**PRUEBA TÉCNICA**

**IT ACADEMY 2021**

**Itinerario:**

**DATA SCIENCE**

Estudiante : Manuel Forcales

Supervisor: Kevin Mamaqi

Septiembre 2021

# Sumario

# Introducción

El objetivo de la prueba técnica final es hacer un análisis y recomendación sobre las precipitaciones de la ciudad de Barcelona basados en sus datos históricos:

Los datos hay que obtenerlos de la siguiente pagina web:
https://datosclima.es/Aemethistorico/Precipisolstad.php
y estructurarlos en formato SQL.

Realizar un plan sobre sobre el tipo de análisis a realizar y cómo se presentarán los resultados.

Utilizar PANDAS, NUMPY, MATPLOTLIB (u otras librerías) junto a un cuaderno JUPYTER para la presentación del trabajo realizado.

Subir el resultado final a un repositorio privado de Github y compartirlo una vez finalizado con el tutor.

Tiempo estimado: 72 horas.

Debe tener:
El plan sobre el análisis a realizar.
Incluir un análisis supervisado y uno no supervisado.

# Plan de Análisis

Vamos a dividir el trabajo en diferentes fases:

## Fase 1-Extracción de los datos

Realizar web scrapping de la web (https://datosclima.es/) mediante el uso de la librería Selenium para analizar los valores históricos de las precipitaciones en Barcelona así como otros valores que puedan resultar de interés para la predicción de lluvia (Temperatura, presión atmosférica, Horas de sol,...). Incorporar los Datos en una Base de Datos estructurados

## Fase 2-Limpieza y preparación de los datos

Limpiar datos, imputar o eliminar valores nulos, agregar nuevos datos si fuera necesario, eliminar “outliers” si fuera necesario,

## Fase 3-Visualización grafica

Análisis Exploratorio (EDA) del histórico de las precipitaciones.
En primer lugar, ver la tendencia de las lluvias a lo largo de estos años. Analizar valores máximos en función del año o el mes. Así mismo analizar el numero de días que llueve anualmente en función de los años y finalmente intentar elucidar si llueve menos actualmente que en el pasado debido a un posible cambio climático.
En segundo lugar, analizar el resto de datos del dataset (presión, días de sol, temperatura,etc..) e intentar relacionar los datos con las precipitaciones.

## Fase 4-Predicción de precipitaciones

Aplicar técnicas de aprendizaje supervisado y no supervisado.
Primero analizaremos mediante aprendizaje supervisado si podemos predecir lluvia o no (problema de clasificación) en función del resto de variables. Como ejemplo se propone utilizar el modelo “decision tree ”o “Nearest Neightbours”.

Finalmente, se aplicara técnicas de “clusterización” para agrupar los datos en función de sus similitudes.

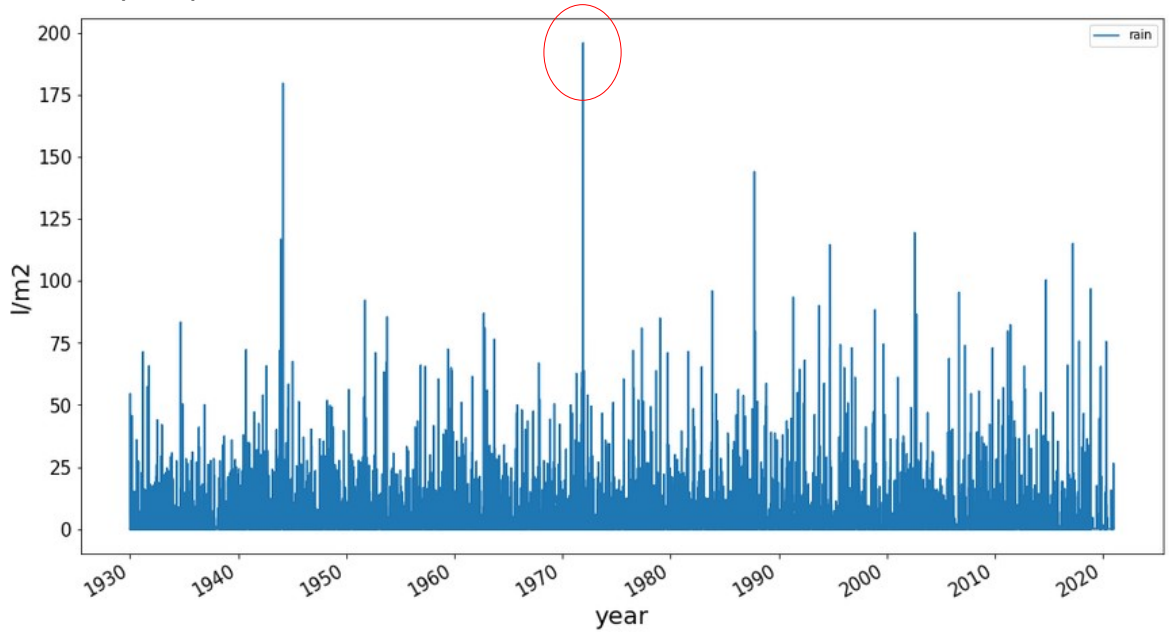
El código de cada apartado se puede encontrar en el GITHUB
link: https://github.com/ManelForcales/prueba_tecnica

