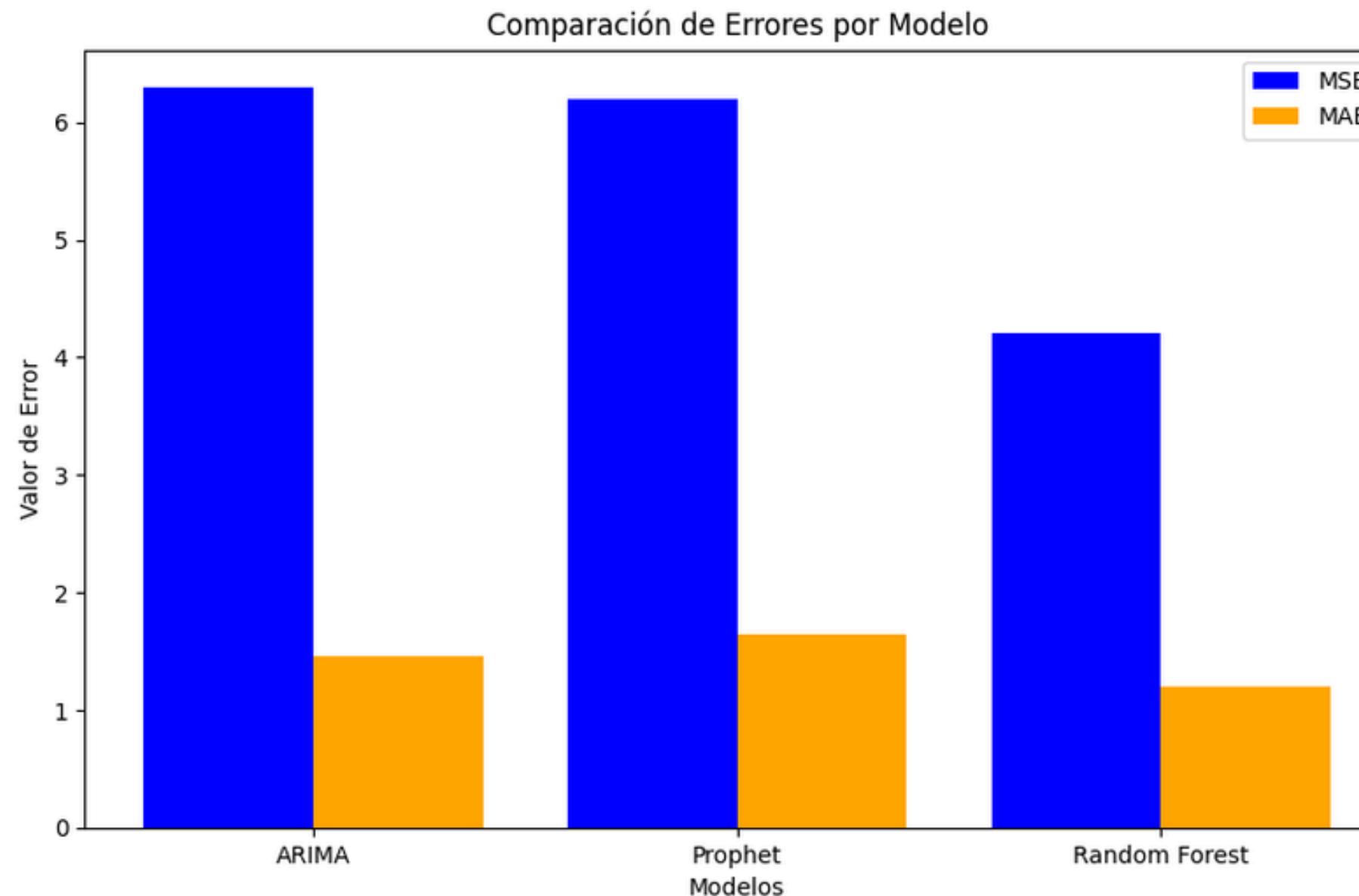
```
6 # Realizar el ADF Test
7 resultado_adf = adfuller(serie_temporal.dropna()) # Asegúrate de que no haya nulos
8
9 # Mostrar resultados
10 print(f'ADF Statistic: {resultado_adf[0]}')
11 print(f'p-value: {resultado_adf[1]}')
12 for key, value in resultado_adf[4].items():
13     print(f'Valor critico {key}: {value}')
14
15 if resultado_adf[1] < 0.05:
16     print("La serie es estacionaria.")
17 else:
18     print("La serie NO es estacionaria.")
19
```

```
ADF Statistic: -5.483131920861664
p-value: 2.2607984367075786e-06
Valor critico 1%: -3.4305391532821132
Valor critico 5%: -2.861623601044006
Valor critico 10%: -2.5668144982138608
La serie es estacionaria.
```

