

MEMORIA

TRABAJO FIN DE MÁSTER

Sandra de la Fuente Cáceres

[Introducción](#)

[Tecnologías aplicadas](#)

[Descripción de la estructura de directorios](#)

[Descripción de los datos de entrada](#)

[Metodología](#)

[Resumen del resultado](#)

[Visualización Dashboard](#)

[Recomendaciones Tácticas & Estratégicas](#)

Introducción

El propósito de este Trabajo Fin de Máster (en adelante TFM) ha sido el de explorar y mostrar las posibilidades que puede ofrecer el análisis de datos, a la hora de tomar decisiones de negocio en el sector turístico español, tan importante para la economía de este país.

Los estudios existentes hasta la fecha sobre Turismo en España, se basan en su gran mayoría en encuestas o entrevistas, por ese motivo para este TFM, los datos utilizados son del Instituto Nacional de Estadística, disponibles en los siguientes enlaces:

- Estadística de movimientos turísticos en frontera.
Frontur http://www.ine.es/dynqs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176996&menu=resultados&secc=1254736195382&idp=1254735576863
- Encuesta de gasto turístico.
Egatur http://www.ine.es/dynqs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177002&menu=resultados&secc=1254736195390&idp=1254735576863

Los datos recopilados son de los últimos 19 meses, periodo entre Octubre del 2015 y Abril del 2017.

Objetivos del proyecto:

1. Análisis descriptivo y predictivo de la actividad turística en España, para predecir el número de turistas por nacionalidades que decidan viajar a un lugar determinado.
2. Segmentación de los turistas según características similares, para poder definir acciones de marketing específicas para cada tipo de turista, acciones publicitarias, así como lanzamiento de promociones especiales dirigidas a un turista objetivo.
3. Conocimiento holístico del turista.

Los estudios existentes hasta la fecha sobre Turismo en España se basan, en su gran mayoría, en encuestas o entrevistas a expertos promovidos por el Ministerio de Industria, Energía y Turismo como principal organización pública o bien por empresas privadas. Esto significa que, en general, el sector no dispone de datos reales de los turistas y sólo pueden extraer muestras de toda la población.

Tecnologías aplicadas

Para este TFM se utilizan las siguientes tecnologías:

- Python (*script_datos.py*) para descomprimir ficheros de los datos obtenidos del INE, versión utilizada la contenida en la máquina virtual.
- R para el proceso Data Engineering:
 - Limpieza de los datos
 - Análisis descriptivo de los datos
 - Análisis predictivo de los datos
 - Librerías de R, se deben instalar previamente en el entorno R-Studio para ejecutar los scripts rmd.
 - library(dplyr)
 - library(ggplot2)
 - library(data.table)
 - library(scales)
 - library(cluster)
 - library(NbClust)
 - library(reshape2)
 - library(effects)
 - library(ROCR)
 - library(stats)
 - library(caTools)
 - library(forecast)
- Tableau para realizar dashboards.

Descripción de la estructura de directorios

A continuación, se describe la estructura de directorios del repo Git con el orden de ejecución de los mismos.

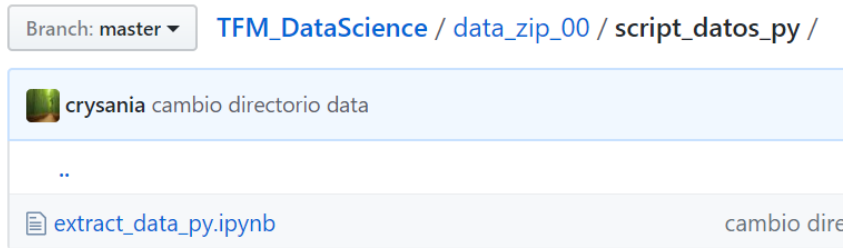
crysania eliminar RoadMap.docx	
data_01	se eliminan los ficheros para añadirlos a otro directorio
data_cleaning_02	Se modifica el nombre del fichero
data_exploring_03	modificacion de la etapa exploring de frontur y generacion de los fic...
data_prediction_04	predicciones series estacionales
data_visualization_05	Primeras visualizaciones
data_zip_00	Se añaden mas ficheros
.RData	Modificacion de exploring y se añade prediction
.Rhistory	Modificacion de exploring y se añade prediction
Memoria.pdf	Avances en la Memoria.pdf
README.md	Update README.md

data_zip_00 → Ficheros zips del INE dentro de cada directorio EGATUR y FRONTUR.

Dentro está el directorio (*script_datos_py*) para descomprimir ficheros de los datos obtenidos del INE.

Branch: master ▼	TFM_DataScience / data_zip_00 /
crysania Se añaden mas ficheros	
..	
Egatur	Se
Frontur	Se
script_datos_py	ca

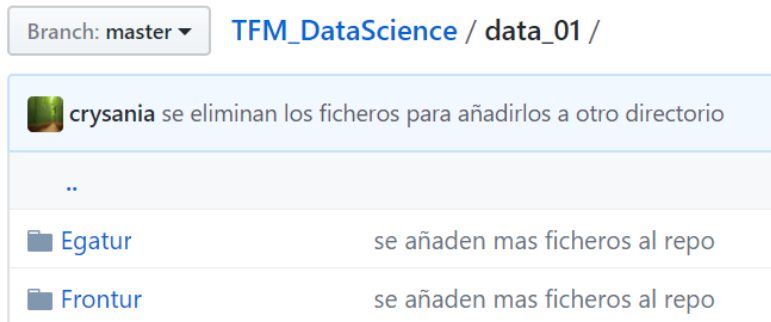
extract_data_py.ipynb → Script de Python para descomprimir los ficheros de los datos obtenidos del INE.



data_01 → Ficheros en formato .txt de Egatur & Frontur por cada “mes año”.

disreg_egatur16.xlsx → Fichero donde se describen los campos de los ficheros egatur.

Diseño_Frontur_OpenData.xlsx → Fichero donde se describen los campos de los ficheros frontur.



data_cleaning_02 → En este directorio están los scripts y ficheros correspondientes a la primera etapa de Data Engineering.

Existen varios ficheros csv con la información de las Comunidades Autónomas y de los Países, correspondientes a Egatur & Frontur.

Branch: master ▼ TFM_DataScience / data_cleaning_02 /	
 crysania Se modifica el nombre del fichero	
..	
 .Rhistory	reorganiza
 CCAA.csv	Se añaden
 CCAAE.csv	Se modific
 Egatur.csv	Actualizaci
 Frontur.csv	Se añaden
 Países.csv	Se añaden
 PaísesE.csv	Cambio no
 dataCleaning_Egatur.Rmd	Actualizaci
 dataCleaning_Egatur.html	Actualizaci
 dataCleaning_Frontur.Rmd	Se añaden
 dataCleaning_Frontur.html	Se añaden















data_exploring_03 → En este directorio están los scripts y ficheros correspondientes a la segunda etapa de Data Engineering, la exploración de los ficheros csv obtenidos anteriormente y se genera los nuevos csv que serán la entrada de los modelos de Machine Learning.

Branch: master ▼ TFM_DataScience / data_exploring_03 /	
crysania modificacion de la etapa exploring de frontur y generacion de los fic... ⋮	
..	
dataExploring_Egatur.Rmd	Revision de scripts para la memoria
dataExploring_Egatur.html	Revision de scripts para la memoria
dataExploring_Frontur.Rmd	modificacion de la etapa exploring de frontur
dataExploring_Frontur.html	modificacion de la etapa exploring de frontur
dsEgatur_Final.csv	generacion de ficheros para predicciones
dsFrontur_Final.csv	modificacion de la etapa exploring de frontur
egaturRFM.csv	se añade dos columnas para el fichero egatur
fronturGLM.csv	modificacion de la etapa exploring de frontur
gasto_fec_ccaa.csv	generacion de ficheros para predicciones
gasto_fec_pais.csv	generacion de ficheros para predicciones
viajeros_fec_ccaa.csv	modificacion de la etapa exploring de frontur
viajeros_fec_pais.csv	modificacion de la etapa exploring de frontur

data_prediction_04 → En este directorio están los scripts y ficheros correspondientes a la tercera etapa de Data Engineering, la predicción de la ocupación turística en España y la segmentación de los turistas por gasto.

Branch: master ▼ TFM_DataScience / data_prediction_04 /	
crysania predicciones series estacionales	
..	
1.Grafico_Clustering_Kmeans_3_CLUSTERS.png	Se añaden ficheros csv y clusters
2.Grafico_Clustering_Kmeans_6_CLUSTERS.png	Se añaden ficheros csv y clusters
RFM_EGATUR_CLUSTERS_6.csv	Modificacion CCAA para visualizacion del tableau
dataPrediction_Egatur.Rmd	modificacion y regeneracion
dataPrediction_Egatur.html	modificacion y regeneracion
dataPrediction_Frontur.Rmd	predicciones series estacionales
dataPrediction_Frontur.html	predicciones series estacionales

data_visualization_05 → En este directorio están los scripts y ficheros correspondientes a la última etapa de Data Engineering, visualización de la predicción de la ocupación turística en España y la segmentación de los turistas por gasto.

 crysania Subida informes finales		
..		
 CCAA.xlsx		Visualizaciones Clustering y Frontur
 Egatur_Clusters.pdf		Subida informes finales
 Egatur_Clusters.twb		Visualizaciones Clustering y Frontur
 Egatur_Clusters.twbx		Visualizaciones Clustering y Frontur
 España_Años.pdf		Subida informes finales
 Frontur_Viajeros.pdf		Subida informes finales
 Frontur_Viajeros.twb		Subida informes finales
 Frontur_Viajeros.twbx		Subida informes finales
 Mapa_Turisticos.pdf		Subida informes finales
 Países.xlsx		Visualizaciones Clustering y Frontur
 RFM_EGATUR_CLUSTERS_6.xlsx		Visualizaciones Clustering y Frontur
 Viajeros_Origen_Destino.pdf		Subida informes finales
 dsFrontur_Final.xlsx		Subida informes finales

Descripción de los datos de entrada

Egatur

TFM_DataScience/data_01/Egatur/disreg_egatur16.xlsx → Fichero donde se describen los campos de los ficheros egatur.