# RESULTADOS

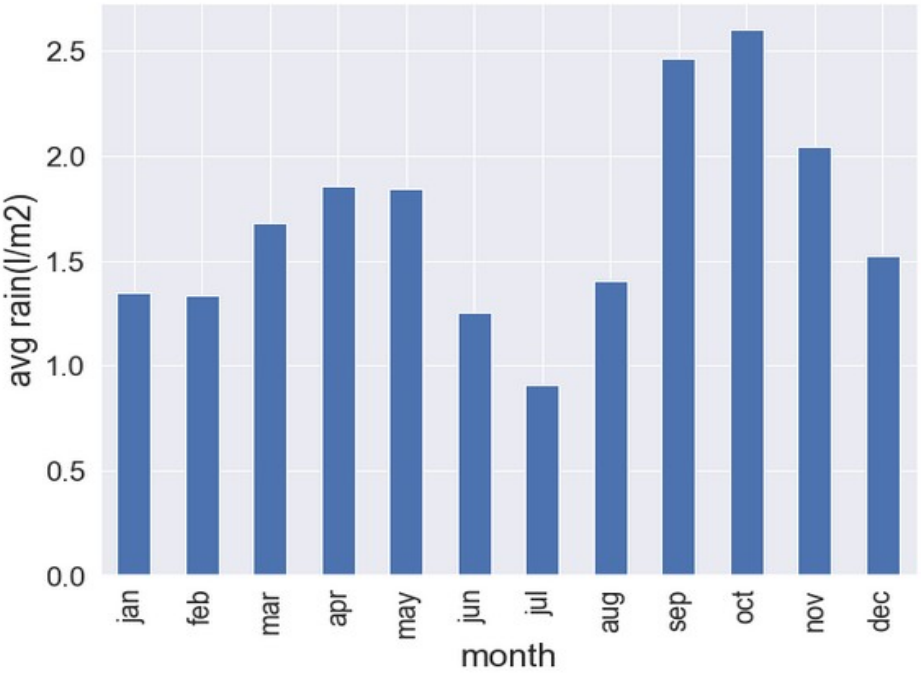
## Visualizaciones

He guardado datos de la estación Meteorológica de Barcelona del Observatorio Fabra (localizado a una altitud de 408 metros) desde el 1 de Enero del 1930 hasta el 31 de Diciembre del 2021.
Las Precipitaciones diarias están registradas en litros/metro cuadrado (l/m²).

