Análisis del Gasto Medio por Turista en España el 2019

Nils Leonardo Vargas Berzina

05 Enero 2022

Periodo: 2019
Tutora de Seguimiento: Lucía Inglada Pérez

Base de datos: elevado_eg_mod_web_tur_2019

Índice

1. Objetivos

Con la realización de este estudio pretendo responder a varias cuestiones con el objetivo de analizar la influencia del turismo en nuestro país durante el año 2019.

Por lo que me gustaría dar respuesta a las siguientes preguntas:

- **¿El gasto medio diario de un turista es mayor que 200 euros?**
- **¿Son los turistas alemanes mejores que los británicos en términos de gasto medio?**
- **¿El gasto medio en agosto es mayor a la media de diciembre?**
- **¿El gasto medio es mayor en la Comunidad de Madrid que en Cataluña?**
- **¿El gasto total es mayor en las Illes Balears o en la Comunitat Valenciana?**

Me apoyaré en los test de diferencia de medias y contraste de varianzas, y analizaré como se comportan los diferentes estadísticos en muestras grandes.

Con este estudio pretendo observar el comportamiento de los turistas según su "gasto medio diario" y comparar el gasto entre comunidades, primero las dos grandes ciudades españolas teniendo en cuenta que la afluencia de turistas es mucho mayor en Cataluña y dos comunidades con una afluencia de turistas similar (Illes Balears y la Comunitat Valenciana); y segundo comparar los gastos medios en dos meses con importante flujo de turismo como lo son Agosto y Diciembre.

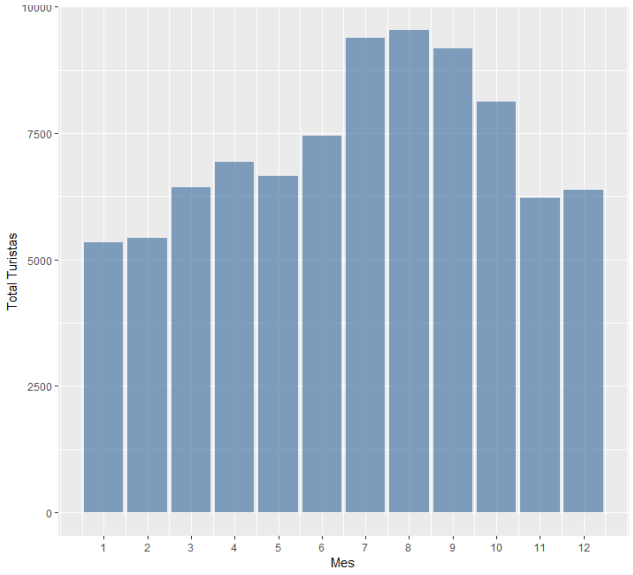


Gráfico 1.1 Afluencia de turistas respecto al mes

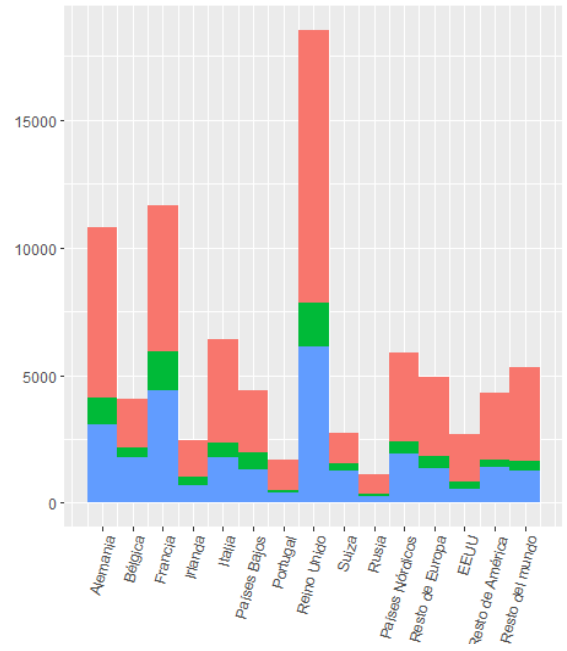


Gráfico 1.2 Afluencia de turistas respecto al país

2. Introducción

La presente investigación proviene de un estudio realizado por el ***INE***, el cual difunde los resultaos obtenidos de las encuestas desde 2015 en adelante para este sector.
La propuesta de estudio se genera en colaboración con con la encuesta de gasto turístico ***EGATUR*** y la estadística realizada por Movimientos Turísticos en Frontera ***FRONTUR*** con objeto de mejora de la eficiencia en obtención de estadísticas de turismo, actualización de datos y en general, a la hora de obtener resultados más precisos.

Entonces, ¿qué es ***EGATUR***? podemos definir este proyecto como "la operación estadística de la Subdirección General de Conocimiento y Estudios Turísticos que recoge datos relativos al gasto que realizan en España los visitantes no residentes en España".
Una vez dicho esto, lo que se pretende estudiar mediante la variable del gasto es dar respuesta a la siguiente cuestión: ¿Cuánto dinero gastan los turistas en España?.
Sabemos por datos del INE que en 2019 llegaron a España **83.7 millones de turistas** y hubo **92.278 millones de euros de gasto** con los turistas provenientes de **Reino Unido** a la cabeza, seguidos por **los Alemanes y Franceses.** Las visitas de estos paises representan la mitad de extranjeros que llegan a España, pero con respecto al 2018 se redujeron debido a competidores como Egipto, Tunez o Turquía que tienen unos precios muy agresivos con los que ganan rápidamente cuota de mercado.

Por otra parte, los denominados 'paises lejanos' incrementan sus viajes a España.
Una tendencia en auge en paises como Estados Unidos y Rusia, sin olvidarnos de Latinoamérica que aumenta sus visitas a ritmo de doble dígito; y por supuesto Asia, de donde recibimos 5.6 millones de turistas en 2019.

Durante los últimos años España está intentando ofrecer una imagen de turismo de calidad y es por lo que quiere apostar el sector. Aumentar el gasto medio por turista es una garantía de ingresos en el país, y se puede conseguir creando un (o varios) prototipo de turista que viene a España (su procedencia, la comunidad autónoma que prefiere, su método de transporte ...).
El objeto de estudio que se propone en el punto anterior me ha resultado interesante ya que España es un país que prácticamente vive de su sector turístico o por así decir, este sector es el que genera gran parte de los ingresos que que se reflejan en nuestro PIB.
Según datos del ***CEOE "En 2019 el turismo suponía el 12,4% del PIB en 2019"*** razón por la que parece lógico pensar que el turismo se muestra como uno de los sectores más dinámicos de nuestra economía.

Volviendo al periodo elegido, año 2019, veremos si se cumplen las hipótesis establecidas sobre el gasto medio por turista, las regiones y los meses comparados. Haciendo uso de RStudio estudiaré el comportamiento de las principales variables y estadísticos propuestos. Por último y de caracter adicional, con las pocas coparaciones que realizaremos trataré de crear un prototipo de turista.

3. Información sobre la Base de Datos

El fichero de microdatos utilizado se encuentra en la página del INE, apartado productos y servicios, concretamente en la encuesta de gasto medio por turista (EGATUR). Se corresponde con una encuesta continua de periodicidad mensual por medio de entrevista personal y los ámbitos tanto poblacional como geográfico se comentan a lo largo del ejercicio.
He elegido todo el año 2019 para hacerme una idea más global de los datos, intentando minimizar el efecto de la estacionalidad y poder identificar mejor los valores atípicos; sabemos que a mayor cantidad de datos mejor inferencia se puede hacer sobre la población ya que al aumentar el tamaño de la muestra se obtienen valores estadísticos que deben parecerse más al valor del parámetro, si todo lo demás sigue igual.

En cada fichero tenemos dos documentos, el primero, "elevado_eg_mod_web_tur_xxXX" con 13 variables y el segundo, "etapas_eg_mod_web_xxXX" que corresponde a los datos de las etapas de los turistas si sus vacaciones en España se componían de más de una comunidad autónoma de destino. Como la primera base de datos tiene en cuenta como destino principal la comunidad con mayor número de pernoctaciones nos vamos a señir a esa interpretación de los datos. Nuestro estudio se centrará en la primera base de datos, donde cada entrada de datos corresponde a un viajero que finaliza su viaje por España.

Aunque nos encontramos con una variable "Factor_Egatur" que el INE utiliza para estimar el gasto turístico medio de la siguiente manera:
"Para estimar el gasto turístico se multiplica la variable 'gastototal' por 'Factor_Egatur'.
El gasto medio por persona se obtendrá como el cociente del gasto turístico entre los turistas, calculados sumando la variable 'Factor_Egatur'.
Para el cálculo de pernoctaciones multiplica la variable 'A13' por 'Factor_Egatur'. Por último, se calcula la variable 'gasto medio diario' dividiendo el gasto turístico total entre las pernoctaciones"

Nuestro acercamiento será de manera más sencilla, dividiremos el 'gasto total' entre 'A13' (número de pernoctaciones) lo que nos dará un gasto medio diario aproximado. En la siguiente tabla podremos encontrar el significado de todas las variables. (Anexo I para ampliar información)

mm_aaaa	Mes y año de referencia
A0	Encuesta de procedencia
A0_1	Identificador de cuestionario
A0_7	Código de cuestionario (TEN)
A1	Vía de salida
pais	País de residencia habitual
ccaa	Comunidad Autónoma de destino principal del viaje
A13	Número de pernoctaciones
aloja	Alojamiento principal
motivo	Motivo prinicpal del viaje
A16	Paquete turístico
gastototal	Gasto total del viaje
factoregatur	Factor de elevación de Egatur

4. Agrupación de Bases de Datos

Script: AgrupaciónBasesDatos.R

La función de este script es agregar en una todas las bases de datos del año 2019. Para esta labor no habría elegido RStudio ya que me parece un programa perfecto para análisis estadístico e inferencial pero para la limpieza y organización de bases de datos habría usado otro (Python con pandas y Numpy o SLQ); pero ya que la tarea de limpieza es relativamente simple y el objetivo de la PEC es aprender a usar RStudio usaré este para la limpieza.

 # Función de Nombres

Primero elegimos la carpeta de donde queremos extraer las bases de datos, para ello usamos la función setwd().
En el paso siguiente cree un data frame para contener todas las bases de datos, lo llamé 'data'.

Como tenemos 12 bases de datos que se llaman igual (elevado_eg_mod_web_tur_XX19), donde lo único que cambia es el número del mes (XX) he creado una función que itera del 1 al 12
cambiando solo el número del mes.
Para los meses menores que 10, en número, agregué un "0" delante.
Cada iteración del loop añadía la base de datos correspondiente al mes 'i' a la base de datos anual.

```
i <- 0
data <- data.frame()

for (i in seq(1:12)) {
  if (i <= 9) {
    b <- paste('elevado_eg_mod_web_tur_','0', toString(i), '19.txt',
sep='')

  } else {

    b <- paste('elevado_eg_mod_web_tur_', toString(i), '19.txt', sep='')
  }

  temp <- read.csv(b, sep=";", header = TRUE, stringsAsFactors = FALSE)

  data <- rbind(data, temp)
}
```

 # Exportar base de datos
Como paso siguiente he exportado la nueva base de datos creada con ayuda de la función write.table().

4.1. Limpieza de Datos

Script: LimpiezaDatos.R

En este script nos vamos a centrar en dejar la base de datos limpia, sin valores nulos y ordenada para poder trabajar de la mejor manera.

 # Paso 1 búsqueda de valores nulos

Primero que nada buscamos si hay algún valor "N.A" en la base de datos, que nos dará problemas para hacer cálculos.

 # Paso 1.1

Creamos una función para detectar las columnas que tienen valores N.A.
Donde pasamos desde la columna 1 a la 9 con la función seq(1:9) y comparamos si la respuesta any(is.na()) es TRUE, entonces sabemos qué columnas tienen valores N.A.

 # Paso 1.2

Tenemos varias maneras de manejar los valores nulos, podíamos asignarles la media/ mediana/moda de la columna (ya que se trata de una variable cuantitativa).
He contado la cantidad de valores nulos en la columna y son 24, al ser mucho menos del 1% de la base de datos (24/87055) he decidido borrarlos de la base de datos.
He convertido el dataframe raw.data.19 a 'tibble' porque tiene muchas más funciones.

 # Paso 2 Eliminación de columnas innecesarias

 # Paso 2.1

Aquí he decidido que en la columna mm_aaaa no nos interesa el año ya que toda la base de datos es del mismo año (2019).
He separado la columna en dos, año y mes, despues he eliminado la columna del año.

 # Paso 2.2

He decidido eliminar la columna A0 también ya que refiriéndose al origen de la encuesta y sabiendo que solo puede ser del EGATUR parece innecesaria.

 # Paso 2.3

Aquí he descubierto que hay id de cuestionario repetidos, porque el resultado de la función es 87008 cuando debería ser 87055. Pero seguramente se deba al hecho que rstudio ha "redondeado" el número de la columna pensando que era una medida y no un id.
Esto no es muy relevante para nuestro estudio así que lo vamos a pasar por alto.

5. Análisis exploratorio de los datos (EDA)

Script: EDA.R
Base de Datos: clean_eg_2019.txt

En este Script nos centraremos en hacer un Análisis Exploratorio de cada variable (EDA, por sus siglas en inglés, Exploratory Data Analysis), lo que significa que analizaremos una a una las variables para determinar la información que podemos sacar de ellas.

 # A0_7 (Variable Dummy, Cualitativa Nominal)

Esta variable nos indica si el Turista se encuentra en tránsito, básicamente sabemos si su destino principal es España o no.

