



Universidad de Sonora
División de Ciencias Exactas y Naturales
Departamento de Física
Física computacional
Análisis Exploratorio de Datos en Python
Actividad 4
Iveth R. Navarro

05 de enero de 2021
Puerto Peñasco, Sonora, México

Resumen

La presente es una práctica de Física Computacional cuyo objetivo es el dar los primeros pasos que se realizan en el Análisis Exploratorio de Datos (EDA, Exploratory Data Analysis) haciendo uso de la información meteorológica de CONAGUA.

1. Introducción

Durante este semestre de la carrera, estamos llevando Estadística a la par de Física Computacional; es por eso que, según los conocimientos adquiridos en estadística, conocemos lo útil del análisis de datos mediante gráficas tales como histogramas, de cajas, de líneas o funciones de densidad de probabilidad. La oportunidad de poder generar estas gráficas mediante un lenguaje de programación resulta muy atractivo. Esta es una de las cosas que se pueden hacer mediante el Análisis Exploratorio de Datos (EDA, Exploratory Data Analysis), cuya exploración es el objetivo de esta práctica.

Los datos a analizar, son los mismos que se han estado manejando desde el comienzo del curso, es decir, la información meteorológica de CONAGUA. Como en las actividades pasadas, la estación a analizar es la de Puerto Peñasco, Sonora.

Muchas de las funciones que hemos utilizado con anterioridad son parte del proceso de Análisis Exploratorio de Datos (EDA), donde buscamos conocer la estructura, contenidos y características del conjunto de datos que se desean analizar.

Entre algunas las características en las nos centramos por ejemplo son:

- Estructuras/patrones en los datos
- Número de datos faltantes
- Detección de datos anómalos y valores extremos
- Extraer y seleccionar variables importantes
- robar alguna teoría de comportamiento de los datos

Entre algunas de las funciones a aplicar para la exploración de un Data Frame son:

1. df.shape : Forma y dimensiones de df.
2. df.types : Tipo de datos de las columnas (numéricas o categóricas)
3. df.head(), df.tail(), df.sample(5) : Despliege de un número de renglones, para ver la estructura de df, nombres de columnas o si los datos tienen sentido a lo esperado.
4. df.info() : Proporciona información general de la estructura y componentes de df
5. df.describe() : Descripción estadística de las variables numéricas.
6. df.describe(include='object') : Descripción estadística de las variables categóricas.
7. df.isnull().sum() : Suma de valores faltantes.

En cuanto al proceso de visualización de datos para mejorar la percepción de los fenómenos o eventos que representan los dato, se utiliza la exploración de datos utilizando las bibliotecas de visualización de Python: Matplotlib y Seaborn.

Matplotlib es la biblioteca de Python desarrollada para la visualización de datos. Es muy potente y permite producir gráficas que deseemos.

Seaborn, se construyó sobre Matplotlib, pero con una estructura menos compleja, de mayor facilidad de manejo.

Es posible combinar ambas bibliotecas para producir una gráfica que satisfaga nuestros propósitos. Primero, comenzamos por analizar las variables numéricas apoyados con gráficas en Matplotlib y Seaborn. Después analizamos variables categóricas.

Se pueden visualizar las características de variables numéricas mediante:

- Gráficas de Histogramas: sns.histplot().
- Funciones de densidad de Probabilidad: sns.kdeplot().
- Gráficas de caja (BoxPlots): sns.boxplot().
- Gráficas de barras: sns.barplot().
- Gráfica de líneas: sns.lineplot().

A continuación se muestra el desarrollo de cada una de las actividades mencionadas haciendo uso del conocimiento expuesto.

2. Desarrollo

Esta actividad se divide en 6 partes:

- **Actividad 4.1:** Crear un nuevo cuaderno de trabajo de Jupyter llamado Actividad4.ipynb en Google Colab. Resumir en una sola celda todas las funciones aplicadas al DataFrame inicial concluyendo con la creación de un nuevo DataFrame para continuar con el trabajo.
Se pide sintetizar las características principales del conjunto de datos que estás analizando, aplicando la siguiente secuencia de funciones de un proceso EDA arriba mencionadas.
- **Actividad4.2:** Crear Histogramas de las variables de Precipitación, Evaporación, Temperaturas Máxima y Mínima de el conjunto de datos que se están analizando (Función: sns.histplot()). Complementar en su caso con las gráficas de la función de densidad de probabilidad correspondiente (Función: sns.kdeplot())
- **Actividad 4.3:** Crear las gráficas de cajas (Boxplot) para la Evaporación, Temperaturas Máxima y Mínima (Función: sns.boxplot()).
- **Actividad 4.4:** Producir las gráficas de barras para la Precipitación agrupado por Años y después por meses (Función: sns.barplot()).

- **Actividad 4.5:** Crear una colección de los últimos 30 años de datos, utilizando condiciones de filtrado por un rango de años. Crear las gráficas de línea de la Precipitación, Temperaturas Máxima y Mínima como funciones del tiempo (Últimos 30 Años). (Función: sns.lineplot()).
- **Actividad 4.6:** Con el conjunto de 30 años de datos, producir diagramas de cajas (Función: sns.boxplot()) para observar la variabilidad de las Temperaturas (Max y Tmin) y la Evaporación agrupados por Mes.