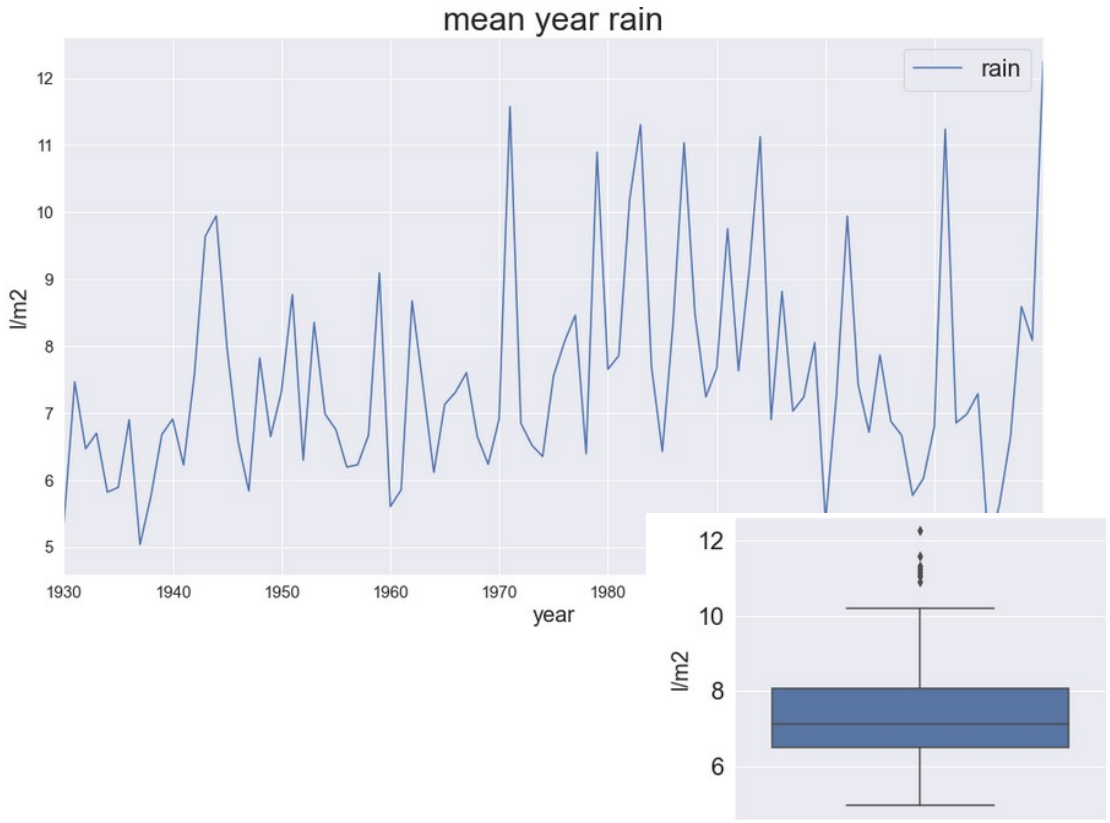
Histórico de precipitaciones diarias



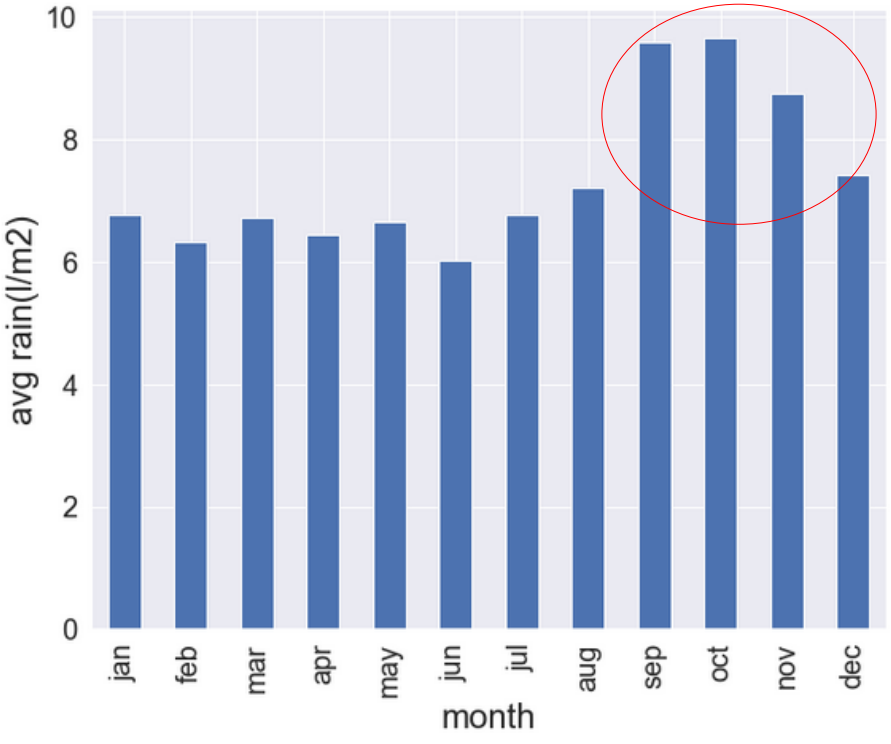

El máximo histórico de lluvias ocurrió el 5 de Diciembre de 1971 con 196 litros/m².
Ha llovido un 22.7 % de tiempo durante los últimos 90 años.
A nivel mensual, la estación de Otoño (21 Sept- 21 Dic) es la más lluviosa respecto a los días que no llueve con un promedio de 2.5 l/m² en Octubre.
.



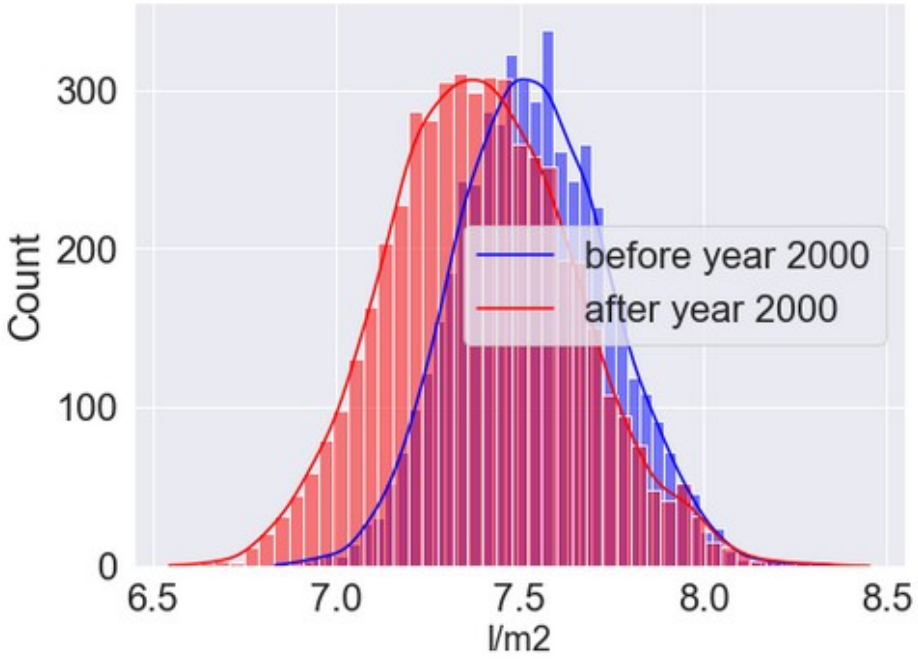
Si miramos el **promedio de las precipitaciones** anuales <u>analizando solamente los días de lluvia</u> observamos una media anual de **7.5 litros/m2**



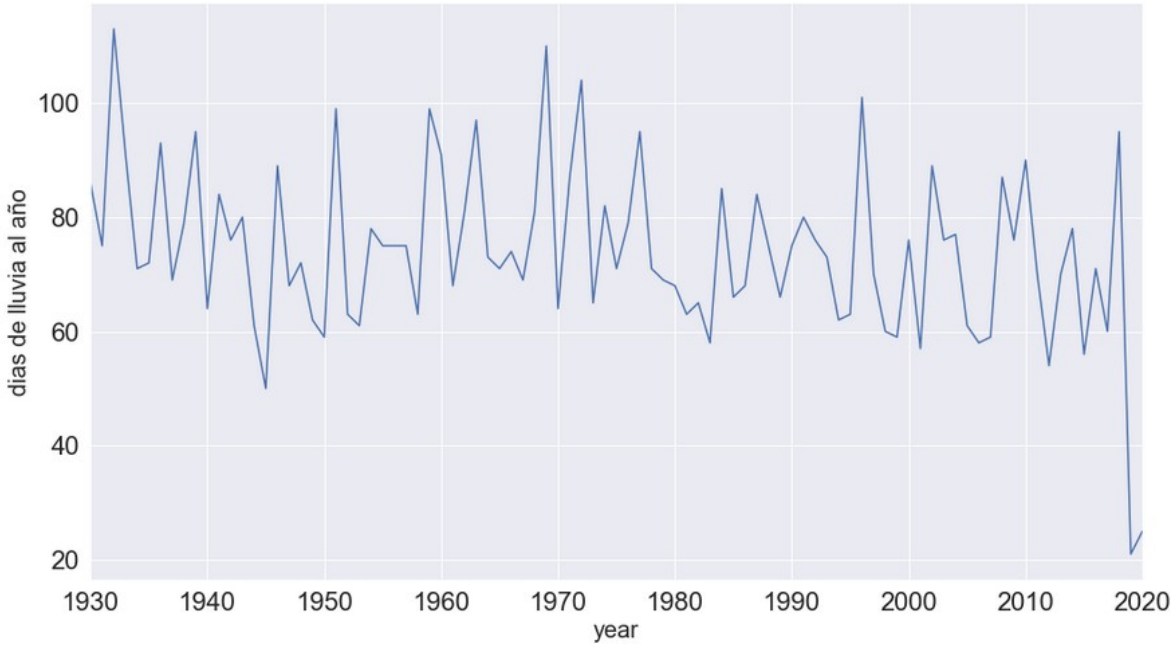

En Otoño ademas las precipitaciones son mas elevadas.