Algo que sabíamos ya era que la mayoría de los turistas venían a España y no estaban de tránsito, por lo que la he comparado con la variable "vía de salida" de los turistas.
Como podemos observar la mayoría de los turistas que vienen a España lo hacen en Avión, algo que no nos debería sorprender; en cambio observamos que la mayoría de turistas en tránsito entran a España por carretera.
Mi suposición es que la mayoría de turistas en tránsito se dirigen a Portugal por lo que tienen que pasar con el coche por territorio español, o van a algún puerto con rumbo a África.

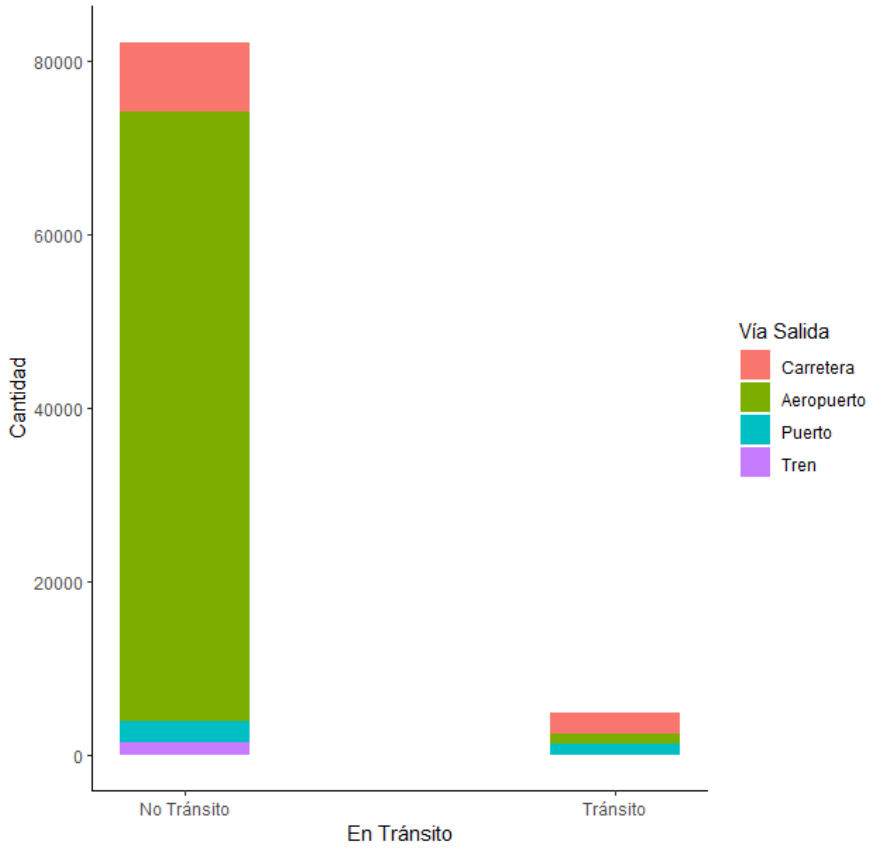


Gráfico 5.1 Turistas en tránsito según Via de Salida

 # A1 (Variable Cualitativa Nominal)

Como habíamos visto en el gráfico anterior, la mayoría de turistas (más del 75%) salió del país en avión, en segundo lugar, con aproximadamente el 16% tenemos carretera como opción de salido de los turistas.

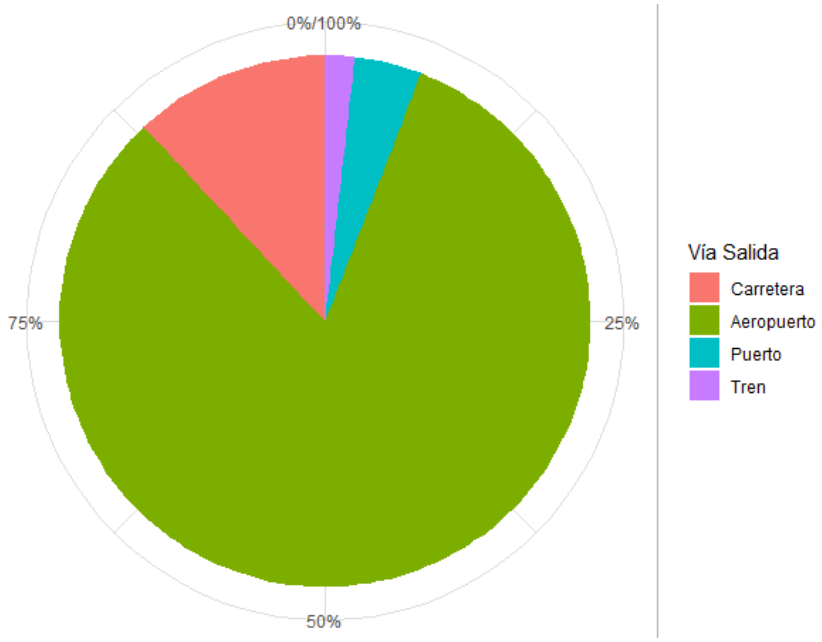


Gráfico 5.2 Porcentaje de Turistas según Via de Salida

 # país (Variable Cualitativa Nominal)

Con el análisis de esta variable podemos observar que casi el 50% de turistas que vinieron en 2019 tienen su residencia habitual en Alemania, Francia o Reino Unido; siendo el primer lugar este último.

Según datos de Statista
https://www.statista.com/statistics/578815/most-visited-countries-united-kingdom-uk-residents/

en 2019 el país más visitado por los británicos fue España con 18.1 millones, sobre todo las Islas Baleares, Canarias y Benidorm; y sus motivos principales son el clima y el bajo coste relativo de los paquetes vacacionales.

Cuando analizamos el tipo de alojamiento según la nacionalidad observamos que la mayoría eligen "Hotel y similar" algo que es comprensible ya que al tratarse de turismo la principal opción para pernoctar es el hotel o similares.
La segunda opción es Alojamiento No Mercado, lo que significa que no se ha hecho una transacción económica por el alojamiento (casa en propiedad, de familiares, amigos …).
Por más que busqué no encontré a qué se refiere el INE con resto de mercado, pero mi idea es que incluye la gente que reserva con plataformas tipo AirB&B o couchsurfing, workaway, o gente que viene con su propio medio tipo autocaravana, furgoneta …

En general todos los turistas, independientemente de la nacionalidad siguen esta pauta:

- Hotel y similar aproximadamente un 60-65%
- Alojamiento no Mercado aprox. 30-35%
- Resto de Mercado aprox. 10%

A excepción de los suizos y los franceses que sus partidas de Hotel y Alojamiento no mercado son prácticamente iguales.

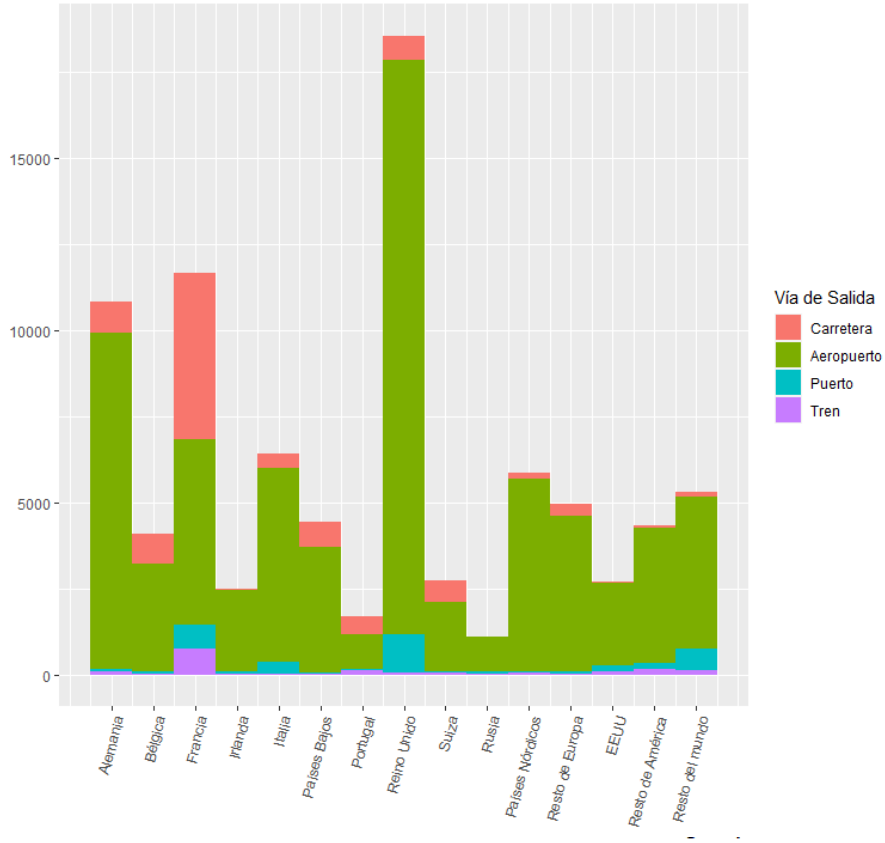


Gráfico 5.3 Turistas según Pais de residencia / Via de Salida

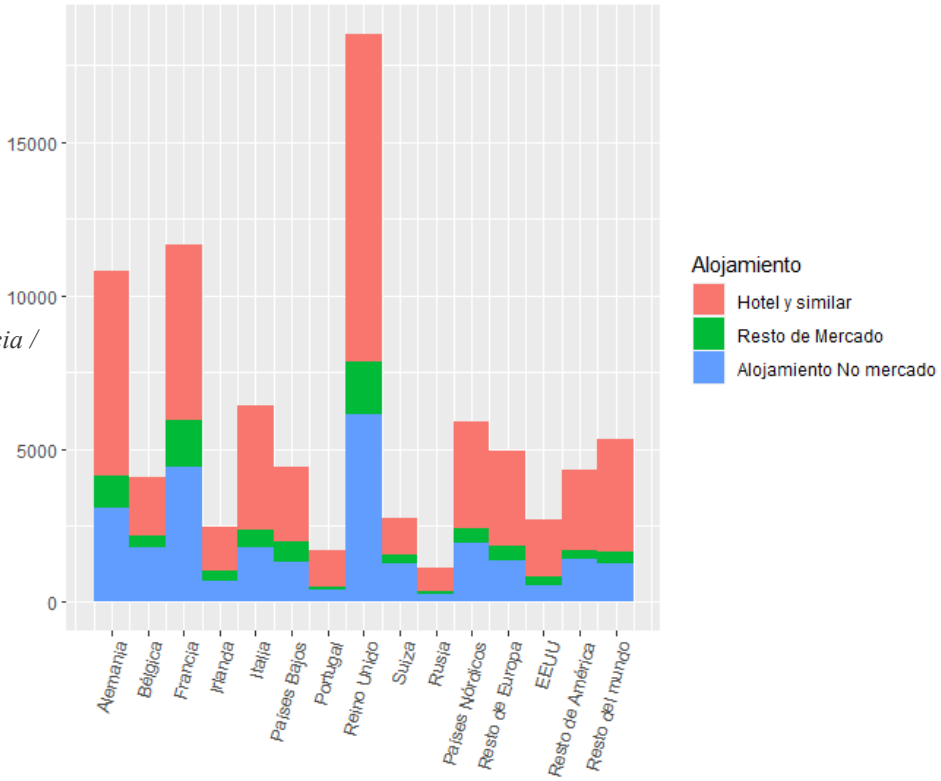


Gráfico 5.4 Turistas según Pais de residencia / Alojamiento

 # ccaa (Variable Cualitativa Nominal)

Aquí podemos comprobar lo que vimos en el apartado anterior, vimos que los turistas de Reino Unido, Alemania y Francia, que fueron la mayoría, elegían como destino las islas (Baleares y Canarias) o la Comunitat Valenciana; estos destinos son de los que más turistas han recibido, pero según los datos el destino principal del 2019 fue Cataluña con casi una tercera parte de los turistas encuestados.
También notamos la importancia de Andalucía y Madrid como destinos turísticos.

Para hacernos una mejor idea de las preferencias de destinos según el país de residencia de los turistas he creado una matriz de correlación entre estas dos variables (5.5). Cada casilla es el la cantidad de turistas del determinado país (Columnas) que visitó el destino específico (Líneas) dividido entre el valor máximo de la fila, así nos podemos hacer una idea de los destinos preferidos según el país de residencia.
Nos apoyamos también en una tabla con los valores totales (5.6).

En esta matriz podemos volver a corroborar que los destinos más elegidos son (Cataluña, Canarias, Baleares, Valencia y Andalucía), pero como ampliación de información vemos que:

- *Reino Unido:* Prefieren las Islas (Tanto Baleares como Canarias) como destino principal, seguido por la Comunitat Valenciana, Andalucía y Cataluña.

- *Alemania :* Prefiere las Islas (Tanto Baleares como Canarias) como destino principal, seguido por Cataluña, la Comunitat Valenciana y Andalucía.

- *Francia:* Prefiere Cataluña como primer destino, seguido por la Comunitat Valenciana, Andalucía y Baleares. Esto se puede ser debido a que Cataluña y Francia comparten frontera y a que la lengua Catalana y Valenciana tienen semejanzas con el idioma francés.

