Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID
ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN
ANÁLISIS DE DATOS MASIVOS (BIG DATA)

TRABAJO FIN DE MÁSTER

Análisis sobre el turismo en España

NOMBRE:
Elizabeth María Toledo Rodríguez
CURSO 2021-2022

TÍTULO: Análisis sobre el turismo en España

AUTOR: Elizabeth María Toledo Rodríguez

TITULACIÓN: Máster Universitario en Análisis de Datos Masivos (Big Data)

DIRECTOR DEL PROYECTO: Joaquín García Onrubia

FECHA: [Octubre] de 2022

RESUMEN

En este proyecto de fin de máster se explora cómo es el turismo en las diferentes comunidades autónomas de España teniendo en consideración el país de origen de estos, la duración media de los viajes y el gasto promedio por día de estancia. Se dispone de dos conjuntos de datos cuya granularidad temporal es a nivel de trimestre desde 2019. El primero contiene datos respecto al número de turistas en cada comunidad autónoma en una fecha determinada y su país de procedencia y el segundo presenta el promedio de gastos diario de estos turistas.

Para realizar el estudio haremos uso un cuaderno de Jupyter donde emplearemos las librerías Pandas, NumPy y Scikit learn. Se incluye una exploración previa de los datos donde se buscará ver cuáles son las comunidades autónomas con más visitantes o cuáles son los turistas que más gastos realizan de media en España, además del entrenamiento de un modelo y la medición de su score.

Durante el desarrollo de este trabajo de fin de máster también exploraremos una de las herramientas más conocidas en lo que a gobierno del dato ser refiere, Collibra. Usaremos, debido a ciertas limitaciones, un conjunto de datos de prueba para mostrar y reflexionar cómo funciona y cómo podría beneficiarnos con nuestros datos en concreto.

Por último, se mostrará un dahsboard construido con la herramienta de business intelligence, Power Bi. En este último paso, usaremos la interfaz de la herramienta para dar lugar a relaciones entre los dos conjuntos de datos principales y uno nuevo que incluye las coordenadas geográficas de cada comunidad autónoma. Además, estudiaremos un nuevo KPI no incluido en los conjuntos de datos originales, el ingreso total que suponen los visitantes a España.

Palabras clave: Turismo, Análisis, Dashboard, Power BI, Business Intelligence, Scikit learn, Pandas.

ABSTRACT

In this master's thesis we will explore how tourism trends behave in different communities in Spain considering the nationality of the tourists, the duration of their stay and the average spending they do.

In order to get to that point, we will be using Jupyter notebook and libraries Pandas, NumPy and Scikit learn. An overview of the data is included where we are going to focus on which communities in Spain have more visitors or which visitors spend more on their stay. Moreover, we will be training a predicting model and measuring the score.

During the development of this thesis, we will also explore a famous data governance tool known as Collibra. However, due to certain limitation, we will use a trial data set and show and reflect on ways of applying this tool to our own data sets.

Lastly, we Will be showing a dashboard built using the business intelligence tool referred as Power BI. In this last step, we'll make use of the tool's interface to create relationships between different tables, adding in this way the geographic coordinates of each community withing Spain. Furthermore, we will study a new KPI that was not included on the original data sets, the total revenue that Spain received thanks to the visitors.

Key words: Tourism, Analysis, Dashboard, Power BI, Business Intelligence, Scikit learn, Pandas.

AGRADECIMIENTOS

Me gustaría a gradecer a mi familia, mi pareja y mis amigos por apoyarme durante los momentos más difíciles en este último año.

Además, quería dar las gracias a mi tutor del TFM que me ha guiado en todo momento tanto como profesor durante el máster como como tutor de este trabajo.

ÍNDICE

Índice de Figuras

Capítulo 1. Introducción

1.1 Antecedentes

La pandemia producida por la Covid-19 ha supuesto un duro golpe al sector turístico, que desde hace muchos años se posiciona como uno de los principales motores de la economía española.

Algunos eventos, como la declaración de España como país de alto riesgo por parte de otras naciones de donde proceden gran cantidad de los turistas, han golpeado fuertemente a este sector.

Nuevo golpe al turismo: Alemania declara a España como territorio de alto riesgo por covid-19

La decisión del país germano obliga a realizar cuarentena a todos los pasajeros procedentes de nuestro país no vacunados excepto si presentan una PCR negativa realizada recientemente. La medida entra en vigor esta medianoche.

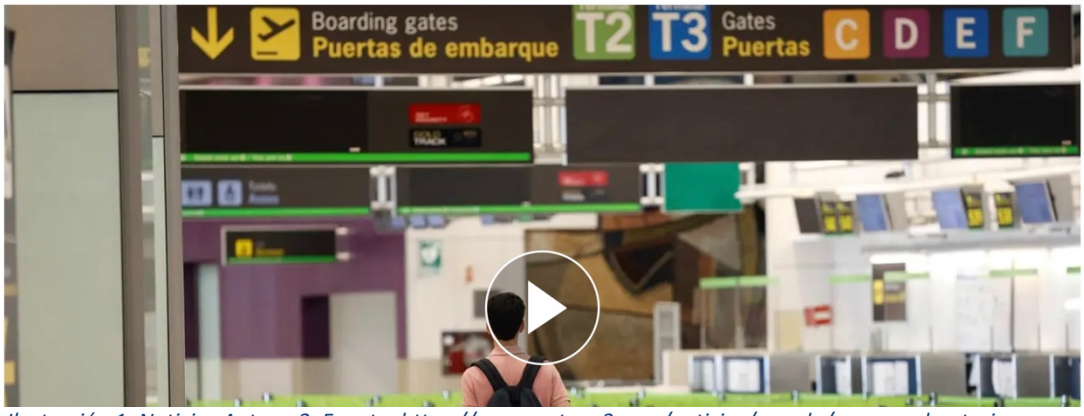

Ilustración 1: Noticias Antena 3. Fuente: https://www.antena3.com/noticias/mundo/nuevo-golpe-turismo-alemania-declara-espana-como-territorio-alto-riesgo-covid19_2021122361c562c3d71c190001d23070.html

No solo fueron las decisiones y pronunciamientos de otros países las que dieron lugar a la difícil situación del sector en 2020, también las decisiones tomadas por España. Por ejemplo, en este informe publicado en una de las webs oficiales del gobierno el veintiséis de marzo de 2020, se declara la suspensión de apertura al público de establecimientos de alojamiento turístico.

MINISTERIO
DE INDUSTRIA, COMERCIO
Y TURISMO

INFORME SOBRE CUESTIONES ACLARATORIAS RESPECTO DE LA APLICACIÓN DE LA ORDEN SND/257/2020, DE 19 DE MARZO, POR LA QUE SE DECLARA LA SUSPENSIÓN DE APERTURA AL PÚBLICO DE ESTABLECIMIENTOS DE ALOJAMIENTO TURÍSTICO.

La Orden SND/257/2020, de 19 de marzo, establece la suspensión de apertura al público de todos los hoteles y alojamientos similares, alojamientos turísticos y otros alojamientos de corta estancia, campings, aparcamientos de caravanas y otros establecimientos similares, ubicados en cualquier parte del territorio nacional.

Ilustración2: Informe de suspensión de alojamientos turísticos. Fuente:https://www.mincotur.gob.es/es-es/COVID-19/turismo/Documents/Informe_suspension_alojamientos_turisticos.pdf

Gracias a la web Statista, unos de los proveedores de datos comerciales más fiables durante alrededor de trece años, podemos ver la fracción tan grande que ha supuesto el turismo en el PIB de España desde 2010.

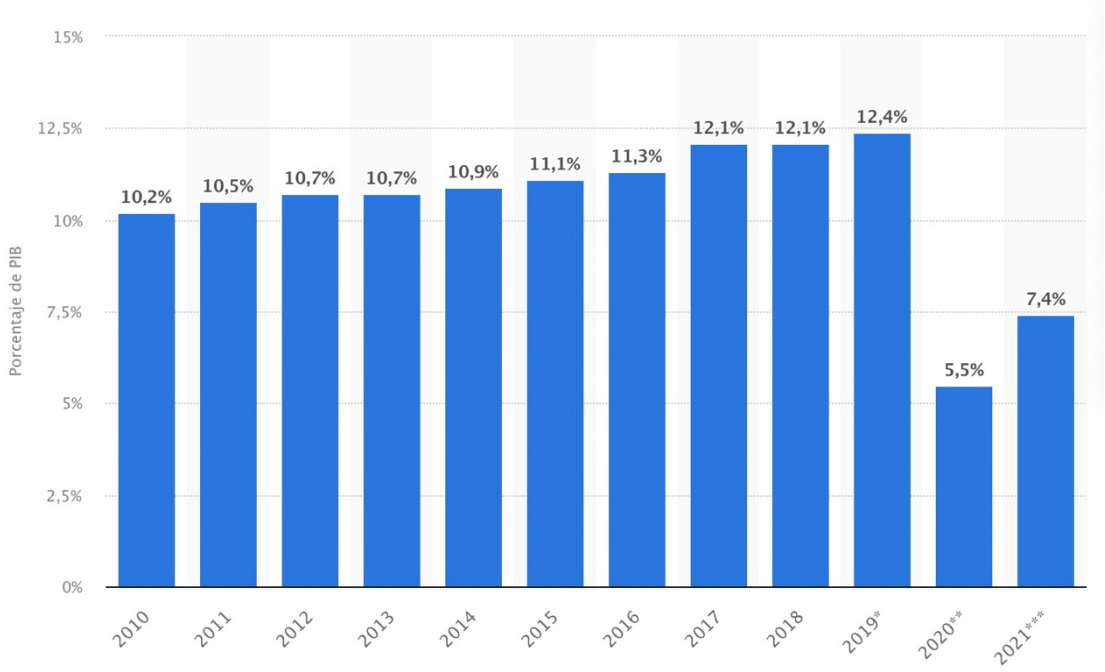

Ilustración 3: Porcentaje del PIB aportado por el sector turístico en España de 2010 a 2021. Fuente: https://es.statista.com/estadisticas/1082929/sector-turistico-porcentaje-del-pib-aportado-espana/

Como podemos observar en la ilustración anterior, el porcentaje es considerable incluso en el año posterior al inicio de la pandemia.

A raíz de esto, la inversión en este sector que trae tantos beneficios a España es primordial.

Por si no fuera motivo suficiente, el turismo crea gran cantidad de puestos de trabajo y es de gran importancia que el tipo de servicios se adecúen al público relevante, con el fin de mantener el estatus actual que tenemos como destino popular. De nuevo, según Statista, podemos ver que alrededor de 2,62 millones de personas trabajaban en el sector turístico en 2019 y con una tendencia al alza que es importante mantener.

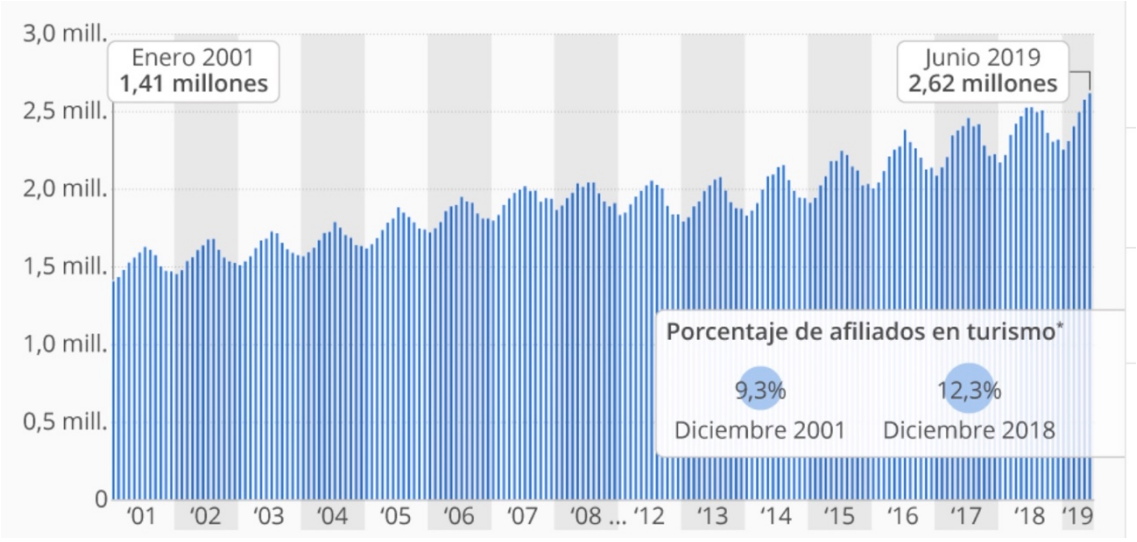

Ilustración 4: Número de trabajadores en el sector turístico. *Fuente: https://es.statista.com/grafico/18875/afiliados-a-la-seguridad-social-en-actividades-turisticas-en-espana/*

1.2 Objeto y alcance

Está claro que interesa mantener esta tendencia. Una buena manera de conseguir mantenerla es entender cuál es el origen de estos turistas y cuales suponen un mayor beneficio económico para así ofrecerles servicios más especializados que les atraigan a nuestro país. Una medida a tomar sería contratar personal en hoteles que conozca sus idiomas y comprender en que zonas abundan más y/o aportan más beneficios los turistas de unas nacionalidades u otras.

Es importante buscar una forma eficiente de realizar dicha inversión a la hora de la contratación de personal o de abrir nuevos puntos de ocio, restauración, hostelería y atención a turistas.

Invertir de manera desorganizada en un punto de información para turistas en Galicia donde contratemos personas que hablen francés podría tener poco sentido si la mayor masa de turistas que recibimos en Galicia es de origen alemán, por ejemplo. De la misma forma abrir restaurantes de lujo en Ibiza si el gasto medio típico de los turistas es muy bajo tampoco sería muy beneficioso y podría llevarnos a la quiebra inminente.

De ahí la relevancia del siguiente análisis que ayudará a comprender de qué forma, en qué épocas del año y dónde es más sensato invertir para obtener el mayor beneficio posible.

De la misma forma, este estudio busca dar visibilidad para zonas menos preparadas para el turismo, que podrían aumentar sus ingresos sabiendo que cerca de ellos hay una gran concentración de turistas. Esto podría hacer interesante para estas zonas invertir en infraestructuras que atrajesen turistas y aumentar así la economía local.

A continuación, veremos un análisis descriptivo del sector turístico en España pre y post-covid-19.

Capítulo 2. Método

Para llevar a cabo el estudio, se ha hecho uso de diferentes herramientas de procesamiento de datos, de visualización y de gobernanza del dato.

- Fuentes: La principal fuente de datos ha sido el INE, de donde se han obtenido datos de gasto medio diario de los turistas por día en cada comunidad autónoma. También, de un data set distinto, se han tomado datos sobre el número de turistas por trimestre en cada comunidad autónoma y el número medio de pernoctaciones de estos. Además, nos hemos servido de la web https://www.antipodas.net para obtener las coordenadas de las diferentes comunidades autónomas.

- Procesamiento: se ha utilizado principalmente Visual Studio Code, donde se ha realizado en análisis en un cuaderno de Jupyter con la ayuda de la librería Pandas para Pyton, entre otras.

- Se llevará a cabo el entrenamiento de un algoritmo simple para predicción del número de turistas futuros.