Llueve actualmente menos que en el pasado?
No existe una clara tendencia que indique que en promedio llueva menos ahora que en el pasado debido a un posible cambio climático.
Se han analizado 2 distribuciones de precipitaciones: la primera corresponde a precipitaciones anteriores al año 2000 y la 2ª para precipitaciones posteriores.
El análisis estadístico nos indica que el **P-value de las dos 2 distribuciones es > 5%** (“alpha”) con lo cual **no podemos rechazar la hipótesis nula** ($H_0$=No hay diferencias en cuanto al volumen de precipitaciones entre el presente y el pasado).

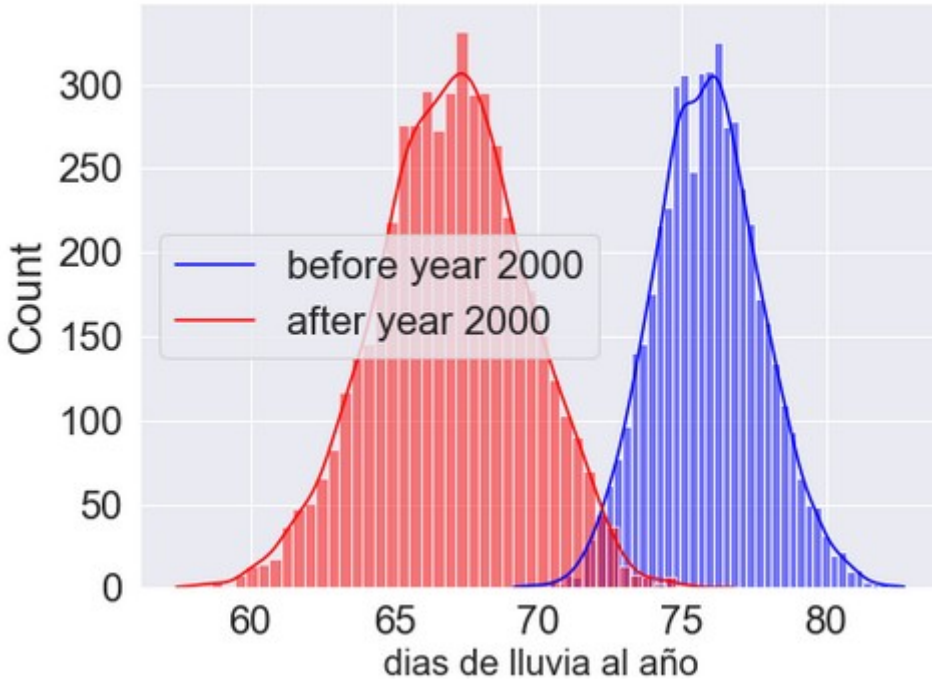



Pero si analizamos los días que han llovido al año, vemos que hay una leve tendencia a disminuir.

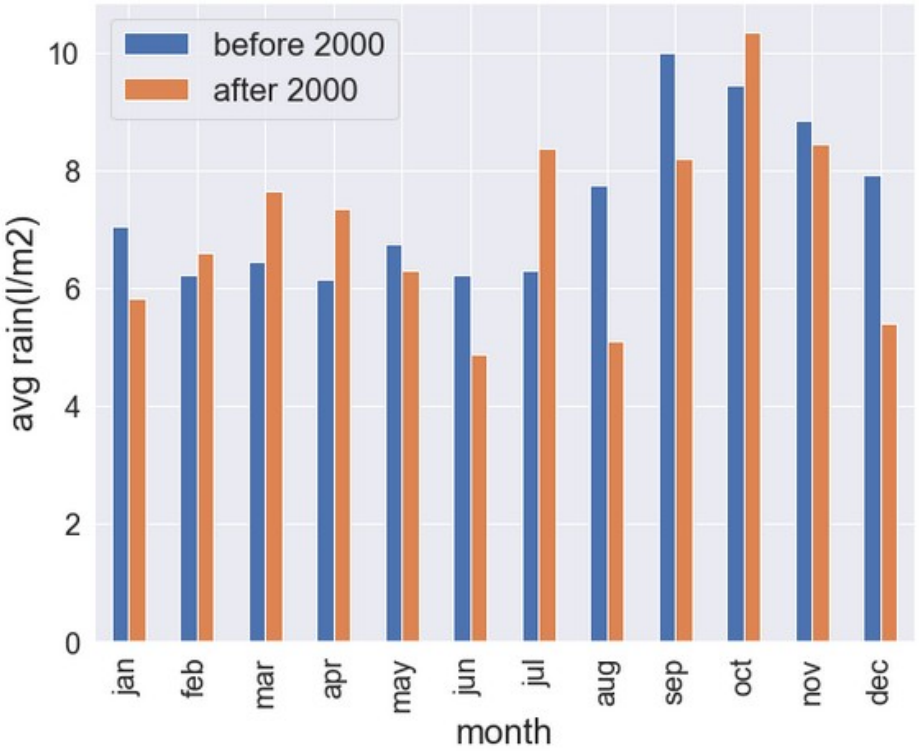



Otro análisis estadístico confirma lo siguiente: **en el pasado llovía más a menudo pero en menos cantidad y actualmente llueve menos veces pero cuando lo hace, llueve**