2.1. Resultados

2.1.1. Actividad 4.1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17602 entries, 0 to 17601
Data columns (total 7 columns):
 #   Column   Non-Null Count  Dtype  
--- 
 0   Fecha     17602 non-null   datetime64[ns]
 1   Precip    17568 non-null   float64 
 2   Evap      9008 non-null   float64 
 3   Tmax      16608 non-null   float64 
 4   Tmin      16648 non-null   float64 
 5   Año       17602 non-null   int64  
 6   Mes       17602 non-null   int64  
dtypes: datetime64[ns](1), float64(4), int64(2)
memory usage: 962.7 KB
```

Figura 1

2.1.2. Actividad 4.2

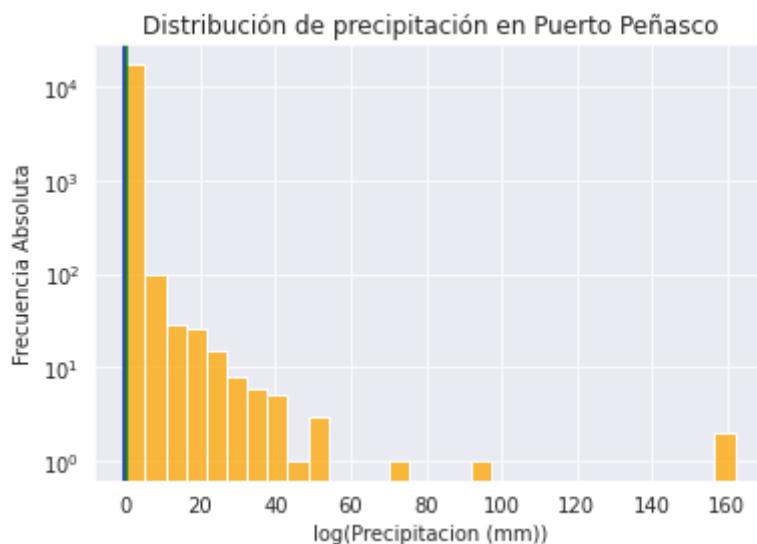
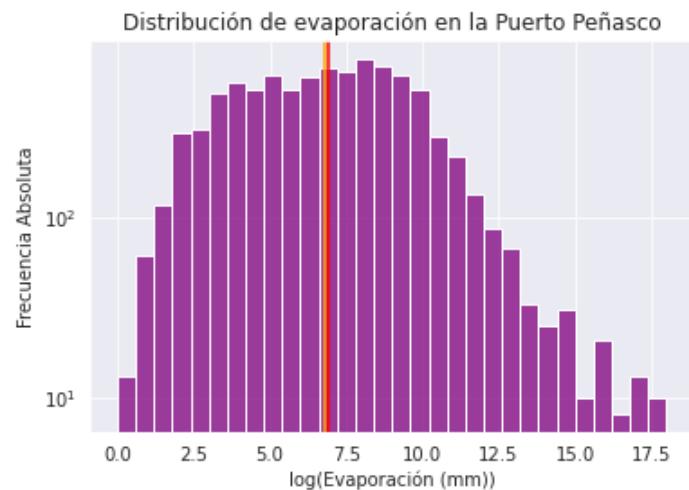
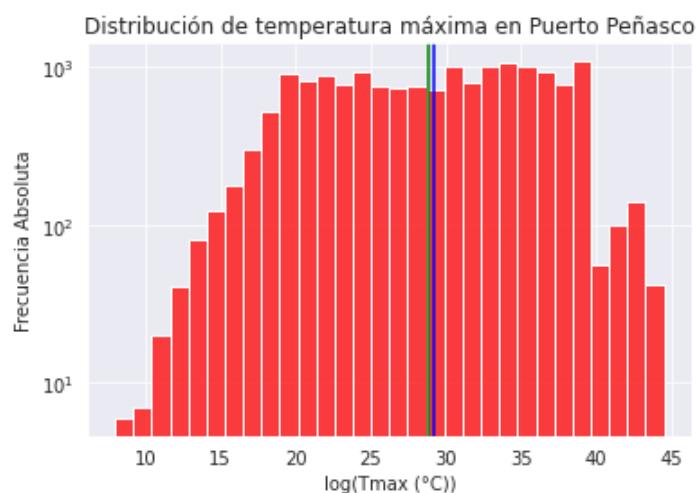