Id <chr>	Tipo_Viajero <fctr>	Via_Salida <fctr>	País <fctr>	Destino_CCAA <fctr>	Pernoctaciones <dbl>	Alojamiento <fctr>	Motivo <fctr>
0115EVA0000106	2	2	14	13	4	1	1
0115EVA0000142	2	2	14	13	18	3	3
0115EVA0000872	2	2	14	13	19	3	3
0115EVA0001326	2	2	14	9	13	3	3
0115EVA0001394	2	2	4	1	7	1	1
0115EVA0001395	2	2	4	1	12	1	1

Alojamiento <fctr>	Motivo <fctr>	Paquete_Turistico <fctr>	Gasto_Total <dbl>	Viajeros <dbl>	Mes <fctr>	Año <fctr>
1	1	6	1464.609	188.80330	1	2017
3	3	6	1690.454	96.94589	1	2017
3	3	6	1890.817	96.94589	1	2017
3	3	6	1671.538	87.97956	1	2017
1	1	6	1009.927	93.57565	1	2017
1	1	1	1602.032	75.96992	1	2017

Frontur

TFM_DataScience/data_01/Frontur/Diseño_Frontur_OpenData.xlsx → Fichero donde se describen los campos de los ficheros frontur.

Id <fctr>	Tipo_Viajero <int>	Via_Entrada <int>	País <int>	Destino_CCAA <int>	Alojamiento <int>	Motivo <int>	Paquete_Turistico <int>
20170102010177	2	2	13	13	8	4	6
20170101010177	2	2	13	13	8	4	6
20170100028497	2	2	13	13	2	1	6
20170101000001	3	2	13	5	0	1	1
20170100020431	2	2	13	5	8	1	6
20170100010177	2	2	13	13	8	4	6

Via_Entrada <int>	País <int>	Destino_CCAA <int>	Alojamiento <int>	Motivo <int>	Paquete_Turistico <int>	Pernoctaciones <int>	Viajeros <dbl>	Mes <int>	Año <int>
2	13	13	8	4	6	4	90.62923	1	17
2	13	13	8	4	6	4	90.62923	1	17
2	13	13	2	1	6	4	103.61073	1	17
2	13	5	0	1	1	1	238.05977	1	17
2	13	5	8	1	6	4	96.91650	1	17
2	13	13	8	4	6	4	90.62923	1	17

rows 14-13 of 12 columns

Metodología

A continuación, se detalla las etapas del Data Engineering, que se ha seguido en este TFM, para cada uno de los grupos de datos Egatur & Frontur.

Data_cleaning_02

En primer lugar, se produce la limpieza de datos, es decir, la decisión sobre qué datos se van a emplear en el análisis, usando criterios relativos a la relevancia para los objetivos, la calidad de los mismos o restricciones técnicas por las técnicas de análisis.

Esta selección a realizar se refiere tanto a los atributos o campos de los registros en los ficheros, como a los registros en sí.

***dataCleaning_Egatur.Rmd* →**

En este script se realiza la limpieza e integración de datos, se enfoca a la combinación de múltiples ficheros de registros, para crear nuevos, con el objetivo final de unir los datos sobre un mismo fichero Egatur.csv.

Los ficheros que se van a limpiar en este script son los relacionados con “Encuesta de gasto turístico. Egatur

Para el TFM se va a utilizar los datos contenidos en los ficheros ../data_01/Egatur/elevado_eg_mod_web_tur_XXXX.txt, se obvian los datos de los ficheros etapas_eg_mod_web_XXXX.txt

disreg_egatur16.xlsx → Fichero donde se describen los campos de los ficheros egatur.

***dataCleaning_Frontur.Rmd* →**

En este script se realiza la limpieza e integración de datos, se enfoca a la combinación de múltiples ficheros de registros, para crear nuevos, con el objetivo final de unir los datos sobre un mismo fichero Frontur.csv.

Los ficheros que se van a limpiar en este script son los relacionados con "Estadística de movimientos turísticos en frontera. Frontur"

Diseño_Frontur_OpenData.xlsx → Fichero donde se describen los campos de los ficheros frontur.

Data_exploring_03

En esta etapa, se va a hacer una exploración y transformación de los datos obtenidos del INE, para poder plantear una solución a partir de los mismos y adecuarlos al problema de modelado que se quiere resolver.

Se procede a partir del fichero csv generado en la etapa anterior "data_cleaning_02", a la generación de nuevos registros o valores transformados de atributos existentes, si fuera necesario, en función de los requerimientos para preparar la entrada a las herramientas de modelado.

Cabe recordar que los datos son la representación simbólica (numérica, alfabética,...) de un atributo o variable cuantitativa o cualitativa.

Se procede a dar formato a los datos, estas transformaciones se refieren a modificaciones sintácticas que se hacen sobre ellos, sin alterar su significado pero que pueden ser requeridas por la herramienta de modelado a utilizar.

Puede que haya requisitos en el orden de los atributos o que los registros estén ordenados según el atributo resultado, mediante la evaluación de la información disponible y refinar la pregunta inicial para evitar resultados ambiguos, sesgos o detectar la necesidad de recopilar nuevos datos.

El período de tiempo que se va a utilizar en el TFM, será desde Octubre del 2015 hasta Abril del 2017 (19 meses).

Esta tarea incluiría, si fuera necesario, la inserción de valores por defecto adecuados o el uso de modelado para estimar los valores ausentes (missing values).

El tratamiento de los datos es fundamental para la correcta interpretación de un análisis, los valores perdidos son aquellos que simplemente se ha perdido algún dato y no sabemos qué valor toma, se representa por el código: NA (Not available).

Hay muchas funciones para identificar estos valores, la más usada es `is.na()`.

dataExploring_Egatur.Rmd →

La variable que identifica de manera única cada registro es `A0_1`.

Para estimar el gasto turístico debe utilizarse la variable 'gastototal' multiplicada por el factor de elevación de cada registro ('Factor_Egatur').

El gasto medio por persona, se obtendrá como el cociente del gasto turístico entre los turistas, calculados sumando la variable 'Factor_Egatur'.

La estimación de las pernoctaciones, resultante de multiplicar la variable 'A13' por 'Factor_Egatur', se utiliza como denominador del gasto medio diario, siendo el numerador el gasto turístico.

Además, se utiliza en la estimación de la duración media del viaje, dividiendo los turistas entre las pernoctaciones estimadas.

El fichero PaísesE.csv se relacionan los códigos de cada país, con el nombre del País de residencia habitual del turista.

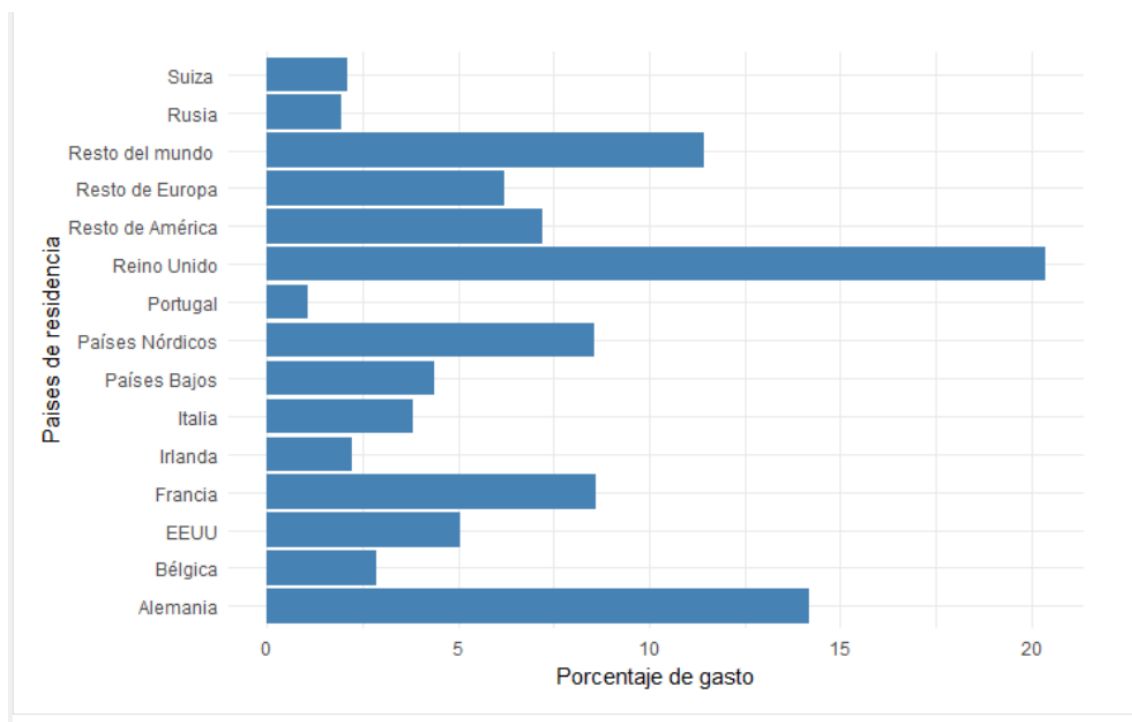
El fichero CCAAE.csv contiene los códigos de las CCAA y su nombre correspondiente como destino principal del viaje.