De hecho si vamos un paso más allá en los gráficos 5.3 y 5.4 del apartado variable país podemos observar que con diferencia los franceses son los turistas que eligen como vía de salida carretera y las CCAA con la tasa más alta de salida por carretera son los destinos de los franceses (Cataluña, C. Valenciana e Andalucía)

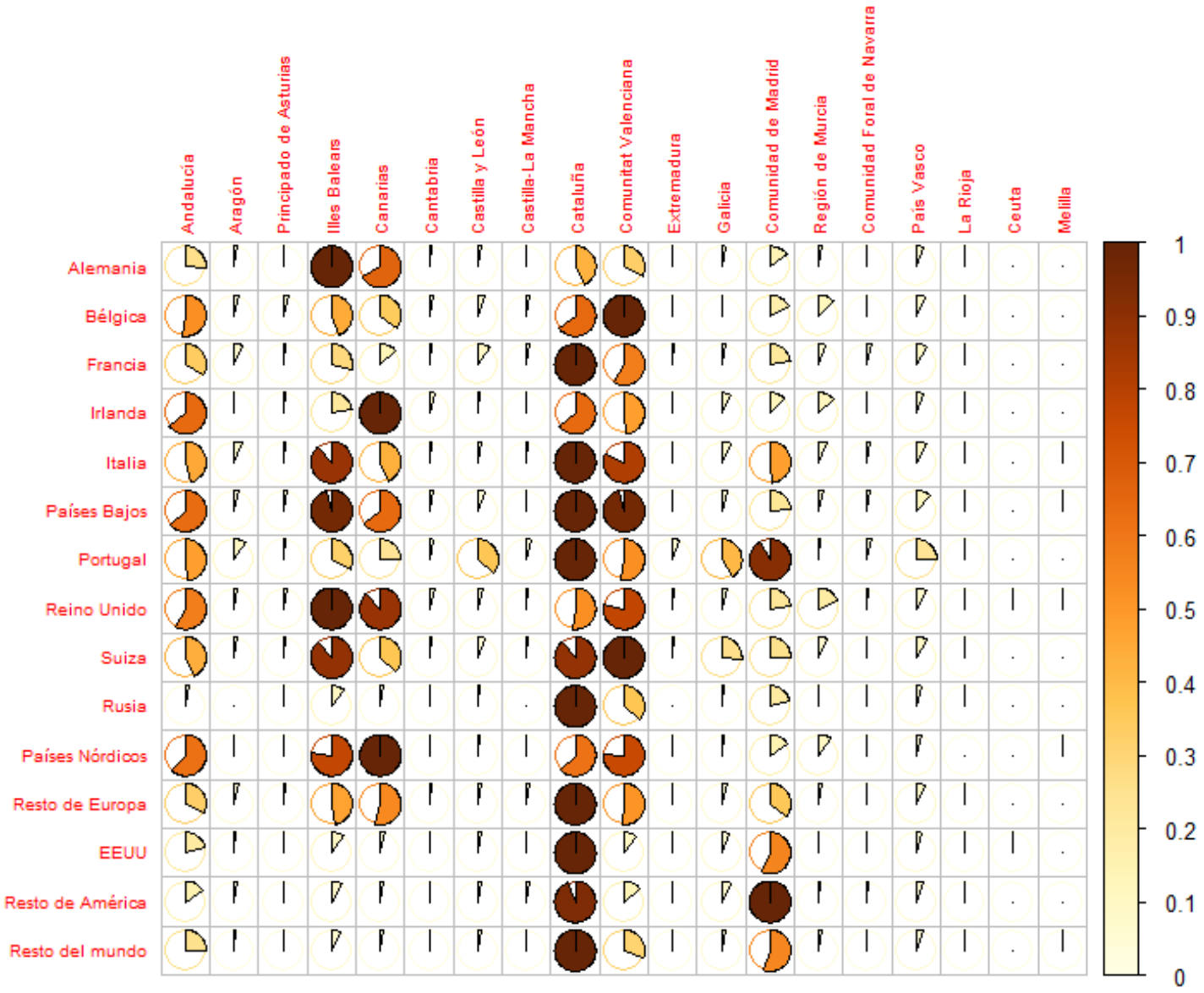


Gráfico 5.5 Matriz relación Pais de Residencia / Comunidad Autónoma de Destino

	Alemania	Bélgica	Francia	Irlanda	Italia	Países Bajos	Portugal	Reino Unido	Suiza
Andalucía	925	603	1312	456	650	581	171	2403	269
Aragón	60	49	254	7	84	35	34	124	15
Principado de Asturias	41	50	57	16	24	26	8	146	15
Illes Balears	3548	512	1107	153	1263	880	114	4109	549
Canarias	2363	401	531	708	620	588	86	3611	226
Cantabria	48	31	90	37	37	34	12	163	12
Castilla y León	97	71	392	13	47	47	128	178	33
Castilla-La Mancha	27	24	72	2	24	9	14	55	10
Cataluña	1517	743	3854	454	1438	914	354	2158	546
Comunitat Valenciana	1156	1146	2238	342	1178	878	186	3203	616
Extremadura	24	9	62	1	13	8	19	48	7
Galicia	133	13	115	54	99	47	144	201	162
Comunidad de Madrid	529	200	844	90	694	212	325	932	158
Región de Murcia	93	142	202	88	90	39	4	784	43
Comunidad Foral de Navarra	21	10	115	6	21	18	11	51	6
País Vasco	211	73	358	36	116	103	86	327	56
La Rioja	13	9	35	5	7	7	3	36	6
Ceuta	0	0	0	0	0	0	0	1	0
Melilla	0	0	0	0	1	1	0	3	0

	Rusia	Países Nórdicos	Resto de Europa	EEUU	Resto de América	Resto del mundo
Andalucía	18	882	459	254	261	568
Aragón	0	10	49	17	31	25
Principado de Asturias	1	10	30	15	17	22
Illes Balears	66	1095	670	119	137	144
Canarias	15	1418	772	39	49	48
Cantabria	1	6	30	6	18	15
Castilla y León	8	24	38	29	53	54
Castilla-La Mancha	0	12	32	8	30	14
Cataluña	607	879	1412	1248	1565	2226
Comunitat Valenciana	224	1074	722	127	242	691
Extremadura	0	4	6	1	6	4
Galicia	11	24	64	70	116	81
Comunidad de Madrid	123	214	504	703	1669	1228
Región de Murcia	5	142	35	4	21	47
Comunidad Foral de Navarra	3	5	13	6	20	13
País Vasco	27	66	104	51	86	112
La Rioja	2	0	14	3	9	7
Ceuta	0	0	0	1	0	0
Melilla	0	1	0	0	0	2

Gráfico 5.6 Tabla relación Pais de Residencia / Comunidad Autónoma de Destino

 # A13 (Variable Cuantitativa Discreta)

Para la variable A13 (Número de Pernoctaciones) he decidido usar un "box plot".
He filtrado el gráfico porque había valores atípicos, he decidido poner como máximo 50 noches ya que la cantidad de datos que "perdemos" es menor del 5% y teniendo una muestra de más de 87000 datos no me pareció grave. Así podemos analizarlo mejor.

Analizándolo vemos que el 50 % de los datos se distribuyen entre 4 noches (Q1) y 9 noches (Q3) con una mediana de 7 noches. Con un mínimo de 2 noches y un máximo de 16, a partir de donde empiezan los "outliers".

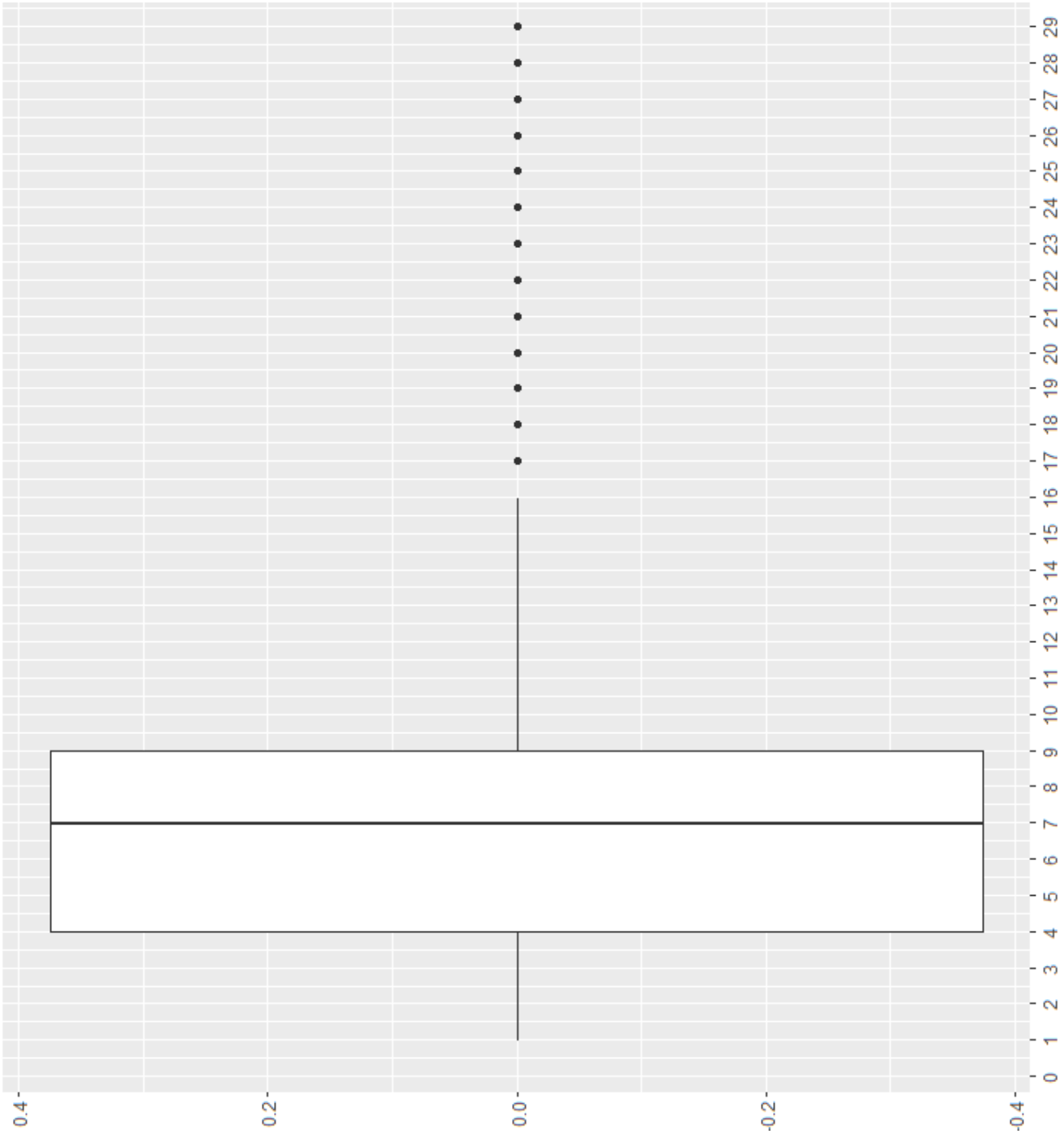


Gráfico 5.7 BoxPlot Número de pernoctaciones

 # aloja (Variable Cualitativa Nominal)

Como era de esperar vemos que más del 60% de los turistas eligen, para su estancia en España, un hotel o similar; aproximadamente un 30% de los turistas se hospedan en un alojamiento de no mercado, ósea, un inmueble de su propiedad o familiar, pero que no hay una retribución monetaria por el alojamiento.
Y el 10% restante se reparte entre el resto de mercado.

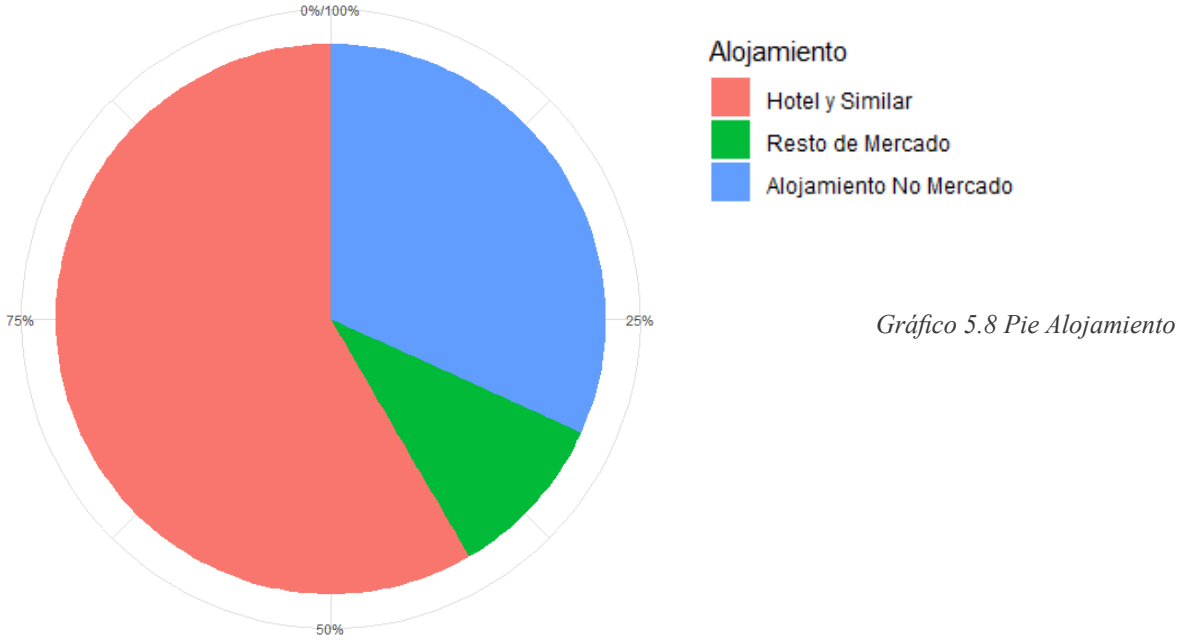


Gráfico 5.8 Pie Alojamiento

 # motivo (Variable Cualitativa Nominal)

Aquí también hay poco que decir, sabemos que en 2019 el turismo aporto un 12,4% del PIB Español.
En el gráfico vemos que poco más del 75% de los turistas vienen a España por Ocio / Vacaciones y aproximadamente un 10% viene por Negocios / Trabajo.