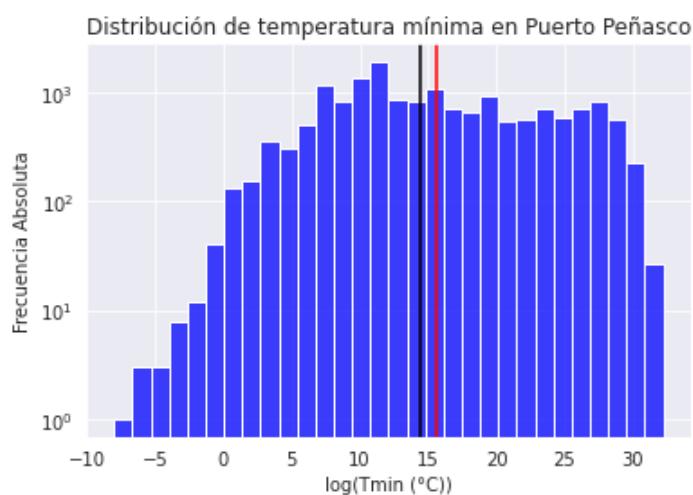
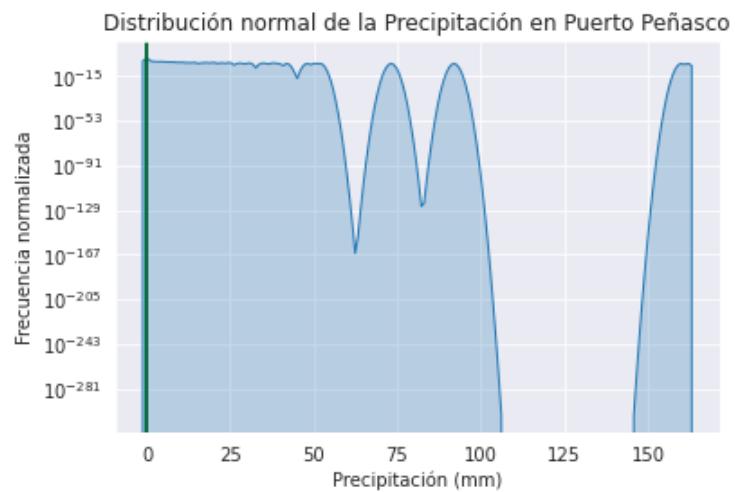
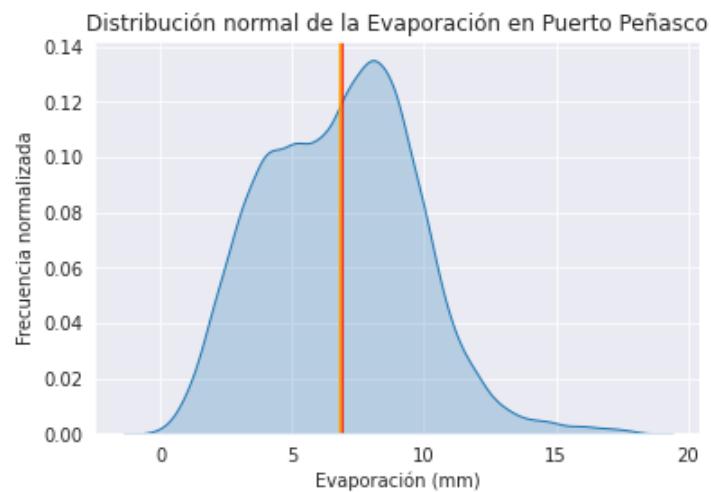
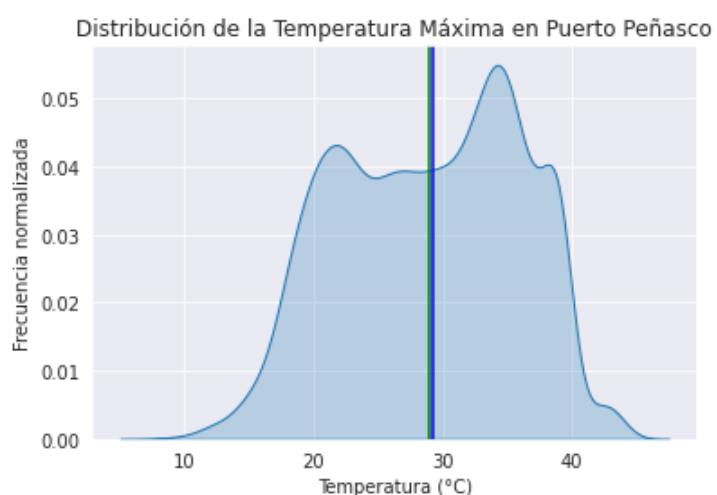
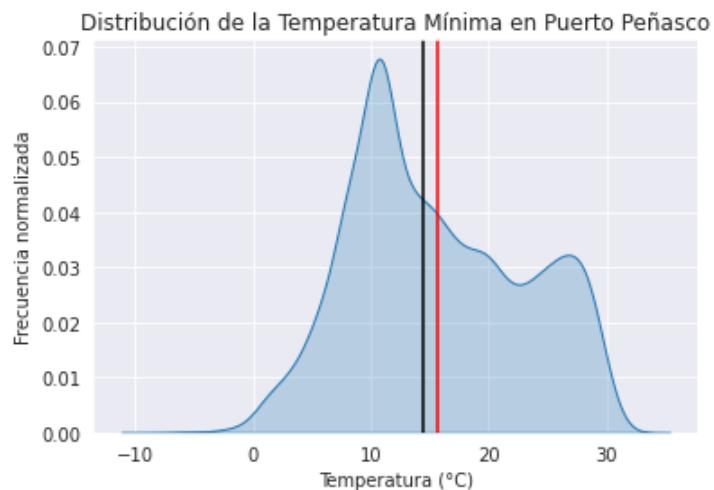
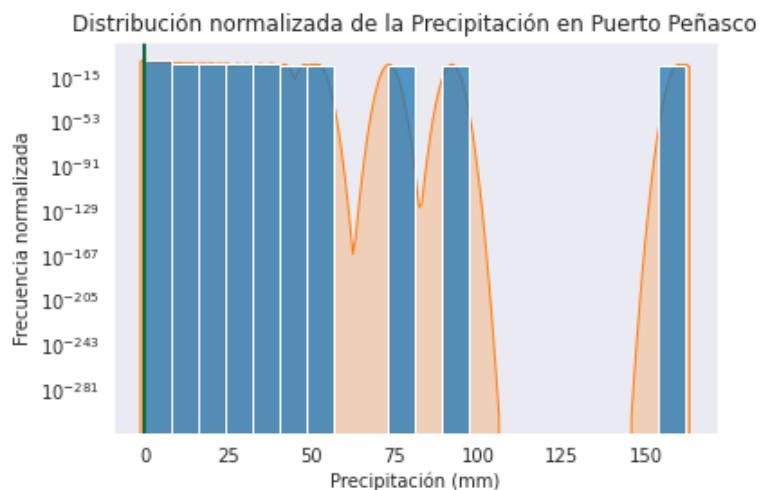
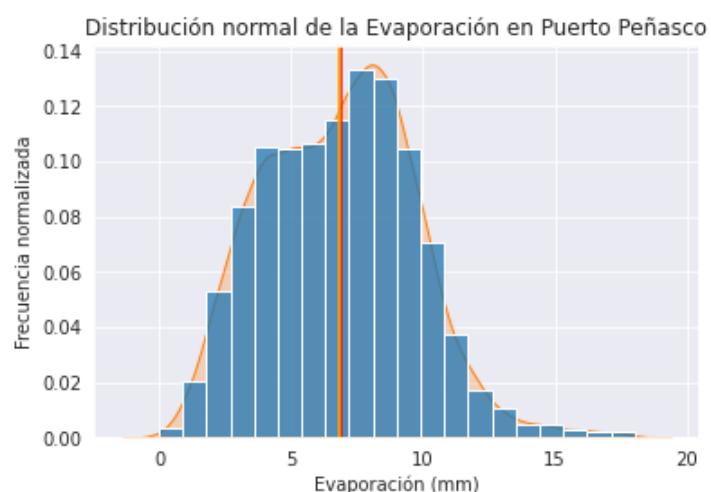
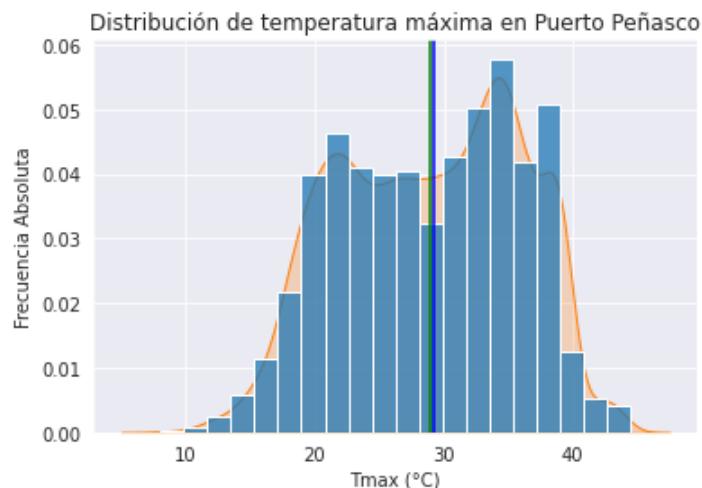
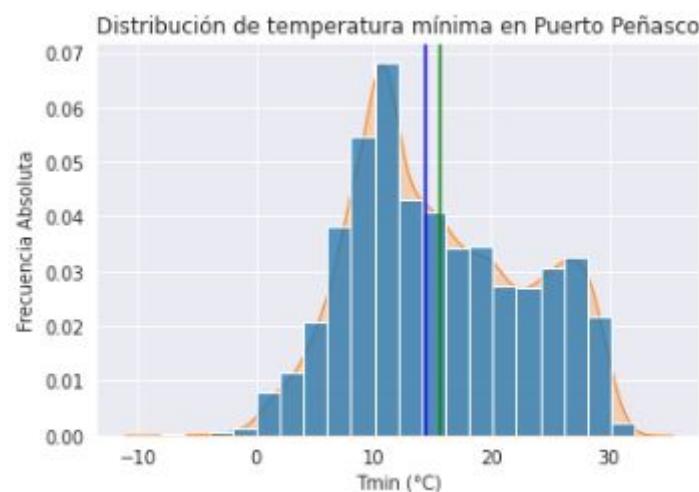


Figura 2

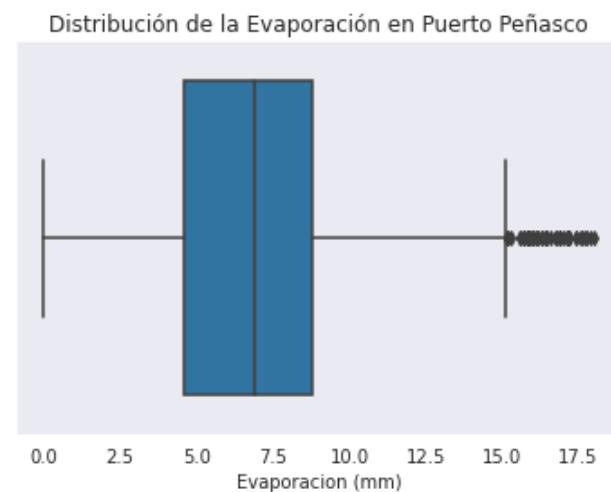
**Figura 3****Figura 4****Figura 5**

**Figura 6****Figura 7****Figura 8**

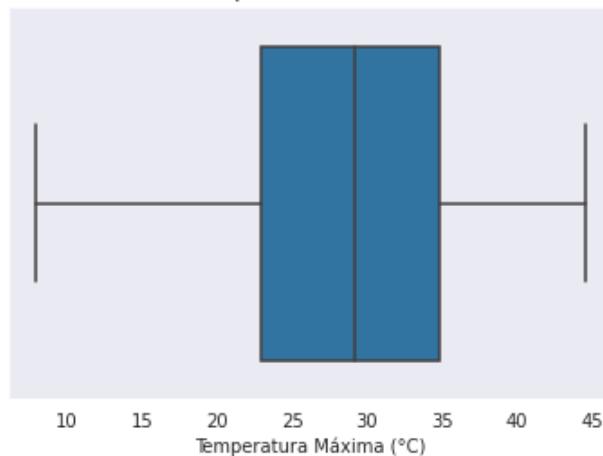
**Figura 9****Figura 10****Figura 11**

**Figura 12****Figura 13**

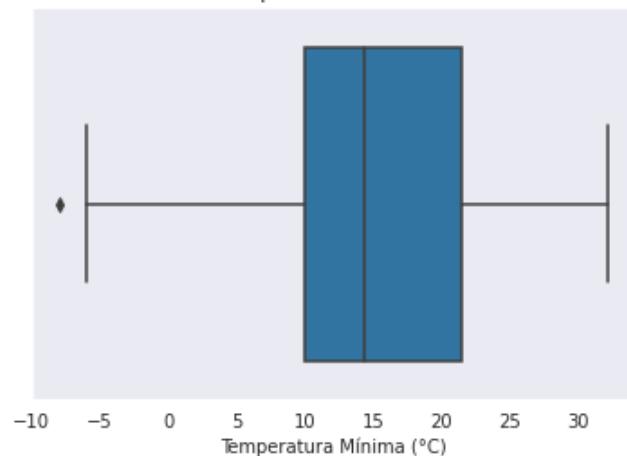
2.1.3. Actividad 4.3

**Figura 14**

Distribución de la Temperatura Máxima en Puerto Peñasco