Se formatea los datos a los tipos adecuados:

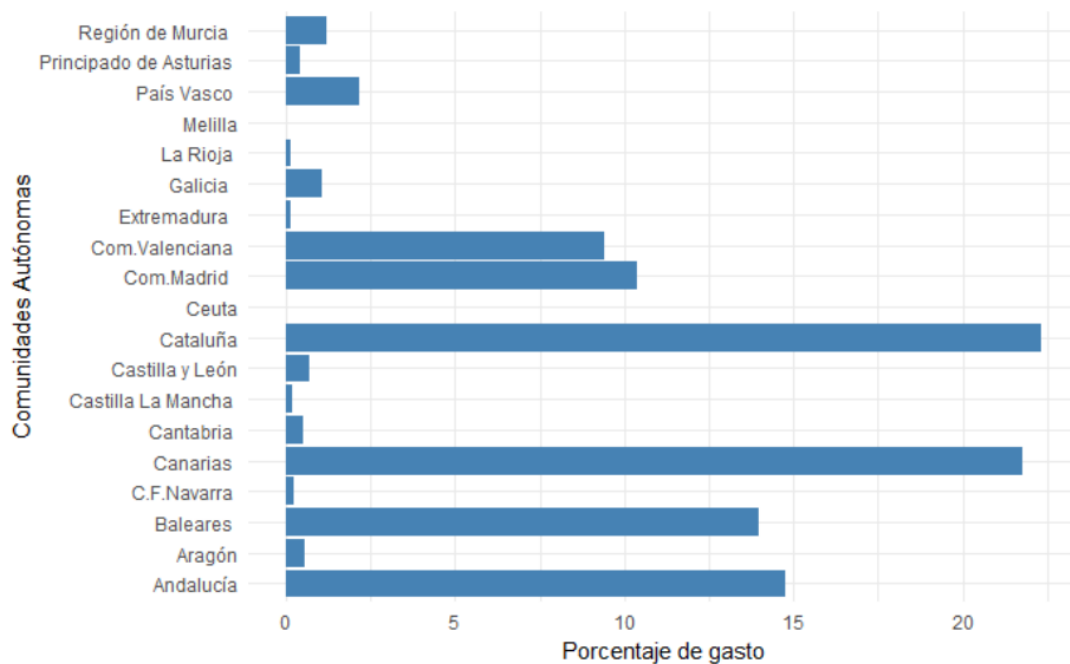
1. Tipos numéricos (Viajeros, pernoctaciones y gasto_total)
2. Tipos alfanuméricos (Id)
3. Tipos Categóricos

Dentro de este script se realiza el siguiente *Análisis Descriptivo* de los datos:

- Los turistas que vienen a España durante el periodo Octubre 2015 a Abril 2017 se gastan **111.760.512.068** euros.
- Se calcula el número total de turistas que vienen a España según los datos de Egatur **109.072.379** de turistas.
- El gasto medio por persona, se obtendrá como el cociente del gasto turístico entre los turistas, calculados sumando la variable 'Viajeros' --> **1024.645** euros.
- Se detecta que los turistas que más gastan son los ingleses (20%), a continuación, los alemanes (14%) y en tercer lugar los extranjeros del resto del mundo (11%).



- **Cataluña** es el principal destino principal de los turistas con un gasto realizado de 24.949.885.499 euros durante el periodo de estudio, lo que representa el 22 % del total, Canarias (21%) y Andalucía (14%) son las siguientes Comunidades Autónomas con más gasto realizado en España.



- De los datos se obtiene que el mes de **Julio** de **2016** hubo el mayor gasto turístico de los viajeros extranjeros en Baleares y en segunda posición Cataluña.
- Por el contrario, en Enero del 2016, fue el periodo con menos gasto realizado, eligiendo Ceuta como destino.
- Los ingleses son los extranjeros que más gastaron en España, durante el mes de Agosto del 2016 -> **2.312.129.733 euros**.
- En Abril del 2016 se ha detectado que el gasto más bajo lo realizaron los portugueses --> **48.219.280 euros**.

***dataExploring_Frontur.Rmd* →**

En el fichero Frontur.csv, cada registro representa a un viajero que finaliza su viaje por España en el mes de referencia.

La variable que identifica de manera única cada registro es Id. Sumando el factor de elevación de cada registro ('Factor') se obtiene la estimación del número de viajeros, que podrá desglosarse en función del resto de variables de clasificación del fichero: turistas-excursionistas, destino principal, país de residencia, vía de acceso y tipo de alojamiento.

El fichero Paises.csv se relacionan los códigos de cada país, con el nombre del País de residencia habitual del turista.

El fichero CCAA.csv contiene los códigos de las CCAA y su nombre correspondiente como destino principal del viaje.

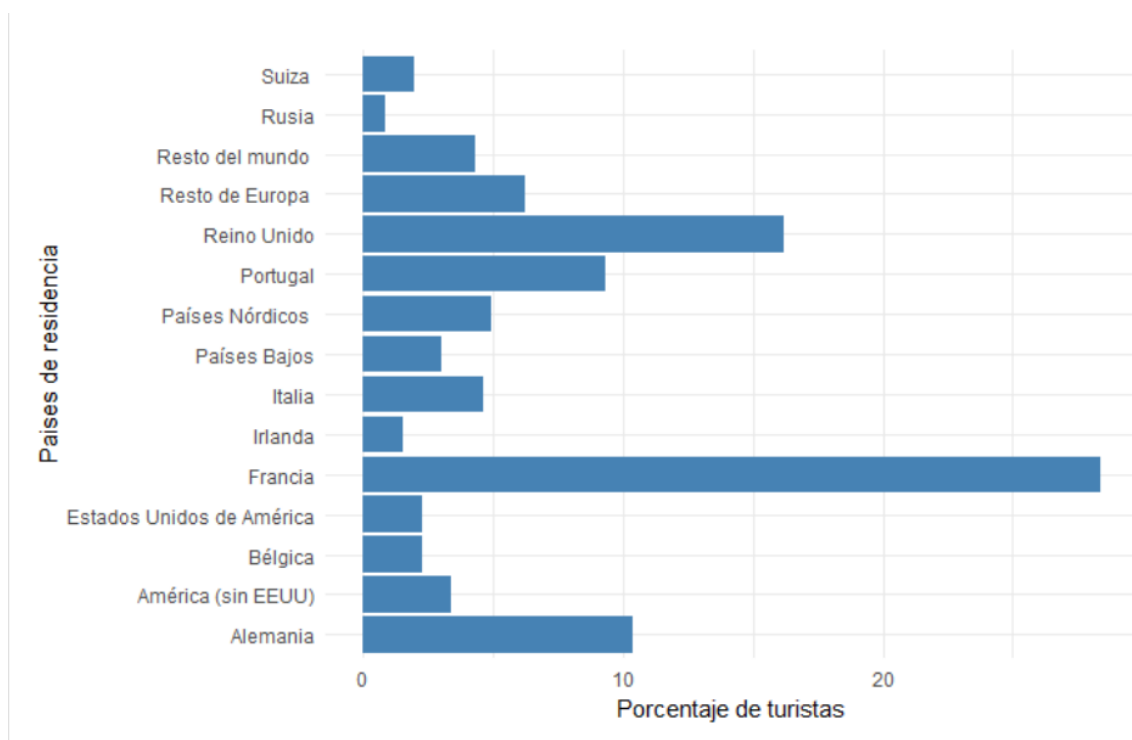
Se formatea los datos a los tipos adecuados:

1. Tipos numéricos (Viajeros)

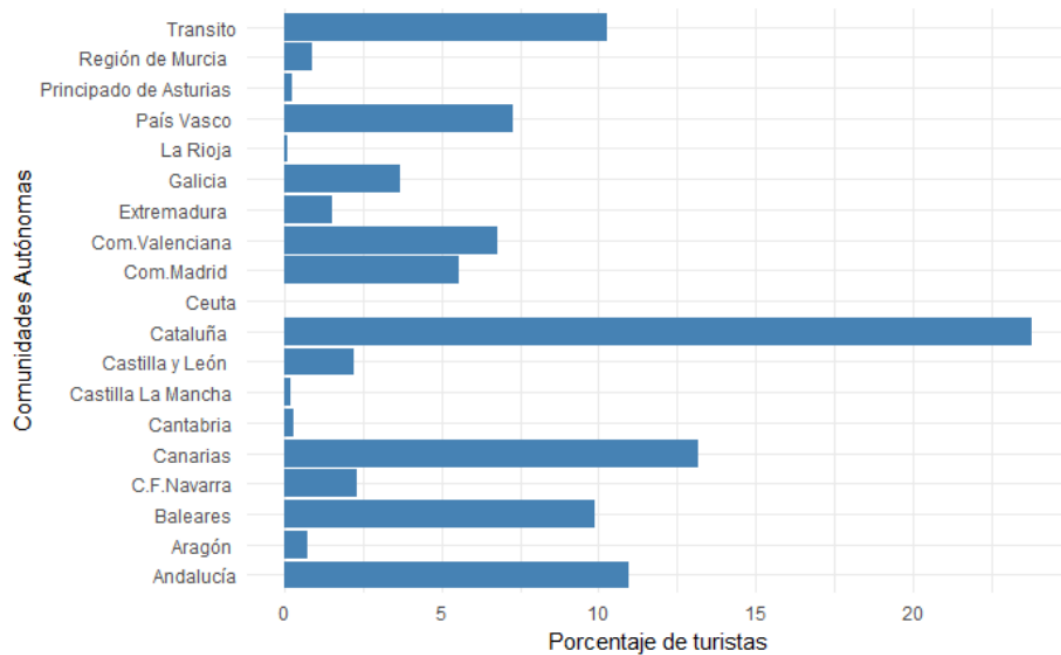
2. Tipos Categóricos (Tipo_Viajero, Via_Entrada, Pais, Destino_CCAA, Alojamiento, Motivo, Paquete_Turistico, Pernoctaciones y Mes).