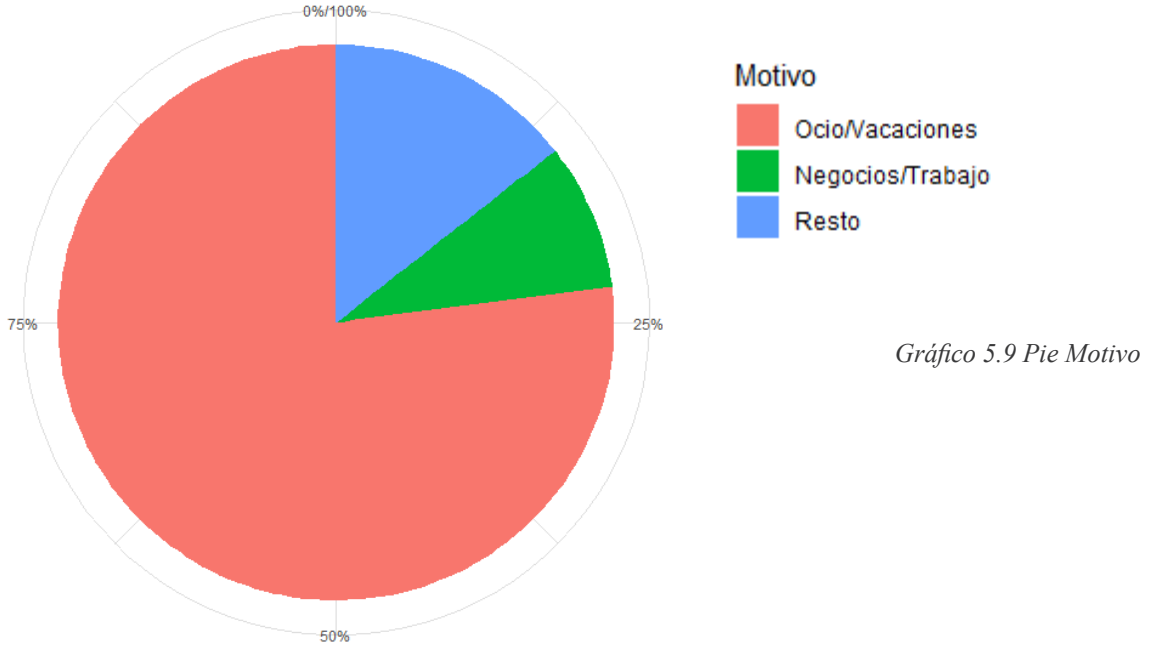


Gráfico 5.9 Pie Motivo

 # A16 (Variable Dummy, Cualitativa Nominal)

En ente gráfico vemos que la norma en todos los turistas es no elegir paquete turístico.

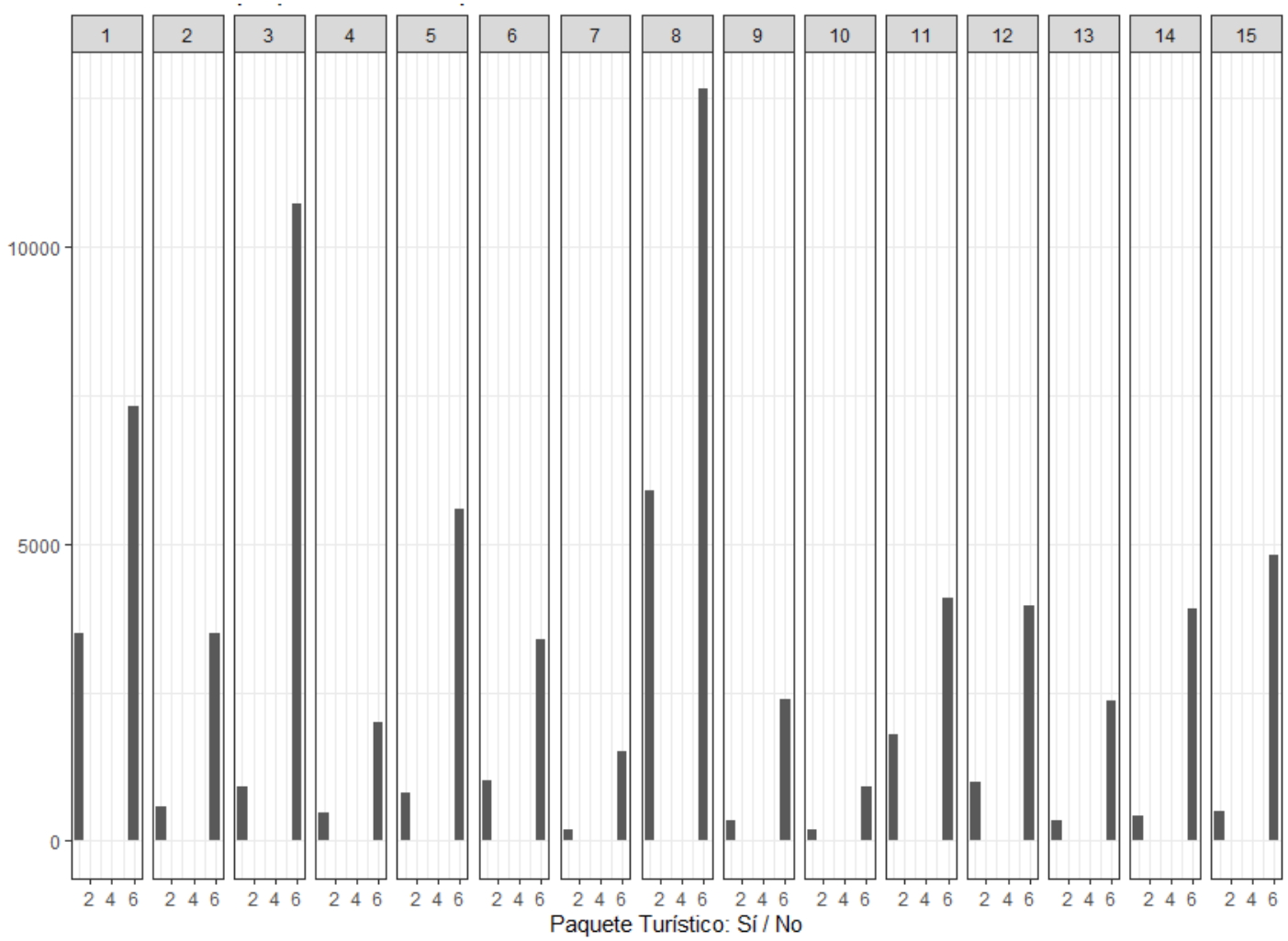


Gráfico 5.10 Barras Paquete Turístico

 # gastototal (Variable Cuantitativa Continua)

Observando los gráficos apreciamos que los países que más han gastado son los que mayor tráfico han generado, y esto es comprensible ya que cuantos más turistas mayor será el gasto.
Para realizar una comparación más adecuada deberemos utilizar el gasto medio.

Se puede realizar una observación, vimos que el segundo país en afluencia de turistas era Francia, seguido de cerca por Alemania; pero si nos fijamos en el gráfico de gasto por país Francia se encuentra por detrás de Alemania.

Esto se puede estar relacionado, como vimos antes, con que los turistas franceses tienen una alta tasa de salida de España por carretera (5.3); esto contrastado con el alto nivel de alojamiento de no mercado de los turistas franceses (5.4) puede llevar a pensar que muchos de ellos optan por unas vacaciones en Autocaravanas, disminuyendo el gasto en alojamiento (que suele ser la mayor partida de gasto de un turista). Pero no hay datos para poder comprobar esta teoría por lo que la comentamos como una curiosidad.

También señalar que aunque los turistas provenientes del resto del mundo y resto de américa representan solo un 11.4 % de los turistas encuestados, sin embargo son responsables del 21.65 % del gasto total.

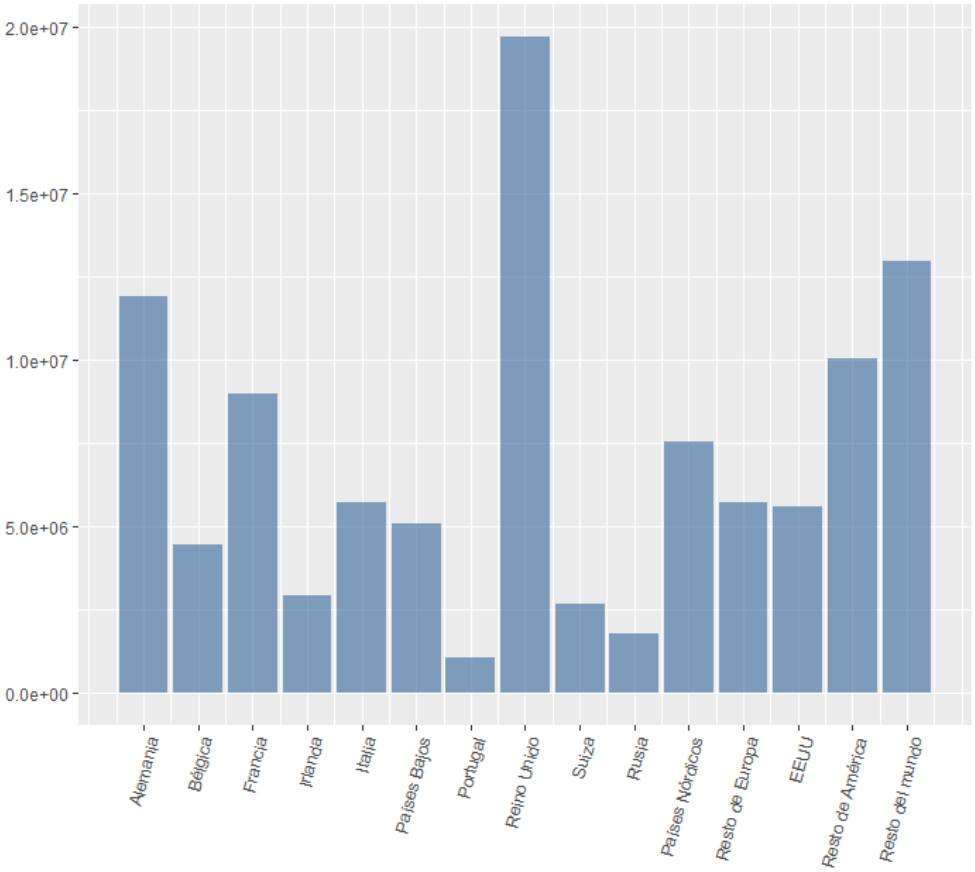
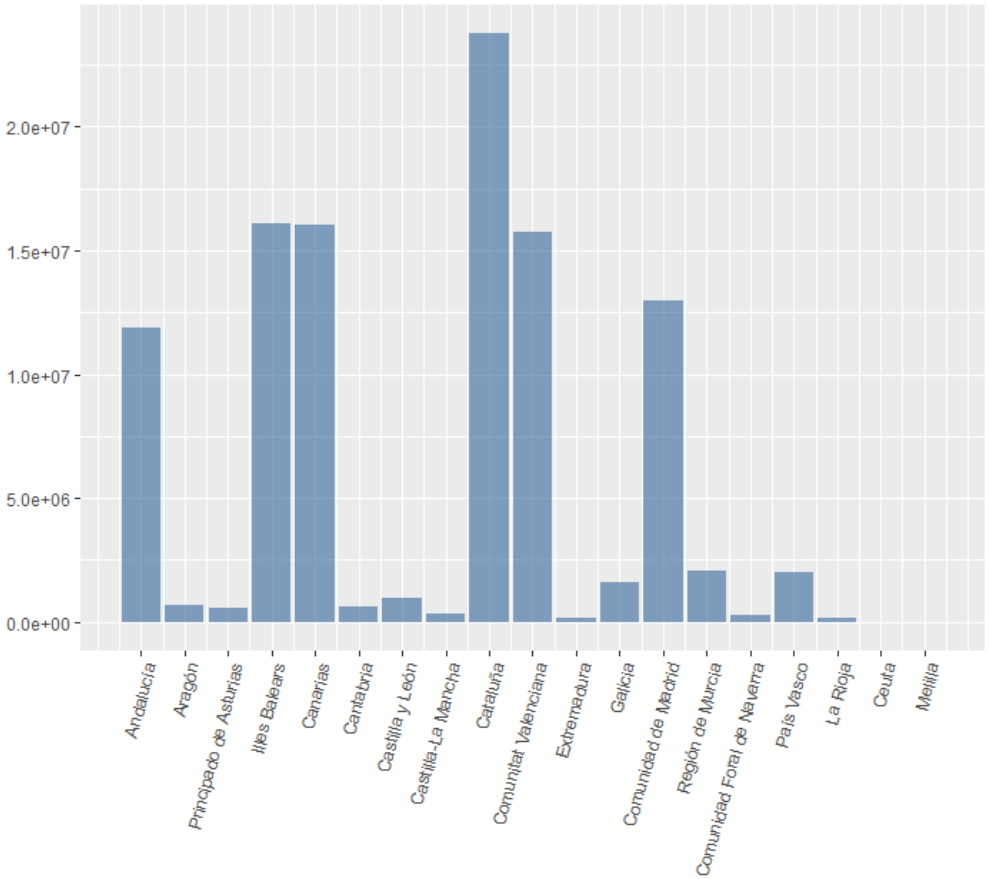


Gráfico 5.11 Gasto total por país

Gráfico 5.12 Gasto total por CCAA



 # mes (Variable Cualitativa Nominal)

Creo que es interesante ver la estacionalidad de la afluencia de turistas, por lo que analizando los datos de la variable mes podemos extraes cosas interesantes.
Del gráfico 5.13 no extraemos gran cosa, podemos notar que los meses de veranos son los más intensos, pero en general vamos una distribución relativamente homogénea, sin picos. Vamos a observar más a fondo.