**con más intensidad ya que el volumen de precipitaciones (l/m2) al final del año es similar.**

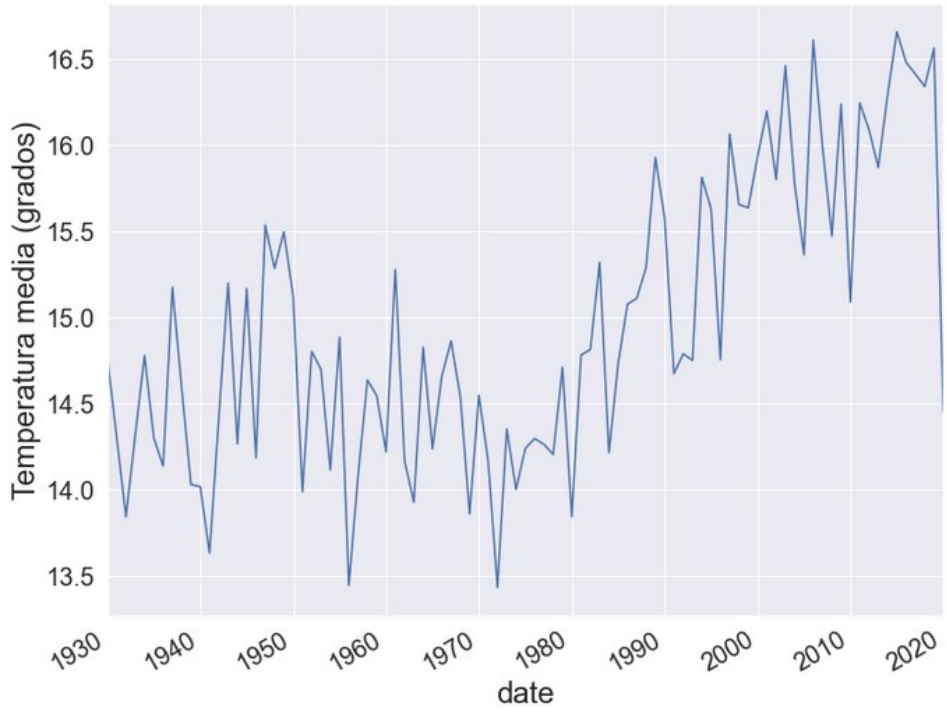



Estas diferencias también se aprecian en las precipitaciones caídas según la época del año o del mes.



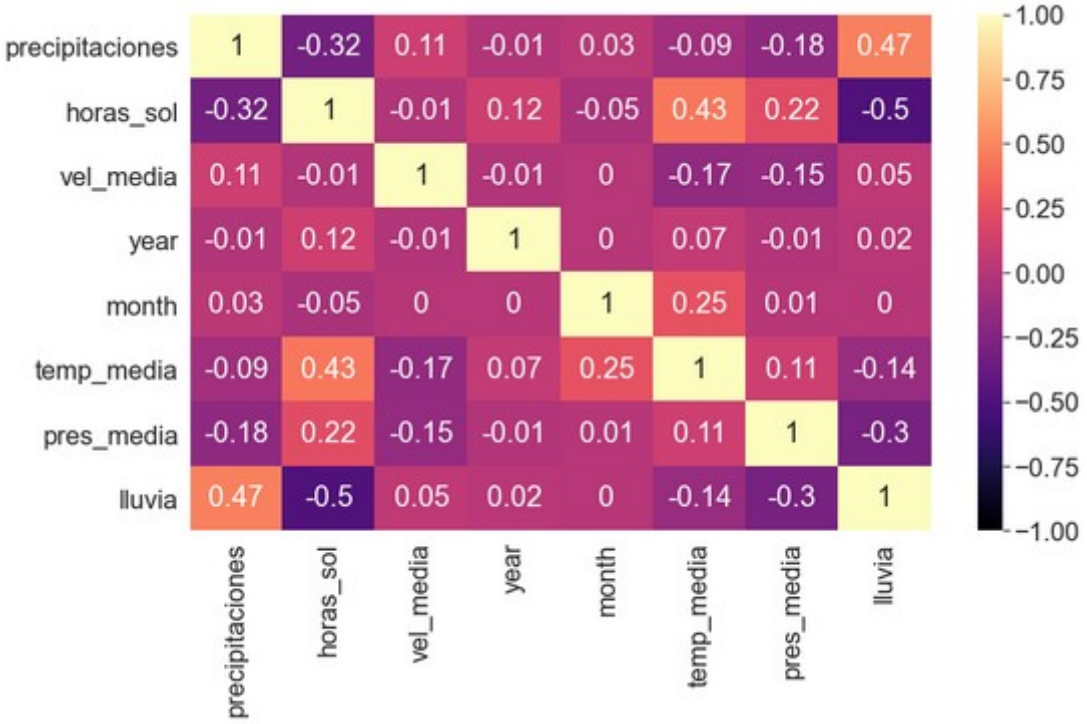

Vamos a analizar el resto de variables del “dataset” y ver que relación tienen:

Se observa por ejemplo que la temperatura media (T_max+T_min/2) ha subido más de un grado respecto a 1980. El resultado es significativo indicando que las temperaturas medias actuales son más altas que en el pasado presuntamente debido al cambio climático.