Dentro de este script se realiza el siguiente *Análisis Descriptivo* de los datos:

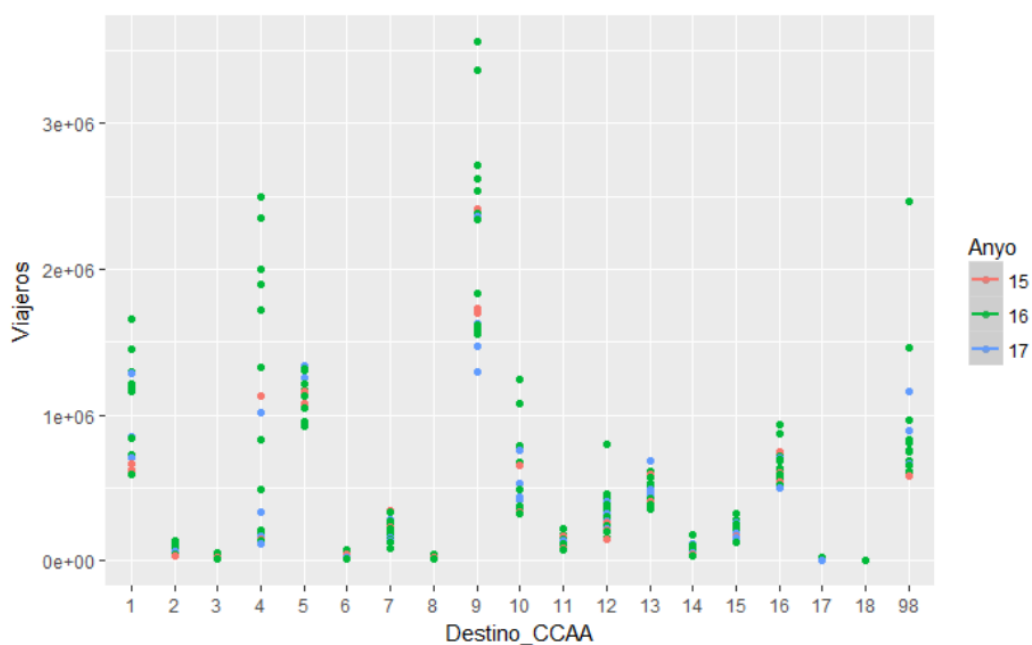
- Se calcula el número total de turistas que vienen a España, durante el periodo Octubre 2015 a Abril 2017, según los datos de Frontur **169.385.169** de turistas.
- **Francia** es el principal país de residencia con 48.070.653 turistas durante el periodo de estudio, lo que representa el 28 % del total, Reino Unido (16%) y Alemania (10%) son los siguientes países de residencia con más turistas que visitan España.



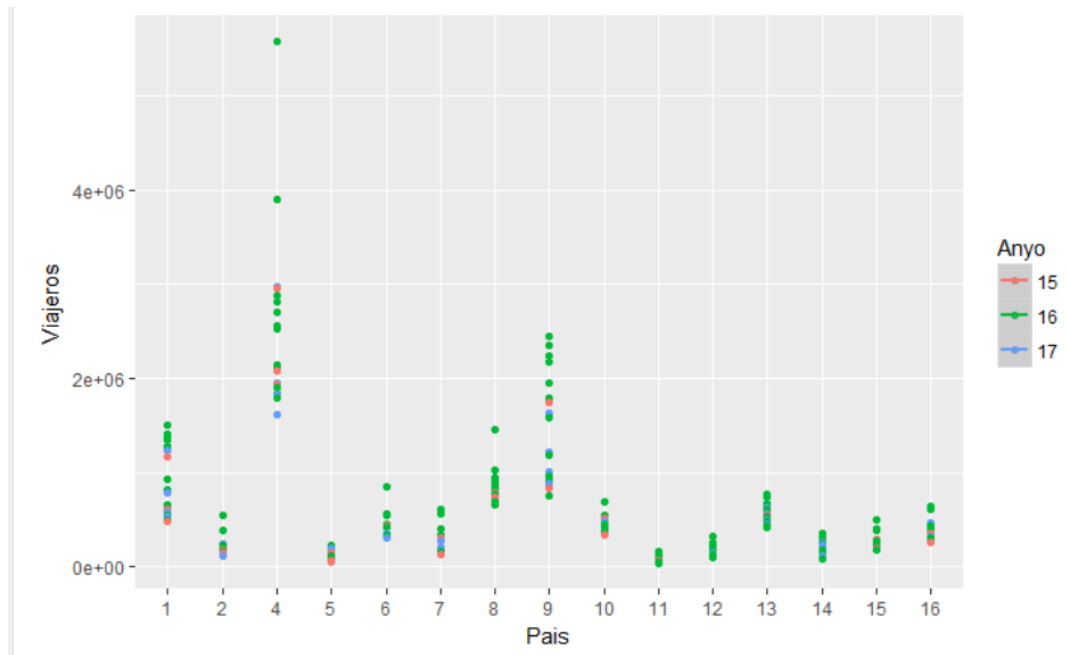
- **Cataluña** es el principal destino principal de los turistas con 40.294.779,77 turistas durante el periodo de estudio, lo que representa el 24 % del total, Canarias (13%) y Andalucía (11%) son las siguientes Comunidades Autónomas con más turistas que visitan España.



- De los datos se obtiene que el mes de **Julio de 2016** hubo la mayor cantidad de viajeros que visitaron **Cataluña** como principal destino turístico nacional.



- Los **franceses** son los extranjeros que más visitaron España durante el mes de **Agosto** del **2016**.



Los algoritmos de aprendizaje automático emplean métodos de cálculo para "aprender" información directamente a partir de los datos sin asumir una ecuación predeterminada como modelo.

***dataPrediction_Egatur.Rmd* →**

En este script se desarrolla el procedimiento para obtener la respuesta a uno de los objetivos de este TFM:

- Segmentación de los turistas según características similares de gasto.

Para ello se utilizará el *Modelo RFM* (Recencia, Frecuencia y Valor Monetario), que consiste en agrupar los registros en distintos clústers o segmentos, de acuerdo a criterios económicos, comportamentales, o de negocio.

La idea es crear distintos segmentos en base a cada una de las 3 variables que hacen al modelo RFM, siguiendo el orden de sus siglas:

- Recencia: Días transcurridos entre la última visita y el final del periodo analizado.
- Frecuencia: Número de visitas realizadas en dicho periodo.
- Monetización: Gastos turísticos realizadas en este periodo.

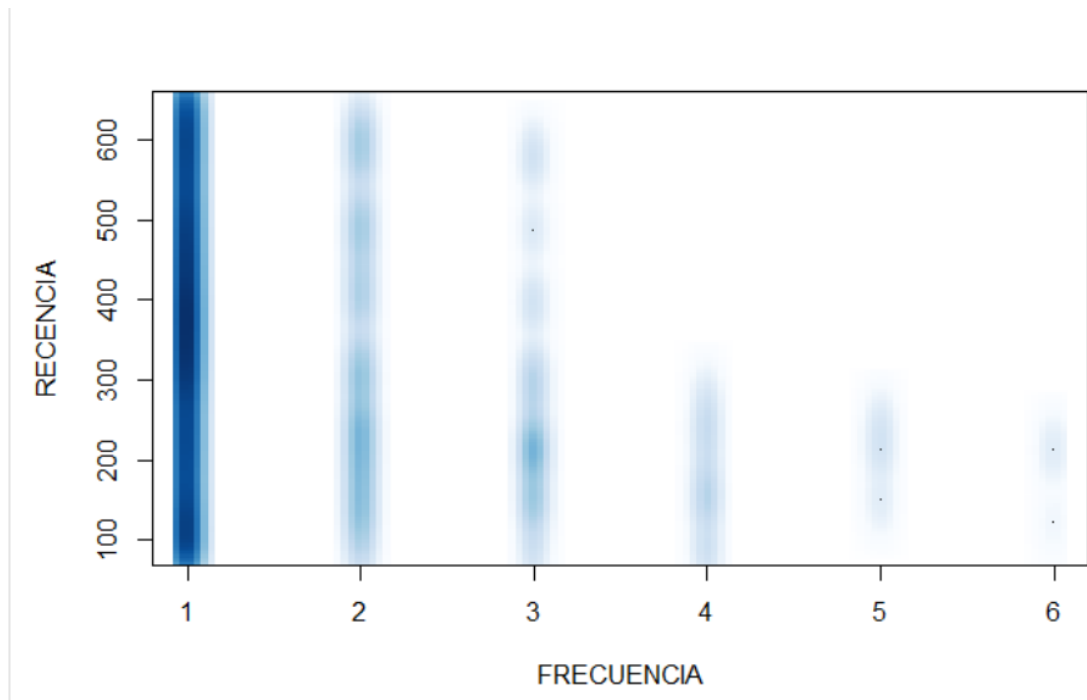
El resultado de la triple segmentación dará un grupo ganador que es el que definirá cuáles son los mejores turistas, es decir, "los turistas más propensos a gastar son aquellos que nos han visitado más recientemente, con más frecuencia y gastan más dinero".