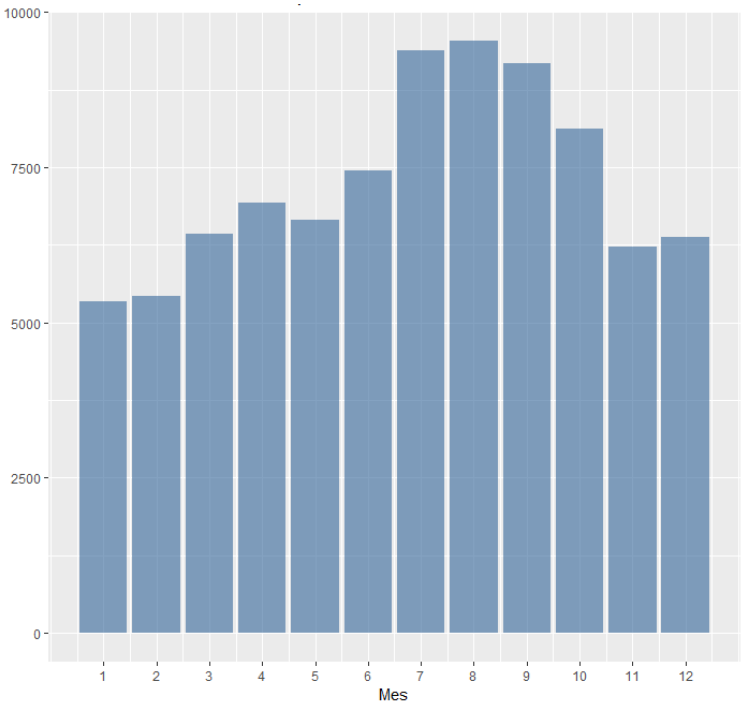


Gráfico 5.13 Afluencia de Turistas dividido por meses

1.1 Creamos una matriz de países vs meses (5.14)

1.2 Con esta función dividimos cada fila de país por su máximo valor de turistas, con esto conseguimos ver el mes favorito para viajar a España de cada país.
Vemos en el gráfico que los meses en el que más turistas vuelan a España es Julio y Agosto; vemos además que los turistas de Portugal, Reino unido, Alemania son bastante homogéneos durante el año, teniendo sus máximos en los meses de verano. No podemos decir lo mismo de los rusos o los franceses, que viajan a España prácticamente solo en los meses de verano.

1.3 Matriz CCAA vs Meses (5.16)

1.4 Con esta función dividimos cada fila de ccaa por su máximo valor de turistas, con esto conseguimos ver el mes favorito para viajar a cada ccaa.
Como decíamos antes los meses de verano en general son los que atraen más turistas; pero podemos observar cosas interesantes como, Aragón tuvo su pico de turistas en Marzo, o las islas canarias recibieron el máximo de turistas en los meses de Marzo y Abril.
Vemos también que la Comunidad de Madrid recibe aproximadamente la misma cantidad de turistas durante todo el año (puede darse porque es el aeropuerto más importante de España, y ya que la mayoría de turistas viene por avión, sería interesante ver las siguientes escalas de los turistas, aunque eso está fuera de esta investigación).

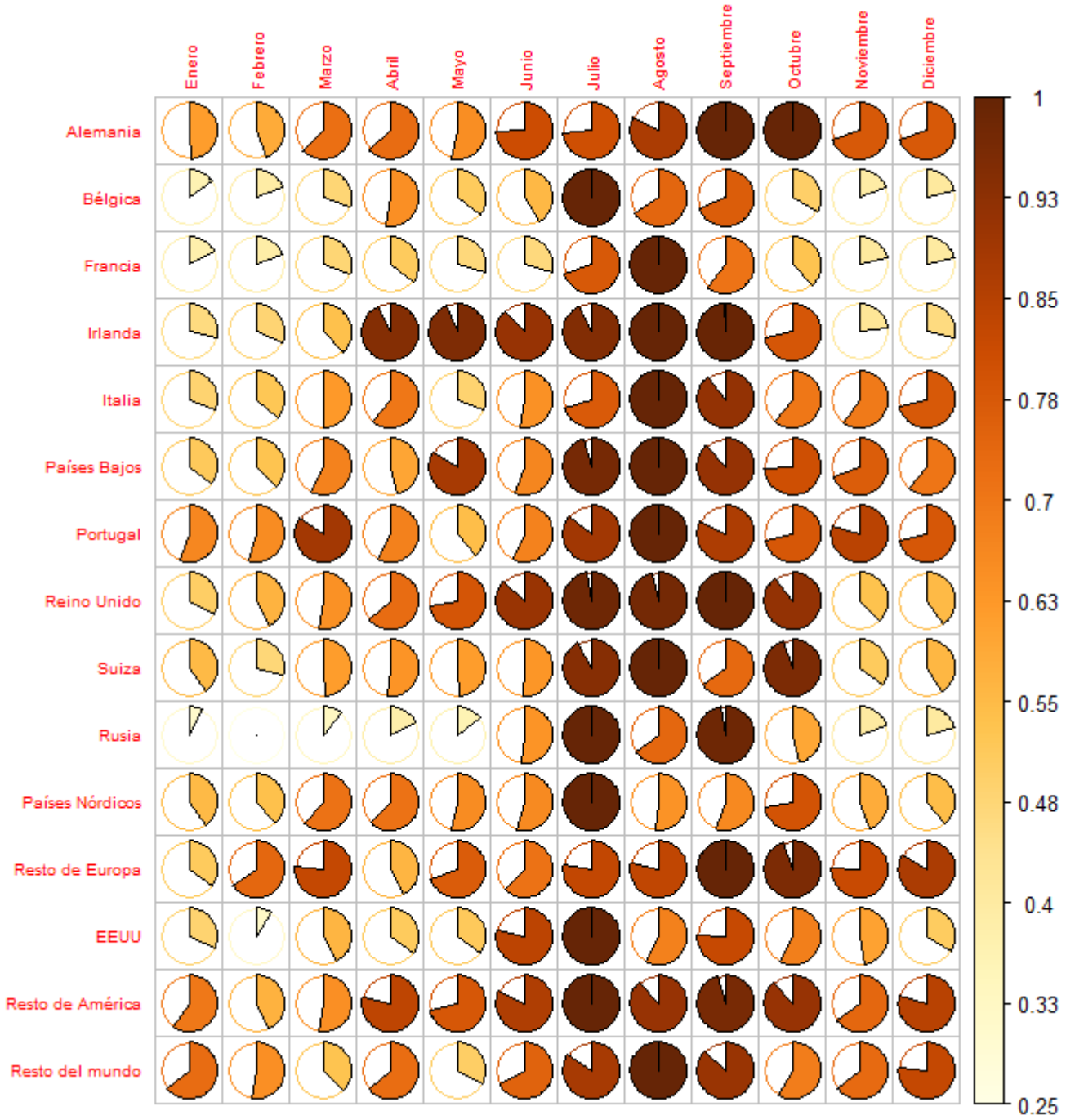


Gráfico 5.14 Matríz Paises / Meses

	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Alemania	714	678	835	842	751	937	930	1009	1156	1156	899	899
Bélgica	221	240	288	390	310	336	602	449	461	301	240	248
Francia	681	702	851	912	839	835	1381	1776	1261	944	734	722
Irlanda	129	134	149	261	264	252	262	277	275	218	118	129
Italia	375	403	480	541	374	494	594	767	704	541	535	598
Países Bajos	254	261	332	293	430	328	475	490	447	395	375	347
Portugal	122	120	162	125	100	124	163	183	159	143	155	143
Reino Unido	1020	1160	1315	1485	1609	1841	2005	1979	2037	1867	1088	1127
Suiza	183	157	205	211	205	209	311	331	244	317	170	186
Rusia	54	44	58	67	63	110	173	128	170	103	70	71
Países Nórdicos	403	392	519	521	475	478	727	466	483	577	426	399
Resto de Europa	270	389	434	296	404	375	436	438	524	501	430	457
EEUU	175	114	202	184	185	301	358	244	294	245	217	182
Resto de América	310	254	287	371	346	383	443	403	427	402	328	376
Resto del mundo	433	385	316	431	299	447	523	594	538	408	436	491

Gráfico 5.15 Tabla Países / Meses

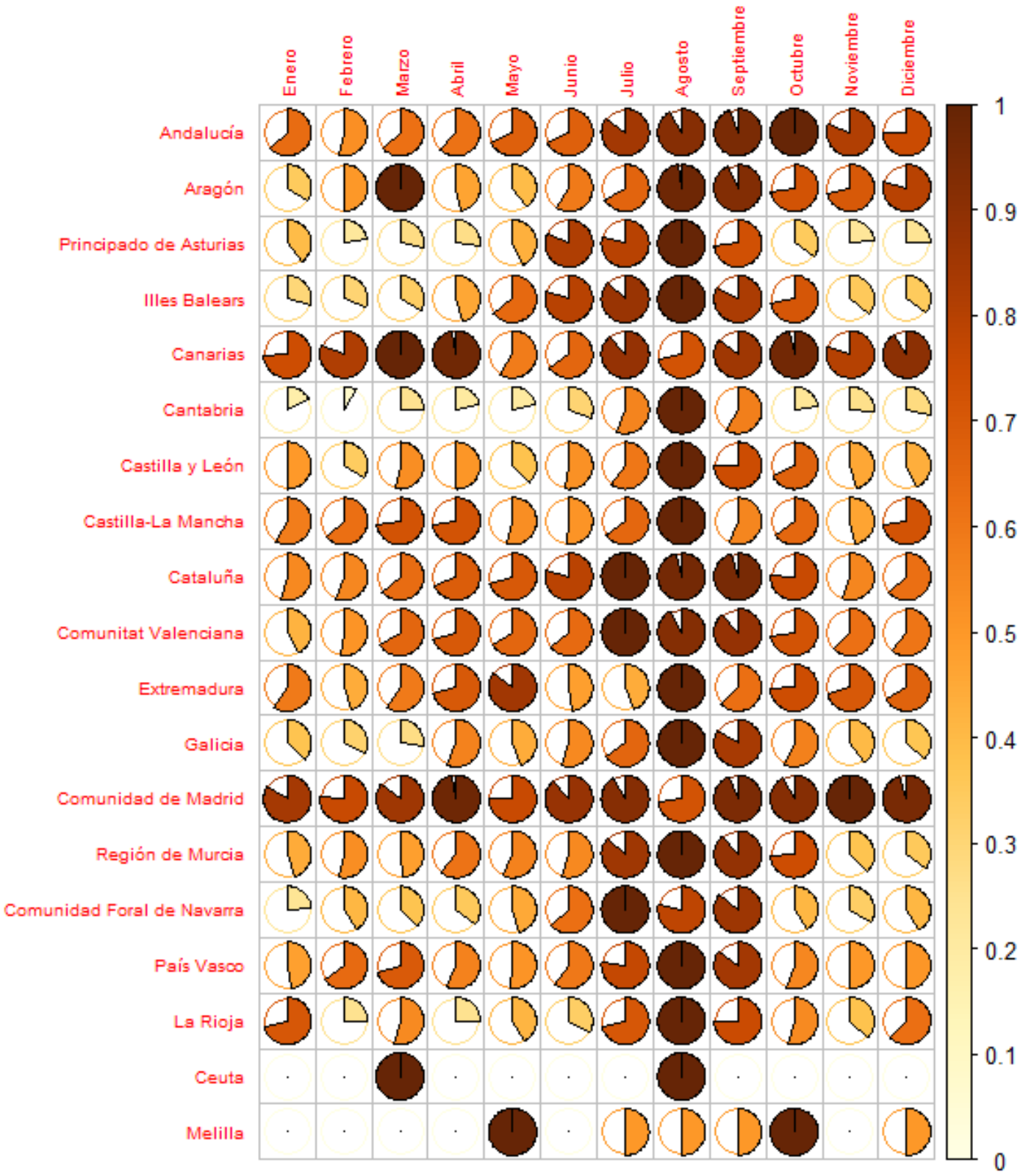


Gráfico 5.16 Matríz CCAA/ Meses

Gráfico 5.17 Tabla CCAA/Meses

	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Andalucía	691	572	677	672	737	735	918	999	1028	1086	881	816
Aragón	34	49	98	46	39	58	65	96	91	71	69	78
Principado de Asturias	33	18	23	22	36	68	66	83	61	29	19	20
Illes Balears	622	643	697	953	1344	1657	1818	2079	1716	1484	733	720
Canarias	858	944	1158	1129	681	759	1020	834	991	1122	932	1047
Cantabria	23	11	33	27	27	40	73	130	75	30	34	37
Castilla y León	90	62	96	92	68	95	109	181	135	122	83	79
Castilla-La Mancha	25	27	31	31	23	22	28	43	24	28	20	31
Cataluña	1247	1252	1443	1556	1589	1795	2270	2185	2162	1721	1265	1430
Comunitat Valenciana	706	864	1107	1180	1100	1079	1679	1551	1480	1209	1048	1020
Extremadura	16	12	16	19	23	13	12	27	17	20	19	18
Galicia	78	67	57	119	93	115	138	210	176	120	84	77
Comunidad de Madrid	674	608	686	782	605	704	735	579	753	736	801	762
Región de Murcia	105	125	113	144	133	129	202	235	208	174	88	83
Comunidad Foral de Navarra	12	21	19	18	23	32	51	40	44	21	17	21
País Vasco	113	152	165	134	121	141	181	236	200	130	119	120
La Rioja	17	6	13	6	10	8	17	24	18	13	9	15
Ceuta	0	0	1	0	0	0	0	1	0	0	0	0
Melilla	0	0	0	0	2	0	1	1	1	2	0	1

 # gastomedio (Variable Cuantitativa Continua)

He creado una nueva columna llamada gasto medio dividiendo la columna "gastototal" entre la columna "A13" (número de pernoctaciones) para tener un gasto medio, aunque no sea ponderado con el factor egatur para ser totalmente preciso, para el supuesto que vamos a hacer nos vale.

