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 turísticos 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
3. Tras el confinamiento, las pernoctaciones decrecieron y en la actualidad han vuelto a las mismas cifras que antes de la pandemia

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 mantenido 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.

Librerías Utilizadas

Las librerías utilizadas han sido:

- Numpy
- Pandas
- Matplotlib.pyplot
- Plotly.express
- Seaborn
- Json

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.

Además, se observa que, algunas provincias tienen dos nombres (uno en español y el otro en su correspondiente idioma) por respeto se mantendrán de esa forma.

Por último, aunque hay datos sobre Álava y Guipúzcoa, no hay información sobre Vizcaya.

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 los meses de abril y mayo 2020, aunque ha habido turistas no se 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 formatear los datos, se han cambiado los datos de turistas y las pernoctaciones a números enteros, y se ha cambiado el formato de mes a un formato de texto para que tengan todos los meses dos dígitos (01, 02, 03… 10, 11, 12) y se mantenga el orden. También se ha cambiado el año a formato de texto para poder concatenarlo con el mes más adelante.

Se ha eliminado datos que no son necesarios y son perjudiciales para el estudio como los totales provinciales y se ha filtrado la información por los totales del continente de origen, seleccionando (América, Asia, Europa, Oceanía y África).

Por último, se han creado dos dataframes, uno con los valores totales anuales y otro en el que se ha creado una columna nueva llamada fecha, que reúne el año y el mes (anteriormente convertidos en texto). En ambas columnas, se ha sacrificado el país de origen para simplificar el estudio.

Análisis Exploratorio de Datos (EDA)

Para poder entender los datos se ha realizado un análisis univariante y un análisis multivariante.

En el análisis univariante, se han calculado los cuartiles, el rango mínimo y el máximo. En la actualidad hay 3328 datos mensuales entre el 07/2019 y el 10/2019. Para facilitar la lectura se resumen en el cuadro siguiente.

	TURISTAS	PERNOCTACIONES	ESTANCIA_MEDIA
Media	1.105.854	8.121.999	8,331430
Desviación Estándar	232.748	1.561.353	2,275669
Mínimo	1.679	20.081	3
Máximo	2.428.342	15.784.772	18
Percentil 25	14.812	125.959,5	7
Percentil 50 (Mediana)	32.256,5	259.137	8
Percentil 75	797.662,5	6.267.038	9

A continuación, se ha confirmado la información a través de histogramas y scatter plots, donde también se ha podido observar que la cantidad de turistas y sus pernoctaciones totales, efectivamente decreció en los años 2020 y 2021, pero en la actualidad se ha recuperado y en algunos casos se ha duplicado. Sin embargo, la media de noches en un destino aumentó en el año 2020 y ha tenido una tendencia a la baja hasta la actualidad. Además, se puede observar que la mayor parte de los turistas internacionales se alojan en España durante una media de 6 noches.

Después se ha realizado un boxplot de los turistas, su estancia media y la totalidad de pernoctaciones que han mostrado que hay un elevado número de outliers en los sectores de turistas y sus pernoctaciones y han confirmado la información anteriormente descubierta.

Por último, se han creado unos mapas interactivos que muestran la fluctuación de la cantidad de turistas, pernoctaciones o estancia media que han acudido a cada provincia mensualmente desde 07/2019 hasta 10/2024. Se puede observar que en julio 2019 las provincias más visitadas han sido Las islas Baleares, Barcelona, Gerona, Málaga, Alicante y Madrid, en marzo 2020 los destinos más populares fueron Barcelona, Madrid, Alicante y Málaga y en julio 2024 los sitios más importantes han sido las islas Baleares, Barcelona, Gerona, Alicante, Málaga y Madrid. Sin embargo, la estancia media más larga en julio 2019 corresponden a Alicante y Murcia, en marzo 2020 las estancias medias más largas fueron en Jaén y en julio 2024 se encuentran distribuidas en 10 provincias diferentes.

Para finalizar se ha realizado un análisis multivariante a través de un heatmap, que ha mostrado una correlación muy fuerte entre la cantidad de turistas y sus pernoctaciones y una correlación negativa entre los turistas y su estancia media y la estancia media y las pernoctaciones.

Conclusión

El sector turístico se ha recuperado volviendo a cifras mucho más altas que las que tenía antes de la pandemia.

Tras el confinamiento, las pernoctaciones totales decrecieron y en la actualidad superan las cifras que antes de la pandemia, sin embargo, analizando las pernoctaciones medias, éstas fueron mucho más altas en el año 2020 y en la actualidad se encuentran en un descenso.

Sin embargo, las provincias más populares entre los turistas internacionales no han variado debido a la pandemia.