- Gobernanza del dato: Demostración de cómo funciona Collibra con data set de prueba y adaptación a los conjuntos de datos de nuestro análisis sobre el turismo. El objetivo es facilitar el uso en el futuro de estos datos de manera más sencilla y organizada.

- Visualización: Para llevar a cabo una visualización más clara y así obtener mejores insights se ha utilizado la herramienta Power BI.

Capítulo 3. Desarrollo del análisis

Para llevar a cabo el análisis empezaremos por cargar los datos a un cuaderno de Jupyter en Visual Studio Code. Nuestra principal herramienta para exploración de los datos serán las librerías Pandas y NumPy en Python.

Pandas es una librería que salió en 2008 enfocada al análisis y manipulación de datos. Su nombre viene de "**Pan**el **Da**ta". Una de sus características más destacadas es la introducción de objetos para manipulación de datos como el "DataFrame" con indexación integrada.

Pandas permite importar datos de muchos formatos distintos, por ejemplo, desde Microsoft Excel o en formato JSON entre otros. La librería Pandas está a su vez construida sobre otra librería anteriormente mencionada, NumPy.

NumPy por su parte, es una librería creada en 2005. Se caracteriza por dar soporte a matrices muy grandes y de varias dimensiones, además de que permite crear vectores. Por otro lado, ofrece muchas herramientas de computación numérica.

Este estudio se basa principalmente en dos data sets obtenidos de la web del Instituto Nacional de Estadística español.

3.1 Primer data set: Número de turistas y pernoctaciones en cada CCAA.

El primero incluye: mes, número de turistas, comunidad autónoma de destino, procedencia y número de pernoctaciones de estos. Para poder realizar el análisis más cómodamente, hemos cambiado el "type" de las columnas relevantes a numérico.

```
df_turistas["Turistas"] = pd.to_numeric(df_turistas["Turistas"])
df_turistas["Pernoctaciones"] = pd.to_numeric(df_turistas["Pernoctaciones"])
df_turistas["Duración media de los viajes"] = pd.to_numeric(df_turistas["Duración media de los viajes"])
df_turistas.head()
```

[353] Python

	Fecha	Turistas	Pernoctaciones	Duración media de los viajes	Pais origen	CCAA
0	2022M05	1076642.0	8160353.0	7.6	Total	Andalucía
1	2022M04	931098.0	7159528.0	7.7	Total	Andalucía
2	2022M03	670275.0	5860862.0	8.7	Total	Andalucía
3	2022M02	490336.0	4320232.0	8.8	Total	Andalucía
4	2022M01	488945.0	4687975.0	9.6	Total	Andalucía

Ilustración 5: Visión preliminar del data set. Fuente: Elaboración propia

Veamos qué aspecto tienen algunas métricas típicas en este data set para cada una de las columnas de tipo numérico:

```
[354]  df_turistas.describe()
                                                                    Python
```

	Turistas	Pernoctaciones	Duración media de los viajes
count	2.419400e+04	2.417700e+04	24177.000000
mean	4.124667e+04	3.431283e+05	8.263110
std	2.939520e+05	2.328101e+06	3.065634
min	3.000000e+01	4.700000e+01	1.000000
25%	1.570000e+02	1.112000e+03	6.200000
50%	8.820000e+02	7.260000e+03	7.800000
75%	6.508500e+03	5.926100e+04	9.700000
max	1.112593e+07	8.871801e+07	30.500000

Ilustración 6: Métricas general del data set. Fuente: Elaboración propia.

Si bien mirar estos resultados puede ser un indicador interesante, tanto las diferentes comunidades autónomas de España como los países de origen de los visitantes presentan una gran diversidad, así que las conclusiones que se puedan sacar de aquí terminarían por ser muy imprecisas. Además, sería interesante conocer que comunidades reciben el mayor número de turistas y de dónde.

Veamos primero algunos hallazgos curiosos al explorar los datos en su totalidad antes de profundizar en cómo se comportó el turismo en España en cada uno de los años de los que tenemos registros.

Por ejemplo, vemos que los visitantes con una duración media de estancia más larga son los que vinieron de Italia a Melilla en septiembre de 2020. Sería curioso investigar en profundidad que pudo ocurrir en 2020 para que durante varios meses tanto visitantes de Suecia como de Italia decidieran permanecer tanto tiempo en Melilla, un lugar no especialmente conocido por ser turístico.

```
[357]  df_turistas.nlargest(n=10, columns = ['Duración media de los viajes'])
                                                                    Python
```

	Fecha	Turistas	Pernoctaciones	Duración media de los viajes	Pais origen	CCAA	media_pernoctaciones
17415	2020M09	106.0	3233.0	30.5	Italia	Melilla	30.500000
32823	2020M01	146.0	4234.0	29.0	Suecia	Melilla	29.000000
32816	2020M08	116.0	3318.0	28.6	Suecia	Melilla	28.603448
9195	2020M04	664.0	18924.0	28.5	Dinamarca	Asturias, Principado de	28.500000
32821	2020M03	151.0	4213.0	27.9	Suecia	Melilla	27.900662
2574	2020M10	5506.0	151966.0	27.6	Alemania	Murcia, Región de	27.600073
33345	2020M04	713.0	19108.0	26.8	Suiza	Madrid, Comunidad de	26.799439
32404	2019M12	1512.0	39766.0	26.3	Suecia	Cantabria	26.300265
2581	2020M03	9700.0	247350.0	25.5	Alemania	Murcia, Región de	25.500000
2265	2020M04	6118.0	155530.0	25.4	Alemania	Canarias	25.421706

Ilustración 7: Top 10 registros por número de pernoctaciones. Fuente: Elaboración propia.

Observemos ahora el top 10 de registros que presentan mayor número de turistas.

```
[359]
df_turistas.nlargest(n=10, columns = ['Turistas'])
```
Python

	Fecha	Turistas	Pernoctaciones	Duración media de los viajes	País origen	CCAA	media_pernoctaciones
11513	2019M08	866912.0	5742799.0	6.6	Francia	Cataluña	6.624431
26738	2019M08	726067.0	5300289.0	7.3	Reino Unido	Balears, Illes	7.300000
26739	2019M07	697806.0	4326397.0	6.2	Reino Unido	Balears, Illes	6.200000
11489	2021M08	587915.0	4057708.0	6.9	Francia	Cataluña	6.901862
26737	2019M09	585943.0	3808630.0	6.5	Reino Unido	Balears, Illes	6.500001
11514	2019M07	559719.0	3046057.0	5.4	Francia	Cataluña	5.442118
2205	2022M05	525881.0	3418227.0	6.5	Alemania	Balears, Illes	6.500001
2237	2019M09	506714.0	3394984.0	6.7	Alemania	Balears, Illes	6.700000
2238	2019M08	505229.0	3688172.0	7.3	Alemania	Balears, Illes	7.300001
26705	2022M05	477809.0	2723511.0	5.7	Reino Unido	Balears, Illes	5.699999

Ilustración 8: Top 10 registros por número de turistas. Fuente: Elaboración propia.

Para resumir más los datos y simplificar el análisis, se creó una nueva columna con la finalidad de agrupar los resultados en trimestres en vez de mensualmente, tras lo cual agrupamos por trimestre, comunidad autónoma de destino y país de origen.

```
[363]
df_turistas['Trimestres']= df_turistas.apply(lambda x: x['Fecha'][:16] + 'T' + ('1' if x['Meses']in ('01', '02', '03') else
'2' if x['Meses'] in ('04', '05', '06') else
'3' if x['Meses'] in ('07', '08', '09') else
'4' if x['Meses'] in ('10', '11', '12') else '0'), axis=1)

df_turistas.head()
```
Python

	Fecha	Turistas	Pernoctaciones	Duración media de los viajes	País origen	CCAA	media_pernoctaciones	Meses	Trimestres
1400	2022M05	126.0	824.0	6.5	Albania	Andalucía	6.539683	05	2022T2
1401	2022M04	145.0	1006.0	6.9	Albania	Andalucía	6.937931	04	2022T2
1402	2022M03	95.0	594.0	6.3	Albania	Andalucía	6.252632	03	2022T1
1403	2022M02	88.0	574.0	6.5	Albania	Andalucía	6.522727	02	2022T1
1404	2022M01	92.0	703.0	7.6	Albania	Andalucía	7.641304	01	2022T1

```
[364]
df_turistas_grouped = df_turistas.groupby(['Trimestres', 'CCAA', 'Pais origen']).sum()
df_turistas_grouped
```
Python

			Turistas	Pernoctaciones	Duración media de los viajes	media_pernoctaciones
Trimestres	CCAA	País origen				
2019T3	Andalucía	Albania	359.0	2019.0	16.9	16.880822
		Alemania	302730.0	3090278.0	30.0	30.035296
		Andorra	2112.0	20830.0	28.9	28.885620
		Armenia	185.0	1602.0	25.1	25.055714
		Austria	50299.0	483578.0	28.7	28.721204
...	...	...	...	...	...	...
2022T2	Rioja, La	Rumanía	387.0	3573.0	18.6	18.596706
		Rusia	0.0	0.0	0.0	0.000000
		Serbia	0.0	0.0	0.0	0.000000
		Suecia	1478.0	15733.0	21.3	21.300093
		Suiza	400.0	2465.0	12.2	12.199264

9895 rows × 4 columns

Ilustración 9: Agrupación del data set por trimestres, CCAA y País de origen. Fuente: Elaboración propia.

Ahora, una vez tenemos el data set estructurado a nuestro gusto, veamos el top 10 de registros donde el número de turistas en mayor por año desde 2019.

Vemos que claramente los dos trimestres donde España tiene más éxito como destino turístico son el tercero y el cuarto. Estos son datos pre-pandemia, veremos a continuación si el comportamiento en cuanto a épocas se mantiene o no durante y después de 2020. Los principales visitantes vienen de Reino Unido, Alemania y Francia. Parece ser que el destino favorito de los británicos en nuestro país son las islas y la costa mediterránea, mientras que los franceses (quizás por la cercanía), se concentran como

visitantes de Cataluña. Los alemanes parece que prefieran principalmente las Islas Baleares y Cataluña.

Trimestres	CCAA	Pais origen	Turistas
2019T3	Balears, Illes	Reino Unido	2009816.0
	Cataluña	Francia	1800139.0
	Balears, Illes	Alemania	1434872.0
	Canarias	Reino Unido	1297050.0
2019T4	Canarias	Reino Unido	1185684.0
2019T3	Andalucía	Reino Unido	1148009.0
	Comunitat Valenciana	Reino Unido	948290.0
	Cataluña	Reino Unido	850047.0
2019T4	Cataluña	Francia	697534.0
2019T3	Cataluña	Alemania	646172.0

Ilustración 10: Top 10 por número de turistas 2019. Fuente: Elaboración propia

En 2020, se puede observar un claro cambio en lo que a estacionalidad se refiere. Debido a la pandemia de la COVID-19, aparecen en el top 5 registros pertenecientes al primer trimestre del año. Esto no ocurría en 2019 ya que la temporada alta de turismo en nuestro país suele coincidir con el verano (tercer trimestre del año). Sin embargo, debido a los cierres de bares, restaurantes y las diversas restricciones ocurridas a partir de abril del 2020, no se pudieron superar en muchos casos el número de turistas recibidos en el primer trimestre.

Solo sería destacable, en el tercer trimestre el turismo en Cataluña, que ha conseguido aparecer en el primer lugar y más veces en el top 10. En general todos los números del top 10 son más pequeños que en 2019 dadas tanto las restricciones anteriormente mencionadas, como la recesión económica que trajo consigo la pandemia.

Trimestres	CCAA	Pais origen	Turistas	Pernoctaciones
2020T3	Cataluña	Francia	898856.0	6644467.0
2020T1	Canarias	Reino Unido	884649.0	6653072.0
	Cataluña	Francia	592089.0	3821663.0
	Canarias	Alemania	564221.0	5576792.0
2020T3	Balears, Illes	Alemania	376404.0	3542745.0
2020T1	Andalucía	Reino Unido	331327.0	2819601.0
	Cataluña	Alemania	304009.0	3127062.0
2020T4	Cataluña	Francia	292078.0	3125143.0
2020T3	Comunitat Valenciana	Francia	290578.0	3395771.0
2020T1	Comunitat Valenciana	Reino Unido	282365.0	2448342.0

Ilustración 11: Top 10 por número de turistas 2020. Fuente: Elaboración propia.

En 2021, todo vuelve a normalizarse poco a poco. Vemos como de nuevo en el top 10 los meses más populares para viajar a España son los del verano y también los últimos tres meses del año, donde las temperaturas en España son más amables que en países del norte de Europa. Si bien se observa que los números no regresaron a niveles de 2019, podemos apreciar una mejora en la situación económica. En 2020 no se llegó a tener un millón de viajeros que viniesen del mismo país a un mismo sitio en un mismo trimestre,

ni siquiera en verano. Sin embargo, en 2021 llegamos a tener más de un millón de turistas, solo de Francia en Cataluña.

Vemos que se mantiene la tendencia de que la mayor parte de nuestros visitantes vienen de Reino Unido, Francia y Alemania. Esto nos da una pista del tipo de habilidades idiomáticas que interesa que aquellos que se dedican al sector turístico y a la hostelería tengan.

			Turistas	Pernoctaciones
Trimestres	CCAA	Pais origen		
2021T3	Cataluña	Francia	1246709.0	8391241.0
2021T4	Canarias	Reino Unido	879152.0	6329462.0
2021T3	Balears, Illes	Reino Unido	769228.0	5610802.0
		Alemania	736865.0	5504340.0
2021T4	Cataluña	Francia	629807.0	4007925.0
2021T2	Cataluña	Francia	463813.0	3052117.0
2021T4	Canarias	Alemania	443873.0	3841120.0
	Balears, Illes	Alemania	428806.0	3085126.0
2021T2	Balears, Illes	Alemania	405508.0	2878958.0
2021T3	Canarias	Reino Unido	404521.0	3417228.0

Ilustración 12: Top 10 por número de turistas 2021. Fuente: Elaboración propia.

Por último, en 2022, en la fecha de realización de este análisis solo tenemos datos de los primeros dos trimestres. Vemos que incluso teniendo en cuenta que no contamos con los datos de la temporada alta, que serían los trimestres tres y cuatro, hay una gran recuperación económica. Vemos que en muchos casos tenemos un mayor número de turistas en la temporada baja que el año anterior en la temporada alta. De nuevo los destinos favoritos son las Islas Baleares, Canarias y Cataluña. Los visitantes principales son aquellos procedentes de Alemania, Reino Unido y Francia.

			Turistas
Trimestres	CCAA	Pais origen	
2022T2	Balears, Illes	Alemania	967224.0
2022T1	Canarias	Reino Unido	942414.0
2022T2	Canarias	Reino Unido	844107.0
	Balears, Illes	Reino Unido	734897.0
	Cataluña	Francia	705025.0
	Andalucía	Reino Unido	544408.0
2022T1	Cataluña	Francia	532057.0
	Canarias	Alemania	529083.0
2022T2	Comunitat Valenciana	Reino Unido	479349.0
	Canarias	Alemania	384859.0

Ilustración 13: Top 10 por número de turistas 2022. Fuente: Elaboración propia.