Elección de variables p y q



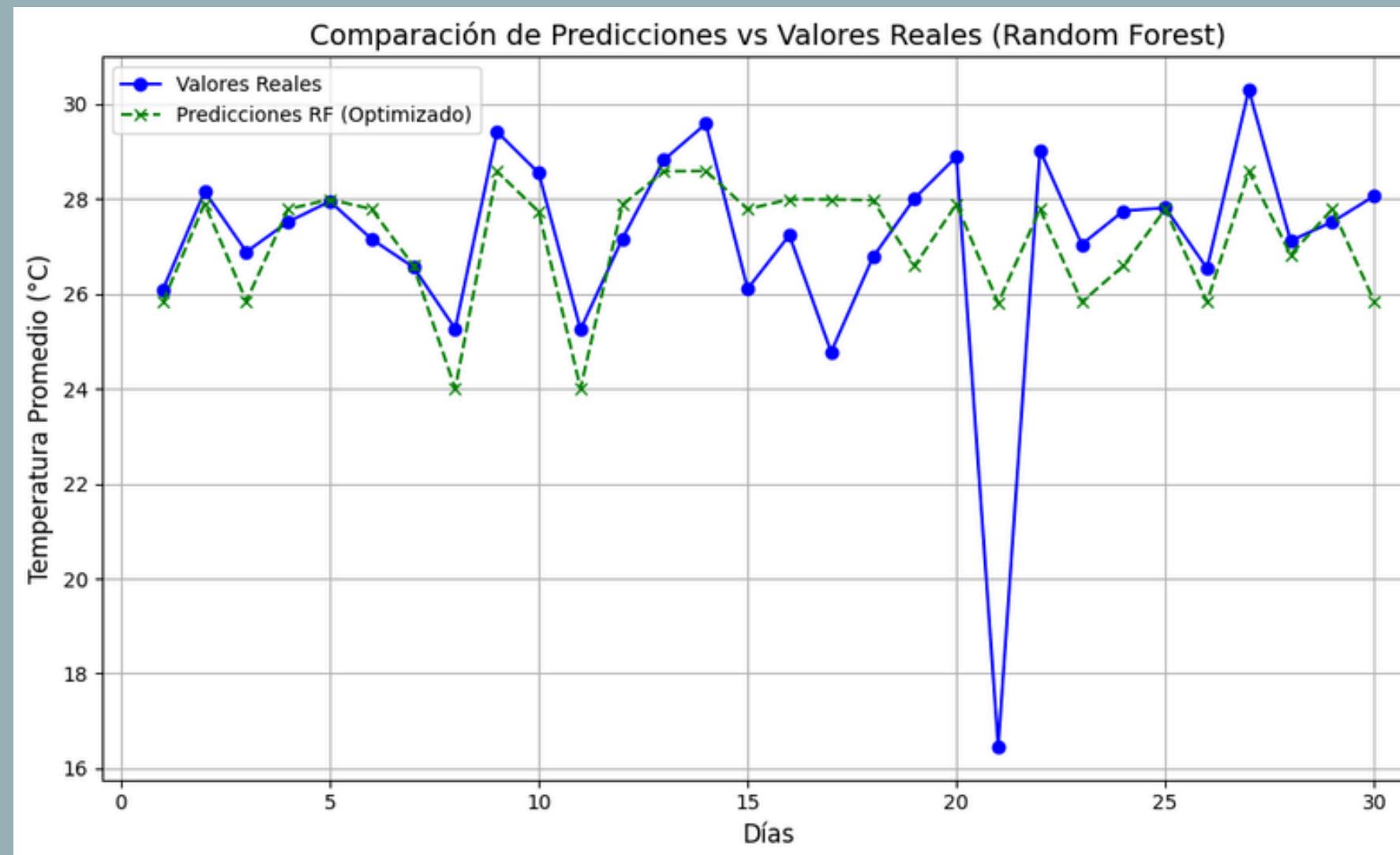
Comparación de modelos



Model	MSE	MAE
0 ARIMA	6.301493	1.457482
1 Prophet	6.202540	1.640123
2 Random Forest	4.211735	1.203485

El modelo con hiperparámetros optimizados mostró una reducción significativa en el error respecto al modelo base de Random Forest y otros modelos como ARIMA y Prophet.