A primera vista se parace mucho al gráfico del gasto total dividido por paises (5.11).
Las principales diferencias son:
- La drástica reducción del gasto de los turistas del "resto del mundo", se puede deber a que abarca una gran cantidad de paises y sus culumnas 'A13' y 'gastototal' son bastante más heterogéneas (sobre todo A13). Lo mismo sucede con los turistas del Resto de América.

- Notamos que aunque los Alemanes tenían un gasto total mayor que los franceses, los franceses gastan de media más que los Alemanes, puede que los franceses pasen menos dias en España comparado con los Alemanes.

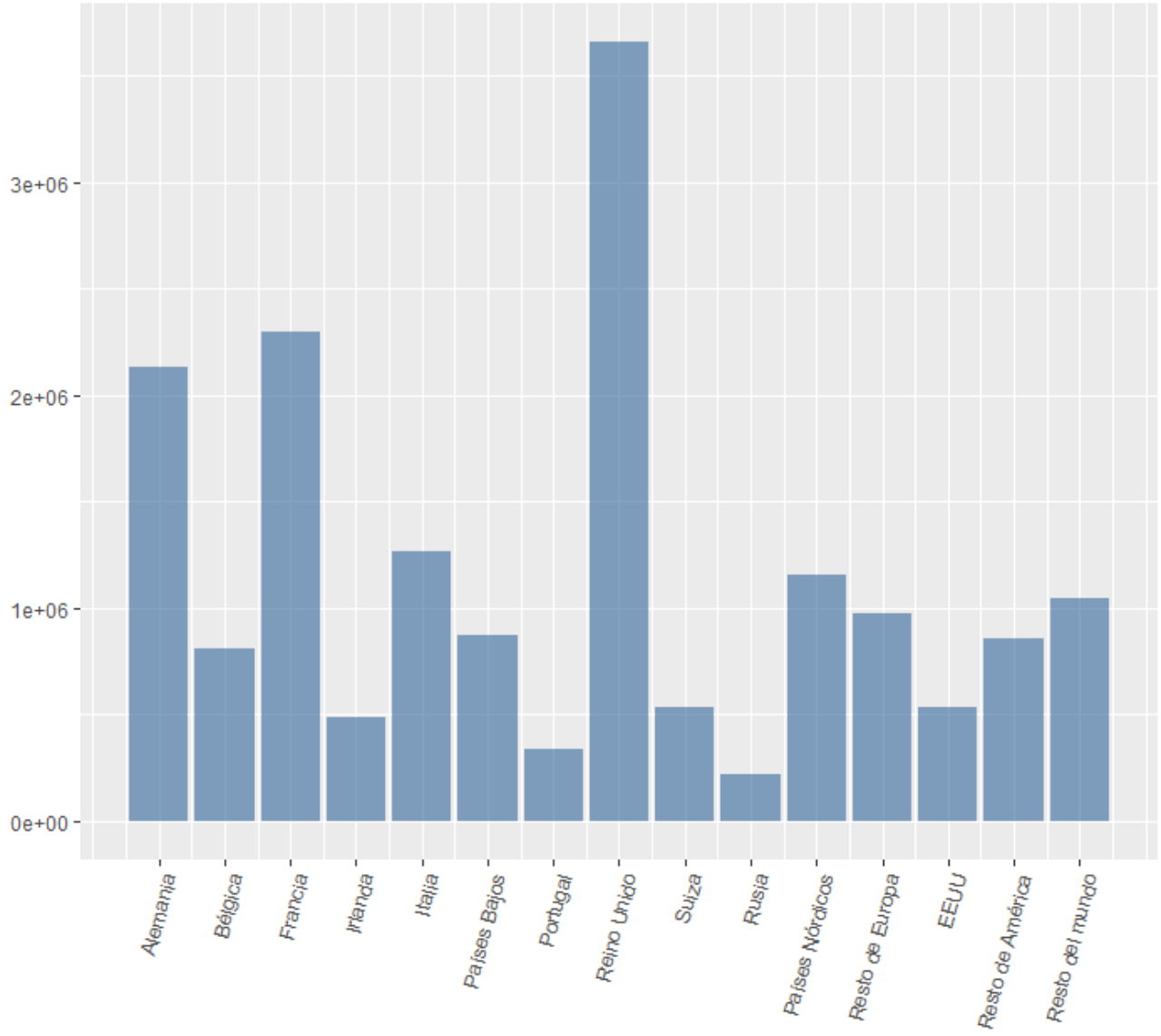


Gráfico 5.18 Gasto Medio Diario dividido por Paises

6. Métodos y Resultados Preeiminares

¿El gasto medio diario de un turista es mayor que 200 euros?

Utilizaremos un contraste de hipótesis para descubrir si el gasto diario de un turista en España es mayor que 200 euros, que es lo que nos dicen los gurus del sector.

¿Son los turistas alemanes mejores que los británicos en términos de gasto medio?

Aquí previamente al contraste de medias debemos realizar un contraste de desviaciones típicas para confirmar que las desviaciones típicas no coinciden; y así decidir que test de medias aplicar.
Según los datos del análisis exploratorio vimos que los británicos tenían un mayor gasto medio. (Gráfico 5.18)

¿El gasto medio en agosto es mayor a la media de diciembre?
Al igual que para los turistas alemanes y británicos en este caso debemos realizar un test de diferencia de desviaciones típicas para aplicar el test correcto de diferencia de medias.

Según vemos en el análisis exploratorio el mes de agosto es el mes de más trafico de turistas lo que llevaría a pensar que es el mes con mayor ingreso por turismo. (Gráficos 5.14 y 5.16)

¿El gasto medio es mayor en la Comunidad de Madrid que es Cataluña?
En el gráfico 5.12 vemos que el gasto total es mayor en Cataluña que en la Comunidad de Madrid y en la matriz 5.5 vemos que los turistas de países con más afluencia hacia España (Reino Unido, Alemania y Francia) prefieren Cataluña a Madrid; lo que podría dar lugar a pensar que el gasto medio será mayor en Cataluña que en la Comunidad de Madrid.

Utilizaremos el test de diferencia de desviaciones típicas y de medias.

¿El gasto total es mayor en las Illes Balears o en la Comunitat Valenciana?
Por últimos tenemos la cuestión del gasto total en Baleares y la Comunitat Valenciana, elegidas por tener una afluencia de turistas similar.

Realizaré un análisis descriptivo de las variables, los principales estadísticos descriptivos que establecen dicho análisis serían la media muestral, varianza, desviación típica, y análisis de regresiones en gráficos, histrogramas y descripción de valores atípicos. Una vez visto esto, se pueden obtener estimaciones por intervalos de confianza para diferentes hipótesis y, también, habría que ver si la distribución que siguen los datos es una normal o no, aplicar un contraste de medias con media y desviaciones típicas conocidas y un test de varianzas para ver si existe o no diferencia entre ellas.

Para la visualización de las variables aprovecharé el paquete “ggplo2” y “GGally” de RStudio diseñado para la realización de gráficos, histogramas y gráficos de dispersión.

7. Aplicación de Inferencia Estadística

Script: AgrupaciónBasesDatos.R

En esta parte del trabajo nos vamos a centrar en responder las preguntas planteadas en la introducción, apoyándome en estadísticos como la media (para detectar en torno a qué valor se agrupan los datos), la desviación estándar (para hallar el grado de dispersión o concentración), varianza y utilizando métodos de inferencia estadística para confirmar o desmentir nuestras hipótesis.

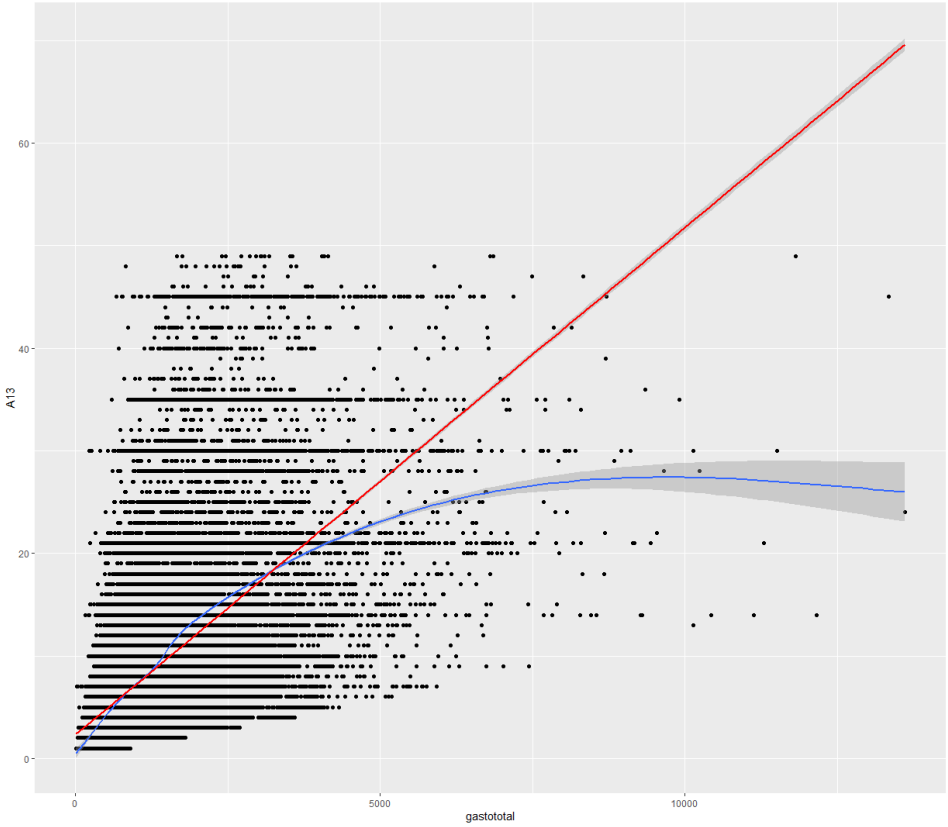
Empezamos con la base de datos "cln.data" que es el resultado del script "LimpiezaDatos.R"
Como vimos en el Análisis Exploratorio (Gráfico 5.7) tenemos en nuestra base de datos varios valores atípicos que nos complicarán e interferirán en los resultados que buscamos.
Para lidiar con estos "outliers" he decidido eliminarlos ya que no representan un parte muy importante de la base de datos (menos del 3%). Con lo dicho he filtrado la base de datos para valores de A13 (número de pernoctaciones) a menos de 50 y gasto medio menor que 900 euros.
Lo que nos da como resultado una base de datos de nombre "f.data".

```
f.data <- filter(cln.data, cln.data$A13 < 50)
f.data <- filter(f.data, f.data$gastomedio < 900)
```

7. 1 Relación entre A13 y Gasto Total

Si nos fijamos en el gráfico de dispersión 7.1 en el que representamos la relación entre el número de pernoctaciones y el Gasto Total de los turistas; podemos darnos cuenta de una relación positiva, cuando aumentan las pernoctaciones aumenta el gasto total, la cual confirmamos con la línea de regresión lineal (Roja) y la línea de regresión polinomial (Azul).

Gráfico 7.1 Dispersión entre A13 y gastototal



Ahora comparamos la variable de gasto medio de la base de datos sin filtrar y la filtrada.

Vemos que la mediana casi no se ha visto afectada mientras que la media ha tenido una reducción de casi 10 euros.

```
>   summary(cln.data$gastomedio)(SIN FILTRAR)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   3.617  112.415  159.939  197.297  227.200 5781.048
>   summary(f.data$gastomedio)(FILTRADA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.617 114.286 160.385 188.987 225.646 899.791
```

Notamos que la desviación típica también se ha visto afectada por los valores atípicos.

```
>   sd(cln.data$gastomedio)
[1] 167.3493
>   sd(f.data$gastomedio)
[1] 121.0687
```

7. 2 Gasto Medio mayor que 200 euros sin y con "outliers"

Realizando un contraste de hipótesis de tipo unilateral derecho, con un nivel de significación por defecto del 5%, y está definido como:

H0 :μ≤200 1-alfa=0.95 ; alfa=0.05 ;
H1:μ>200 s =121.0687

```
> t.test(gastomedio, mu= 200, stdev= sd(gastomedio), conf.level= 0.95, al-
ternative= "less")
     One Sample t-test

data:  gastomedio
t = -26.532, df = 85066, p-value < 2.2e-16
alternative hypothesis: true mean is less than 200
95 percent confidence interval:
     -Inf 189.6693
sample estimates:
mean of x
 188.9866
```

Con un valor cercano a 0 del p-value nos vemos obligados a rechazar la hipótesis nula, por lo que no podemos decir que el gasto medio de un turista en España en el 2019 sea mayor que 200 euros.

