Actividad 2: Aprendizaje no supervisado en series temporales: clubes de convergencia.

Juan Smoke es Data Analyst en un grupo multinacional que fabrica y vende labores de tabaco en todo el mundo. Así, como primer proyecto de Inteligencia de Negocio, se le ha encomendado el análisis detallado del mercado de tabacos español. Para ello, ha recurrido a las estadísticas que publica el Comisionado para el Mercado de Tabacos español en cuanto a las ventas provinciales de cigarrillos desde enero de 2005 hasta diciembre de 2021. En este sentido, Juan ha obtenido un dataset que contiene las siguientes variables:

* Provincias: representa la zona geográfica de la que proviene la observación.
* Year: hace referencia al año al que corresponde la observación.
* Mes: en número (del 1 al 12), representa el mes.
* Día: hace referencia al día del mes (esta variable es orientativa, se debe utilizar la serie temporal mensual).
* Fecha: es la concatenación de las variables anteriores (en formato año-mes-día).
* Población: representa la población adulta (más de 18 años) total de cada provincia y período.
* Unidades: son las ventas en cajetillas de cigarrillos.
* Euros: son las ventas en euros corrientes.
* Precio: esta variable es el precio medio de las cajetillas vendidas en cada provincia y período (se calcula como el cociente entre Euros y Unidades).
* UdesxCapita: son las ventas per cápita en cada provincia y mes (se calcula como el cociente entre Euros y Población.

El dataset se llama **dataset\_provincias.csv**. En base a este conjunto de datos, la dirección general de la tabaquera pide a Juan, antes de iniciar el proyecto de Inteligencia de Negocio, que conteste a las siguientes preguntas:

* ¿De qué tipo de variables se compone el dataset? Realiza **una representación gráfica** que permita conocer el comportamiento de la serie temporal de las ventas de cigarrillos en cada una de las provincias. Describe brevemente los hallazgos más relevantes tras este conocimiento inicial.
* Haz **una representación gráfica** que permita conocer, de forma ordenada, cuál es el ranking de las **provincias de más a menos venta (toma como referencia la media de UdesxCapita)**.
* La dirección general quiere llevar a cabo acciones diferenciadas en las distintas provincias españolas. Por ello, se le pregunta, **¿podemos decir que existe convergencia absoluta** (las ventas por provincia se comportan de manera homogénea) en las ventas de cigarrillos? En base a los resultados alcanzados, ¿qué podemos contestar a la dirección general? ¿Se pueden llevar a cabo acciones diferenciadas?
* ¿En el caso de que no exista convergencia absoluta, en **cuántos grupos (clubes) podemos dividir las provincias** en función al comportamiento de las series temporales? ¿Existe algún patrón de comportamiento geográfico y/o relacionado con el ranking elaborado en el segundo apartado?
* Haz una **predicción a 12 meses para, como mínimo, una provincia de cada club**, ¿Hay clubes cuyas series temporales sean más fácilmente predecibles que otras? Justifica tu respuesta.

**Rúbrica:**

|  |  |  |  |
| --- | --- | --- | --- |
| Actividad final | Descripción | Puntuación máxima  (puntos) | Peso  % |
|  |  |  |  |
| Criterio 1 | Es capaz de describir hallazgos en cuanto a las diferencias entre las ventas de las distintas provincias. | 2 | 20% |
| Criterio 2 | Realiza la representación y responde a la pregunta de manera razonable (aunque los gráficos sean mejorables). | 2 | 20% |
| Criterio 3 | Es capaz de hacer la prueba de convergencia absoluta e interpreta correctamente el resultado alcanzado. | 2 | 20% |
| Criterio 4 | Aunque no sean correctos los resultados alcanzados, habla sobre patrones de comportamiento. | 2 | 20% |
| Criterio 5 | Aunque las predicciones no sean correctas, es capaz de diferencias entre la predicción realizada para cada grupo. | 2 |  |
|  |  | **10** | **100 %** |