Al realizar los cálculos anteriores, hicimos una agrupación (group by) de las filas usando los campos “Trimestres”, “CCAA” y “País de origen” y la operación realizada fue la suma. Sin embargo, no se incluye en el cálculo el número de pernoctaciones ya que, al usar la suma como operación, nos daría un número equivocado. A continuación, se presentan el número de medio de pernoctaciones para los top 10 anteriores por cada año:

- 2019

```
#AÑO 2019
df_turistas_19 = df_turistas[df_turistas.Trimestres.str[2:16] ==' 2019']
#df_turistas_19.head()
df_turistas_grouped_19 = df_turistas_19.groupby(['Trimestres', 'CCAA', 'Pais origen']).mean()
df_turistas_grouped_19.sort_values(by=['Turistas'], ascending=False)

df_turistas_grouped_19 = df_turistas_grouped_19.drop(['Pernoctaciones'], axis=1)
df_turistas_grouped_19 = df_turistas_grouped_19.drop(['media_pernoctaciones'], axis=1)

df_turistas_grouped_19 = df_turistas_grouped_19[['Duración media de los viajes', 'Turistas']]

df_turistas_grouped_19.nlargest(n=10, columns = ['Turistas'])
```

[73] ✓ 0.1s

			Duración media de los viajes
Trimestres	CCAA	Pais origen	
2019T3	Balears, Illes	Reino Unido	6.666667
	Cataluña	Francia	5.900000
	Balears, Illes	Alemania	6.833333
	Canarias	Reino Unido	7.866667
2019T4	Canarias	Reino Unido	7.033333
2019T3	Andalucía	Reino Unido	7.700000
	Comunitat Valenciana	Reino Unido	7.866667
	Cataluña	Reino Unido	6.100000
2019T4	Cataluña	Francia	6.000000
2019T3	Cataluña	Alemania	7.833333

Ilustración 14: Número medio de pernoctaciones 2019. Fuente: Elaboración propia.

- 2020

```
#AÑO 2020
df_turistas_20 = df_turistas[df_turistas.Trimestres.str[2:16] ==' 2020']
#df_turistas_19.head()
df_turistas_grouped_20 = df_turistas_20.groupby(['Trimestres', 'CCAA', 'Pais origen']).mean()
df_turistas_grouped_20.sort_values(by=['Turistas'], ascending=False)

df_turistas_grouped_20 = df_turistas_grouped_20.drop(['Pernoctaciones'], axis=1)
df_turistas_grouped_20 = df_turistas_grouped_20.drop(['media_pernoctaciones'], axis=1)

df_turistas_grouped_20 = df_turistas_grouped_20[['Duración media de los viajes', 'Turistas']]

df_turistas_grouped_20.nlargest(n=10, columns = ['Turistas'])
```

[74] ✓ 0.2s

			Duración media de los viajes
Trimestres	CCAA	Pais origen	
2020T3	Cataluña	Francia	7.366667
2020T1	Canarias	Reino Unido	7.600000
	Cataluña	Francia	6.800000
	Canarias	Alemania	9.966667
2020T3	Balears, Illes	Alemania	10.033333
2020T1	Andalucía	Reino Unido	8.600000
	Cataluña	Alemania	10.433333
2020T4	Cataluña	Francia	11.300000
2020T3	Comunitat Valenciana	Francia	11.700000
2020T1	Comunitat Valenciana	Reino Unido	8.733333

Ilustración 15: Número medio de pernoctaciones 2020. Fuente: Elaboración propia.

- 2021

```
#AÑO 2021
df_turistas_21 = df_turistas[df_turistas.Trimestres.str[2:16] =='          2021']
#df_turistas_19.head()
df_turistas_grouped_21 = df_turistas_21.groupby(['Trimestres', 'CCAA', 'Pais origen']).mean()
df_turistas_grouped_21.sort_values(by=['Turistas'], ascending=False)

df_turistas_grouped_21 = df_turistas_grouped_21.drop(['Pernoctaciones'], axis=1)
df_turistas_grouped_21 = df_turistas_grouped_21.drop(['media_pernoctaciones'], axis=1)

df_turistas_grouped_21 = df_turistas_grouped_21[['Duración media de los viajes', 'Turistas']]

df_turistas_grouped_21.nlargest(n=10, columns = ['Turistas'])
```

✓ 0.1s

			Duración media de los viajes
Trimestres	CCAA	Pais origen	
2021T3	Cataluña	Francia	6.700000
2021T4	Canarias	Reino Unido	7.200000
2021T3	Balears, Illes	Reino Unido	7.300000
		Alemania	7.500000
2021T4	Cataluña	Francia	6.500000
2021T2	Cataluña	Francia	6.966667
2021T4	Canarias	Alemania	8.666667
	Balears, Illes	Alemania	7.400000
2021T2	Balears, Illes	Alemania	7.266667
2021T3	Canarias	Reino Unido	8.766667

Ilustración 16: Número medio de pernoctaciones 2021. Fuente: Elaboración propia.

- 2022

```
#AÑO 2022
df_turistas_22 = df_turistas[df_turistas.Trimestres.str[2:16] =='          2022']
#df_turistas_19.head()
df_turistas_grouped_22 = df_turistas_22.groupby(['Trimestres', 'CCAA', 'Pais origen']).mean()
df_turistas_grouped_22.sort_values(by=['Turistas'], ascending=False)

df_turistas_grouped_22 = df_turistas_grouped_22.drop(['Pernoctaciones'], axis=1)
df_turistas_grouped_22 = df_turistas_grouped_22.drop(['media_pernoctaciones'], axis=1)

df_turistas_grouped_22 = df_turistas_grouped_22[['Duración media de los viajes', 'Turistas']]

df_turistas_grouped_22.nlargest(n=10, columns = ['Turistas'])
```

✓ 0.2s

			Duración media de los viajes
Trimestres	CCAA	Pais origen	
2022T2	Balears, Illes	Alemania	6.650000
	Canarias	Reino Unido	7.150000
	Balears, Illes	Reino Unido	5.700000
	Cataluña	Francia	5.700000
2022T1	Canarias	Reino Unido	7.600000
2022T2	Andalucía	Reino Unido	6.700000
	Comunitat Valenciana	Reino Unido	6.850000
	Canarias	Alemania	9.650000
2022T1	Cataluña	Francia	6.533333
	Canarias	Alemania	9.933333

Ilustración 17: Número medio de pernoctaciones 2022. Fuente: Elaboración propia.

Podemos ver que la duración típica de la estancia de los turistas es de alrededor de siete días. Conocer el dato de duración media de las estancias nos va a ser muy útil para poder cuantificar los beneficios de las visitas una vez empecemos a explorar el segundo data set, donde veremos el gasto medio diario de los turistas.

Pero antes de llegar allí, ya habiendo visto que las comunidades autónomas con más turistas en general son las Islas Baleares, Cataluña y Canarias, veamos en más profundidad, sin influencia de datos de otras comunidades, la procedencia principal de los visitantes de estas tres comunidades desde 2019, donde la actividad económica por el turismo es más destacada.

- Islas Baleares.

Pais origen	Turistas
Alemania	5632111.0
Reino Unido	4468081.0
Francia	1643219.0
Italia	1059535.0
Países Bajos	982395.0
Suiza	570583.0
Bélgica	410000.0
Austria	389081.0
Suecia	364240.0
Dinamarca	285659.0

Ilustración 18: Procedencia de los visitantes Islas Baleares. Fuente: Elaboración propia.

- Cataluña.

Pais origen	Turistas
Francia	8184029.0
Alemania	3185014.0
Países Bajos	2905685.0
Reino Unido	2340732.0
Italia	1767833.0
Bélgica	1106328.0
Suecia	968957.0
Polonia	782559.0
Suiza	587720.0
Andorra	549035.0

Ilustración 19: Procedencia de los visitantes Cataluña, Fuente: Elaboración propia.

- Canarias.

Pais origen	Turistas
Reino Unido	6718180.0
Alemania	3601016.0
Francia	1218372.0
Países Bajos	1056599.0
Bélgica	814238.0
Italia	803221.0
Irlanda	598951.0
Suecia	582355.0
Polonia	500924.0
Noruega	480860.0

Ilustración 20: Procedencia de los visitantes Canarias .Fuente: Elaboración propia.

Está claro, ya sin influencia de otras comunidades, que el mayor número de visitantes a zonas costeras de España son de Reino Unido, Alemania y Francia.

3.2 Segundo data set: Gasto medio diario de los turistas.

Ahora comencemos a explorar nuestro segundo data set que incluye el gasto diario de los turistas según su procedencia, la CCAA de destino y el trimestre en que vinieron desde 2019.

Empecemos viendo de manera general, sin discernir por años, orígenes ni destino, que registros son del top 10 teniendo en cuenta el gasto diario medio por persona.

País de origen	Fecha	CCAA	gasto_medio_diario_persona
Rusia	2022T1	Navarra, Comunidad Foral de	448.58
Suiza	2022T1	Navarra, Comunidad Foral de	311.59
Francia	2022T1	Navarra, Comunidad Foral de	269.39
Andorra	2022T1	Navarra, Comunidad Foral de	266.21
Suiza	2020T3	Madrid, Comunidad de	254.36
	2020T4	Madrid, Comunidad de	235.81
	2021T4	Madrid, Comunidad de	235.19
	2021T2	Madrid, Comunidad de	228.57
Noruega	2020T4	Madrid, Comunidad de	225.78
Alemania	2019T3	Aragón	225.67

Ilustración 21: Gasto medio diario. Fuente: Elaboración propia.

Observamos un hallazgo curioso: el número uno en este top 10 corresponde a visitantes rusos, durante el primer trimestre de 2022 gastando alrededor de 450 euros diarios en la Comunidad Foral de Navarra. Veremos más adelante, cuando exploremos que ocurre año por año y no en general si es una tendencia de los visitantes rusos el gastar grandes

cantidades en esta comunidad autónoma o si es algo exclusivo del primer trimestre de 2022.

De momento, empecemos por ver en qué comunidades autónomas los visitantes tienen un mayor gasto medio diario en cada uno de los años de los que tenemos registros.

- 2019.

CCAA	gasto_medio_diario_persona
País Vasco	105.761500
Madrid, Comunidad de	103.826625
Cataluña	88.224500
Navarra, Comunidad Foral de	85.185500
Castilla y León	84.328000
Castilla - La Mancha	80.607625
Aragón	79.116500
Rioja, La	78.140000
Cantabria	73.656125
Andalucía	72.026375

Ilustración 22: Comunidades con más gasto medio diario de los visitantes 2019. Fuente: Elaboración propia.

- 2020.

CCAA	gasto_medio_diario_persona
Madrid, Comunidad de	79.685500
País Vasco	67.249125
Navarra, Comunidad Foral de	58.167000
Canarias	55.629375
Balears, Illes	55.018375
Cataluña	54.692875
Aragón	53.371250
Castilla - La Mancha	51.482875
Castilla y León	50.634375
Comunitat Valenciana	48.681750

Ilustración 23: Comunidades con más gasto medio diario de los visitantes 2020. Fuente: Elaboración propia.

- 2021.

CCAA	gasto_medio_diario_persona
Madrid, Comunidad de	112.708375
País Vasco	95.106000
Navarra, Comunidad Foral de	88.466750
Aragón	81.437125
Cantabria	81.294125
Cataluña	79.321875
Balears, Illes	79.245000
Castilla y León	77.833125
Rioja, La	77.376250
Castilla - La Mancha	77.188625

Ilustración 24: Comunidades con más gasto medio diario de los visitantes 2021. Fuente: Elaboración propia.

- 2022.

CCAA	gasto_medio_diario_persona
País Vasco	116.090000
Madrid, Comunidad de	115.971000
Navarra, Comunidad Foral de	101.514359
Castilla y León	83.610000
Balears, Illes	82.643000
Cataluña	79.580500
Comunitat Valenciana	79.004000
Castilla - La Mancha	78.058500
Cantabria	76.447500
Andalucía	75.665000

Ilustración 25: Comunidades con más gasto medio diario de los visitantes 2022.Fuente: Elaboración propia.

De nuevo, como vimos con el primer data set y el número de turistas que viajaban a España, en este data set se pueden ver las variaciones del poder adquisitivo de los turistas a partir de su gasto medio diario en cada uno de los años estudiados. Vemos como en 2020, coincidiendo con la toma de medidas con respecto a la pandemia, el gasto medio diario de los turistas se desplomó en comparación con 2019. En 2021, en parte por la recuperación económica y en parte por la inflación, vemos como el gasto de los visitantes aumenta progresivamente.

Exploremos ahora en cada año, el top 3 de las comunidades donde los visitantes gastaron más diariamente. Nos interesa saber la procedencia de los turistas que más gastan en nuestro país.

- 2019.

País de origen	CCAA
Suiza	País Vasco
Andorra	País Vasco
Austria	País Vasco
Alemania	País Vasco
Noruega	País Vasco
Portugal	País Vasco
Total	País Vasco
Bélgica	País Vasco
Reino Unido	País Vasco
Dinamarca	País Vasco

Ilustración 27: Top 1, País Vasco (2019). Fuente: Elaboración propia.

País de origen	CCAA
Suiza	Madrid, Comunidad de
Andorra	Madrid, Comunidad de
Reino Unido	Madrid, Comunidad de
Rumanía	Madrid, Comunidad de
Total	Madrid, Comunidad de
Rusia	Madrid, Comunidad de
Alemania	Madrid, Comunidad de
Austria	Madrid, Comunidad de
Portugal	Madrid, Comunidad de
Otros Países Europeos	Madrid, Comunidad de

Ilustración 26: Top 2, Comunidad de Madrid (2019). Fuente: Elaboración propia.

País de origen	CCAA
Suiza	Cataluña
Alemania	Cataluña
Austria	Cataluña
Andorra	Cataluña
Rumanía	Cataluña
Bélgica	Cataluña
Total	Cataluña
Rusia	Cataluña
Francia	Cataluña
Irlanda	Cataluña

Ilustración 15: Top 3, Cataluña (2019). Fuente: Elaboración propia.

- 2020.

País de origen	CCAA
Suiza	Madrid, Comunidad de
Andorra	Madrid, Comunidad de
Rumanía	Madrid, Comunidad de
Suecia	Madrid, Comunidad de
Rusia	Madrid, Comunidad de
Noruega	Madrid, Comunidad de
Total	Madrid, Comunidad de
Portugal	Madrid, Comunidad de
Alemania	Madrid, Comunidad de
Otros Países Europeos	Madrid, Comunidad de

Ilustración 30: Top 1, Comunidad de Madrid (2020). Fuente: Elaboración propia.

País de origen	CCAA
Suiza	País Vasco
Andorra	País Vasco
Francia	País Vasco
Total	País Vasco
Alemania	País Vasco
Noruega	País Vasco
Austria	País Vasco
Bélgica	País Vasco
Portugal	País Vasco
Italia	País Vasco

Ilustración 28: Top 2, País Vasco (2020). Fuente: Elaboración propia.

País de origen	CCAA
Suiza	Navarra, Comunidad Foral de
Rusia	Navarra, Comunidad Foral de
Francia	Navarra, Comunidad Foral de
Total	Navarra, Comunidad Foral de
Rumanía	Navarra, Comunidad Foral de
Alemania	Navarra, Comunidad Foral de
Portugal	Navarra, Comunidad Foral de
Dinamarca	Navarra, Comunidad Foral de
Noruega	Navarra, Comunidad Foral de
Bélgica	Navarra, Comunidad Foral de

Ilustración 29: Top 3, Comunidad Foral De Navarra (2020). Fuente: Elaboración propia.