**Figura 15**

Distribución de la Temperatura Mínima en Puerto Peñasco

**Figura 16**

2.1.4. Actividad 4.4

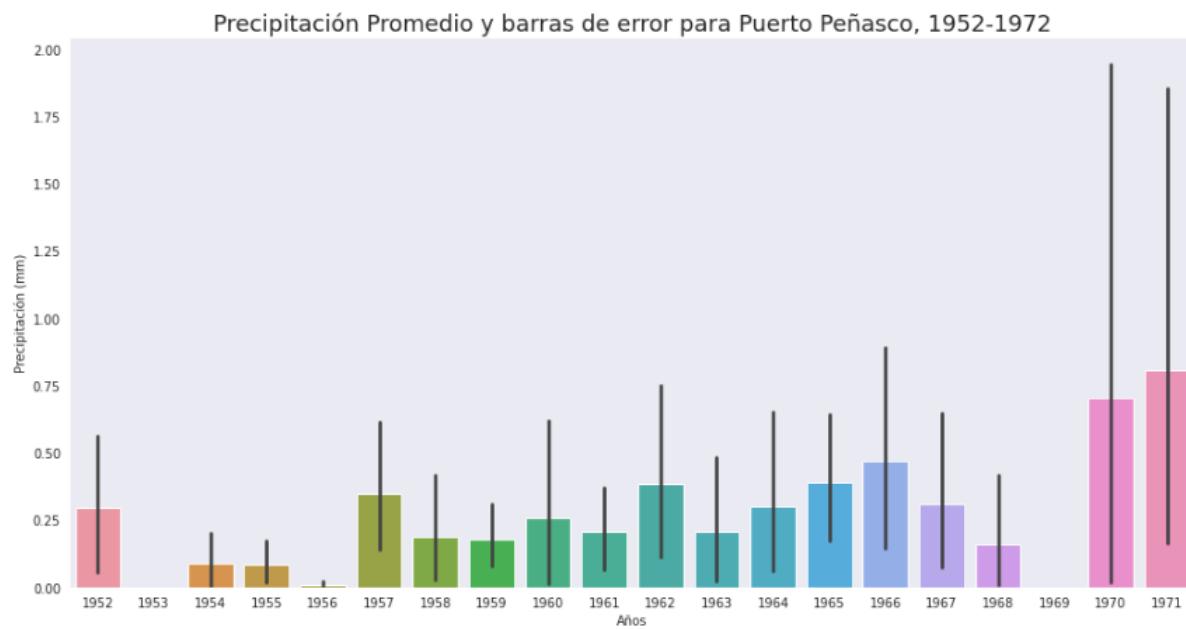


Figura 17

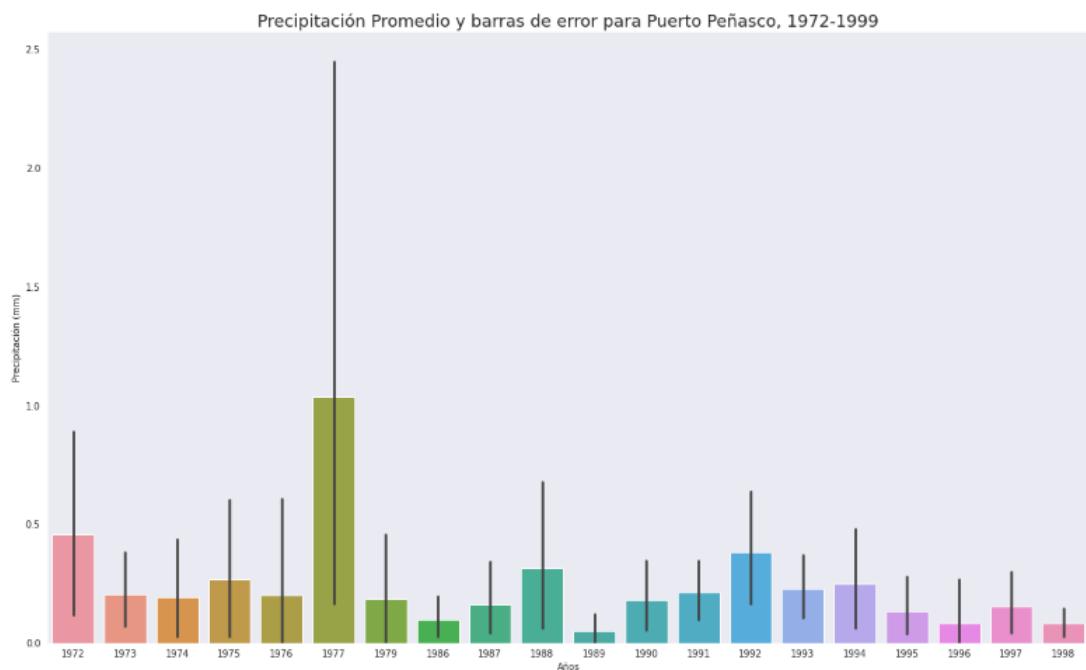
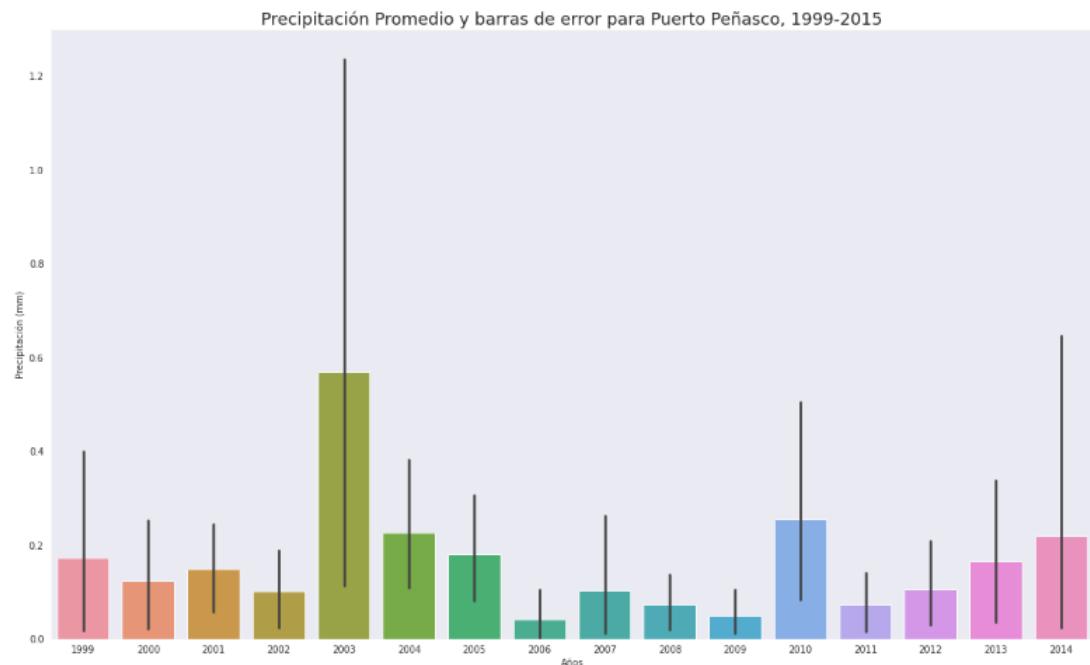
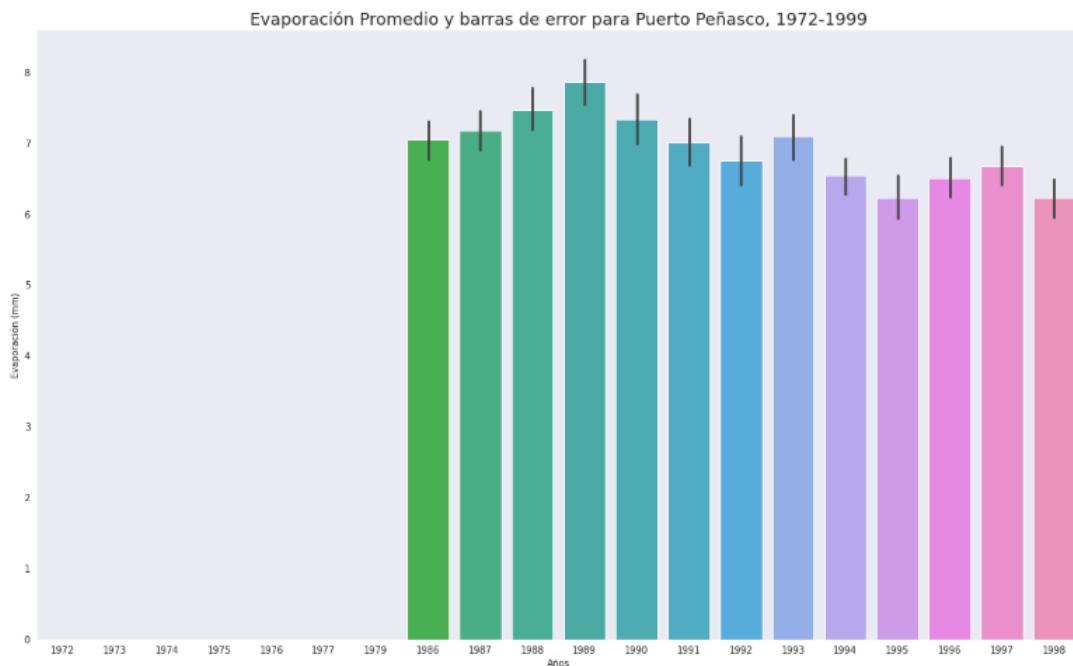
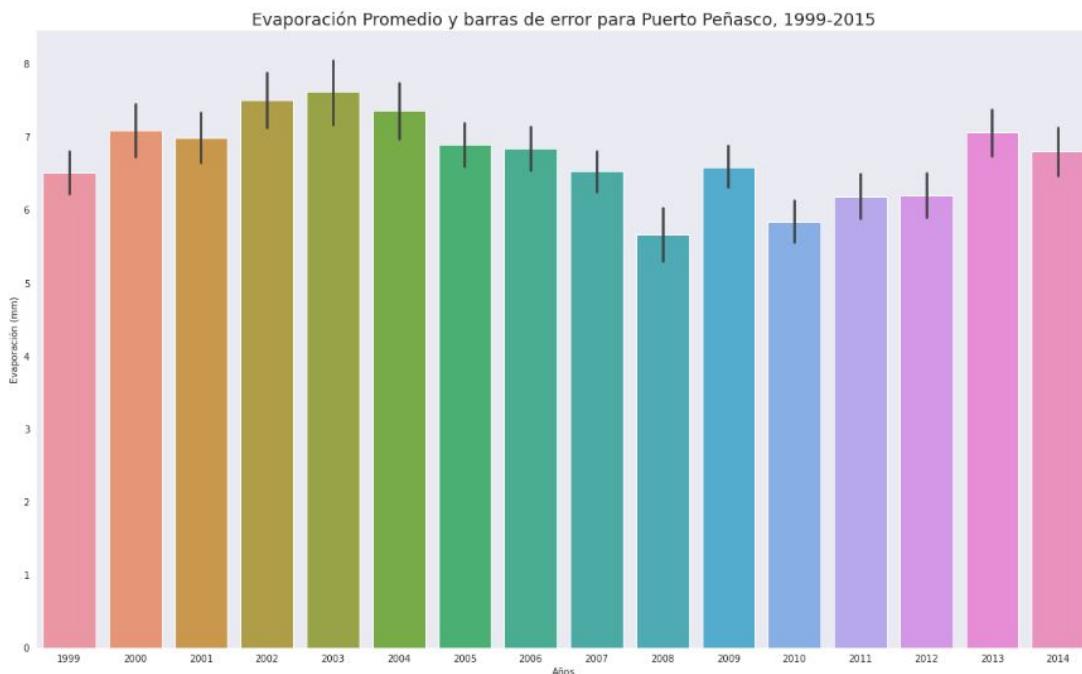
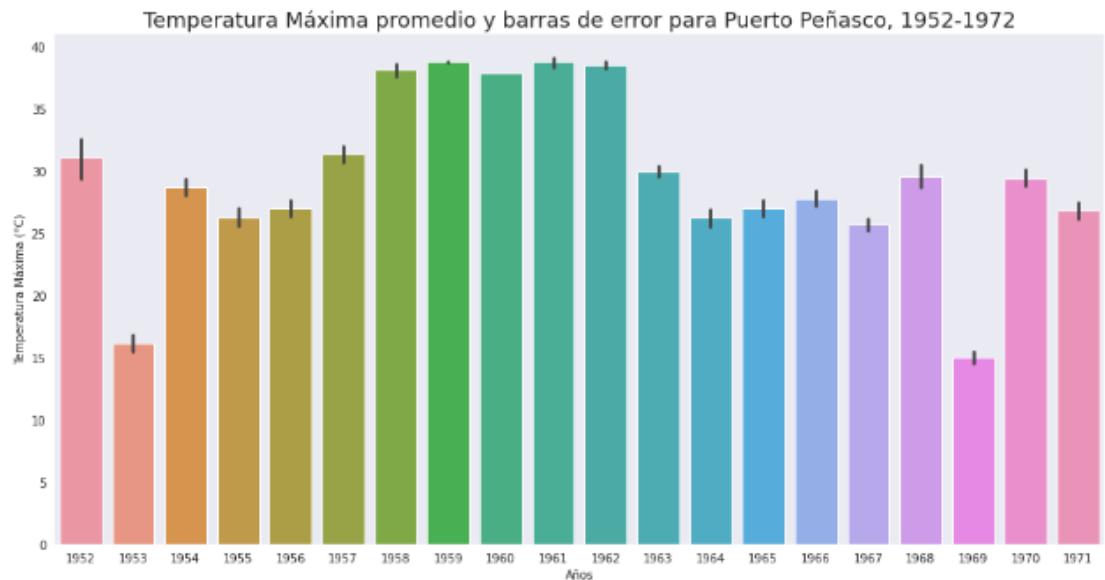
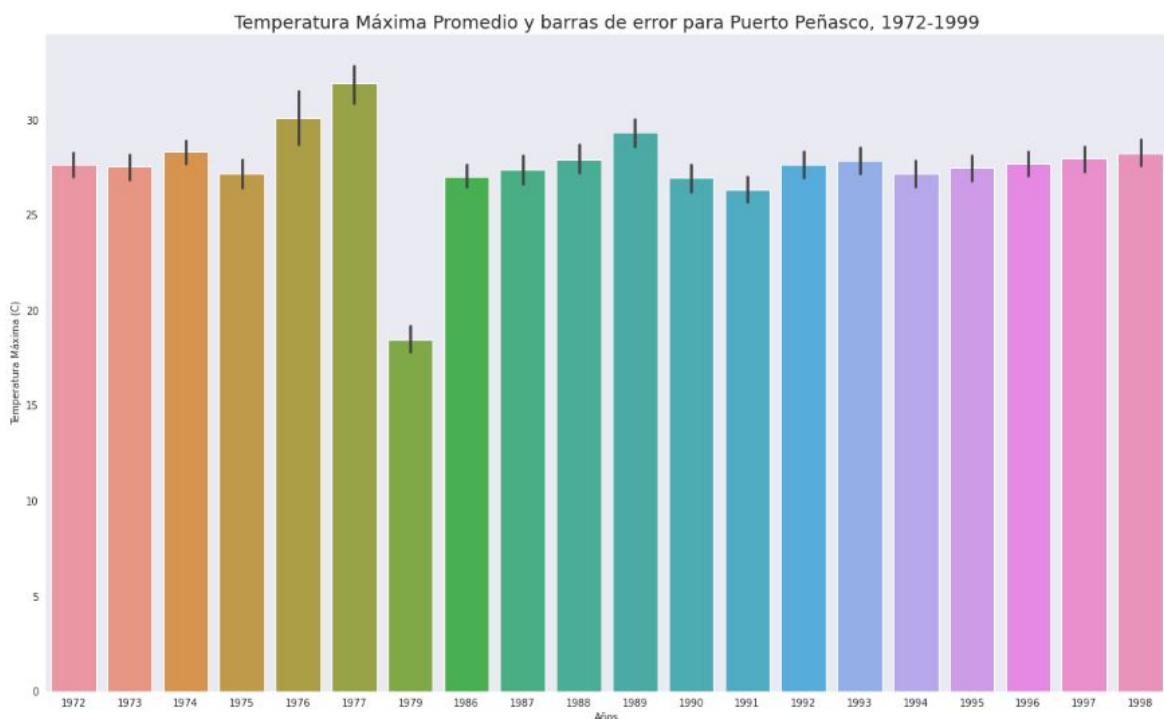
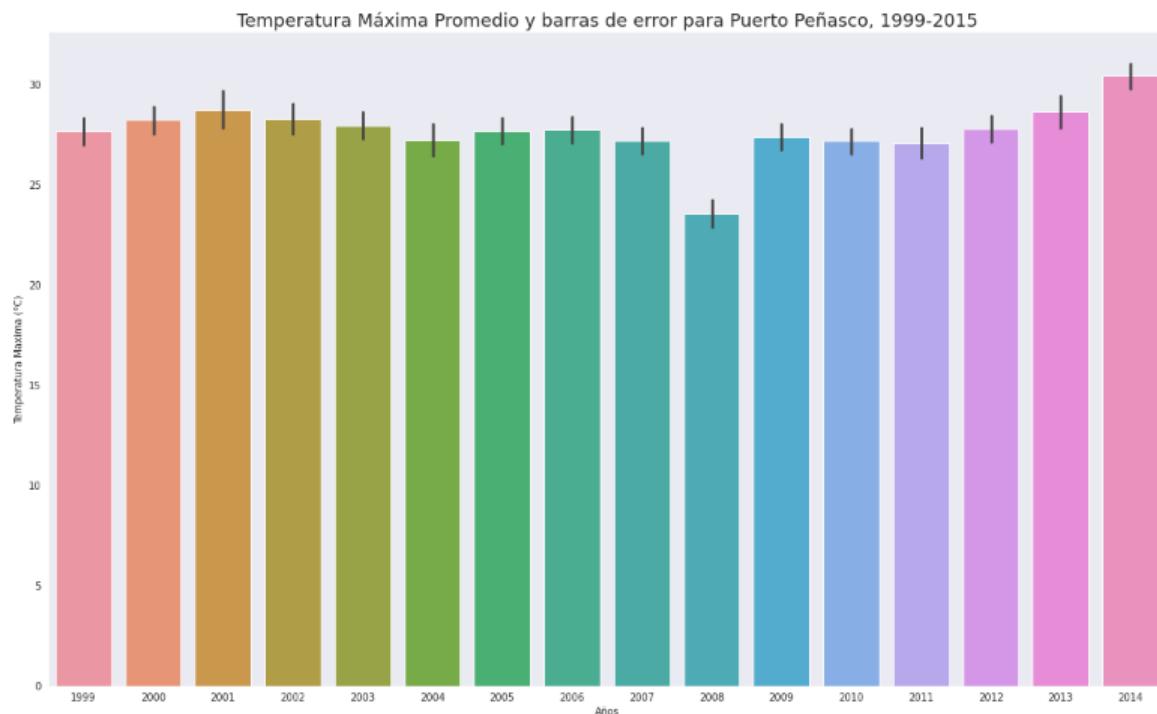
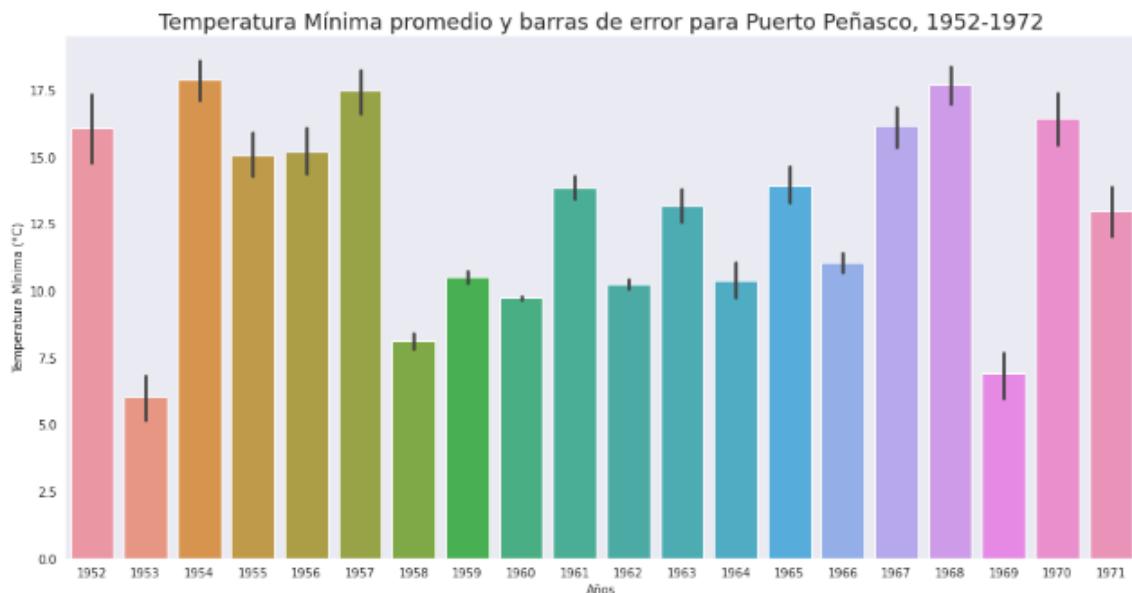


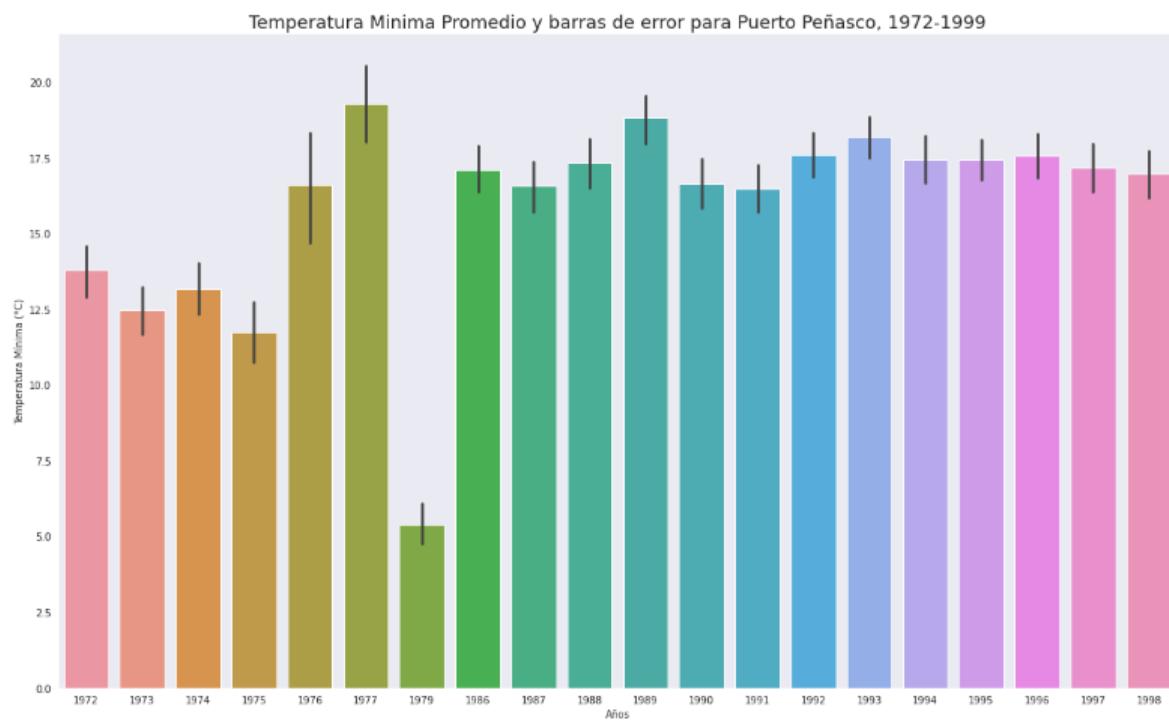
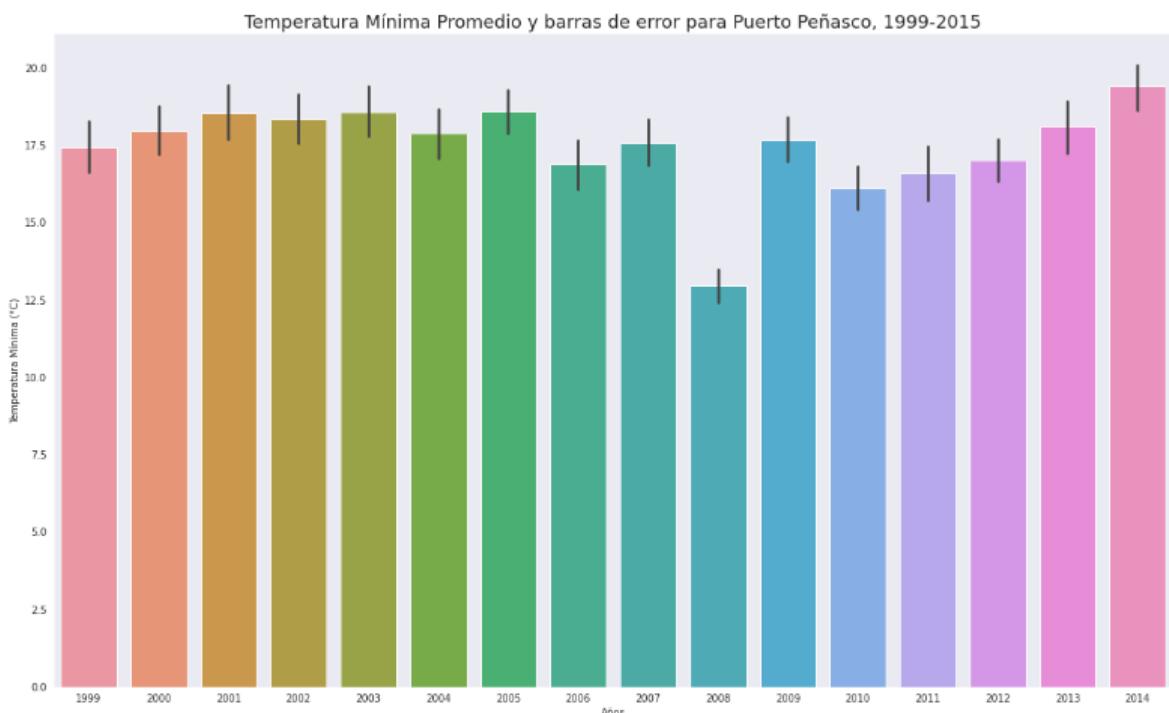
Figura 18

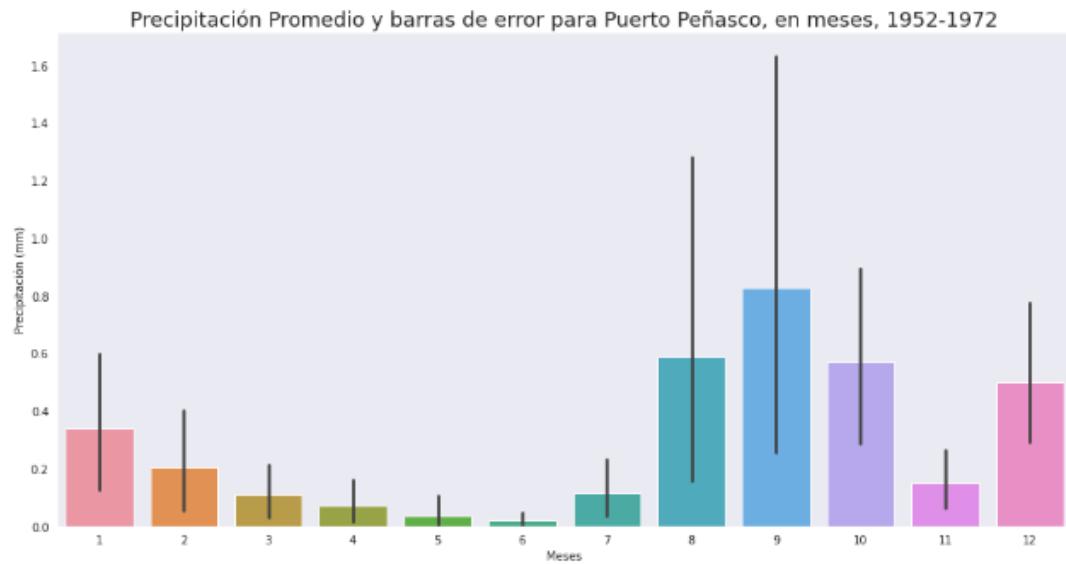
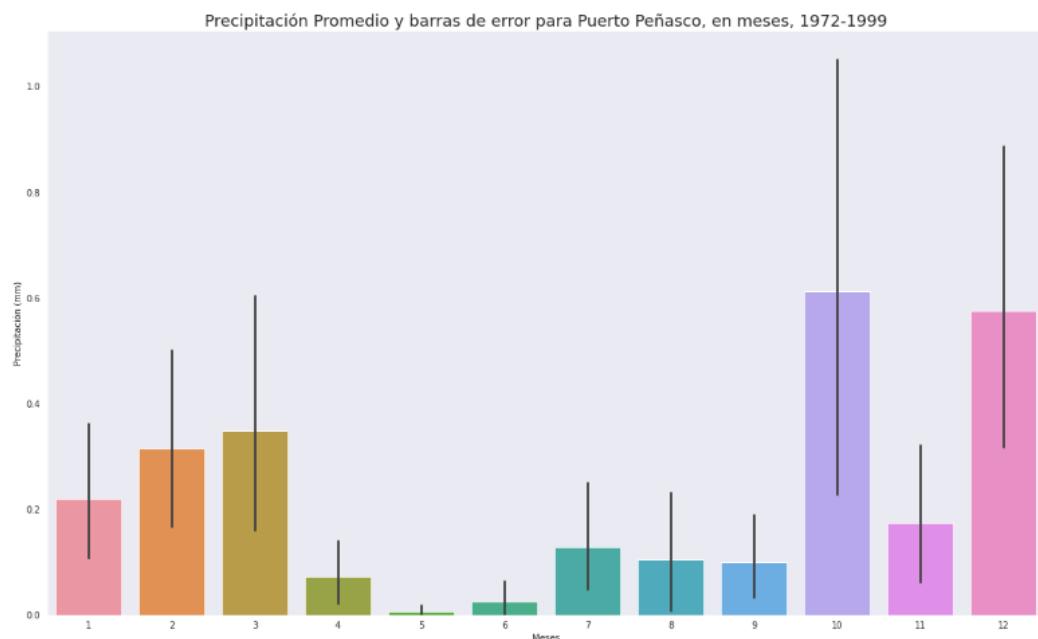
**Figura 19****Figura 20**

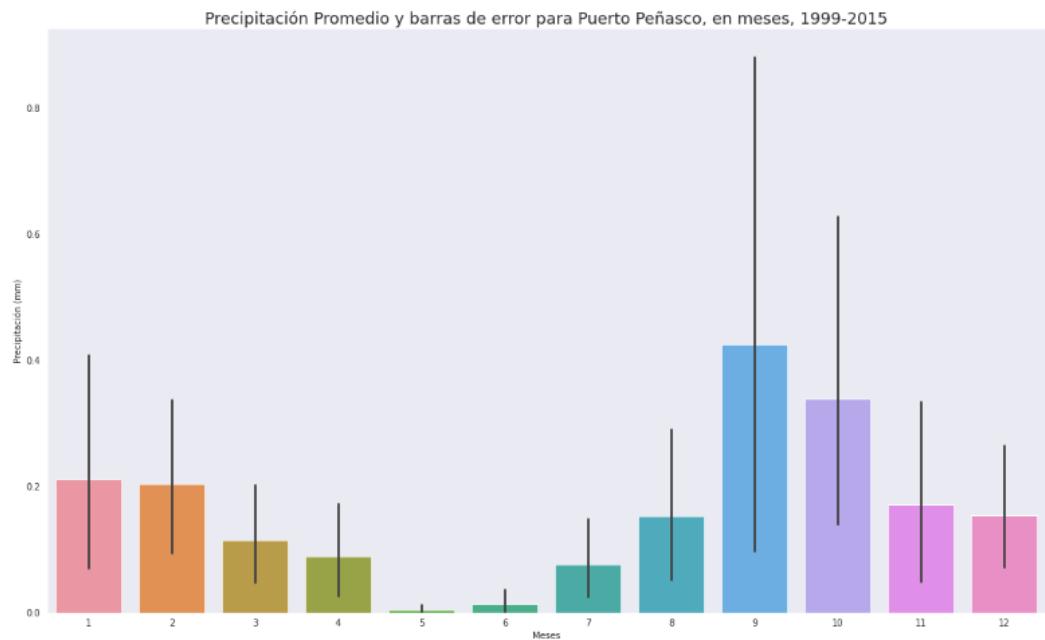
