Análisis de Datos – Turismo Internacional en España 2019-2024

Tema

En el siguiente trabajo se ha realizado un análisis de las fluctuaciones turísticas internacionales a nivel provincial en España. Para ello, se tendrán en cuenta datos desde antes del confinamiento provocado por la pandemia de COVID-19.

Hipótesis

1. El sector turístico se ha recuperado volviendo a las mismas cifras que tenía antes de la pandemia.
2. A lo largo de 2020 provincias menos frecuentados ganaron popularidad frente a provincias principalmente turísticas.

Los datos provienen de la página web www.dataestur.es. Dispone de datos relacionados con la procedencia de los turistas extranjeros que visitan España a nivel provincial. Los datos seleccionados son desde el 07/2019 hasta el 10/2024 y se han incluido todas las provincias de España, los continentes y país de origen, la pernoctación total y la estancia media.

Los datos permitirán estudiar el impacto que ha tenido la pandemia y su respectivo confinamiento en el sector turístico de España, así como si el sector turístico se ha recuperado en la actualidad. Además, permitirá mostrar si otras provincias menos visitadas han conseguido mantener los niveles de turistas y pernoctaciones desde la pandemia.

Obtención de Datos

Para ello, se ha descargado los datos desde la página oficial turística “Dataestur” y se ha utilizado “Visual Studio Code” para poder cargar la información, limpiarla y analizarla con el fin de conseguir unas conclusiones que permitan constatar o refutar las hipótesis planteadas.

Para poder interpretar la información se ha tenido que indicar que el separador utilizado es “;”, el código está en “latin1” y que el símbolo decimal es “,”. Debido a ello, se podrán detectar tildes y caracteres alfabéticos como la “ñ” a lo largo del trabajo.

Primera Exploración

Tras cargar las librerías y realizar una copia del dataframe se ha procedido a realizar una primera exploración que muestra que está compuesto por 190578 filas y 8 columnas.

Las columnas están compuestas por 2 int64 (año y mes), 3 objetos (provincia de destino, continente de origen y país de origen) y 3 float64 (turistas, pernoctaciones y estancia media), todas están escritos en “Screaming Snake Case”. La información ocupa más de 11,6 MB de memoria. Se observa que estancia media tiene 190493 datos de 190578 (85 null) que corresponde con 0,04% de la información, sin embargo, no dispone de datos duplicados.

Por otra parte, al realizar un estudio de los datos únicos del “dataframe”, se puede ver que los años estudiados se encuentran entre 2019 y 2024, los meses están representados en números del 1 al 12 y que la provincia destino, el continente de origen y el país de origen tienen al menos una fila que resume la información en un “Total” que van a causar problemas a la hora de realizar el estudio.

Limpieza y Preprocesado

En primer lugar, se ha procedido a analizar y corregir los datos faltantes (null). Para ello se ha buscado dónde se encuentran los datos. Los datos corresponden con abril y mayo 2020, aunque ha habido turistas no han realizado ninguna pernoctación en el destino, por lo que se le ha asignado el valor 0 a la estancia media de cada turista.

Después, se ha procedido a tratar con los Outliers, a través del filtrado de los datos deseados, eliminando los totales ad hoc.(Pendiente de terminar)

Análisis Exploratorio de Datos (EDA) Exploratorio (He llegado hasta este punto)

Conclusión (Pendiente de realizar)