- 2021.

País de origen	CCAA
Suiza	Madrid, Comunidad de
Andorra	Madrid, Comunidad de
Rumanía	Madrid, Comunidad de
Rusia	Madrid, Comunidad de
Total	Madrid, Comunidad de
Portugal	Madrid, Comunidad de
Noruega	Madrid, Comunidad de
Suecia	Madrid, Comunidad de
Países Bajos	Madrid, Comunidad de
Otros Países Europeos	Madrid, Comunidad de

Ilustración 33:Top 1, Comunidad de Madrid (2021). Fuente: Elaboración propia.

País de origen	CCAA
Suiza	País Vasco
Andorra	País Vasco
Francia	País Vasco
Noruega	País Vasco
Total	País Vasco
Alemania	País Vasco
Dinamarca	País Vasco
Bélgica	País Vasco
Rusia	País Vasco
Austria	País Vasco

Ilustración 32: Top 2, País Vasco (2021). Fuente: Elaboración propia.

País de origen	CCAA
Suiza	Navarra, Comunidad Foral de
Francia	Navarra, Comunidad Foral de
Total	Navarra, Comunidad Foral de
Alemania	Navarra, Comunidad Foral de
Rumanía	Navarra, Comunidad Foral de
Portugal	Navarra, Comunidad Foral de
Austria	Navarra, Comunidad Foral de
Andorra	Navarra, Comunidad Foral de
Bélgica	Navarra, Comunidad Foral de
Dinamarca	Navarra, Comunidad Foral de

Ilustración 31: Top 3, Comunidad Foral De Navarra (2020). Fuente: Elaboración propia.

- 2022.

País de origen	CCAA
Total	País Vasco

Ilustración 36: Top 1, País Vasco (2022). Se observa que en 2022 en el País Vasco no se recogieron datos de las procedencias de sus visitantes para este data set. Fuente: Elaboración propia.

País de origen	CCAA
Suiza	Madrid, Comunidad de
Rusia	Madrid, Comunidad de
Andorra	Madrid, Comunidad de
Reino Unido	Madrid, Comunidad de
Rumanía	Madrid, Comunidad de
Total	Madrid, Comunidad de
Irlanda	Madrid, Comunidad de
Portugal	Madrid, Comunidad de
Otros Países Europeos	Madrid, Comunidad de
Noruega	Madrid, Comunidad de

Ilustración 35: Top 2, Comunidad de Madrid (2022). Fuente: Elaboración propia.

País de origen	CCAA
Rusia	Navarra, Comunidad Foral de
Suiza	Navarra, Comunidad Foral de
Francia	Navarra, Comunidad Foral de
Andorra	Navarra, Comunidad Foral de
Total	Navarra, Comunidad Foral de
Rumanía	Navarra, Comunidad Foral de
Portugal	Navarra, Comunidad Foral de
Bélgica	Navarra, Comunidad Foral de
Dinamarca	Navarra, Comunidad Foral de
Noruega	Navarra, Comunidad Foral de

Ilustración 34: Top 3, Comunidad Foral De Navarra (2022). Fuente: Elaboración propia.

Con respecto al comentario sobre los visitantes rusos en la Comunidad Foral de Navarra, vemos que no es una tendencia que se ha repetido a la largo de los años, sino algo ocurrido solo en 2022 y gracias al top 10 sin discernir por años ni CCAA mostrado al principio de esta sección, observamos que el pico ocurrió en el primer trimestre del año.

Otra observación curiosa, una vez explorados ambos data sets por separado es que las comunidades autónomas de destino y aquellas en las que los visitantes gastan más dinero no coinciden. De la misma manera, no coinciden siempre los orígenes de los turistas que más gastan en nuestro país con el origen de aquellos que vienen en mayor número.

Para comprender qué tipo de turistas más nos interesa atraer y hacer sentir cómodos para que vuelvan, habría que explorar un nuevo parámetro que sería el producto del número de turistas, el gasto diario de los mismos y la duración media del viaje, según origen y comunidad autónoma de destino.

Con este objetivo, la mejor estrategia sería hacer un "join" de ambos data sets para poder multiplicar las columnas pertinentes. Sin embargo, con el fin de obtener a partir de esto unas gráficas lo más ilustrativas posibles, realizaremos este paso en Power BI más adelante.

Otra conclusión destacable durante el procesamiento y análisis de estos dos conjuntos de datos es que este proceso hubiese sido aún más fácil teniendo un buen glosario de datos. Además, para el futuro análisis de estos mismos datos, sería beneficioso contar con una guía y en general, un buen sistema de gobernanza de datos.

Veremos una herramienta para conseguir dicha organización en el capítulo 5.

Capítulo 4. Aprendizaje Automático

Un objetivo muy interesante de este análisis es poder predecir el número de turistas que tendremos en el futuro de un determinado país de origen en una determinada comunidad autónoma.

Este objetivo puede resultar un tanto ambicioso ya que a los datos históricos a los que se han tenido acceso no son muy extensos. Nuestros datos más antiguos van hasta el 2019. De todas formas, exploraremos hasta que nivel de precisión podría llegar una regresión lineal.

La regresión lineal es un algoritmo de aprendizaje supervisado. ¿Qué significa esto? Significa que es un método de análisis que aprende iterativamente de los datos. Se supone que partimos de un data set que ha sido etiquetado previamente o del que se conoce el valor objetivo a predecir de una cantidad de datos que se utilizarán para entrenar un modelo de predicción.

En el caso del aprendizaje no supervisado, se parten de datos no etiquetados previamente.

Usamos la librería de Python scikit-learn y en concreto la función incluida en esta librería, LinearRegression. Esta función hace un modelo lineal con coeficientes w = (w1, …, wp) para minimizar la suma de los cuadrados de la diferencia entre el resultado ya observado en el data set y el resultado que predice esta aproximación lineal.

El modelo lineal puede ser expresado de la siguiente manera según "*The Elements of statistical learning*":

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$$

En esta ecuación, la Y representa lo que se suele llamar como "target" ya que es aquello que queremos predecir. En nuestro data set del número de turistas, nuestro objetivo sería predecir el valor de la columna "Turistas", que indica el número de turistas en un trimestre determinado desde 2019, en una comunidad autónoma dada y de un país de origen concreto.

Por otra parte, las X's representan las variables que tendremos en cuenta para entrenar un modelo de regresión lineal que nos ayude a predecir como será el "target", es decir, como será el número de turistas. En nuestro caso, tendremos tres X's: la comunidad autónoma, el trimestre (fecha) y el país de origen de los visitantes.
Por último, hablemos de las β_i; $i = 0, .., i$. Estas β_i son los parámetros del modelo (o coeficientes). Básicamente, estos parámetros miden la influencia que tiene cada una de las variables que usamos para el entrenamiento a la hora de predecir el "target".

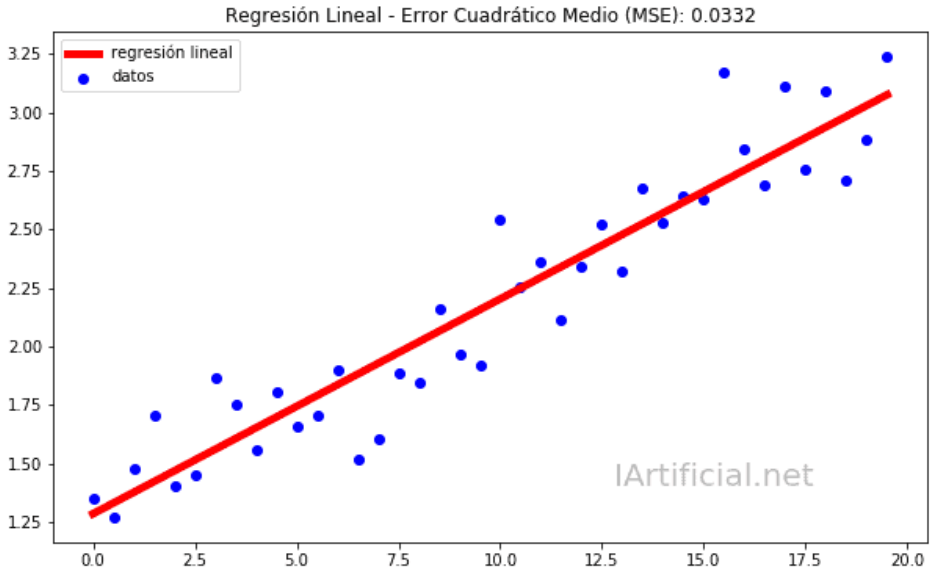

Ilustración 37: Ejemplo de regresión lineal simple. $Y = \beta_0 + \beta_1 X_1$. Fuente: https://www.iartificial.net/regresion-lineal-con-ejemplos-en-python/

Como se observa en la imagen de esta regresión lineal de tan solo una variable X, los puntos serían los datos reales que tenemos y la recta sería la aproximación que nos da la regresión. Nos interesa que, al calcular la distancia entre cada punto y la línea, la media de los cuadrados de estas distancias sea lo más pequeña posible. Cuanto más pequeña, mejor será nuestro modelo ya que más se acerca la regresión a los datos regales que tenemos como "target".

El objetivo de la regresión lineal es conseguir una combinación de parámetros β_i, tales que el "mean squared error" o la suma de los cuadrados de la diferencia entre el resultado ya observado en el data set y el resultado que predice esta aproximación lineal, sea lo menor posible.

La limitación a tener en cuenta con la regresión lineal es que solo podemos utilizar variables numéricas para entrenar nuestro modelo. Sin embargo, las variables que queremos utilizar en nuestro caso no son numéricas. Tendremos que hacer uso de una de las técnicas para codificar variables categóricas más conocida, el codificador "one hot".

Gracias al codificador ya mencionado, transformaremos los diferentes valores de la columna "CCAA" en columnas con valores 0 o 1. Empecemos codificando solamente esta columna y viendo el nivel de precisión solo tomando en consideración para el entrenamiento la comunidad autónoma y el año correspondiente.

A menudo, a la hora de codificar variables categóricas, resulta beneficioso realizar una codificación ordinal. Esto ocurre cuando los valores de nuestra variable categórica se pueden ordenar. Por ejemplo, si hablamos de un data set que muestra datos académicos

sobre alumnos de un colegio, es lógico asumir que una de las columnas hace referencia a su grado educativo.

Algunos de los valores posibles para esta variable serían “Primero de primaria”, “Segundo de primaria”, “tercero de la ESO”, etc. Estos grados se pueden ordenar fácilmente. Para eso existe una función llamada “Ordinal Encoder”.

Aunque esta función es muy conveniente para muchos casos, en el nuestro no existe una componente de ordinalidad en las variables que queremos codificar. Por ello usaremos un método más adecuado, el codificador “one hot”.

Debido a que no teníamos una columna de “Año” en nuestro data set original, deberemos crearla como paso previo al codificador a partir de la columna que nos indica los trimestres y que incluye en la misma el año al que pertenece el susodicho trimestre.

```
# Creamos las variables binarias
dummies = pd.get_dummies(df_turistas['CCAA'], drop_first = True)
dummies.head()
# Añadimos las variables binarias al DataFrame
df_turistas = pd.concat([df_turistas, dummies], axis = 1)

# Eliminamos la vairable original
df_turistas = df_turistas.drop(columns=['CCAA'])
df_turistas.head()
df_turistas['Año']= df_turistas.Fecha.str[12:16]



targets = df_turistas["Turistas"]
variables = df_turistas.drop(["Turistas", "Pernoctaciones", "Duración media de los viajes", "media_pernoctaciones", "Fecha", "Pais origen", "Meses", "Trimestres"], axis=1, inplace=False)

df_turistas.head()
```

✓ 0.1s

	Fecha	Turistas	Pernoctaciones	Duración media de los viajes	Pais origen	media_pernoctaciones	Meses	Trimestres	Aragón	Asturias, Principado de	...	Comunitat Valenciana	Extremadura	Galicia
1400	2022M05	126.0	824.0	6.5	Albania	6.539683	05	2022T2	0	0	...	0	0	0
1401	2022M04	145.0	1006.0	6.9	Albania	6.937931	04	2022T2	0	0	...	0	0	0
1402	2022M03	95.0	594.0	6.3	Albania	6.252632	03	2022T1	0	0	...	0	0	0
1403	2022M02	88.0	574.0	6.5	Albania	6.522727	02	2022T1	0	0	...	0	0	0
1404	2022M01	92.0	703.0	7.6	Albania	7.641304	01	2022T1	0	0	...	0	0	0

5 rows × 27 columns

Ilustración 38: One hot encoding de la columna CCAA en el data set que nos indica el número de turistas. Además de hacer la codificación, añadimos dos nuevas variables: "targets" y "variables". Fuente: Elaboración propia.

Como se observa en el código, a parte de la codificación, creamos dos nuevas variables con la finalidad del posterior entrenamiento del algoritmo:

- La variable “targets” hace referencia a aquello que nos gustaría predecir que, como mencionamos previamente, no es más que la columna “Turistas” del DataFrame.

- La variable “variables”, indica el resto de las columnas que utilizaremos para la predicción. Si nos fijamos, para este primer experimento solo utilizaremos el año en que se produce la visita de los turistas y la comunidad autónoma.

Este codificador, lo que hace es transformar los campos categóricos. El objetivo es obtener un array de números (0 o 1). De esta manera, si los registros de una de nuestras

filas del DataFrame original tiene como valor en la columna de comunidad autónoma "Aragón", tras hacer la codificación, en esa fila, en la columna creada correspondiente a Aragón, pondrá un 1 y en las columnas correspondientes al resto de comunidad autónomas pondrá un 0.

Tradicionalmente, las nuevas variables binarias que se crean a partir de los valores que puede tomar la variable original que queremos codificar, se conocen como "dummies".

Como se aprecia en la Ilustración 38, se puede llevar a cabo esta técnica usando la función de pandas "get_dummies". Esta función crea las variables binarias "dummies", como el nombre de la función indica, de manera automática. Como vemos en la siguiente ilustración, ya no tenemos las columnas originales de este dataset, sino que tenemos muchas más debido al codificador.

```
[37]  variables.columns

...  Index(['Pernoctaciones', 'Duración media de los viajes',
        'media_pernoctaciones', 'Año', '    Aragón',
        '    Asturias, Principado de', '    Balears, Illes', '    Canarias',
        '    Cantabria', '    Castilla - La Mancha', '    Castilla y León',
        '    Cataluña', '    Ceuta', '    Comunitat Valenciana',
        '    Extremadura', '    Galicia', '    Madrid, Comunidad de',
        '    Melilla', '    Murcia, Región de',
        '    Navarra, Comunidad Foral de', '    País Vasco', '    Rioja, La',
        'Alemania', 'Andorra', 'Armenia', 'Austria', 'Belarús',
        'Bosnia y Herzegovina', 'Bulgaria', 'Bélgica', 'Chipre', 'Croacia',
        'Dinamarca', 'Eslovenia', 'Estonia', 'Finlandia', 'Francia', 'Georgia',
        'Gibraltar', 'Grecia', 'Hungría', 'Irlanda', 'Islandia', 'Italia',
        'Letonia', 'Liechtenstein', 'Lituania', 'Luxemburgo', 'Macedonia',
        'Malta', 'Moldavia', 'Montenegro', 'Mónaco', 'Noruega',
        'Otros países o territorios de  Europa', 'Países Bajos', 'Polonia',
        'Portugal', 'Reino Unido', 'República Checa', 'República Eslovaca',
        'Rumanía', 'Rusia', 'Serbia', 'Suecia', 'Suiza', 'Ucrania',
        '           2019T4', '           2020T1', '           2020T2',
        '           2020T3', '           2020T4', '           2021T1',
        '           2021T2', '           2021T3', '           2021T4',
        '           2022T1', '           2022T2'],
       dtype='object')
```

Ilustración 39: Columnas nuevas añadidas al crear las variables categóricas. Fuente: Elaboración propia.