Utilizar este modelo predictivo, consigue mejores resultados para las campañas de marketing por 2 motivos principales:

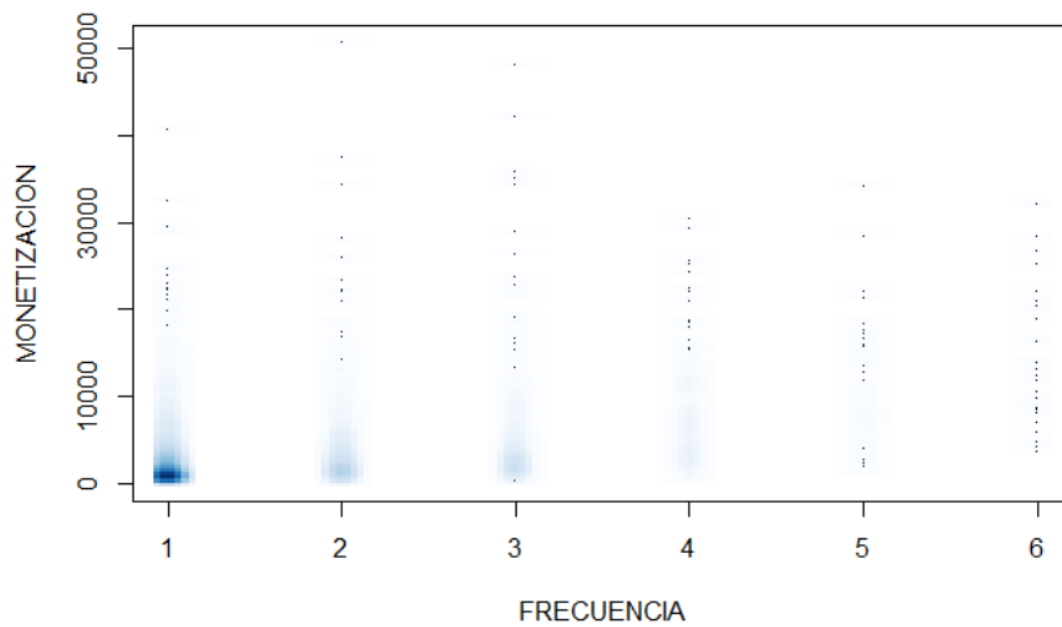
- Permite dirigirse a cada grupo de turistas con mensajes y ofertas diferentes para optimizar los programas de comunicación para cada segmento y maximizar los resultados económicos de las campañas.
- Conocimiento holístico de los turistas.

Se realizan las siguientes gráficas para mostrar la densidad de los turistas a través del Modelo RFM.

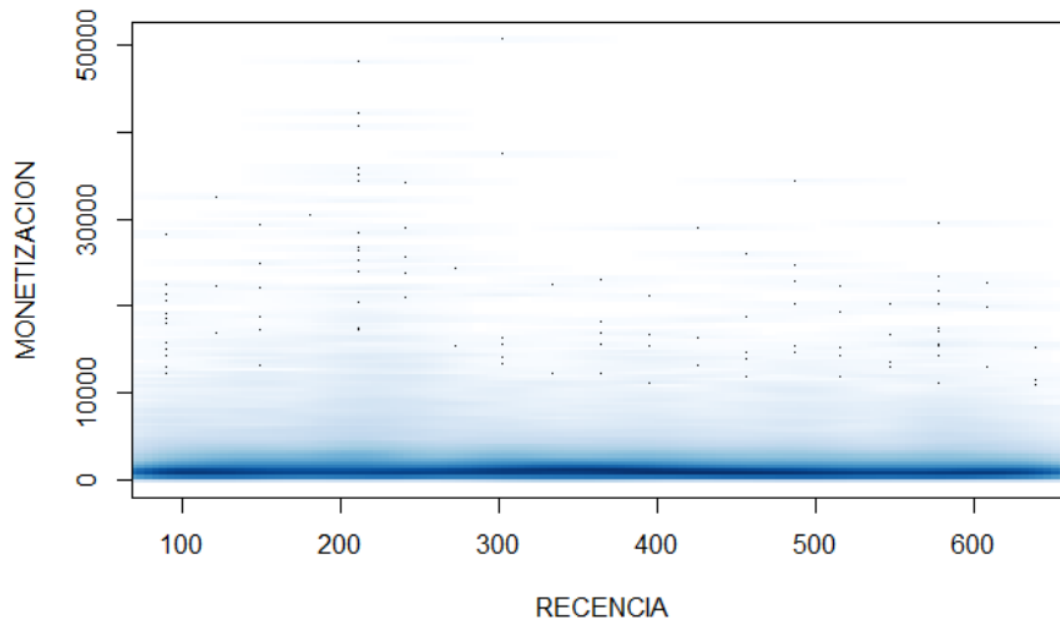
1. Gráfico de Densidad Modelo RFM_EGATUR\$FRECUENCIA - RFM_EGATUR\$RECENCIA



2. Gráfico de Densidad Modelo RFM_EGATUR\$FRECUENCIA - RFM_EGATUR\$MONETIZACION



3. Gráfico de Densidad Modelo RFM_EGATUR\$RECENCIA - RFM_EGATUR\$MONETIZACION



Método Elbow

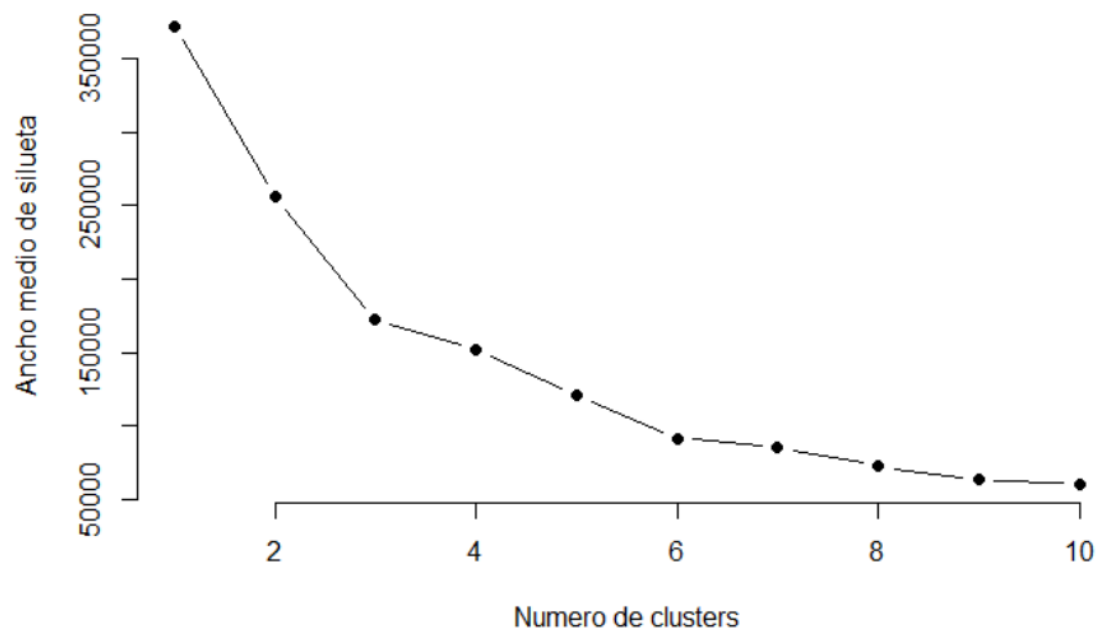
Uno de los objetivos de este TFM, es clasificar en grupos a los turistas según el gasto realizado, para implementar el cálculo óptimo en la selección del número de Clusters, se ha utilizado el método Elbow (método del codo).

Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de Clusters (desde 1 a N Clusters), siendo la inercia la suma de las distancias al cuadrado de cada objeto del Cluster a su centroide.

Una vez obtenidos los valores de la inercia tras aplicar el K-means de 1 a N Clusters, se representa en una gráfica lineal la inercia respecto del número de Clusters.

En esta gráfica se debería de apreciar un cambio brusco en la evolución de la inercia, teniendo la línea representada una forma similar a la de un brazo y su codo.

El punto en el que se observa ese cambio brusco en la inercia, dirá el número óptimo de Clusters a seleccionar para ese dataset, dicho de otra manera, el punto que representaría al codo del brazo será el número óptimo de Clusters para ese dataset.



- ***dataPrediction_Frontur.Rmd*** →

En este script se desarrolla el procedimiento para obtener la respuesta a uno de los objetivos de este TFM:

- Análisis predictivo de la actividad turística en España.

Para ello se van a utilizar las técnicas de Análisis de Series Temporales:

- Gráficos temporales.
- Series estacionarias y no estacionarias.
- Predicción.

Una serie temporal es una sucesión de observaciones de una variable tomadas en varios instantes de tiempo. Por lo tanto, interesa estudiar los cambios en esa variable con respecto al tiempo, es decir, predecir sus valores futuros.

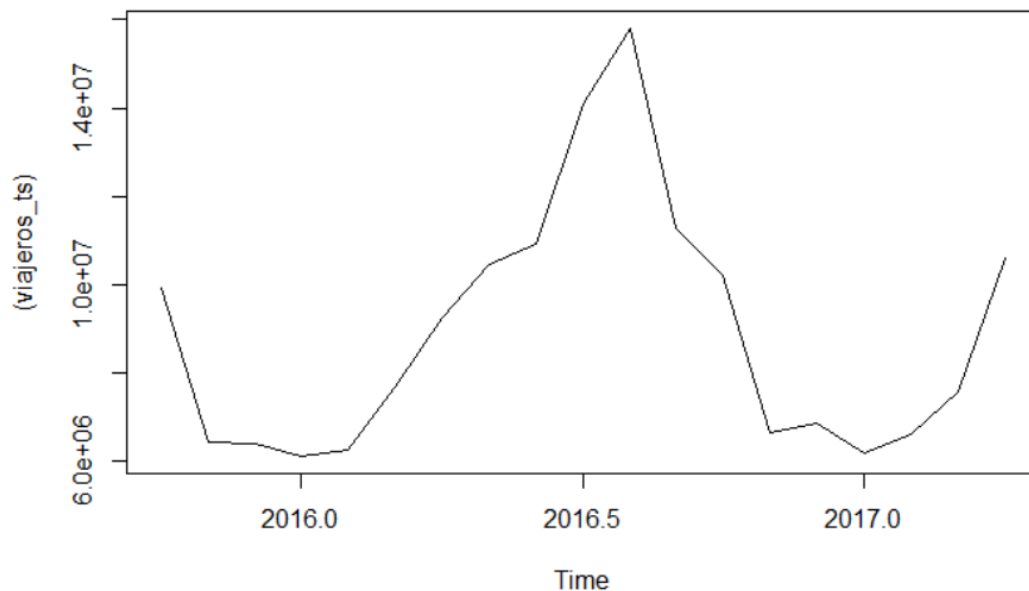
Clasificación de series temporales:

- Una serie es estacionaria si la media y la variabilidad se mantienen constantes a lo largo del tiempo.