Modelo Elegido



Random Forest

Rendimiento:

- Mean Squared Error (MSE): 4.211735
- Mean Absolute Error (MAE): 1.203485

Mejores Hiperparámetros Encontrados:

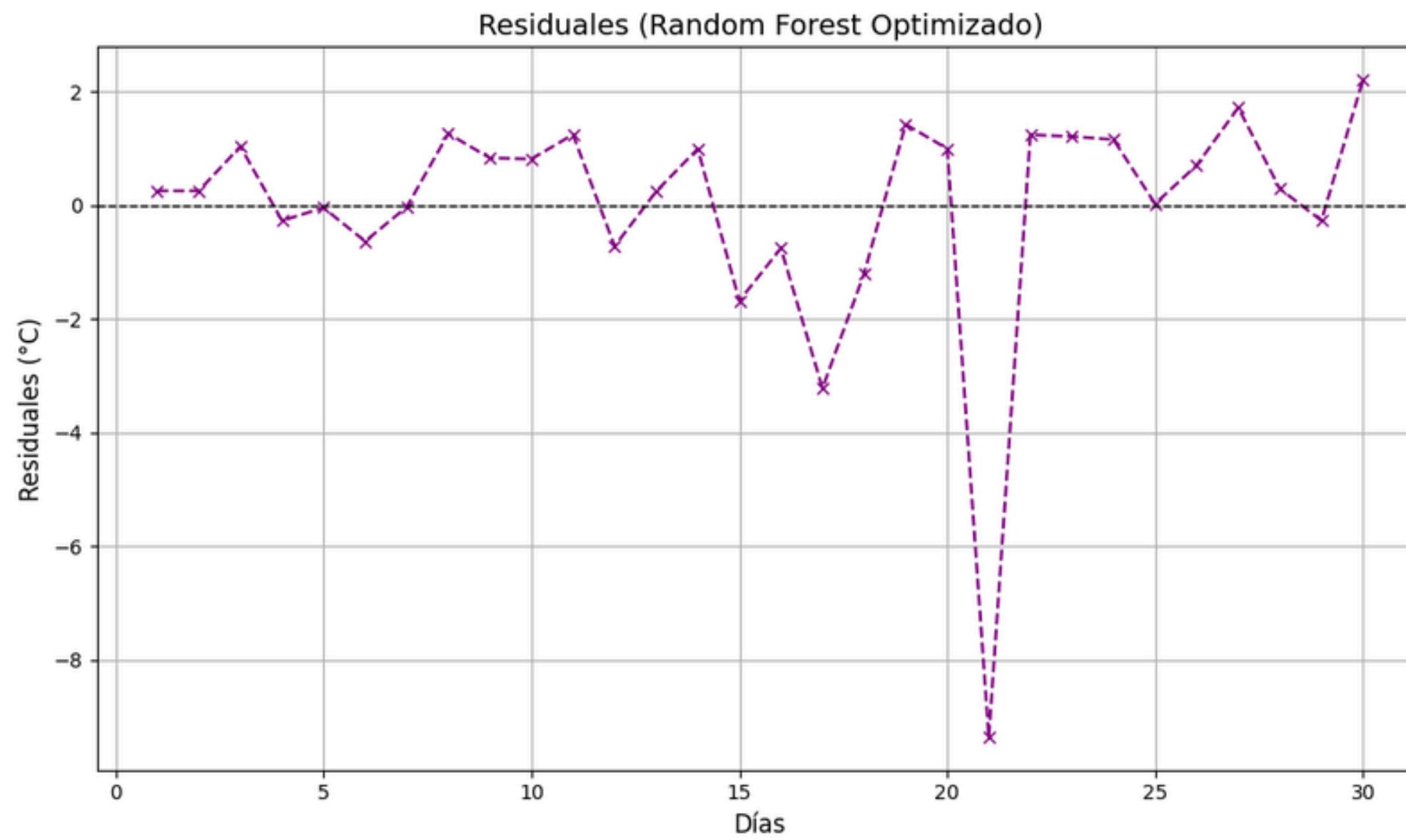
- max_depth: 10
- max_features: auto
- min_samples_leaf: 4
- min_samples_split: 2
- n_estimators: 200