Probemos el modelo con nuestras nuevas variables binarias como se muestra en la imagen siguiente.

```
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score

X_train, X_val, y_train, y_val = train_test_split(
    variables, targets, test_size=0.4, random_state=0)



linr = Pipeline([('std', StandardScaler()), ('lr', LinearRegression())])
linr.fit(X_train, y_train)
score = linr.score(X_val, y_val)


print(f"{score}")
print(mean_absolute_error(y_val, linr.predict(X_val)))
```

```
[34]  ✓ 0.2s
...  0.04890598596870277
     6381.620530962915
```

Ilustración 40: Al usar una regresión lineal solo considerando CCAA y año, vemos que el score es muy bajo (0.0489), lo que indica que nuestro modelo es muy malo. Fuente: Elaboración propia.

En vistas de este resultado, optaremos por añadir para el entrenamiento más variables. Añadiremos para el entrenamiento el país de origen y los trimestres. De nuevo, dado que se tratan de variables de tipo texto y la regresión lineal solo admite valores numéricos, tendremos que usar el codificador one hot.

Como se puede apreciar en esta última ilustración, el score mejora considerablemente, dando lugar a un modelo que nos ofrece mucha más precisión con un “mean absolute error” menor.

```
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score

X_train, X_val, y_train, y_val = train_test_split(
    variables, targets, test_size=0.4, random_state=0)



linr = Pipeline([('std', StandardScaler()), ('lr', LinearRegression())])
linr.fit(X_train, y_train)
score = linr.score(X_val, y_val)


print(f"{score}")
print(mean_absolute_error(y_val, linr.predict(X_val)))
```

```
[43]
...  0.9651256496104113
     1579.4407561150433
```

Ilustración 41: Modelo de regresión lineal considerando como variables para el entrenamiento las columnas de CCAA, país de origen y trimestre. Fuente: Elaboración propia.

Experimentemos a continuación con una regresión logística. Las regresiones logísticas están más enfocadas a obtener "targets" categóricos donde hay un número limitado de posibilidades. Es probable que obtengamos peores resultados que con la lineal ya que realmente nuestro "target", el número de turistas, no es una variable categórica.

La ecuación de la regresión logística es la siguiente según "*The Elements of statistical learning*":

$$Y = \frac{1}{1 + e^{x}}$$

Y tiene la siguiente forma:

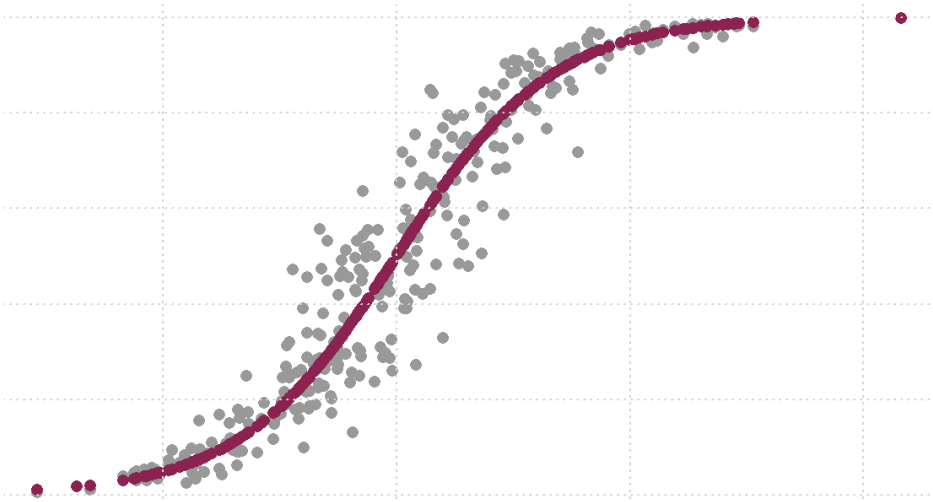

Ilustración 42: Regresión logística. Fuente: https://www.statdeveloper.com/wp-content/uploads/2020/02/funcion-logistica.png

A continuación, veamos cómo queda el modelo de regresión logística al hacer el entrenamiento con nuestros datos.

Como era de esperar, el error es mayor y el score es bastante más bajo. Si nos referimos al modelo de regresión lineal previamente expuesto, nuestro score era de alrededor 0.96, quedando muy lejos del 0.27 que obtenemos con el modelo de regresión logística. Esto tiene sentido ya nuestro variable "target" no se trata de una variable categórica.

```
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score

X_train, X_val, y_train, y_val = train_test_split(
    variables, targets, test_size=0.4, random_state=0)



log = Pipeline([('std', StandardScaler()), ('lg', LogisticRegression(max_iter = 1000))])
log.fit(X_train, y_train)
score = log.score(X_val, y_val)


print(f"{score}")
print(mean_squared_error(y_val, log.predict(X_val)))
```

```
41]
0.2684389545291801
61816871.047708556
```

Ilustración 43:Regresión logística. Fuente: Elaboración propia.

Como conclusión, por la naturaleza de nuestro "target" es mucho mejor optar por la regresión logística.

Capítulo 5. Gobierno del dato

5.1 ¿Qué es?

Tras esta exploración y para facilitar el uso futuro de las fuentes de datos que hemos utilizado, es importante establecer una buena organización. Según la DAMA (International Data Management Association), definimos el gobierno del dato como:

"El Gobierno de Datos es el ejercicio de la autoridad y el control (planificación, el seguimiento y la aplicación) a través de la gestión de los activos de datos. La función de Gobierno de Datos guía de cómo se llevan a cabo todas las demás funciones de gestión de datos. El Gobierno de Datos es de alto nivel y administra ejecutivamente los datos."

Como se puede apreciar, implantar un sistema de gobierno puede ser tremendamente beneficioso. Uno de los pasos más básicos hacia un buen gobierno del dato pasa por clarificar el linaje de los datos, entender las fuentes y como han sido usados posteriormente, además de tener un glosario de los datos que explique qué información nos está dando cada tabla y cada columna.

Para el uso de nuestras tablas por usuarios futuros, que no conozcan en profundidad la naturaleza de los datos, es muy beneficioso tener recogido en un mismo sitio no solo qué información nos da cada una de las columnas, sino también qué tipo de variables son ("float", "string", etc.) y que herramientas se han creado usando estos datos.

Por ejemplo, a lo mejor un usuario quiere usar los data sets expuestos previamente, unirlos usando una primary key, y calcular el producto entre "número de turistas", por "duración media de los viajes", por "gasto medio por visitante", con el fin de obtener los ingresos que proporcionan un grupo de visitantes, en una fecha determinada en una comunidad autónoma concreta.

Con un buen glosario de datos, este usuario podría ver directamente qué campos usar para establecer relaciones entre los distintos data sets o incluso, creando un linaje de los datos, podría averiguar si alguno de sus colegas ya ha establecido esas relaciones antes que él e incluso ya hay una herramienta interactiva creada, por ejemplo, un dashboard, para así obtener la información que necesita sin duplicar esfuerzos. Veremos un ejemplo de dashboard como el mencionado en el próximo capítulo.

En nuestro caso, un linaje de los datos se vería tal que así:

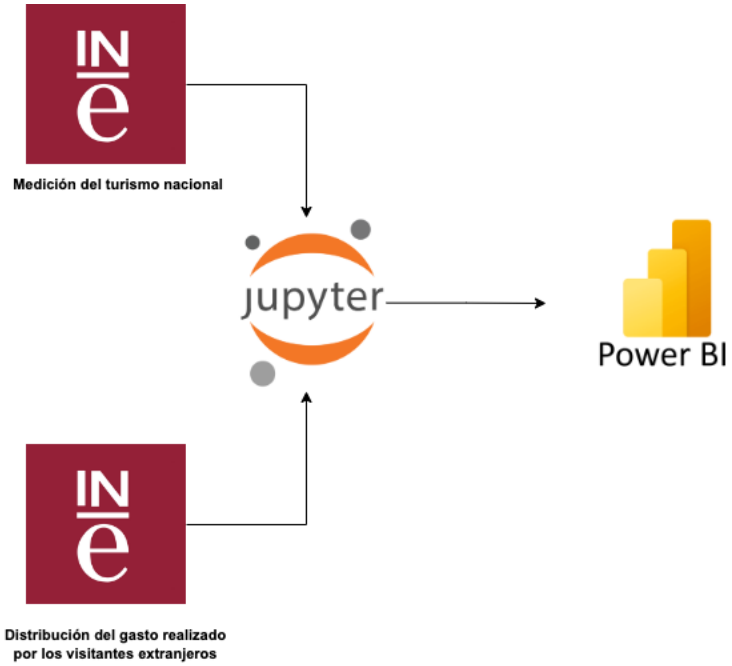

Ilustración 44: Ejemplo de linaje de los datos. Fuente: Elaboración propia.

Si bien en este caso el esquema es bastante sencillo, la elaboración de este tipo de diagramas puede ahorrar mucho tiempo y esfuerzos duplicados. Para construir un linaje de los datos y un glosario existen múltiples herramientas de pago como Collibra o azure purview que pueden enlazarse directamente con la fuente donde estén recogidos los datos y que elaborarán por nosotros tanto el diagrama para el linaje como el glosario.

Estas herramientas identificarían automáticamente el tipo de dato que es cada campo y los clasifican en una categoría. Además, son capaces de identificar duplicado, registros vacíos y ayudar con el procesamiento y limpieza de estos.

Si bien los datos en si son muy importantes, los metadatos lo son tanto o más. De ahí que se elaboren diccionario de datos cuyo propósito es informar a los usuarios sobre el tipo de datos, la longitud permitida, el linaje, las transformaciones, etc.

Esta información ayuda a los analistas a entender como unir diferentes tablas, como reportar estos datos y cómo procesarlos, además suele indicar la granularidad de los mismos. Por ejemplo, un diccionario de datos para nuestros data sets debe indicar que la granularidad en cuanto a fecha es que los datos son trimestrales.

Es importante no confundir el diccionario de datos con el catálogo de datos. El catálogo de datos tiene como objetivo facilitar la búsqueda de los datos. Ayuda a localizar la información que se busca y nos da un mapeo de los diccionarios de datos.

Según DAMA: “Un catálogo de datos puede presentarse de diversas maneras, como una especie de ‘marketplace’ de datos empresariales”.

5.2 5.2 Ejemplo práctico con Collibra

Para ver en práctica los conceptos que hemos explorado anteriormente, vamos a ver cómo funciona la herramienta de gobierno del dato Collibra. Collibra es una empresa de data intelligence fundada en 2008. Su plataforma para el gobierno del dato es uno de sus productos de mayor éxito.

Haremos uso de los data sets de prueba ya que con la versión de prueba no se nos permite cargar nuestros datos a la herramienta, pero con nuestros data sets se vería de manera muy similar como iremos explicando a continuación.

Una de las tablas de datos de prueba nos muestra datos de recursos humanos de los empleados de una empresa. Si un analista quisiera explorar los salarios de los empleados en los diferentes departamentos de la empresa, tendría que buscar primero la tabla apropiada usando Collibra.

Para encontrar lo que buscamos rápidamente, basta con buscar “income” y filtrar los resultados con el filtro “table”, ya que es justo lo que buscamos, una tabla que contenga datos sobre los salarios de los empleados.

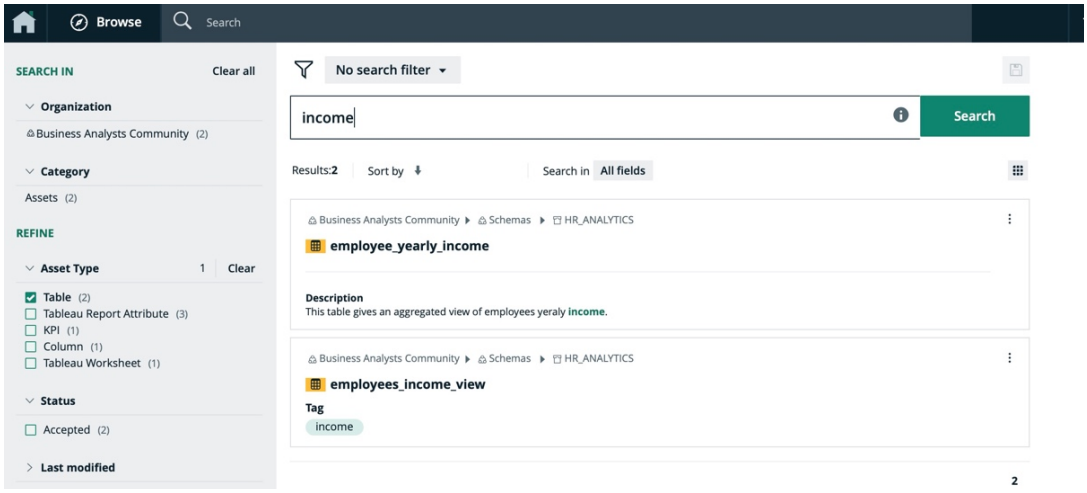

Ilustración 45: Como vemos, podeos filtrar a la derecha si buscamos una columna, una tabla un reporte, etc. Facilitando encontrar los datos necesarios sin necesidad de saberlo previamente. Fuente: https://www.collibra.com/us/en

Con nuestros data sets sobre el turismo en diferentes puntos de España, bastaría con buscar algo como “turistas” y nos aparecerían todas las tablas con datos relacionados con el turismo.

En este caso vamos a mirar la segunda tabla que nos aparece en la ilustración anterior, llamada “employees_income_view”. Cuando abrimos esa tabla, veremos información relevante como el “owner” de la tabla, las columnas, una descripción de la tabla. Además, se ve cuál es la fuente, el datamart, etc.