- Una serie es no estacionaria si la media y/o la variabilidad cambian a lo largo del tiempo.
- Series no estacionarias pueden mostrar cambios de varianza.
- Series no estacionarias pueden mostrar una tendencia, es decir que la media crece o baja a lo largo del tiempo.
- Además, pueden presentar efectos estacionales, es decir que el comportamiento de la serie es parecido en ciertos tiempos periódicos en el tiempo.

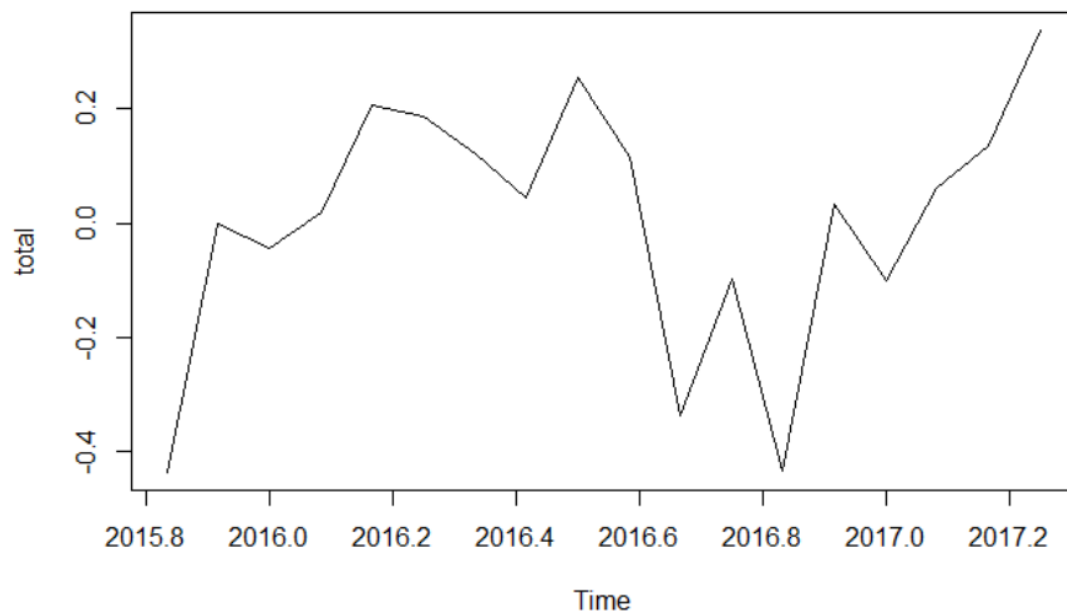
En este script se han realizado los siguientes pasos:

- Se realiza una agrupación por mes y año de la suma de los Viajeros.
- Se ordena los resultados por año.
- Se genera la serie temporal definiendo el comienzo 2015-10 y periodicidad 12 meses.
- Se dibuja la serie temporal.

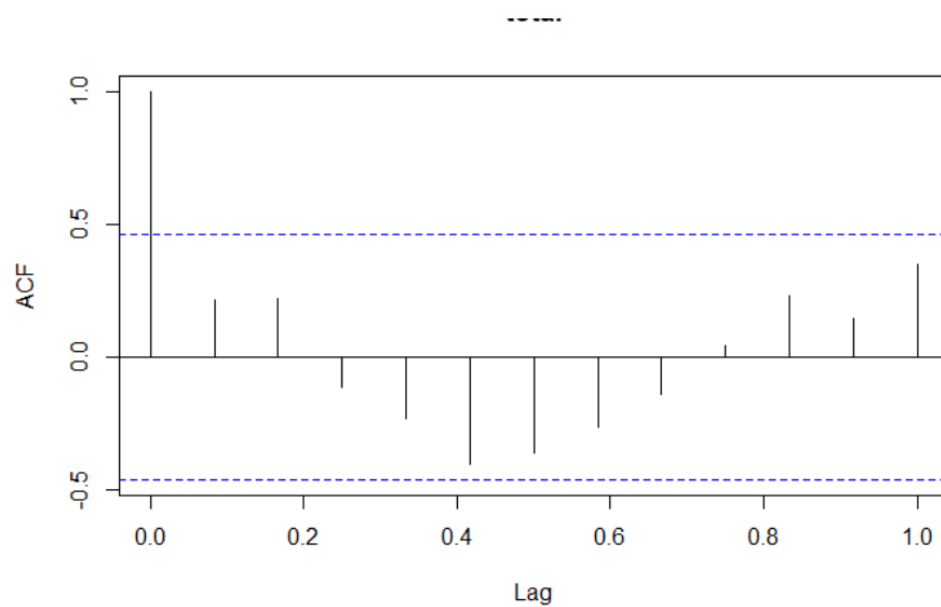


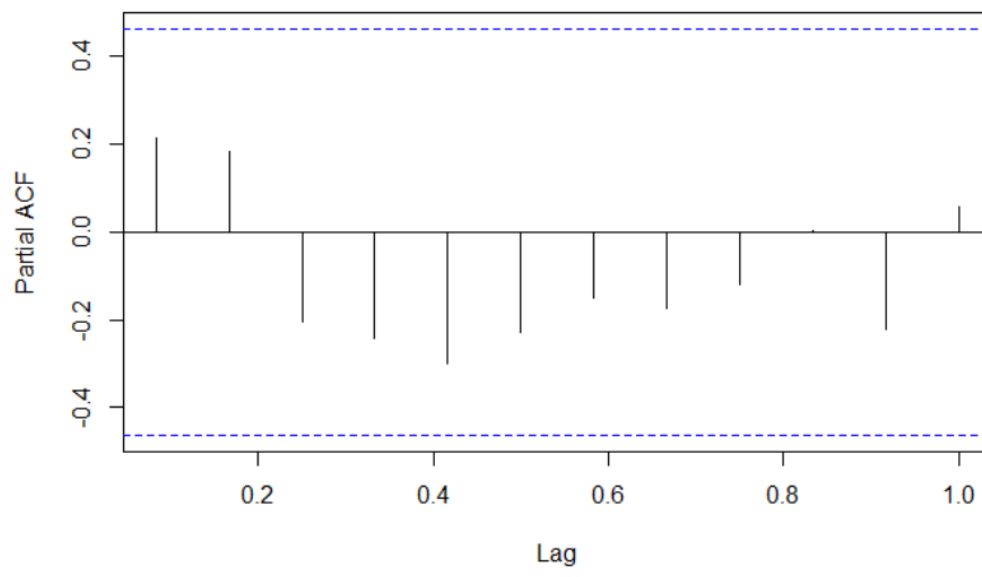
- Se considera que los datos no presentan forma creciente(tendencia), pero por el contrario si existe influencia en un periodo de tiempo por lo que se puede intuir cierta estacionalidad.
- No aparecen outliers (observaciones extrañas o discordantes).
- Se realiza la predicción, mediante la observación de los valores de la serie, se pretende normalmente predecir el futuro.
- Estos modelos pueden estimarse fácilmente en R mediante la función ARIMA().
- En este TFM se realiza la predicción de la Serie Temporal utilizando el Modelo ARIMA, para realizar predicciones se usa predict().
- No se ha usado forecast() porque daba error por no tener más de dos periodos.
- A continuación, se dibuja la gráfica de predicción del número de turistas en los dos próximos años, con un incremento significativo.

- La serie es obviamente no estacionaria y la serie diferenciada tampoco, como puede observarse en el siguiente gráfico:



- Los correlogramas para la serie diferenciada:





Resumen del resultado

El TFM ha permitido obtener información sobre los siguientes aspectos:

- Origen principal de los visitantes.
- Origen de los turistas que optan por cada una de las Comunidades Autónomas.
- Zonas donde prefieren alojarse los visitantes extranjeros.
- Gasto medio diario y gasto acumulado a lo largo de toda la estancia.

Egatur → Análisis RFM

Este nos ha permitido estudiar detalladamente el gasto de los turistas que recibe España desde Octubre de 2015 hasta Abril de 2017 y segmentarlos en grupos específicos con unas características similares.

El objetivo era hacer una segmentación de los turistas según características de gasto similares, para poder definir acciones de marketing específicas para cada tipo de turista, acciones publicitarias, así como lanzamiento de promociones especiales dirigidas a un turista objetivo.

La idea era crear distintos segmentos en base a cada una de las 3 variables que hacen al modelo RFM, siguiendo el orden de sus siglas:

Recencia: Días transcurridos entre la última visita y el final del periodo analizado (01-07-2017)

Frecuencia: Número de visitas realizadas en dicho periodo.

Monetización: suma de los gastos turísticos realizadas en este periodo.