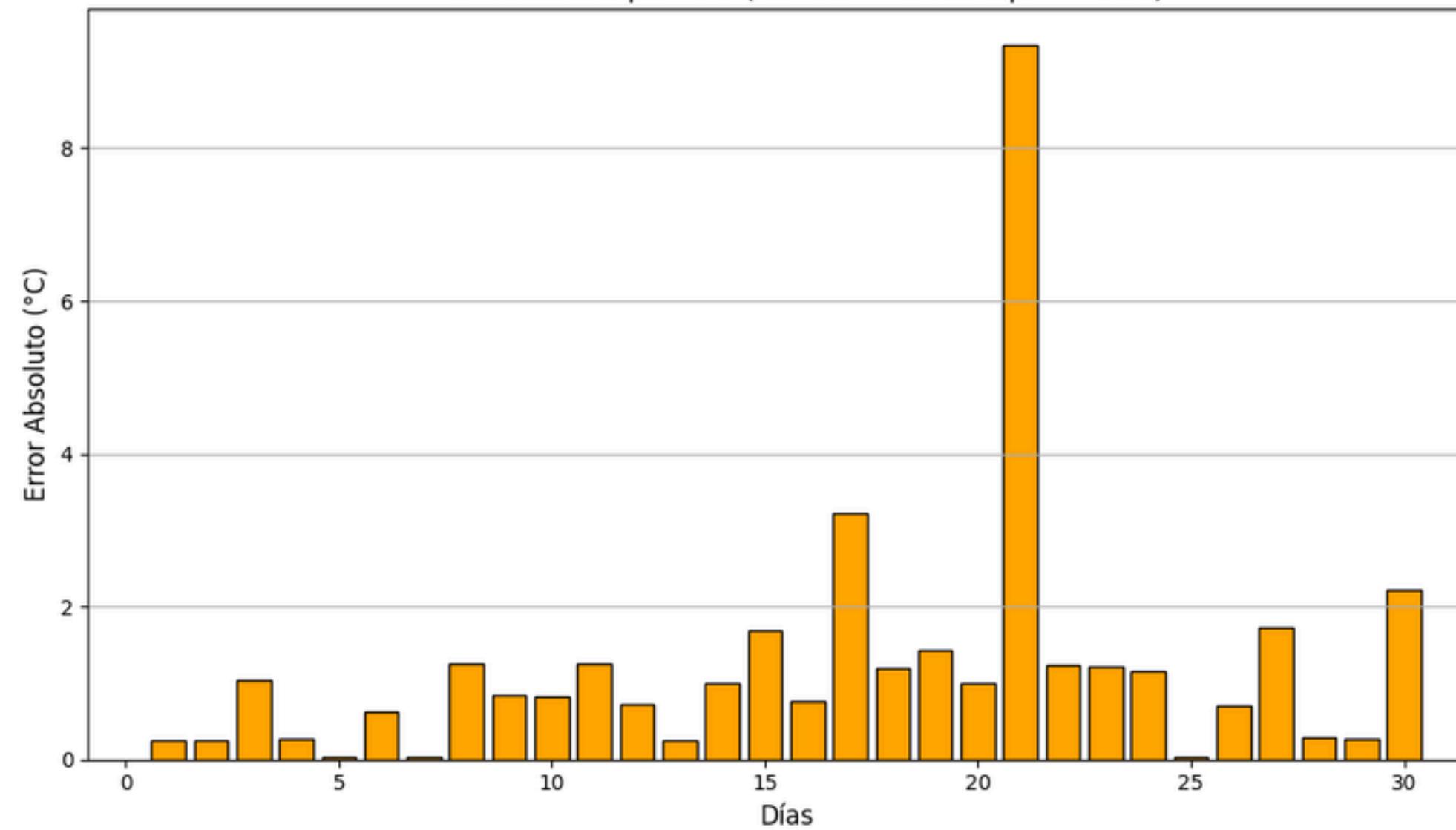
Utilizamos Grid Search



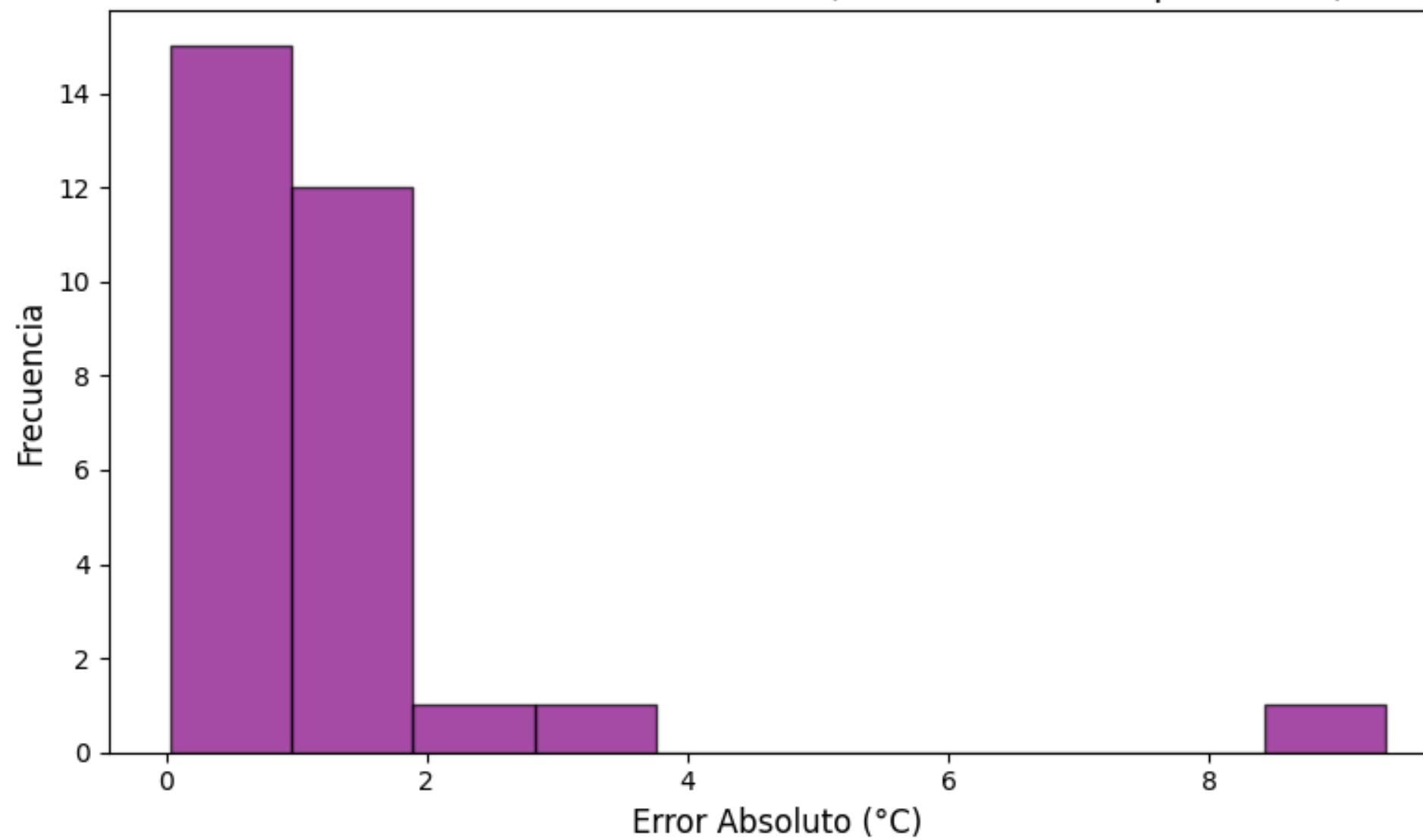
Evaluación y Visualización



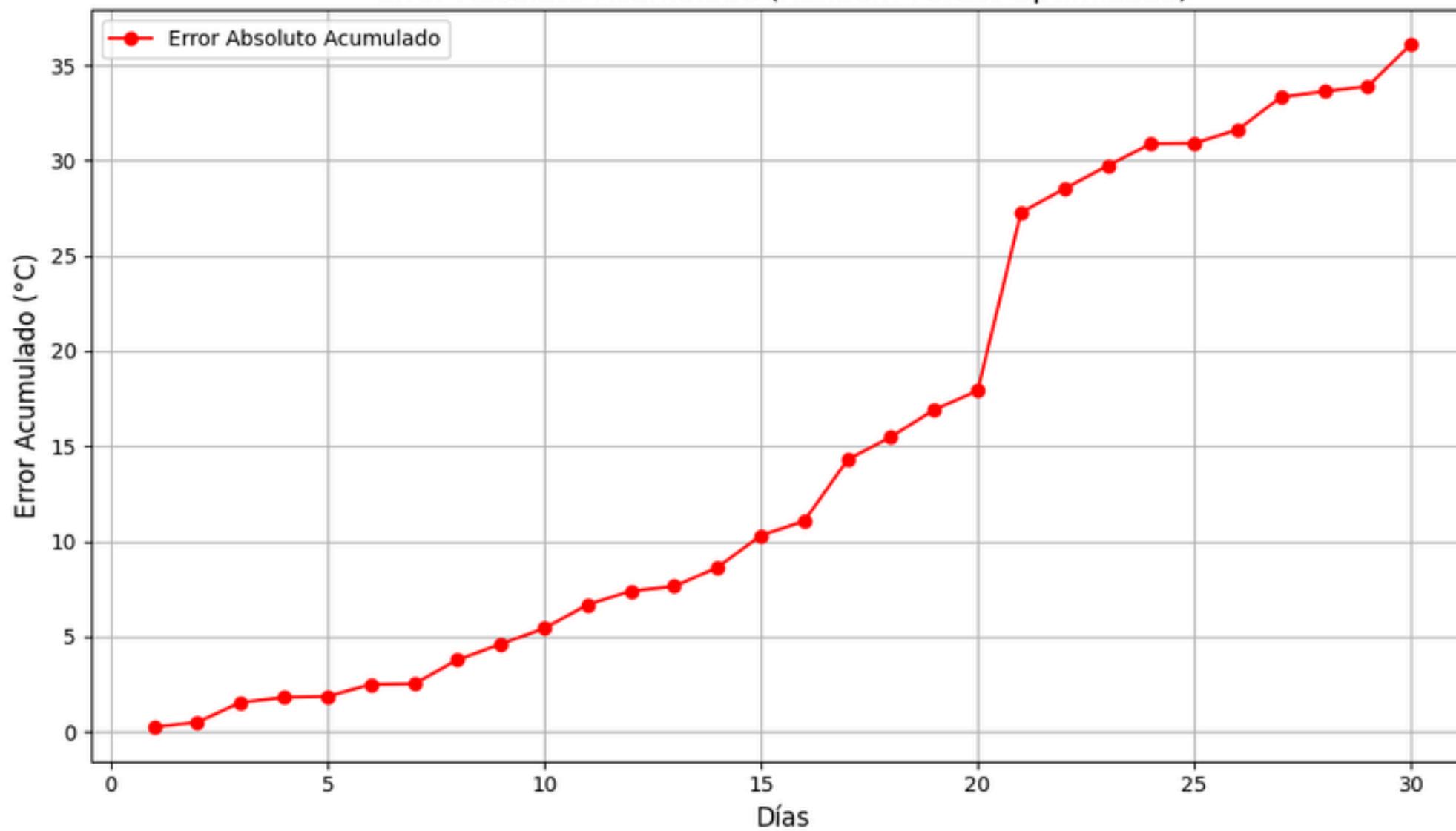
Error Absoluto por Día (Random Forest Optimizado)