**Figura 21****Figura 22**

**Figura 23****Figura 24**

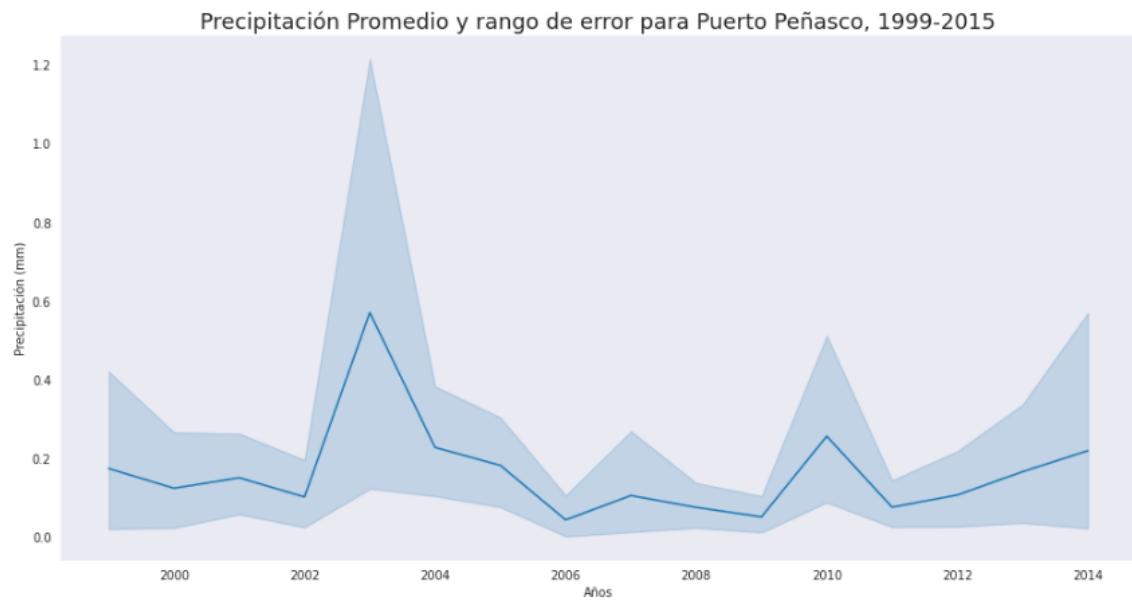
**Figura 25****Figura 26**

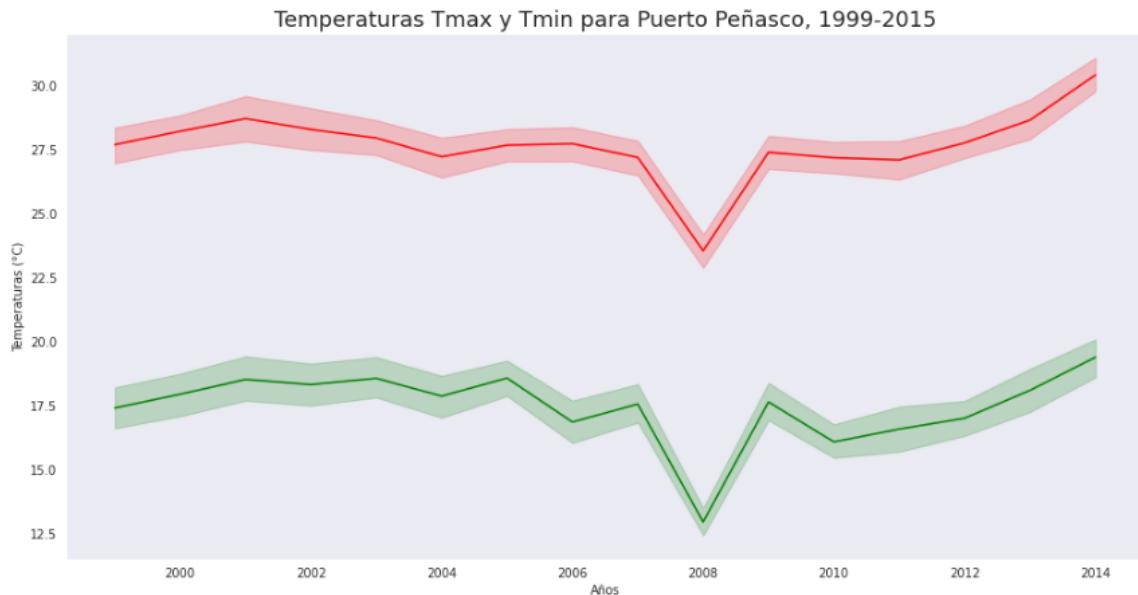
**Figura 27****Figura 28**

**Figura 29****Figura 30**

**Figura 31**

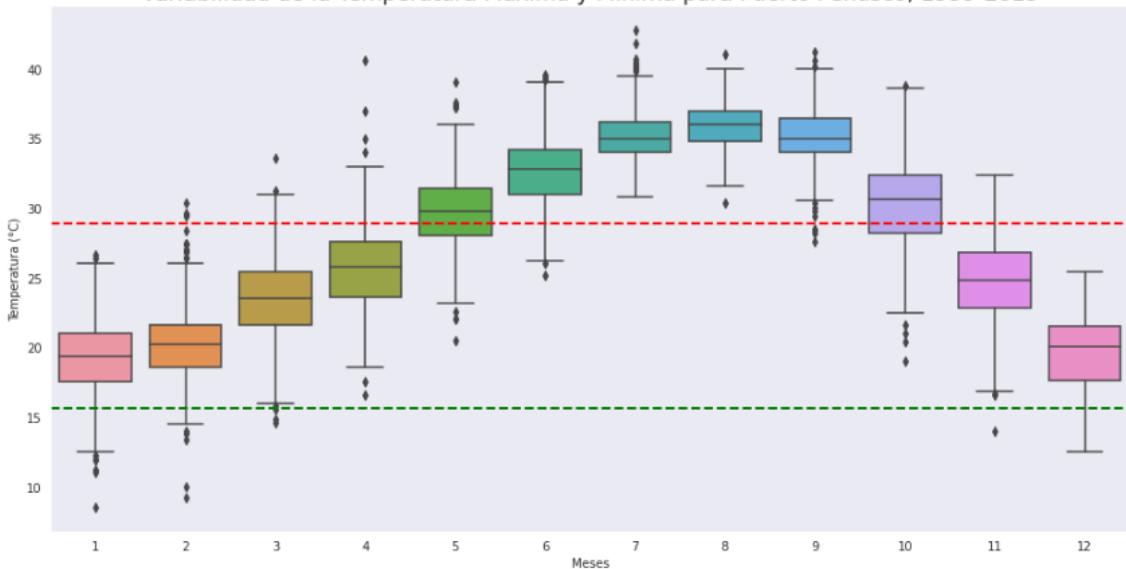
2.1.5. Actividad 4.5

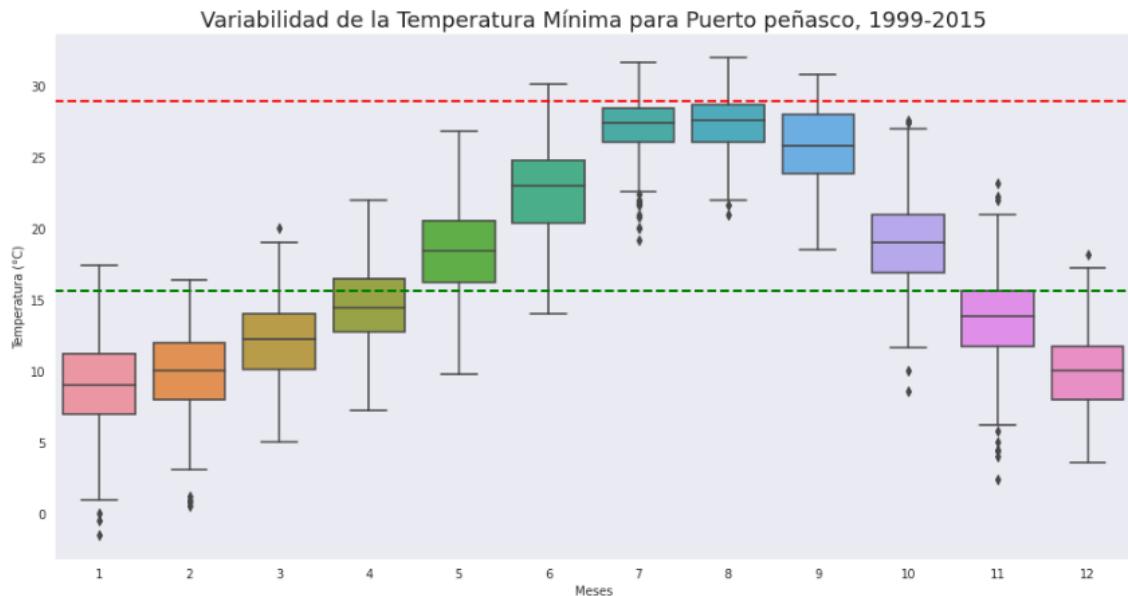
**Figura 32**

**Figura 33**

2.1.6. Actividad 4.6

Variabilidad de la Temperatura Máxima y Mínima para Puerto Peñasco, 1999-2015

**Figura 34**

**Figura 35**

3. Conclusión

En general, a pesar de la falta de datos recavados (que se vio reflejado principalmente en una gráfica vacía: Figura 20), se puede observar como es que la precipitación, la