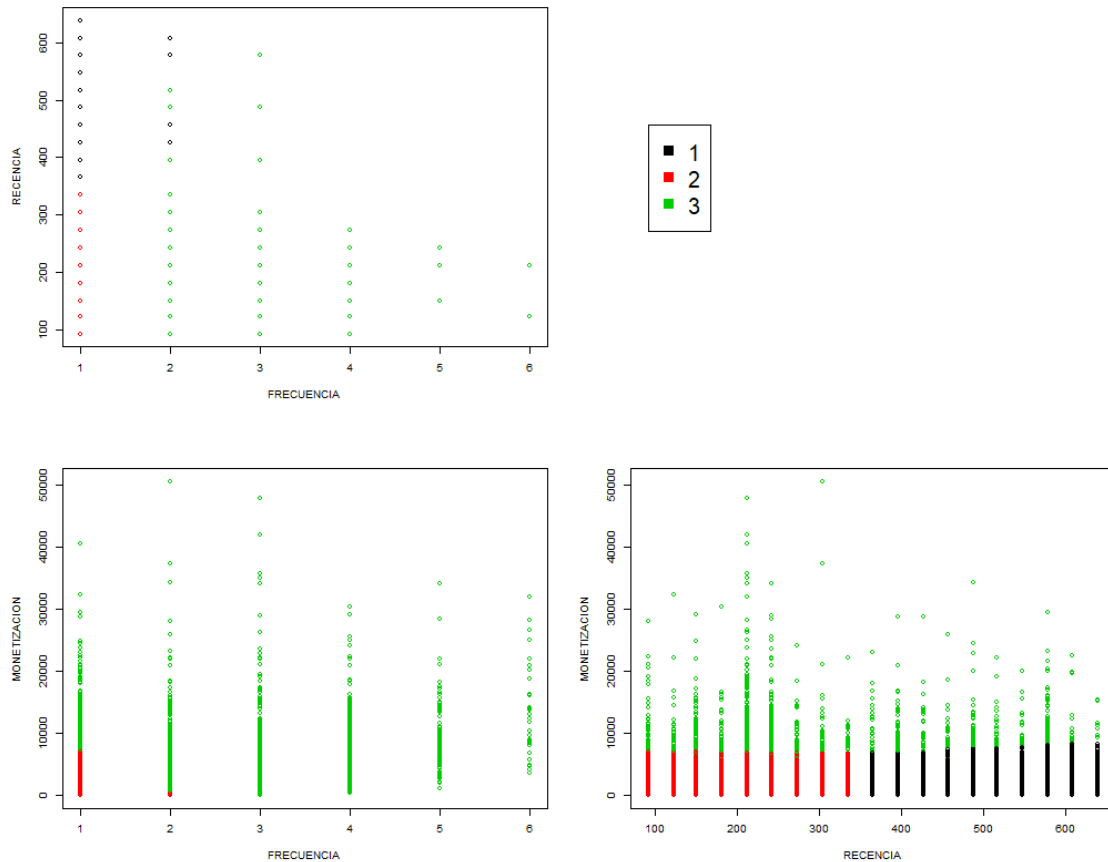
Existen diferencias entre los grupos obtenidos según el número de Clusters, se ha querido realizar un análisis más detallado de cada perfil de los turistas y por ello se ha realizado una comparativa de resultados de 3 y 6 Clusters.

3 Clusters

Existe el grupo 3 definido como los "turistas vips", son aquellos que han realizan un gasto alto en España, que han visitado España recientemente y con mayor frecuencia.

Los grupos 1 y 2 son los más numerosos, tienen características similares en cuanto al gasto realizado y la frecuencia, pero por el contrario el grupo 1 lleva ms tiempo sin visitar España.

Clusterización kmeans de clientes mediante Modelo RFM

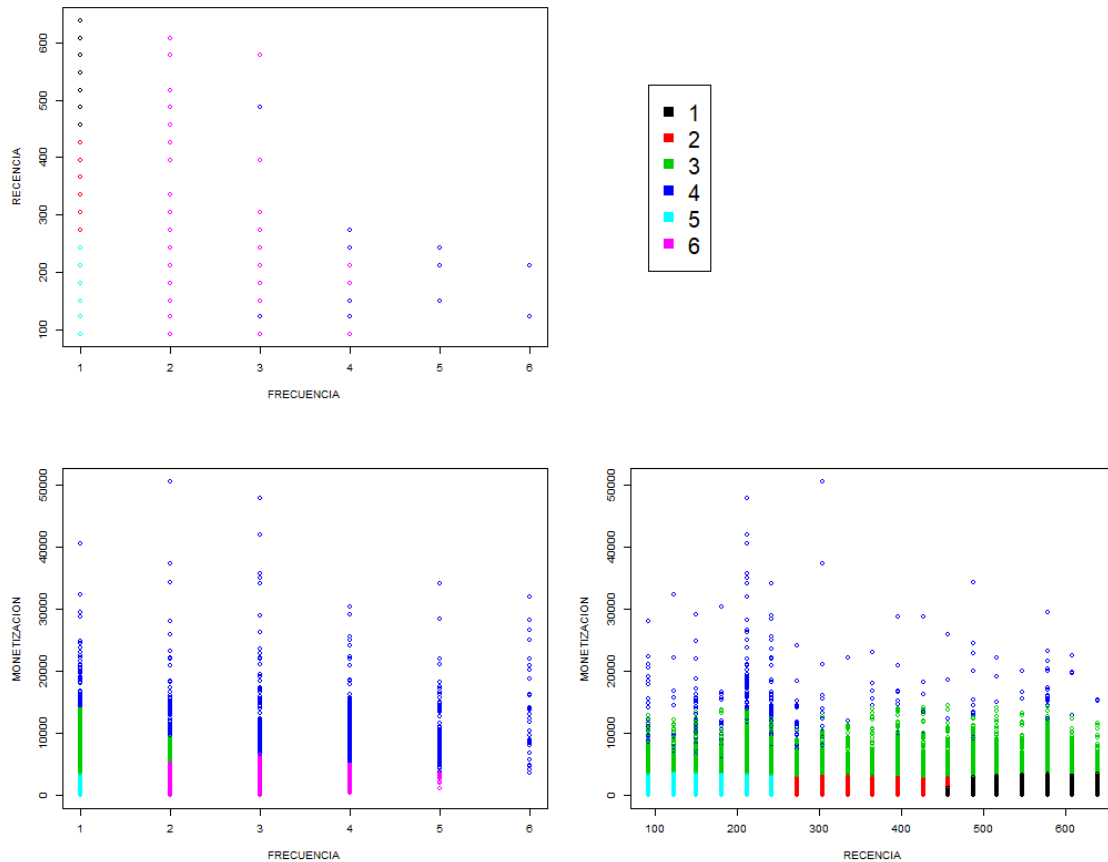


6 Clusters

Del resultado de la segmentación en los 6 Clusters, se observa dos grupos (4 y 6) definidos como mejores turistas, es decir, “los turistas más propensos a gastar son aquellos que nos han visitado más recientemente, con más frecuencia y gastan más dinero”, es decir, recencia baja, frecuencia alta y monetización alta, pero por el contrario son los menos numerosos.

En cambio, el grupo 1 es objetivo de campañas de marketing, promociones u ofertas puesto que el gasto es muy bajo, la frecuencia también y ha pasado bastante tiempo sin visitar España.

Clusterización kmeans de clientes mediante Modelo RFM



Frontur → Series temporales

Para conseguir uno de los objetivos del TFM, realizar un análisis predictivo de la actividad turística en España, para predecir el número de turistas que decidan viajar a un lugar determinado en España, se ha utilizado el método de predicción de Serie Temporal.

Se realiza la predicción de la Serie Temporal utilizando el Modelo ARIMA, para realizar predicciones se usa `predict()`.

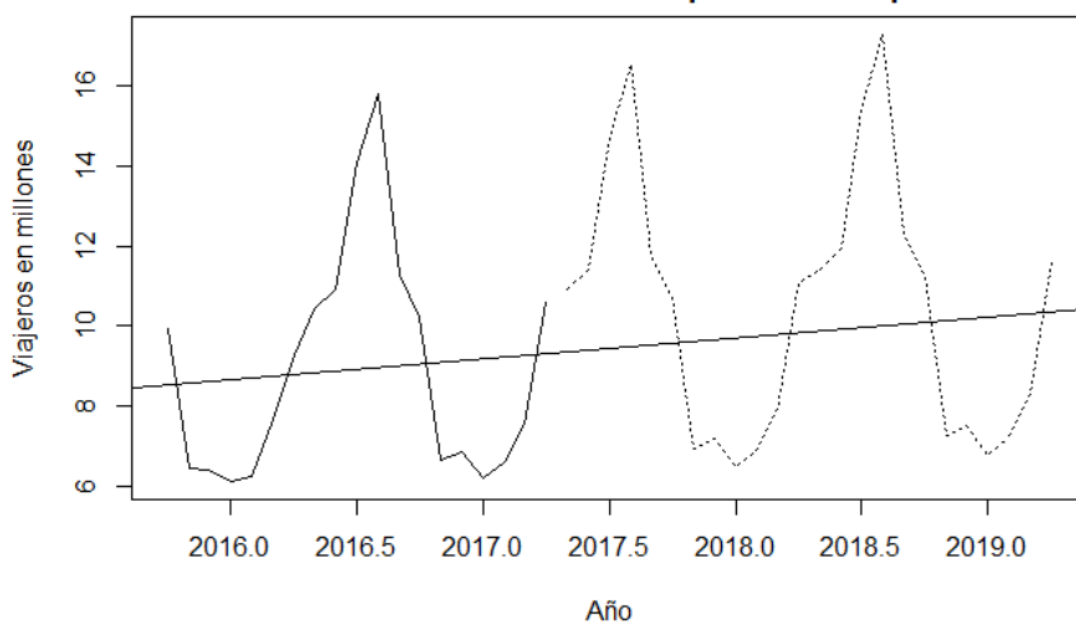
No se ha usado `forecast()` porque se obtiene un error, por no tener más de dos periodos.

Se observa una tendencia a incrementar el número de viajeros que visitarán España, sobre todo durante la estacionalidad del verano.