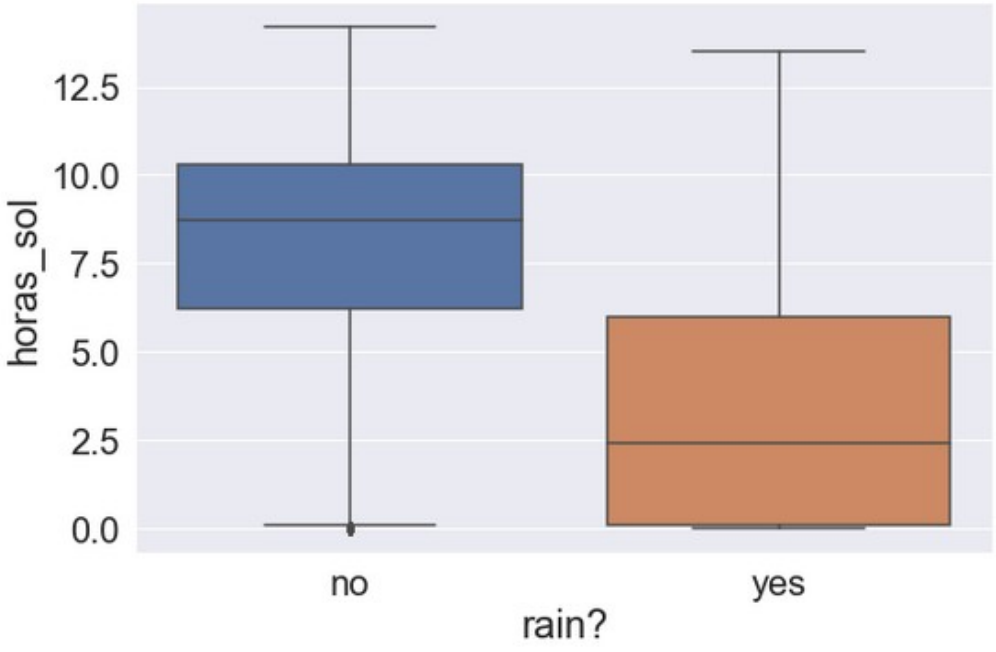


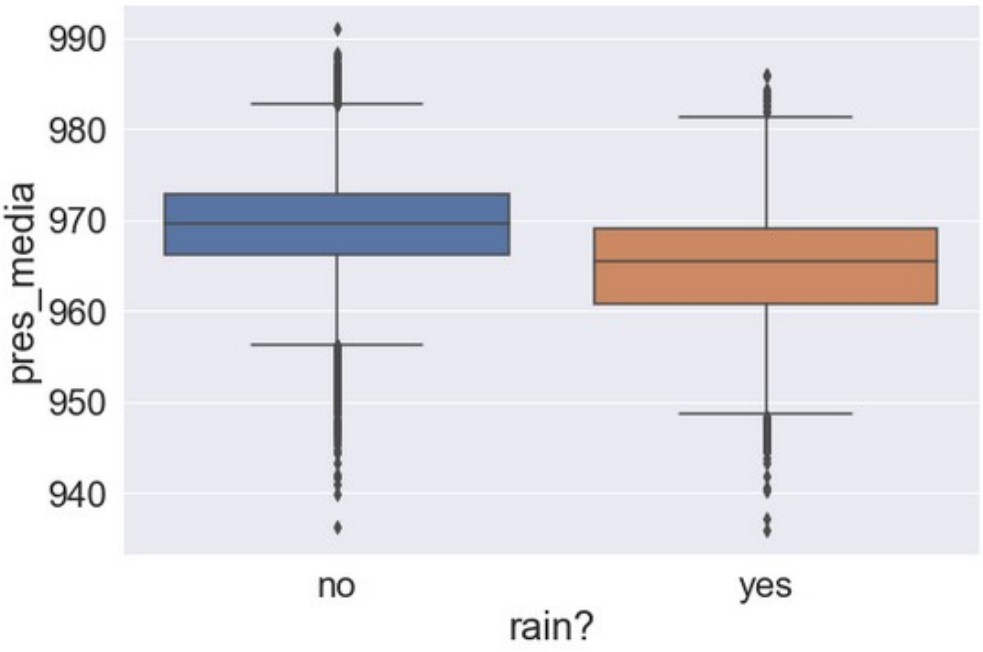

La tabla de correlaciones muestra que las variables más correlacionadas son:

- Horas de sol con temperaturas media (0.43)
- Precipitaciones con horas de sol (-0.32)
- Temperatura media con mes (0.25)
- Presión media con las Horas de sol (0.22)



Estos datos se pueden resumir en la última fila: se observa una correlación entre lluvia (si/no o 1/0) y un menor numero de horas de sol (día nublado puede ser indicador de lluvia) y por tener una menor presión atmosférica (baja presión puede ser indicador de lluvia).
En los gráficos siguientes se observa lo comentado arriba