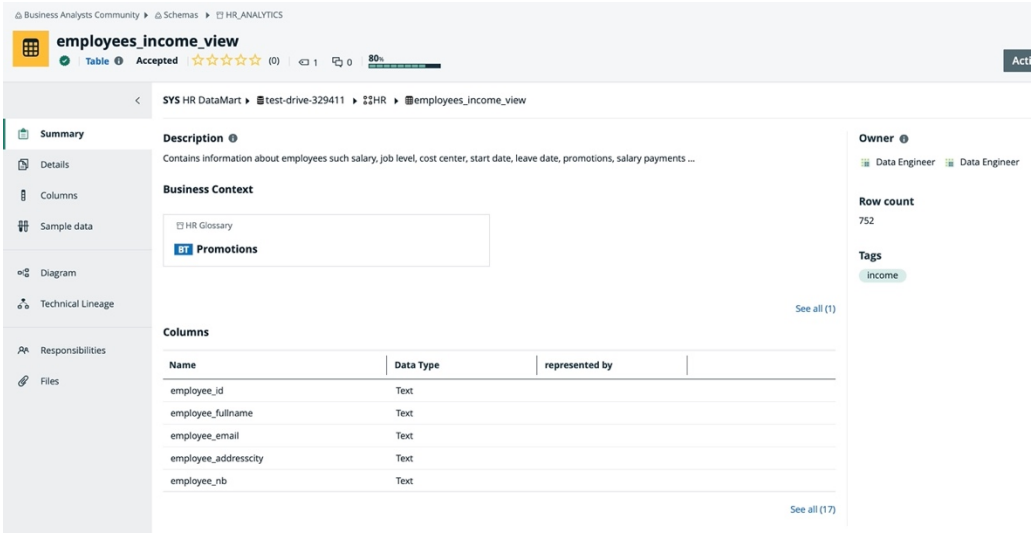

Ilustración 46:Interfaz Collibra con metadatos de la tabla "employees_income_view". Fuente: https://www.collibra.com/us/en

Si hacemos click en "see all" podremos ver todas las columnas con el data type e incluso con algunas representaciones gráficas de las mismas.

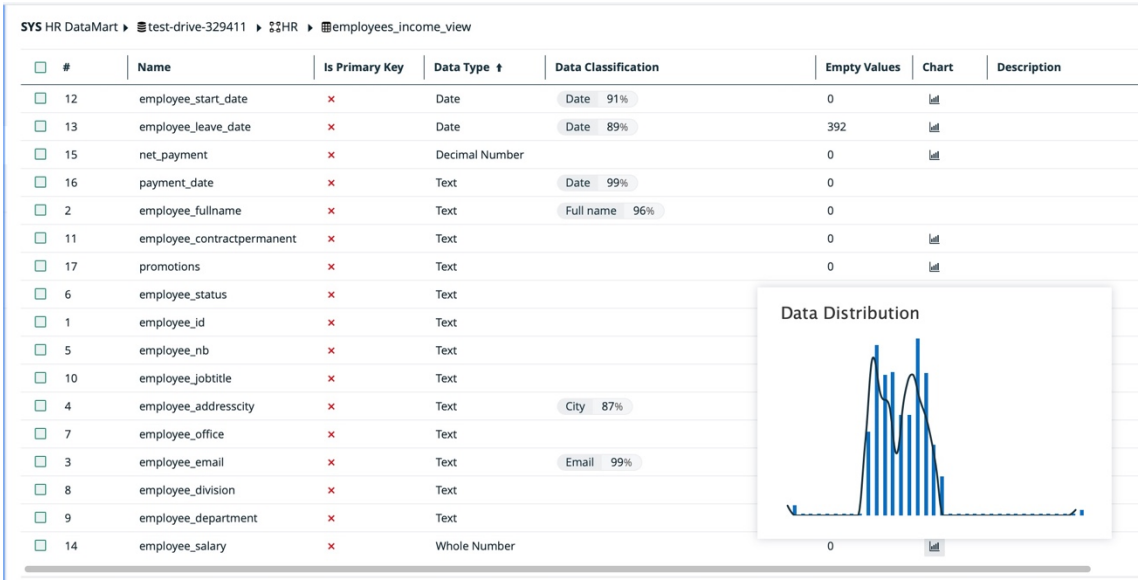

Ilustración 47: Columnas con el data type y representaciones gráficas relevantes. Fuente: https://www.collibra.com/us/en

Como se observa, podemos ver si cada una de las columnas es de tipo text, whole number, decimal number. Además, muchas de las variables tienes charts que indica la distribución de las mismas. En la imagen se muestra la distribución de los datos de la columna "employee_salary". En esta representación el eje x representa el salario y. el eje y el número de empleados con ese salario. Por otro lado, también se nos ofrece información respecto a si las columnas tienen valores vacíos y cuántos.

Collibra ofrece una plataforma más “user friendly”. Por ejemplo, con nuestro primer data set que contiene información sobre el número de turistas, podemos ver en el jupyter notebook cuantos valores vacíos hay, pero desde luego, no es tan “user friendly” como se nos presenta en Collibra.

Ilustración 48:Número de valores vacíos en la columna "Turistas", que indica el número de turistas, en el DataFrame df_turistas. Fuente: Elaboración propia.

En collibra, se crea automáticamente un linaje de cada una de las tablas para comprender si vienen a su vez de otras tables a establecer relaciones entre diferentes campos. En nuestro ejemplo podríamos ver como se construye la tabla “employees_income_view”.

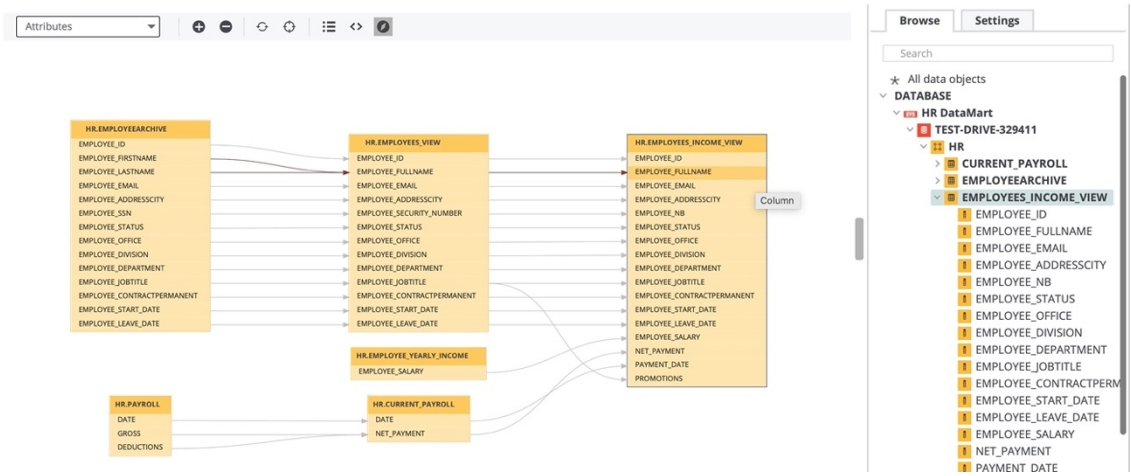

Ilustración 49: Linaje de la tabla “employees_income_view”. Como se aprecia en la ilustración, podemos situar el puntero sobre una de las tablas y se resalta las flechas que indican las relaciones entre los diferentes campos de otras tablas que dieron lugar a la columna en la que nos estamos enfocando. Fuente: https://www.collibra.com/us/en

Si subiésemos ala interfaz de Collibra nuestros data sets sobre el número de turistas y el gasto medio, se podría construir un linaje que relacionase ambas tablas para construir un nuevo campo que sería el ingreso que proporcionan los turistas a España (gasto medio por duración del viaje, por número de turistas).

Otra ventaja de esta herramienta es que, además, podemos ver si una tabla determinada ha sido utilizada en algún reporte. Esto puede ser muy útil para evitar duplicar esfuerzos por desconocimiento de ciertas herramientas de reporting ya creadas dentro de una empresa.

Ilustración 50: Información sobre reportes ya creados donde se ha utilizado un determinado conjunto de datos. Fuente: https://www.collibra.com/us/en

En nuestro caso, si la información sobre los conjuntos de datos sobre turismo que hemos utilizado estuviese en Collibra, podríamos ver si ya se ha creado un dashboard utilizando dichos data sets, como el que veremos más adelante. En este caso accedemos a un reporte hecho en Tableau sobre el salario medio de los empleados por departamento.

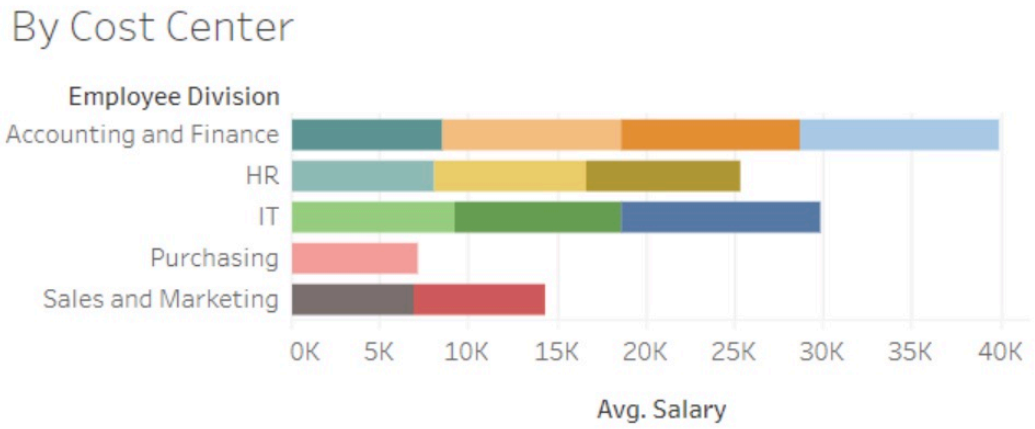

Ilustración 51: Visualización sobre salario medio de los empleados según la división en la que trabajan dentro de la compañía. Fuente: https://www.collibra.com/us/en

Como conclusión: tener en cuenta una herramienta de gobierno del dato es muy beneficioso ya que da lugar a un gran ahorro de tiempo y energía al poder tener una mejor comprensión de los recursos y los datos a los que se tiene acceso. Otro añadido es que podemos ver incluso como otros usuarios han hecho uso de los datos disponible para no repetir el esfuerzo realizado por otros o para comprender mejor el potencial que tienen los datos y poder usarlos de la manera más práctica y eficiente.

Capítulo 6. Visualización en Power BI

6.1 Importancia de una buena visualización y contexto.

La visualización de datos es algo que la humanidad lleva haciendo durante siglos. Se podría decir que una de las representaciones más antiguas de ello son la creación de mapas. Este tipo de visualización no solo se lleva haciendo durante mucho tiempo, sino que no pasa de moda, ya que en la actualidad seguimos viendo muy a menudo mapas para representar todo tipo de eventos.

Por ejemplo, el siguiente mapa, publicado en el periódico "El País" en mayo de 2021 es un indicativo de que es un tipo de representación muy útil e ilustrativa.

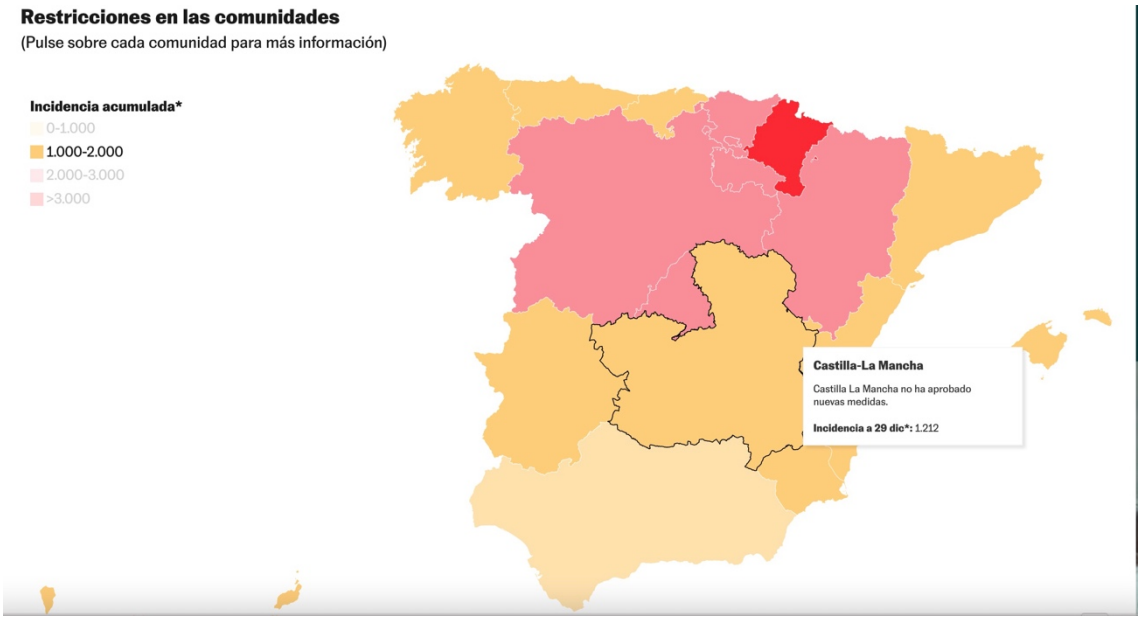

Ilustración 52: Mapa de la incidencia acumulada causada por la covid-19. Publicado por el periódico "El País" en mayo de 2021. Fuente: https://elpais.com/sociedad/2021-05-27/el-mapa-de-las-restricciones-por-la-covid-en-espana-y-la-incidencia-acumulada-de-coronavirus-en-cada-comunidad.html

Este mapa además no solo nos muestra la incidencia de las diferentes comunidades autónomas usando diferentes colores, sino que además es interactivo. Al pasar el ratón por encima de una comunidad autónoma nos informa de si se han tomado nuevas medidas y si hay nuevas restricciones impuestas por esa comunidad autónoma.

Pero como hemos mencionado, esto no es nada nuevo. Una de las representaciones de ello más sonada es del siglo XIX y lleva la cuenta de las vidas perdidas dentro de ejército de Napoleón en su marcha hasta Rusia.

Cincuenta años después de este viaje realizado por el ejército francés, conocido como la "trágica marcha sobre Moscú", Charles Minard creó un registro de las vidas perdidas

en el viaje. En su representación de esto se incluye las temperaturas en diferentes épocas del año y el número de muertos que hubo en ese momento en el tiempo.

Minard optó por representar en un mapa con dos colores la cantidad de hombre que iban y los que regresaban.

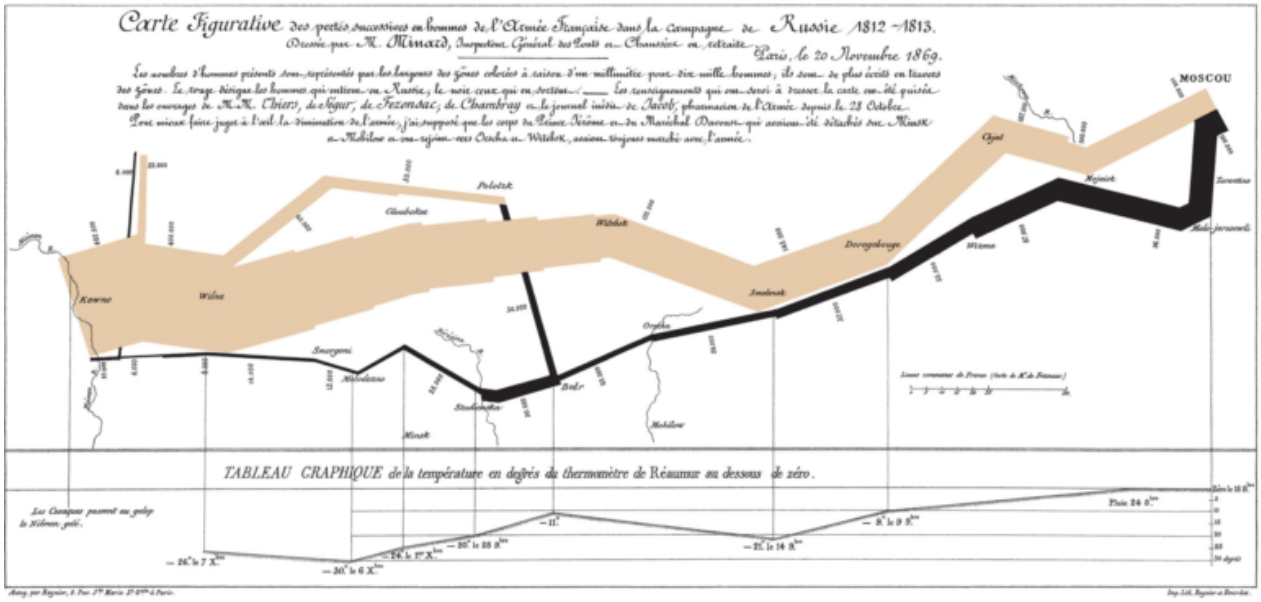

Ilustración 53: Mapa elaborado por Charles Minar en el que se observan el número de hombre que marcharon a Moscú en beige y cuántos regresaron en negro. En la parte inferior se añade además un registro de las temperaturas a lo largo del camino de este ejército. Fuente: Friendly, M. (2002). Visions and Re-Visions of Charles Joseph Minard. Journal Of Educational And Behavioral Statistics, doi: 10.3102/10769986027001031.