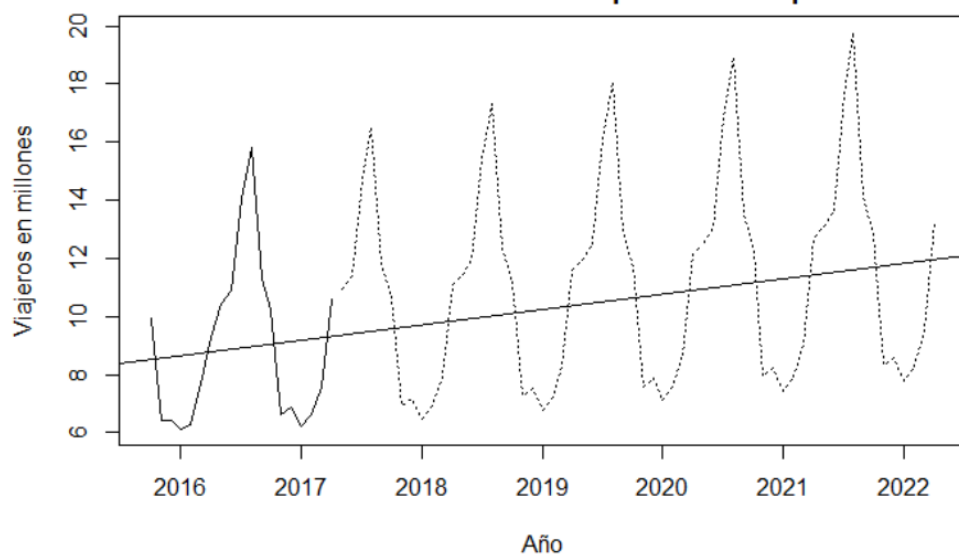
Se denomina tendencia de una serie temporal, a su comportamiento o movimiento a largo plazo. Por ejemplo, las gráficas que aparecen a continuación, muestran la tendencia en el sector turístico español en los próximos dos y cinco años.

La línea recta, en ambas gráficas, representa la tendencia (creciente).

Evolución del nº de turistas que visitan España



Evolución del nº de turistas que visitan España



Visualización Dashboard

Data_visualization_05

Para realizar el análisis de los datos y visualizar los resultados, se realiza un dashboard interactivo, la herramienta seleccionada es Tableau (debe estar instalado previamente).

Los datos a visualizar, parten de la exportación de los mismos en ficheros con formato csv, pero por problemas de detección en Tableau los “.” en los decimales, se ha decidido generar.xlsx cambiando las cantidades decimales a “,” y generar nuevos ficheros que serán la entrada de Tableau.

EGATUR

Egatur_Clusters.twbx →

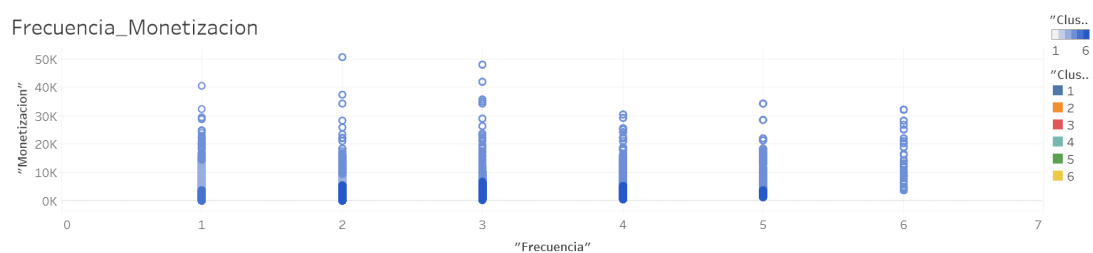
En este apartado se procede a analizar los datos generados en un dashboard interactivo, para ello se parte de un fichero csv con los datos calculados en el Modelo RFM utilizado, mostrando el Clúster al que pertenece el turista y el país de procedencia.

Se añade una variable al dataset RFM_EGATUR, con el número de Cluster de cada instancia, el país de origen y el destino CCAA.

Se va a utilizar el fichero RFM_EGATUR_CLUSTERS_6.xlsx para visualizar los datos en Tableau y generar Dashboards no cargaba correctamente los valores.

Egatur_Clustering

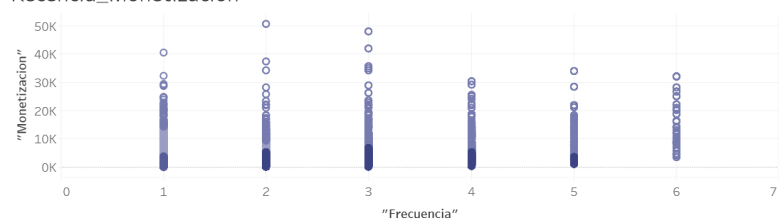
Frecuencia_Monetizacion



Clusters_Monetiz-
acion



Recencia_Monetizacion



FRONTUR

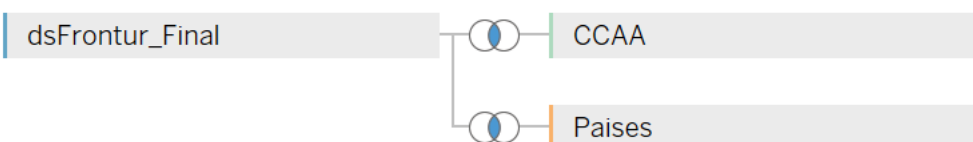
Frontur_Viajeros.twb →

Se va a utilizar el fichero dsFrontur_Final.xlsx para visualizar los datos en Tableau y generar Dashboards no cargaba correctamente los valores.

Se ha creado en Tableau un campo calculado Fecha, es el resultado de la concatenación 01 + Mes + Anyo.

También se realiza un join con los ficheros Países.xlsx y CCAA.xlsx, para obtener los nombres de cada uno.

dsFrontur_Final (dsFrontur_Final)



Relación Viajeros por Origen - Destino

Viajeros por CCAA



Nombre

- Alemania
- América
- Bélgica
- Estados Unid..
- Francia
- Irlanda
- Italia
- Países Bajos
- Países Nórdic..
- Portugal
- Reino Unido
- Resto de Eur..
- Resto del mu..
- Rusia

Viajeros por Origen



Ccaa

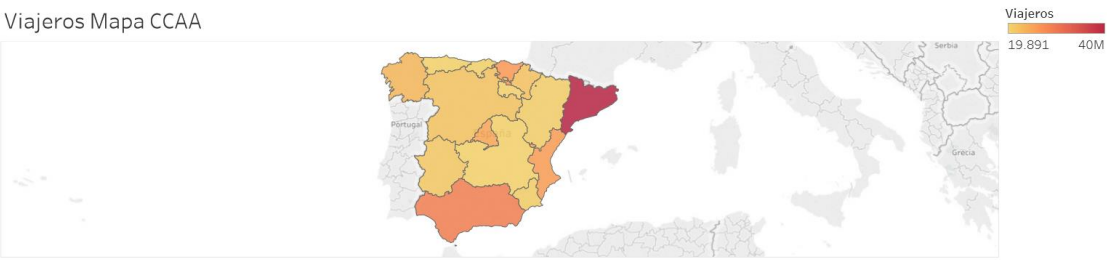
- Andalucía
- Aragón
- Baleares
- Canarias
- Cantabria
- Castilla La M..
- Castilla y Leon
- Cataluña
- Ceuta

Viajeros Origen - Destino



Mapas Turisticos

Viajeros Mapa CCAA

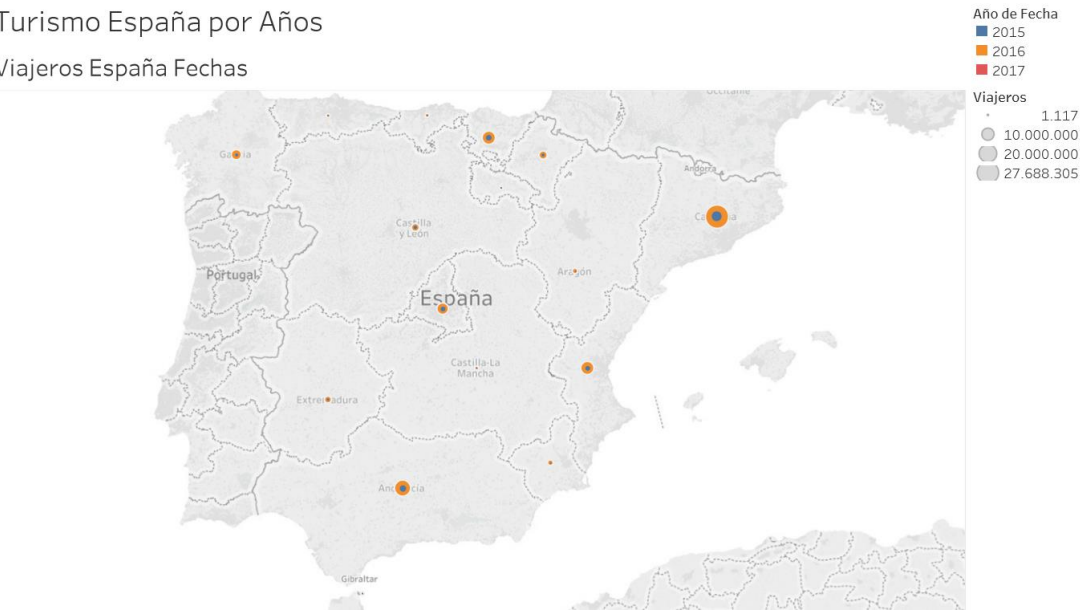


Viajeros Mapa Origen



Turismo España por Años

Viajeros España Fechas



Recomendaciones Tácticas & Estratégicas

A partir de las conclusiones que ofrece el análisis de datos realizado en este TFM, se concluye el estudio con una serie de recomendaciones tácticas y estratégicas dirigidas a los gestores hoteleros.

Estas recomendaciones se enfocan a:

- Aumentar la captación de clientes y determinar en qué países es recomendable focalizar la acción comercial.
- Determinar las Comunidades Autónomas donde se realizan más transacciones comerciales, para emprender nuevos negocios turísticos.
- Garantizar un producto atractivo y adaptado a las verdaderas necesidades de los turistas, por ejemplo, realizar una oferta complementaria o duración óptima de los paquetes de estancias, según nacionalidades.
- Dado el volumen de turistas, realizar acciones de captación en países con menor número de visitantes.
- Ubicación/expansión: áreas de interés en función de las nacionalidades.
- Área tecnológica: campañas SEM / SEO, idiomas en los que la web debería estar traducida y presencia en webs especializadas o intermediarios específicos.