Si realizamos el test con la base de datos sin filtrar vemos que el p-value sigue estando muy cercano al 0, por lo que no hay duda de que se rechaza la hipótesis nula:

```
> t.test(gastomedio, mu= 200, stdev= sd(gastomedio), conf.level= 0.95, al-
ternative= "less")

      One Sample t-test

data:  gastomedio
t = -4.7664, df = 87054, p-value = 9.393e-07
alternative hypothesis: true mean is less than 200
95 percent confidence interval:
     -Inf 198.2295
sample estimates:
mean of x
 197.2966
```

7. 3 Los turistas Británicos son mejores que los Alemanes en cuanto a gasto medio

Ahora nos disponemos a comprobar si los turistas Alemanes (1) tienen un mayor gasto medio que los Británicos(8), nuestra hipótesis es que los Británicos tienen mayor gasto que los Alemanes, pero primero partimos con la premisa de que el gasto medio de ambos es igual:

H0 : μ1-μ8=0; de lo que μ1=μ8 1-alfa=0.95
H1: μ1- μ8≠0 alfa=0.05;

Creamos un subconjunto de datos para realizar el test, donde filtramos y nos quedamos solo con los turistas provenientes de Alemania y Gran Bretaña en un data frame llamado "f.data.i.a"

```
> f.data.i.a <- subset(f.data, pais == 1 | pais == 8)
```

Damos un vistazo rápido a los gráficos de cajas del gasto medio entre las dos nacionalidades sin preocuparnos por valores atípicos porque trabajaremos con la base de datos filtrada.

```
> aggregate(gastomedio, by=list(f.data.i.a$pais), mean)
  Group.1        x
1       1 158.5621
2       8 155.5308
> aggregate(gastomedio, by=list(f.data.i.a$pais), sd)
  Group.1        x
1       1 75.91273
2       8 75.07832
```

Podemos observar que las medias son muy parecidas y las desviaciones típicas casi coinciden, así que ya podemos darnos una idea del resultado. De todas formas vamos a revisar los gráficos de ambas nacionalidades para estár más seguros.

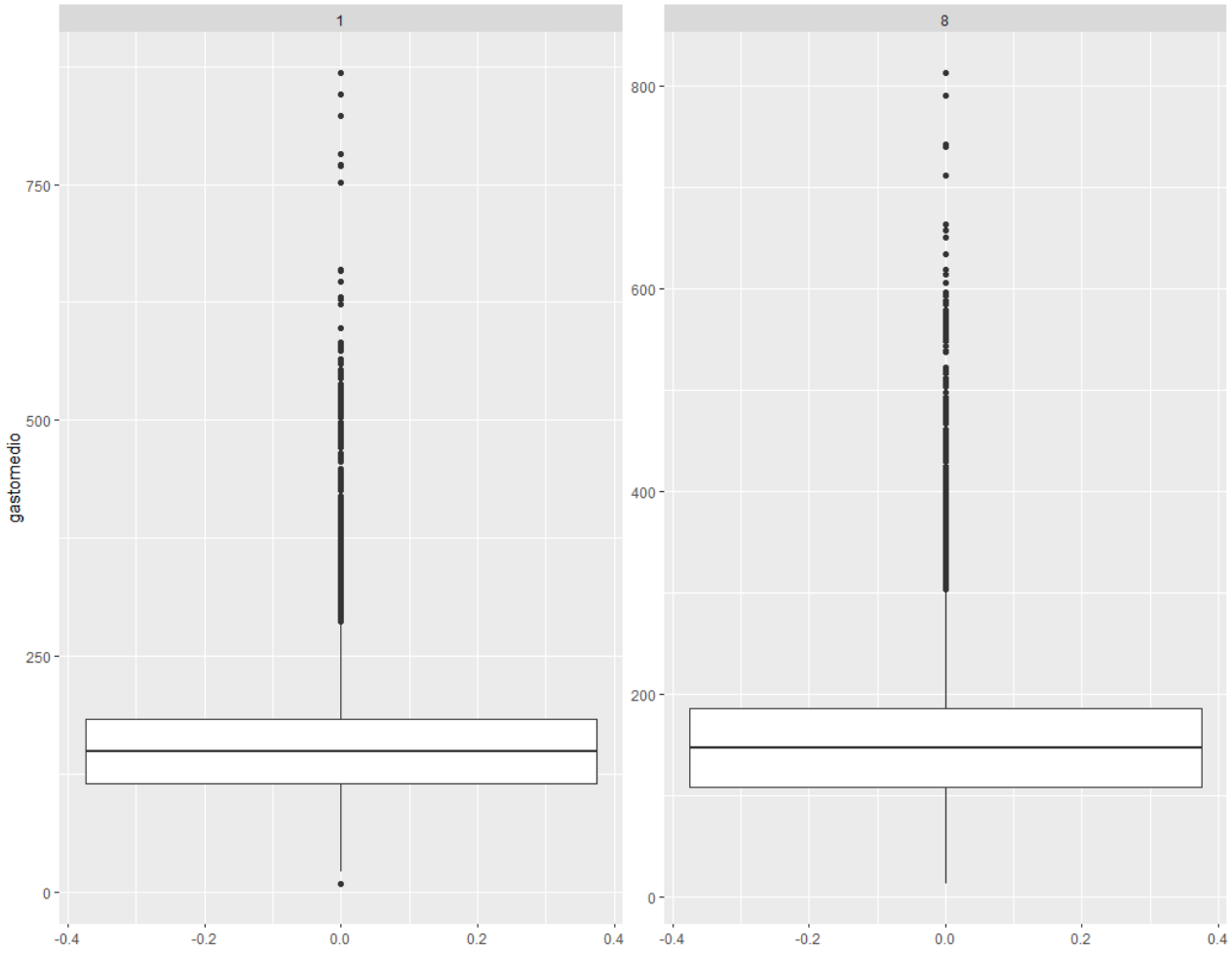


Gráfico 7.2 Box plot Gasto medio Alemania y Gran Bretaña

De todas maneras, para asegurarnos realizamos un test de diferencia de desviaciones típicas para estar seguros de que método de diferencia de medias utilizar.

```
> var.test(f.data.i.a$gastomedio~f.data.i.a$pais, alternative="two.sided",
conf.level=0.95)

     F test to compare two variances

data:  f.data.i.a$gastomedio by f.data.i.a$pais
F = 1.0224, num df = 10700, denom df = 18156, p-value = 0.1987
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9884639 1.0575585
sample estimates:
ratio of variances
          1.022351
```

Con un p-value de 0.1987 no podemos rechazar la hipótesis nula, por lo que deducimos que las desviaciones estándar deben ser iguales. Ni siquiera para un nivel de confianza del 99%.

```
> var.test(f.data.i.a$gastomedio~f.data.i.a$pais, alternative="two.sided",
conf.level=0.99)

     F test to compare two variances

data:  f.data.i.a$gastomedio by f.data.i.a$pais
F = 1.0224, num df = 10700, denom df = 18156, p-value = 0.1987
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
 0.9780607 1.0688837
sample estimates:
ratio of variances
          1.022351
```

Ahora que sabemos que las desviaciones estándar son iguales nos aseguramos que la opción var.equal del test de diferencia de medias se encuentre en TRUE.

```
> t.test(f.data.i.a$gastomedio~f.data.i.a$pais, alternative="two.sided",
conf.level=0.95, var.equal = TRUE)

     Two Sample t-test

data:  f.data.i.a$gastomedio by f.data.i.a$pais
t = 3.2993, df = 28856, p-value = 0.0009703
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.230492 4.832146
sample estimates:
mean in group 1 mean in group 8
       158.5621        155.5308
```

Con un p-value de 0.0009703 rechazamos la hipótesis nula de que ambos grupos de turistas tienen el mismo gasto medio, y de hecho estábamos equivocados con nuestra suposición, los Alemanes tienen un gasto medio mayor que los Británicos, aunque sea por muy poco (3 euros). De hecho, si hacemos la prueba (cambiamos mu=3) podemos ver que es así (obtenemos un p-value casi de 1):

```
> t.test(f.data.i.a$gastomedio~f.data.i.a$pais, alternative="t",
conf.level=0.95, var.equal = TRUE, mu=3)

     Two Sample t-test
data:  f.data.i.a$gastomedio by f.data.i.a$pais
t = 0.034088, df = 28856, p-value = 0.9728
alternative hypothesis: true difference in means is not equal to 3
95 percent confidence interval:
 1.230492 4.832146
sample estimates:
mean in group 1 mean in group 8
158.5621        155.5308
```

7. 4 El gasto medio de agosto es mayor que el de diciembre

Ahora nos disponemos a comprobar si el gasto medio del mes de agosto, el que recibió más turistas, es mayor que el gasto medio de diciembre, que es un buen mes pero ha recibido menos turistas:

H0 : μ8-μ12=0; de lo que μ8=μ12 1-alfa=0.95
H1: μ8- μ12≠0 alfa=0.05;

Creamos dos subconjuntos de datos:

f.data.agosto : Para los turistas que llegaron en agosto
f.data.diciembre : Para los turistas que llegaron en diciembre

```
f.data.agosto <- subset(f.data, mes == 8)
f.data.diciembre <- subset(f.data, mes == 12)
```

Comprobamos las medias de los conjuntos de datos filtrados y sin filtrar.

```
> mean(f.data.agosto$gastomedio)
[1] 172.2139
> mean(f.data.diciembre$gastomedio)
[1] 196.9606
> aggregate(cln.data$gastomedio, by=list(cln.data$mes), mean)
   Group.1        x

8        8 177.5086

12      12 204.2899
```

Vemos que ambas medias han disminuido debido al filtrado.
Comprobamos la medias de pernoctaciones.

```
> aggregate(f.data$A13, by=list(f.data$mes), mean)
   Group.1        x

8        8 9.505894
12      12 7.621488
```

Como podemos comprobar la media de pernoctaciones es más alta en agosto que en diciembre, esto seguramente se deba al tiempo meteorológico de ambos periodos. Mientras que en agosto es verano y los turistas deciden quedarse más tiempo para aprovecar el buen tiempo, en diciembre hace frio lo que puede condicionar a la baja el tiempo de estadía en el país.

También debemos tener en cuenta que en verano se suelen dar mayor tiempo de vacaciones, en la escuela, en verano las vacaciones son de 3 meses mientras que las vacaciones de navidad suelen ser 15 dias.

Todo esto son por supuesto suposiciones para intentar explicar la diferencia entre las medias no tenemos datos concretos para poder comprobar las suposiciones.

El paso siguiente es realizar un test de diferencia de desviaciones típicas para saber que método de diferencia de medias utilizar.

```
> var.test(f.data.agosto$gastomedio, f.data.diciembre$gastomedio, alterna-
tive="two.sided", conf.level=0.95)

     F test to compare two variances

data:  f.data.agosto$gastomedio and f.data.diciembre$gastomedio
F = 0.75933, num df = 9416, denom df = 6086, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7254090 0.7946648
sample estimates:
ratio of variances
         0.7593336
```

Con un p-value de prácticamente 0 rechazamos la hipótesis nula que las desviaciones son iguales y nos aseguramos que la opción var.equal del test de diferencia de medias se encuentre en FALSE.

```
> t.test(f.data.agosto$gastomedio, f.data.diciembre$gastomedio, alterna-
tive="two.sided", conf.level=0.95)

     Welch Two Sample t-test

data:  f.data.agosto$gastomedio and f.data.diciembre$gastomedio
t = -12.5, df = 11704, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.62730 -20.86611
sample estimates:
mean of x mean of y
 172.2139  196.9606
```

Con un p-value de prácticamente 0 rechazamos la hipótesis nula que las medias son iguales.
Comprobamos que la media en agosto es mucho más baja que en diciembre, vimos antes que la media de A13 en agosto era más alta, esto definitivamente ha influido en el resultado.
Otra cosa que puede haber influido es que en Diciembre se suelen hacer las compras de Navidad, lo que equivale a un mayor gasto, pero como he dicho antes son todo suposiciones.

```
> t.test(f.data.agosto$gastototal, f.data.diciembre$gastototal, alterna-
tive="two.sided", conf.level=0.95)

data:  f.data.agosto$gastototal and f.data.diciembre$gastototal
t = 10.718, df = 12590, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 120.9571 175.1012
sample estimates:
mean of x mean of y
 1273.944  1125.915
```

Con el gasto total ocurre lo mismo debemos rechazar la hipotesis nula con un p-value cercano a 0.