Hoy en día existen muchas herramientas de visualización: Power BI, Cartos, Tableau, etc. Estas herramientas nos permiten crear dashboards interactivos.

En una primera etapa de la visualización, cuando se carecían de las tecnologías actuales, solo se buscaba representar una "captura" de una serie de parámetros sin que fuese ni dinámico ni modificable.

Hoy en día tenemos herramientas como Cartos, que nos permite elaborar mapas que cambian con el paso de tiempo y nos permite visualizar esos cambios a modo de animación.

También contamos con herramientas como Tableau y Power Bi que nos permiten establecer filtros accesibles para los usuarios de nuestras visualizaciones. Estos filtros crean una nueva posibilidad de customización de la visualización acorde con las necesidades del usuario.

Si Charles Minar hubiese tenido estas herramientas podría haber creado una visualización donde los trazos de color beige y negro se fueran abriendo paso sobre un mapa a medida que pasaba el tiempo que tardo el ejército de Napoleón en recorrer el territorio representado.

Además, por hacerlo más sofisticado, podría haber añadido filtros para poder ver cuántos hombres de diferentes rangos de edad morían. Por ejemplo, seleccionar ver solo las líneas para aquellos miembros del ejército de entre 20 y 30 años.

En cualquier caso, con la tecnología actual se podría caer en la tentación de añadir demasiados elementos visuales a nuestra representación y dar lugar a una imagen poco clara.

6.2 Visualización de nuestro análisis. Creación de un dashboard.

Para tener una mejor perspectiva de los datos y poder conseguir una serie de conclusiones significativas, utilizaremos Power BI como herramienta de visualización.

El objetivo es que, con esta herramienta, logremos fabricar un dashboard que sea de utilidad para los usuarios y que puedan customizar ellos mismos mediante diferentes filtros.

Empezaremos añadiendo ambas fuentes de datos a Power BI y haciendo algunas transformaciones con la herramienta Power Query. Además de las dos fuentes de datos ya conocidas, necesitamos un nuevo data set que será de gran relevancia para una de las visualizaciones de la que hablaremos más adelante.

Este nuevo data set, contiene las coordenadas geográficas de cada una de las comunidades autónomas de España y las de Ceuta y Melilla.

Este nuevo data set incluye una columna con el nombre de las comunidades autónomas que utilizaremos como primary key para relacionar con los otros dos conjuntos de datos. Sin embargo, nos encontramos con que el nombre de las comunidades esta escrito de forma distinta a como está en los otros dos data sets, por este motivo usaremos Power Query para arreglarlo.

El lenguaje nativo de esta herramienta es DAX y mediante las siguientes líneas de código podremos cambiar el nombre de aquellas comunidades autónomas que no coincidan para que tengan el nombre pertinente:

```
“= Table.AddColumn(#"Tipo cambiado", "CCAA_KEY", each if [CCAA] = "Islas Baleares" then "Balears, Illes" else if [CCAA] = "Principado de Asturias" then "Asturias, Principado de" else if [CCAA] = "Comunida de Madrid" then "Madrid, Comunidad de" else if [CCAA] = "Castilla la Mancha" then "Castilla - La Mancha" else if [CCAA] = "Comunidad Valenciana" then "Comunitat Valenciana" else if [CCAA] = "Región de Muercia" then "Murcia, Región de" else if [CCAA] = "Comunidad Foral de Navarra" then "Navarra, Comunidad Foral de" else [CCAA])”
```

Como vemos en la siguiente imagen, gracias a esas líneas de código, se crea una nueva columna llamada “CCA_KEY” que nos ayudará a establecer las relaciones necesarias.

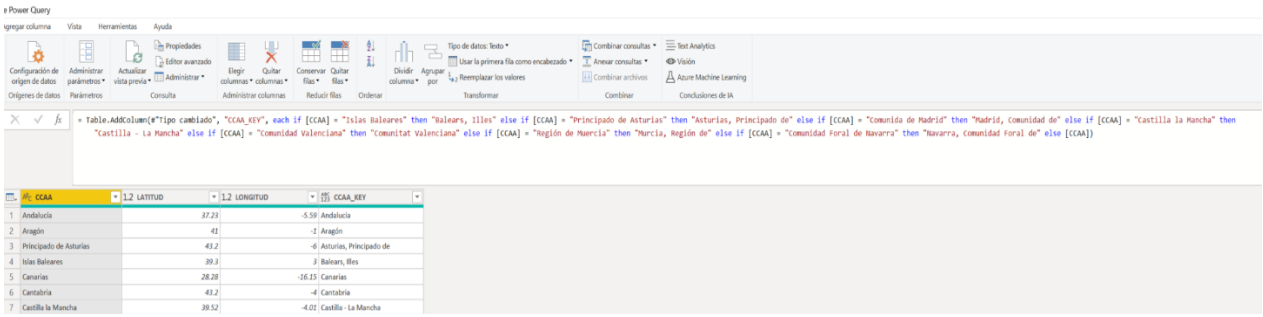

Ilustración 54: Líneas de código Dax junto a la tabla con la nueva columna "CCAA_KEY". Fuente: Elaboración propia.

Tras obtener esta columna podemos empezar a establecer las relaciones entre las tablas. En la interfaz de Power BI existe una pestaña llamada “Modelo”. Aquí podemos ver de manera clara como se organizan nuestros datos y sus relaciones. A la hora de añadir las relaciones manualmente podemos administrarlas nosotros mismos. Se trata de una interfaz sencilla en la que podemos seleccionar las columnas que coinciden en ambas tablas para establecer así la relación entre ambas.

Por ejemplo, la siguiente imagen muestra cómo se realiza la relación del tipo “Varios a uno” entre la tabla que indica el número de turistas por comunidad autónoma y las coordenadas de estas. La relación es de varios a uno ya que la tabla que indica el número de turistas por comunidad autónoma tiene montones de registros en diferentes fechas y haciendo referencia a turistas de distinta procedencia, en cambio, la tabla de las coordenadas geográficas solo tiene diecinueve filas ya que solo contiene el nombre de las comunidades autónomas (más Ceuta y Melilla), la latitud y la longitud.

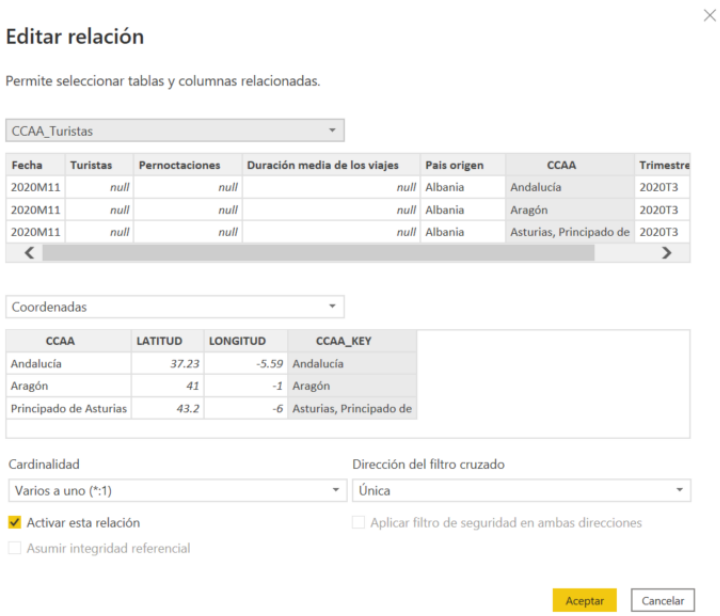

Ilustración 55: Ventana para configuración de la relación entre dos tablas. Como se observa, ambas tablas se unen gracias a la columna que indica el nombre de las comunidades autónomas. Fuente: Elaboración propia.

Una vez establecidas todas las relaciones podemos observar un esquema de estas en la pestaña de “Modelo”.

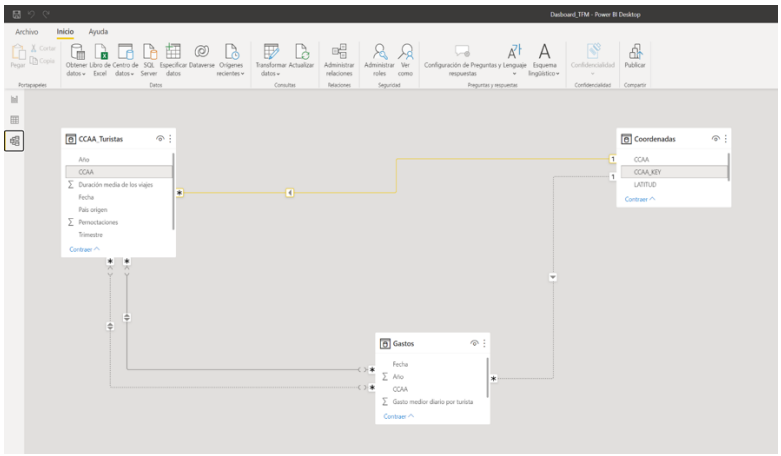

Ilustración 56: Esquema de relaciones entre las tablas. Fuente: Elaboración propia.

Además de por la columna de las comunidades autónomas, establecemos relaciones entre los conjuntos de datos que indican el número de turistas y el gasto de estos usando los campos de fecha y procedencia de los turistas.

Ahora que hemos hecho todos los “joins” pertinentes podemos comenzar con la visualización. Teniendo en cuenta que hay una componente geográfica en este estudio, sería de gran utilidad poder visualizar los datos sobre un mapa. Con frecuencia, la representación sobre un mapa puede ser confusa. Para mayor claridad diseñamos un tooltip personalizado que de una imagen más clara y precisa de los datos de gasto medio de los visitantes en cada comunidad autónoma y del número de visitantes con el top 5 de países de procedencia. Basta con poner el cursor encima de la comunidad autónoma de interés para ver claramente esta información.

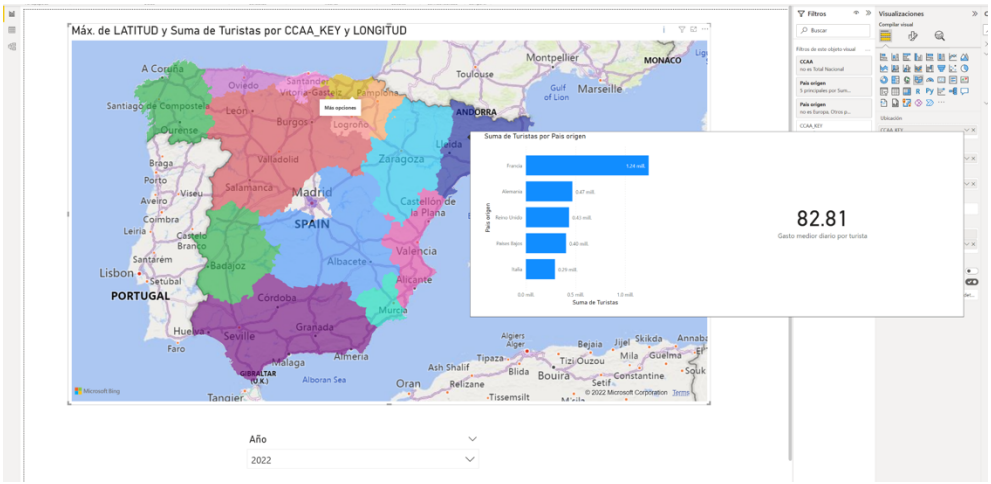

Ilustración 57: Mapa de España interactivo. Se muestra el tooltip al pasar el cursor por encima de Cataluña. Fuente: Elaboración propia.

Se optó por mostrarlo de esta manera ya que mostrar toda esta información de todas las comunidades autónomas a la vez sobre el mapa podría dar lugar a confusión.

Si bien este mapa interactivo ayuda a llevarnos una idea general de cada comunidad autónoma, existe un KPI muy importante del que no hemos hablado aún ni siquiera el primer capítulo de exploración de los datos. Este KPI es los ingresos proporcionados por el turismo en cada comunidad autónoma en una determinada fecha. Esto supone multiplicar la duración media de los viajes, por el número de turistas, por el gasto medio diario de cada uno de estos turistas.

Nótese que este producto incluye campos de dos fuentes distintas, lo que implica que es necesario hacer la operación "join" para poder hacer dicho producto. Esta operación se puede realizar fácilmente tanto usando código en la librería Pandas como con la interfaz de Power BI, que es lo que hemos hecho en este caso.

Como se ha mostrado anteriormente, la creación y configuración de ese esquema de "relaciones" entre tablas gracias a determinadas columnas que nos servían de primary key, no son más que la realización de un "join".

De todas formas, no acaba todo ahí. Una vez establecidas estas relaciones, necesitamos hacer el cálculo del producto anteriormente expuesto. La manera de hacer dicho cálculo es mediante otra opción de transformación de datos a parte de Power Query que aparece al hacer click en la pestaña "Datos". Tendremos que crear una nueva medida y utilizaremos una vez más el lenguaje nativo de esta aplicación, DAX.

Mediante la fórmula que se muestra en la imagen obtenemos un nuevo campo disponible para añadir a nuestras visualizaciones con este KPI tan importante.

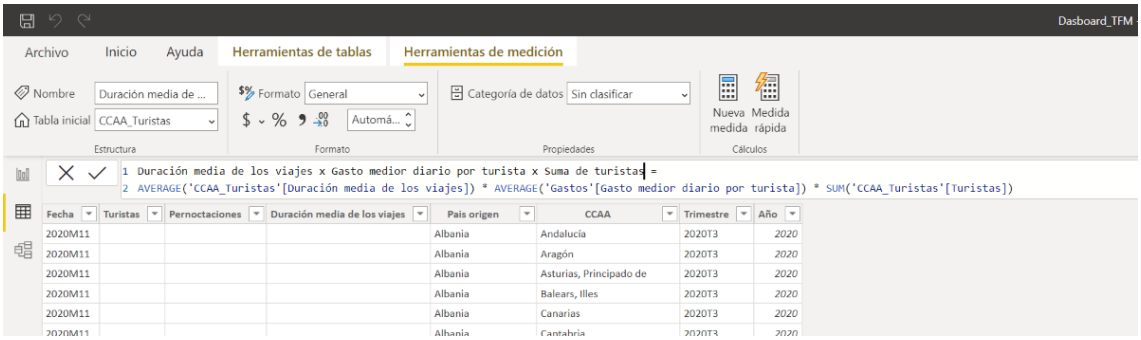

Ilustración 58: Interfaz Power BI. Como se muestra en la imagen, creamos una nueva medida llamada "Duración media de los viajes x Gasto medio diario por turista x Suma de turistas". Fuente: Elaboración propia.