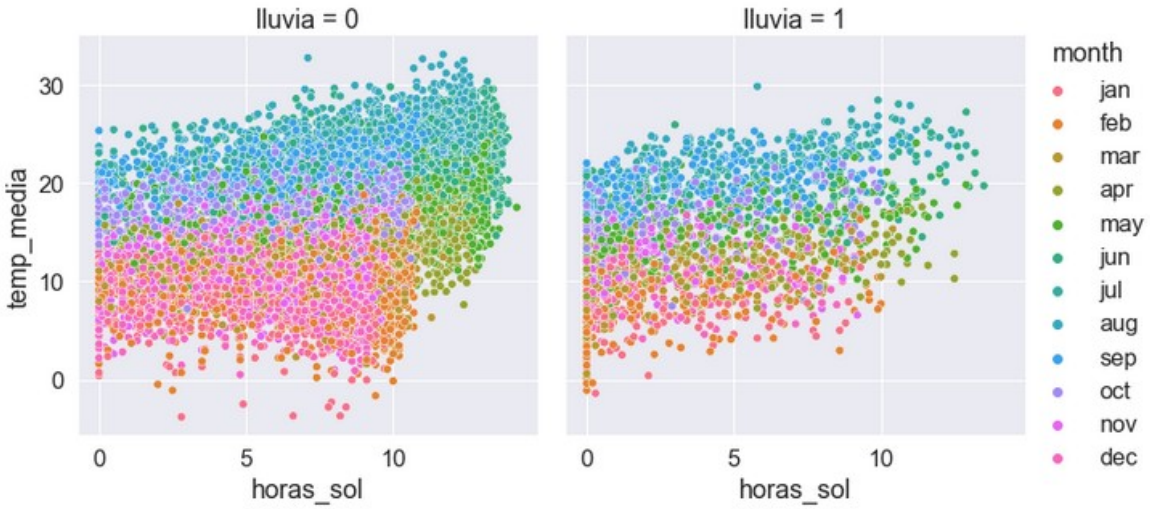




# Predicciones

## Aprendizaje Supervisado

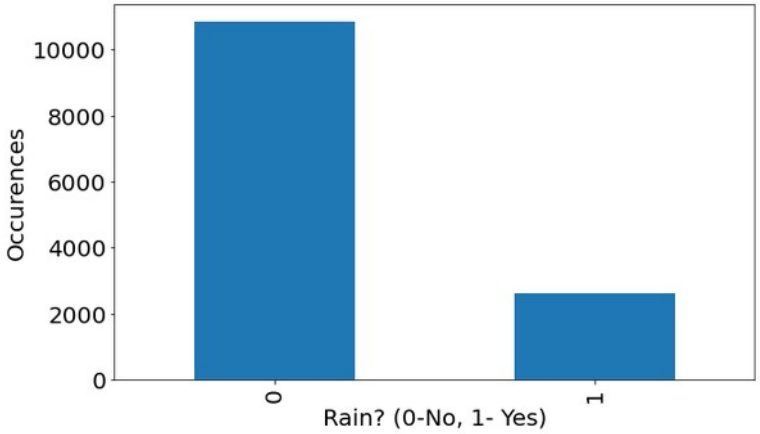
Para predecir si va a llover o no, estudiaremos el “dataset” a partir del año 1984 en adelante. Esto es debido a que para fechas anteriores, muchos valores están incompletos y no es posible inferir su valor.

La nube de puntos de la grafica (horas_sol vs temp_media)  nos indica que no hay una clara separación entre los días que ha llovido y los que no. Parece que será un reto obtener un buen clasificador utilizando estas variables.



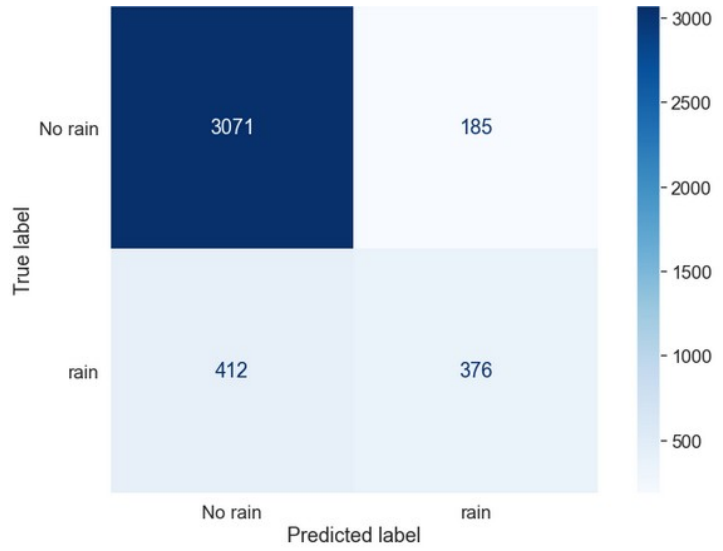

Otro ejemplo lo podemos ver cuando utilizamos la presión_media vs temperatura_media como variables para predecir lluvia. No existe una clara separación entre los puntos


El Dataset tiene más  de 10000 datos (80% del total) donde no existe lluvia y alrededor de 2600 datos (20% del total) donde existe lluvia

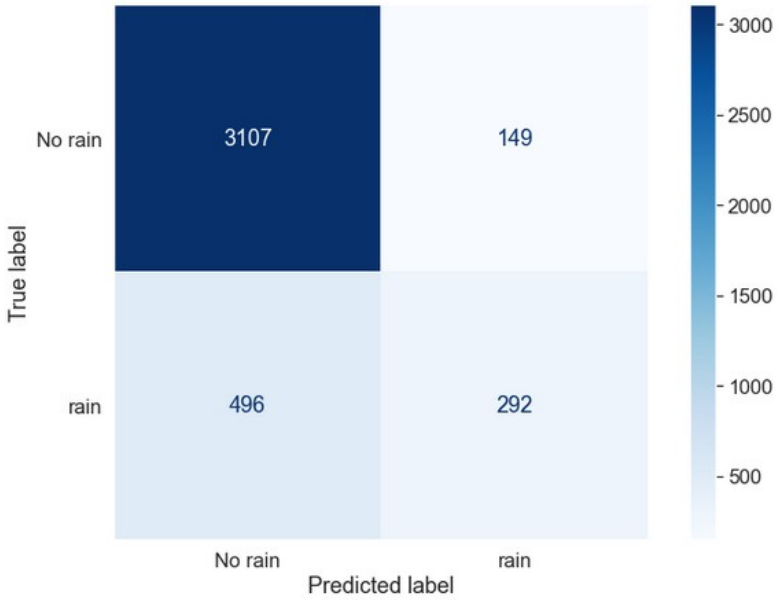


Por tanto tenemos un dataset no balanceado (aka Imbalance). Nuestro modelo debe predecir por encima del 80% sino no es mejor que un modelo puramente aleatorio.

Aplicando “decision tree” obtenemos un puntuación de 86% pero la matriz de confianza muestra que no podemos predecir bien los días de lluvia