Distribución de Errores Absolutos (Random Forest Optimizado)

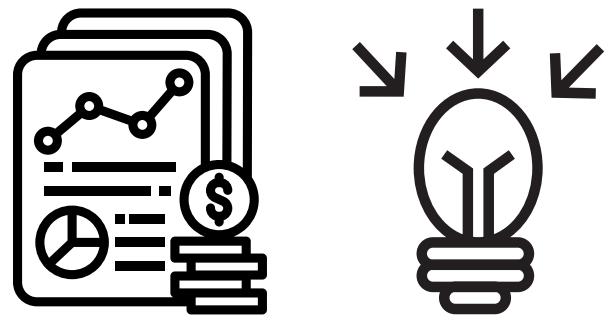


Error Absoluto Acumulado (Random Forest Optimizado)





Conclusiones



- Modelo con mejor rendimiento: Random Forest
- Posibles mejoras:
 - Predecir más días.
 - Verificar si la serie es. estacionaria o no con todo el dataset.
 - Ampliar el modelo a otros países.





ITBA



2do cuatrimestre - 2024

TRABAJO PRÁCTICO FINAL
ANALISIS DE BIG DATA CON SPARK



Integrantes

- Manuel Saul Hanono
- Dana Nabel
- Nicole Reiman
- Sol Winkel