Ahora que esta medida esta creada, tenemos todo lo necesario para añadir esta nueva página interactiva de nuestro dashboard donde para una fecha dada, podemos ver los ingresos producidos en la comunidad autónoma que pinchemos. Además, vemos en un diagrama de barras ordenado las comunidades que más turistas han recibido y dos líneas que representan el gasto medio diario por turista u la duración media de las estancias de los mismos.

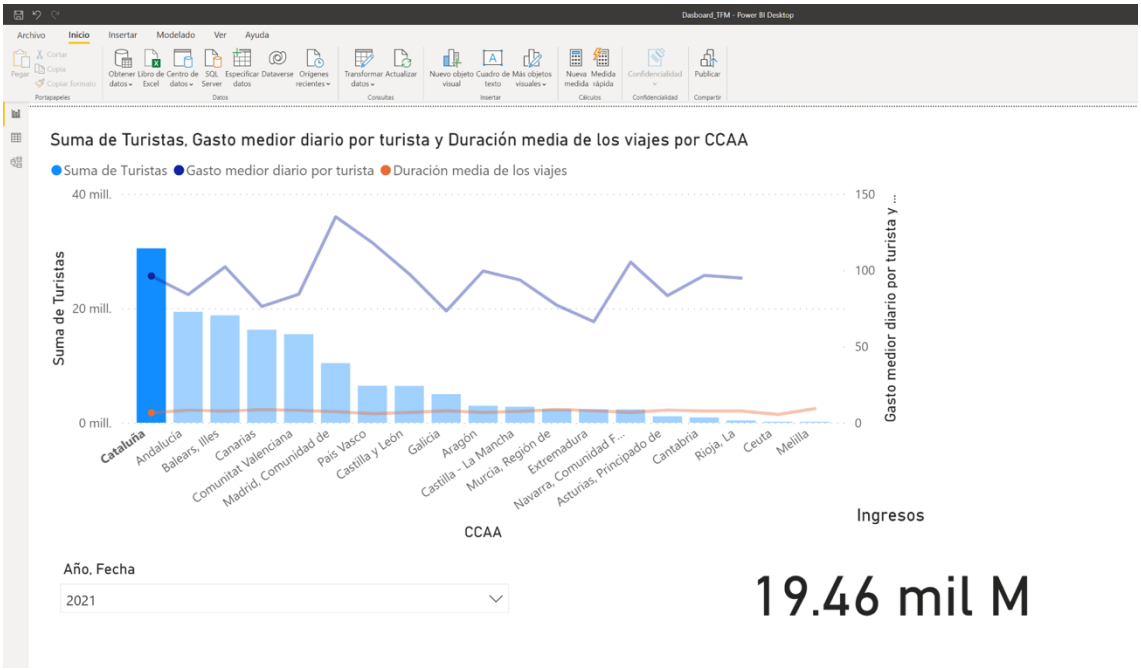

Ilustración 59: En la gráfica de la imagen se muestra para Cataluña, los ingresos totales que proporcionaron los visitantes durante 2021. Además, si hacemos click sobre las líneas podremos ver tanto la duración media de las estancias (línea naranja), como el gasto medio por visitante (línea azul). Fuente: Elaboración propia.

Para además tener otra representación visual que nos permita comparar a la vez la duración media de la estancia, el número de turistas y el promedio de gasto medio en cada comunidad autónoma, podemos crear un diagrama de burbujas donde el tamaño de estas corresponda al gasto promedio. Como añadido, gracias a las relaciones que hemos establecido entre columnas de diferentes tablas con anterioridad, si hacemos click en una de las burbujas también obtendríamos en las representaciones de la ilustración anterior los ingresos totales de esa comunidad como KPI y los datos del gráfico de barras para esa comunidad autónoma concreta.

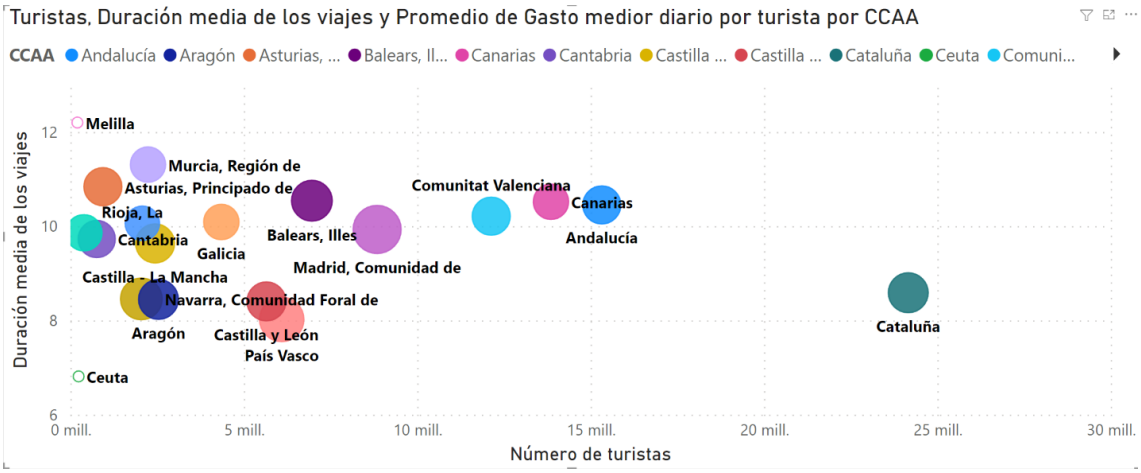

Ilustración 60: Diagrama de burbujas. Fuente: Elaboración propia.

Imaginemos que ahora queramos ver la evolución a través de los años del gasto medio de los diferentes turistas según su procedencia en España. ¿Y si quisiéramos verlo para una comunidad autónoma y para un país de procedencia concretos?

Con los datos que tenemos es fácil ver esta evolución temporal con un diagrama de líneas y un par de filtros como se observa en la siguiente imagen.

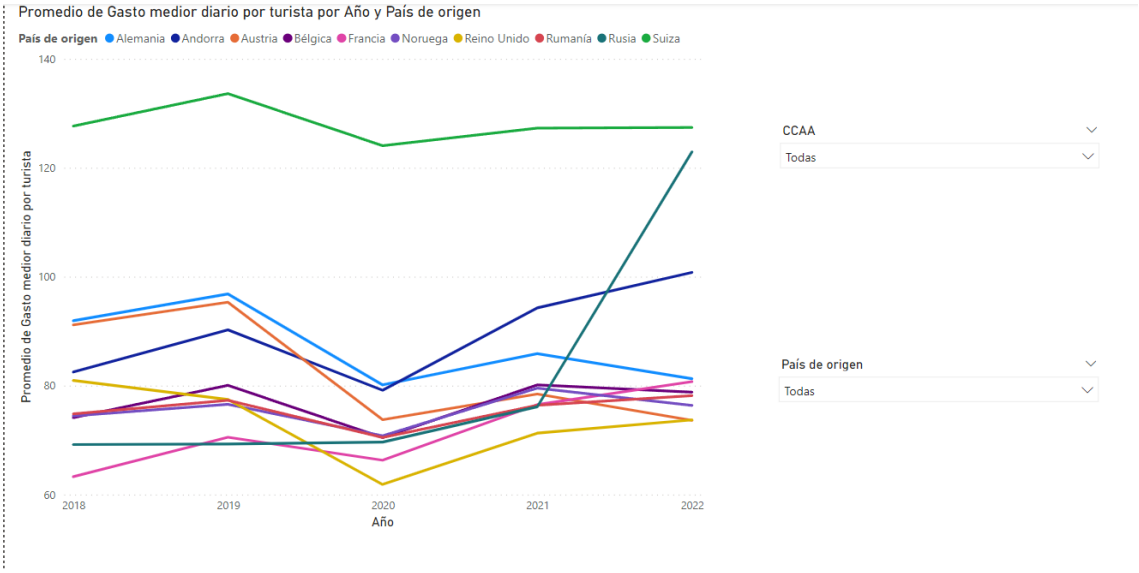

Ilustración 61: Evolución temporal del gasto promedio de los turistas según procedencia y comunidad autónoma de destino.

Capítulo 7. Conclusiones y futuras líneas de trabajo

En este análisis hemos sido capaces de observar cómo ha evolucionado el turismo en España en los últimos años que, por si fuera poco, han sido unos años bastante peculiares. Haber apreciado las diferencias en las cifras antes, durante y después de las medidas restrictivas que se tomaron durante la pandemia, hace que sea más interesante y que las conclusiones que se puedan sacar con respecto al futuro del turismo sean más relevantes que nunca. Los hallazgos se presentan optimistas ya que se ve una recuperación del sector tras un duro 2020.

Además, queda claro tras las diversas representaciones expuestas que las comunidades autónomas que más turistas han recibido en los últimos años son Cataluña, las Islas Baleares y Canarias. Esto nos sugiere que son comunidades en los que vale la pena invertir en el sector turístico. Sin embargo, cabe hacer una mención especial a la Comunidad de Madrid ya que, aunque el número de turistas sea más bajo, el gasto medio de los mismo es más alto. Para conseguir mayores ingresos, incentivar el turismo a Madrid es sin duda una muy buena inversión.

Gracias al dashboard pudimos observar como a pesar de que Cataluña, teniendo más del doble de los turistas que tiene la Comunidad de Madrid en 2021, tiene unos ingresos dejados por los turistas de 19.46 millones de euros, que son menos del doble que los ingresos que recibió la Comunidad de Madrid (30.36 millones de euros). De esto deducimos que es más rentable invertir en aumentar el turismo en la Comunidad de Madrid que en Cataluña. Esta tendencia además se mantiene a principios de 2022.

Elaborar un dashboard como herramienta de business intelligence ha sido vital para poder ver con más comodidad la evolución mencionada previamente. Además, es un ejemplo claro de cómo convertir una cantidad muy grande y a veces incomprensible de datos, en algo claro, informativo y accesible para gran cantidad de usuarios, puede facilitar la obtención de “insights”.

A pesar de no haber podido utilizar la herramienta Collibra, expuesta en el capítulo 5, con los conjuntos de datos sobre turismo, explorarla con los datos de prueba ha sido muy significativo para comprender la utilidad de la misma. En nuestro caso, los conjuntos de datos eran solo tres en total, pero normalmente una organización puede llegar a contar con muchas tablas y fuentes de datos distintas. Una herramienta que ayude a dar orden a todo ello ahorra el tiempo y aumenta la productividad además de evitar la replicación de esfuerzos.

Como línea de trabajo futura, explorar la versión de pago de Collibra con nuestros datos para ver cómo funciona en nuestro caso sería un siguiente paso a dar lógico.

Además, durante la realización del modelo de aprendizaje automático, solo se usa la regresión lineal como modelo válido ya que los datos que teníamos disponibles solo

eran de 2019 en adelante además de tener pocas columnas. Sería interesante entrenar modelos distintos que puedan ser mucho más exactos en las predicciones con data sets que así lo permitieran. Un data set con más datos como por ejemplo la edad de los visitantes, podría ayudar a dar una predicción con un score aún más alto.

Debido a que el data set con el que contábamos ha dado lugar a algunas limitaciones a la hora de obtener buenos resultados nos hemos limitado a entrenar un modelo de regresión lineal y a medir el score. Para un estudio futuro, contar con más datos históricos e incluso mayor granularidad a nivel temporal sería idóneo para hacer predicciones de mayor calidad.

BIBLIOGRAFÍA

INE. 2022. *INE. Instituto Nacional de Estadística*. [online] Available at: <https://www.ine.es> [Accessed 22 September 2022].

Epdata.es. 2022. *La ocupación de los hoteles, en datos y gráficos*. [online] Available at: <https://www.epdata.es/datos/ocupacion-hotelera-hoteles-datos-graficos/94/espana/106> [Accessed 22 September 2022].

DataScience Made Simple. 2022. *Home - DataScience Made Simple*. [online] Available at: <https://www.datasciencemadesimple.com> [Accessed 22 September 2022].

Pandas.pydata.org. 2022. *pandas - Python Data Analysis Library*. [online] Available at: <https://pandas.pydata.org> [Accessed 22 September 2022].

Statista. 2022. *Turismo: porcentaje del PIB en España | Statista*. [online] Available at: <https://es.statista.com/estadisticas/1082929/sector-turistico-porcentaje-del-pib-aportado-espana/> [Accessed 22 September 2022].

Statista Infografías. 2022. *Infografía: El 'boom' del empleo en el turismo en España*. [online] Available at: <https://es.statista.com/grafico/18875/afiliados-a-la-seguridad-social-en-actividades-turisticas-en-espana/> [Accessed 22 September 2022].

Numpy.org. 2022. *NumPy*. [online] Available at: <https://numpy.org> [Accessed 22 September 2022].

Scikit-learn.org. 2022. *scikit-learn: machine learning in Python — scikit-learn 1.1.2 documentation*. [online] Available at: <https://scikit-learn.org/stable/> [Accessed 22 September 2022].

Collibra.com. 2022. *Collibra: Data Catalog, Data Governance & Data Quality*. [online] Available at: <https://www.collibra.com/us/en> [Accessed 22 September 2022].

Henderson, D. and Earley, S., n.d. *DAMA-DMBOK*.

DAMA Spain.org. 2022. Retrieved 22 September 2022, from https://www.damaspain.org/data-governance-diferencias-entre-business-glossary-data-dictionary-and-data-catalog

Collibra.com. 2022. *Collibra: Data Catalog, Data Governance & Data Quality*. [online] Available at: <https://www.collibra.com/us/en> [Accessed 22 September 2022].

La historia de la visualización de datos - Brandwatch. (2022). Retrieved 22 September 2022, from https://www.brandwatch.com/es/blog/historia-de-la-visualizacion-de-datos/#:~:text=El%20inicio%20de%20la%20visualizaci%C3%B3n,mantiene%20vigente%20a%20la%20fecha.

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The Elements of statistical learning*. New York: Springer.

McKinney, W. *Python for data analysis*.

Heydt, M. (2015). *Learning pandas*. Birmingham, UK: Packt Publishing.

Friendly, M. (2002). Visions and Re-Visions of Charles Joseph Minard. *Journal Of Educational And Behavioral Statistics*, *27*(1), 31-51. doi: 10.3102/10769986027001031

(2022). Retrieved 25 September 2022, from https://www.mincotur.gob.es/es-es/COVID-19/turismo/Documents/Informe_suspension_alojamientos_turisticos.pdf

Rincón, S. (2022). Nuevo golpe al turismo: Alemania declara a España como territorio de alto riesgo por covid-19. Retrieved 25 September 2022, from https://www.antena3.com/noticias/mundo/nuevo-golpe-turismo-alemania-declara-espana-como-territorio-alto-riesgo-covid19_2021122361c562c3d71c190001d23070.html

Antípodas y coordenadas GPS. (2022). Retrieved 27 September 2022, from https://www.antipodas.net/