si aplicamos Nearest Neighbors Classification, la matriz de confianza mejora

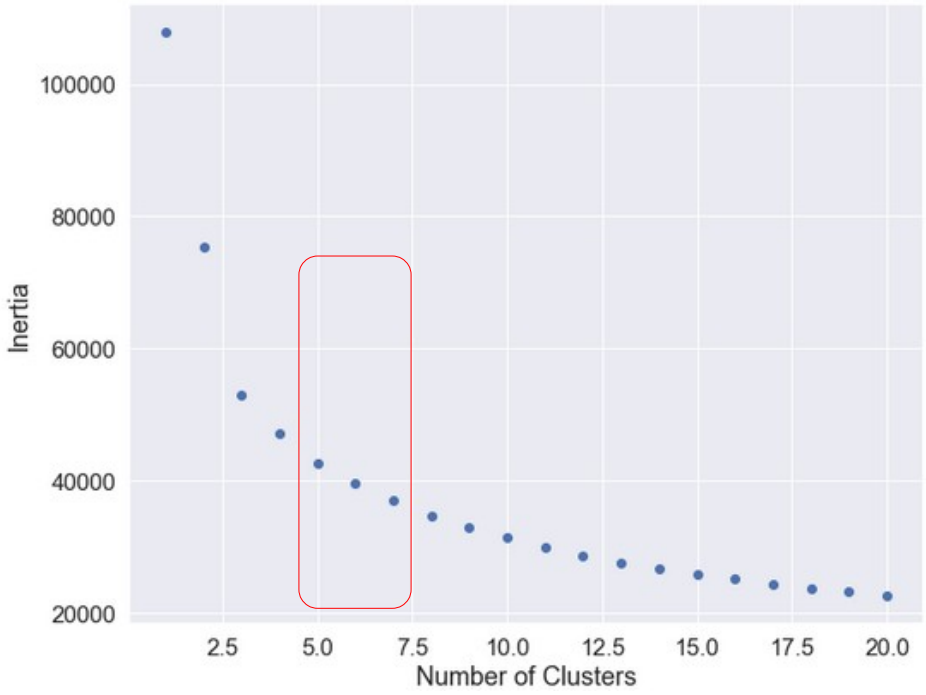



He aplicado técnicas de estandarización/normalización de variables, pero aun así los resultados no mejoran mucho respecto al 86% inicial.

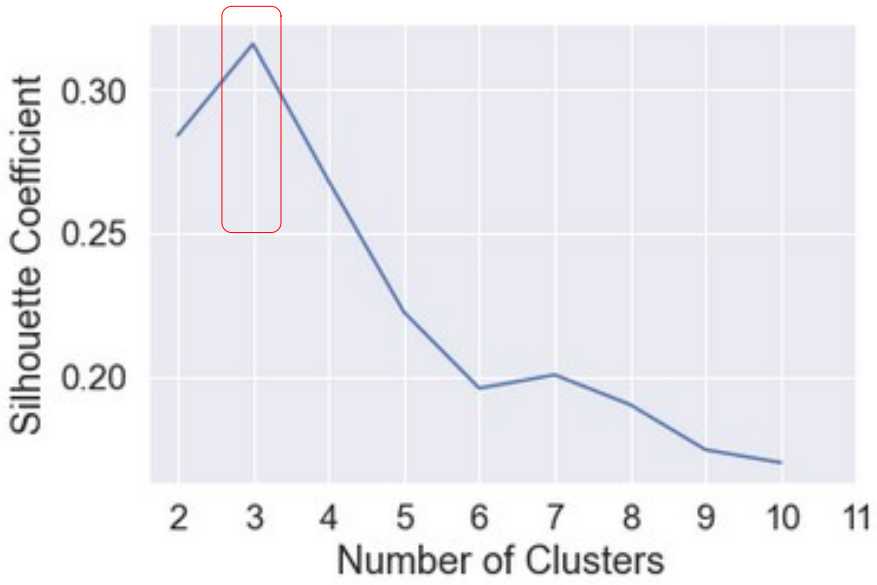
## Aprendizaje No Supervisado

Para el proceso de clusterización he quitado las variables referentes a precipitaciones y lluvias así como el mes o año para hacerlo mas sencillo. La idea es determinar el numero de clúster donde estas variables tienden a agruparse debidos a punto en común. He usado la librería Kmeans ya que es una de las mas comunes y bastante directa de implementar.

Primero determino el numero de los diferentes clúster y veo que la solución más optima es alrededor de 5- 7 clúster con el método del “elbow”

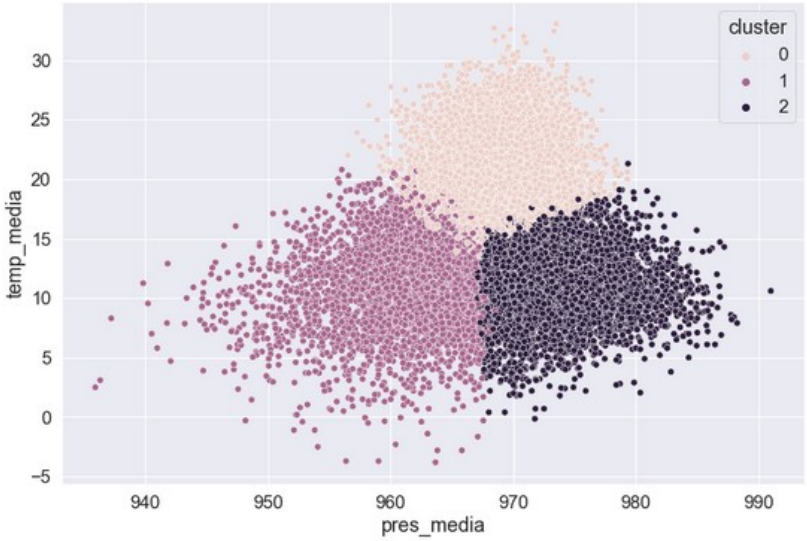



Mientras que usando el método de Silhouette, el valor más optimo son 3 clúster




Elijo 3 clúster y evalúo los resultados mirando que valores tienen asignados cada clúster (0,1,2)
En la siguiente grafica se muestra un ejemplo de los datos iniciales (sabiendo cuales son los que tienen precipitaciones y los que no) y los resultados derivados de la clusterización. Básicamente los puntos del clúster n.º 1 són los que aglutinan la mayor parte de datos que contienen las precipitaciones pero como no hay una clara separación entre puntos muchos de ellos quedan mis-clasificados.


lluvia
0
1
temp_media
30
25
20
15
10
5
0
−5
940
950
960
970
980
990
pres_media


cluster
0
1
2
temp_media
30
25
20
15
10
5
0
−5
940
950
960
970
980
990
pres_media