evaporación y las temperaturas varían a lo largo de los años, además de que son de acuerdo a la región que se está analizando.

Siendo un área árida y seca, se esperaría que la precipitación sea baja, al igual que las temperaturas sean altas; ambas esperanzas son confirmadas por los datos. Por otro lado, si solo se comentara que es un lugar árido y seco, se esperaría que la evaporación sea mínima también, sin embargo, podemos ver en los gráficos que la evaporación es mayor a la precipitación; esto se puede deber a que Puerto Peñasco se encuentra en una costa y podemos asumir que la evaporación es casi completamente del mar vecindante.

También podemos ver un pseudo-patrón en cuanto a que las temperaturas en general están comenzando a aumentar. Esto se lo podemos atribuir al calentamiento global, y como lugares aridos como este se están haciendo aún más áridos.

Esta actividad me pareció interesante, ya que, a pesar de la falta de datos recavados por la estación, pude apreciar como es el comportamiento del clima en mi ciudad natal. Pudiendo observar así de una forma más formal y científica lo que he estado experimentando toda mi vida empíricamente, tal como el que en esta ciudad es muy poco común que llueva, algo que es reflejado por las gráficas anteriormente mostradas.

Por otro lado, pude aprender mucho más acerca del manejo de datos usando un lenguaje de programación, lo que es algo que me servirá en el futuro, sin duda, aún cuando no se trate de datos climatológicos. Sin embargo, se que aún queda mucho por aprender en este aspecto.

La actividad en general me parecio muy bien planeada. El profesor nos ayudo con cada parte de ella en las clases y ademas me parece que las instrucciones son claras en el portal, ambas cosas ayudaron a que la actividad se pudiera llevar a cabo de una manera simple y sencilla, aunque considero que ha tenido una dificultad mayor que las anteriores. Una observación es que esta actividad me pareció más larga que el resto; combinado con el hecho de que tuve algunos problemas personales a lo largo de la semana que me impidieron trabajar a la par que el profesor, resultaron en una entrega tardía.