7. 5 Gasto medio es menor en Comunidad de Madrid que Cataluña

Comprobamos si el gasto medio de la Comunidad de Madrid es menor que en Cataluña:

H0 : μ9-μ13=0; de lo que μ9=μ13 1-alfa=0.95
H1: μ9-μ13≠0 alfa=0.05;

Creamos un subconjunto de datos para realizar el test, donde filtramos y nos quedamos solo con los turistas que eligieron como destino la Comunidad de Madrid o Cataluña en un data frame llamado "f.data.m.c"

```
f.data.m.c <- subset(f.data, ccaa == 9 | ccaa == 13)
```

Realizamos un test de diferencia de desviaciones típicas para estar seguros de que método de diferencia de medias utilizar.

```
> var.test(f.data.m.c$gastomedio~f.data.m.c$ccaa, alternative="two.sided",
conf.level=0.95)

      F test to compare two variances

data:  f.data.m.c$gastomedio by f.data.m.c$ccaa
F = 0.69084, num df = 19451, denom df = 8041, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6658221 0.7166042
sample estimates:
ratio of variances
         0.6908448
```

Con un p-value de prácticamente 0 rechazamos la hipótesis nula que las desviaciones son iguales y nos aseguramos que la opción var.equal del test de diferencia de medias se encuentre en FALSE.

```
> t.test(f.data.m.c$gastomedio~f.data.m.c$ccaa, alternative="less", conf.
level=0.95, var.equal = FALSE)

      Welch Two Sample t-test

data:  f.data.m.c$gastomedio by f.data.m.c$ccaa
t = -37.118, df = 12857, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -76.97928
sample estimates:
 mean in group 9 mean in group 13
        224.5514         305.1004
```

Con un p-value de prácticamente 0 rechazamos la hipótesis nula que las medias son iguales.
Contrariamente a lo que pensabamos hemos visto que los resultados arrojan que el gasto medio en la Comunidad de Madrid es 81 euros mayor que en Cataluña.

7. 6 Gasto Total es mayor en Baleares que en Comunitat Valenciana

Comprobamos si el gasto total de las Illes Balears es mayor que el de la Comunitat Valenciana:

H0 : μ4-μ10=0; de lo que μ4=μ10 1-alfa=0.95
H1: μ4-μ10≠0 alfa=0.05;

Creamos un subconjunto de datos para realizar el test, donde filtramos y nos quedamos solo con los turistas que eligieron como destinolas Illes Balears o la Comunitat Valenciana en un data frame llamado "f.data.b.v"

```
f.data.b.v <- subset(f.data, ccaa == 4 | ccaa == 10)
```

Antes que nada revisamos los datos de las medias del gasto total en ambas comunidades.

```
> aggregate(f.data.b.v$gastototal, by=list(f.data.b.v$ccaa), mean)
  Group.1        x
1       4 1089.118
2      10 1013.842 = 4 | ccaa == 10)
```

También sus desviaciones típicas.

```
> aggregate(f.data.b.v$gastototal, by=list(f.data.b.v$ccaa), sd)
  Group.1        x
1       4 525.3589
2      10 782.3397
```

Realizamos un test de diferencia de desviaciones típicas para estar seguros de que método de diferencia de medias utilizar.

```
> var.test(f.data.b.v$gastototal~f.data.b.v$ccaa, alternative="two.sided",
conf.level=0.95)

      F test to compare two variances

data:  f.data.b.v$gastototal by f.data.b.v$ccaa
F = 0.45094, num df = 14378, denom df = 13538, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4362156 0.4661595
sample estimates:
ratio of variances
         0.4509429
```

Con un p-value de prácticamente 0 rechazamos la hipótesis nula que las desviaciones son iguales y nos aseguramos que la opción var.equal del test de diferencia de medias se encuentre en FALSE.

```
> t.test(f.data.b.v$gastototal~f.data.b.v$ccaa, alternative="greater", conf.
level=0.95, var.equal = FALSE)

      Welch Two Sample t-test

data:  f.data.b.v$gastototal by f.data.b.v$ccaa
t = 9.3801, df = 23488, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 62.07558      Inf
sample estimates:
 mean in group 4 mean in group 10
        1089.118         1013.842
```

Con un p-value de prácticamente 0 rechazamos la hipótesis nula que las medias son iguales.
Algo que habíamos ya intuido mirando los datos, podemos asegurar que el Gasto total es mayor en Las islas Baleares que en la Comunidad Valenciana.

8. Conclusiones

Script: Inferencia.R

La investigación sobre este periodo de datos finaliza haciendo uso de los paquetes gráficos propuestos por las librerías de R-Studio.
Como tenemos variables cuantitativas, aplicando en la consola los correspondientes comandos descritos en el script se generan varios conjuntos gráficos de los que podemos extraer las siguientes conclusiones:

Primero que nada presentaré una matriz de correlaciones de las variables cuantitativas de la base de datos (A13, gasto total y gasto medio):

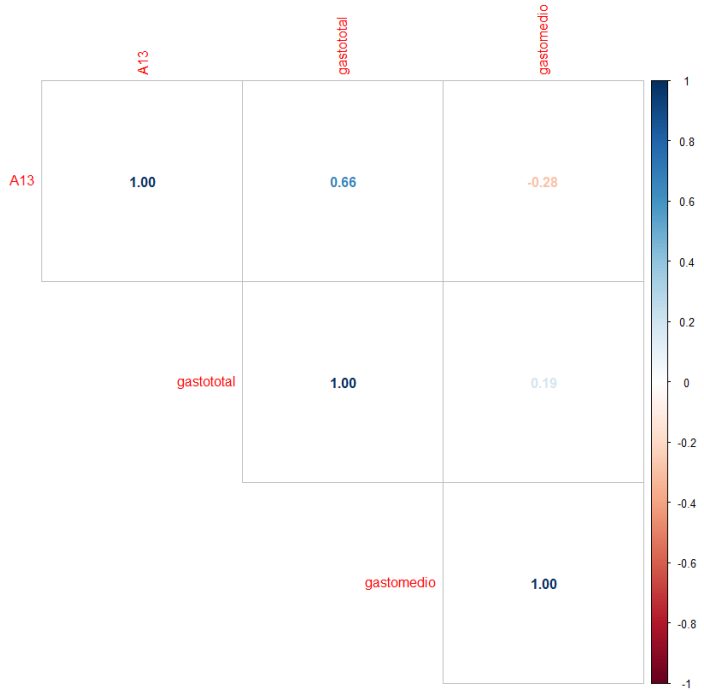


Gráfico 8.1 Matríz de Correlaciones Variables Cuantitativas

En la matriz de correlaciones podemos observar la gran correlación positiva de A13 con el gasto total, lo que tiene sentido porque a mayor número de pernoctaciones la posibilidad de realizar más gasto aumenta, contando con el gasto de alojamiento (el mayor gasto para la mayoria de turistas).

Por otra parte vemos que la relación entre A13 y el gasto medio es negativa, algo evidente ya que para un mismo valor de gasto total menor va a ser el gasto medio a medida que aumentamos las pernoctaciones (Gasto Medio = Gasto Total / A13).

Por último tenemos la relación positiva, pero menor, de Gasto Total con Gasto Medio. Sabemos que a mayor gasto total más posibilidades de que aumento el gasto medio, pero siempre tenemos que tener en cuenta las pernoctaciones (A13).

8.1 Gasto Medio Alemania y Gran Bretaña

Gráfico 8.2 Alemanes vs Británicos

En el gráfico 8.2 tenemos las correlaciones de las variables pais, A13, gastototal y gastomedio con respecto a dos nacionalidades, los turistas alemanes y los británicos.
Podemos observar que ambos paises coinciden en que la mayor parte de los turistas decide quedarse en España durante 7 dias y las medias del gasto medio y total coinciden aproximadamente.
De hecho son dos demografías bastante similares en casi todos los arpectos de la gráfica, la unica diferencia notable es el volumen de viajeros, que es superior por parte de los británicos.

Pasamos a analizar las correlaciones entre A13 y el gasto total. Vemos que la correlación de los turistas alemanes (0.593) es bastante mayor a la de los británicos (0.493) lo que conlleva a pensar que los turistas germanos son más propensos a aumentar su gasto a medida que su estancia de alarga.
Pasando ahora a la correlación gasto medio A13 es prácticamente igual en ambas nacionalidades (-0.541 Alemania y -0.553 Gran Bretaña) ligeramente superior la de los británicos.

8.2 Gasto medio Comunidad de Madrid y Cataluña

Gráfico 8.3 Cataluña vs Comunidad de Madrid

En el gráfico 8.3 tenemos las correlaciones de las variables ccaa, A13, gastototal y gastomedio con respecto a dos comunidades autónomas, Comunidad de Madrid y Cataluña.
Como primera observación podemos decir que las variables A13, gasto medio y total tienen una mayor dispersión para la C.M que para Cataluña aunque la cantidad de viajeros que recibe cada comunidad es cláramente mayor en Cataluña.

Prosiguiendo con las correlaciones son bastante similares para ambas comunidades, menos la correlación A13 con el gasto total que es bastante más alta la de la C.M que la de Cataluña (0.479 y 0.362 respectivamente). Podemos concluir lo mismo que con los turistas Alemanes, para un aumento de una pernoctación en ambas comunidades es más probable que un turista gaste más dinero en la Comunidad de Madrid que en Cataluña.

8.3 Gasto Total Islas Baleares y Comunidad Valenciana

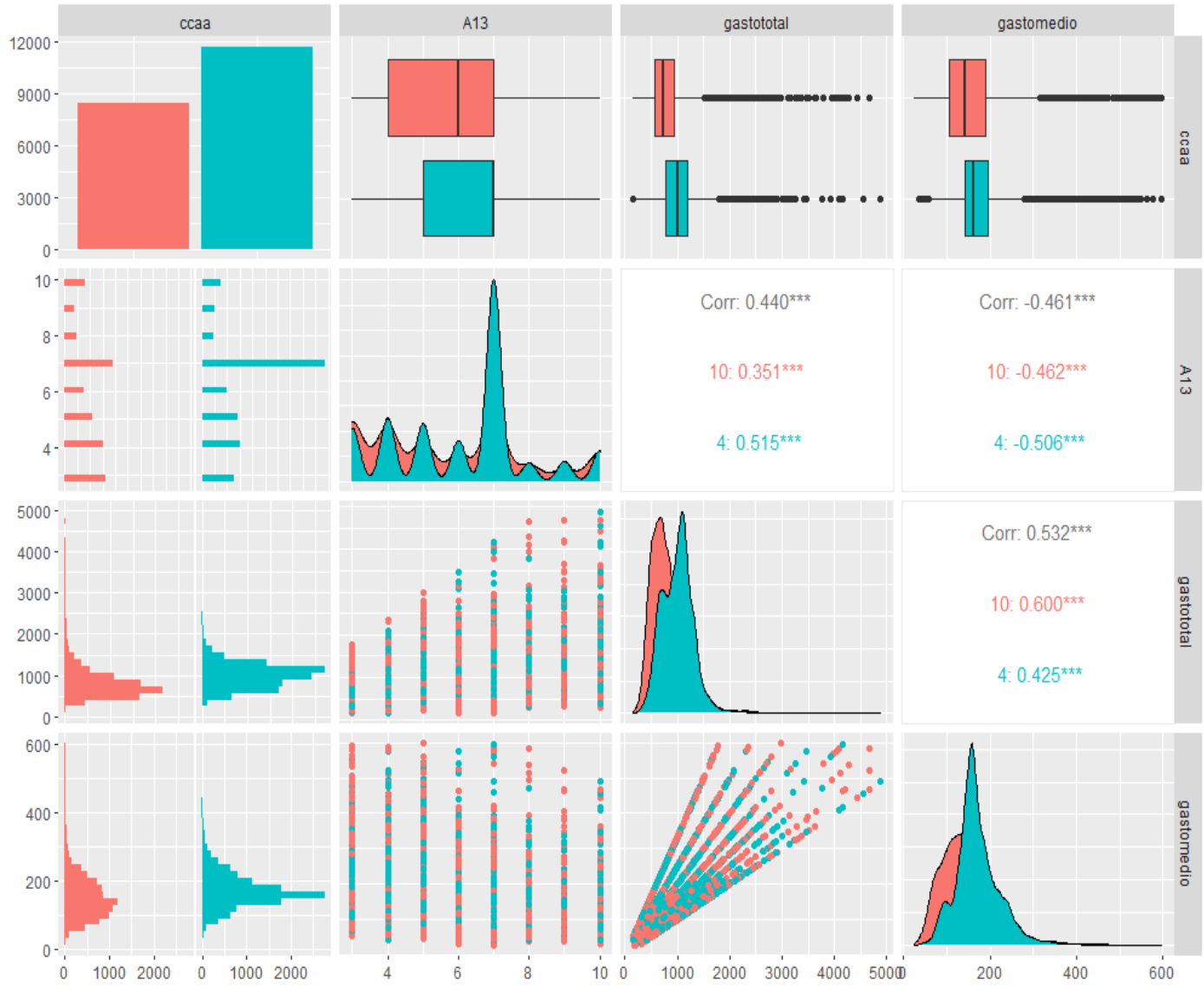


Gráfico 8.4 Baleares vs Valencia

En el gráfico 8.4 tenemos las correlaciones de las variables ccaa, A13, gastototal y gastomedio con respecto a dos comunidades autónomas, Illes Balears y Comunitat Valenciana.
Dando un vistazo rápido se puede observar que las variables de ambas comunidades tiene una dispersión similar, menos A13 en la que la comunidad Valenciana claramente supera a las islas.
Si hablamos de medias de las variables nos damos cuenta que la C.V está siempre por detrás de las Baleares (lo podemos ver en la primera fila).

Continuando con las correlaciones vemos que las Islas tienen una mayor correlacion A13-GastoTotal (0.2 veces superior) y también es mayor su correlación negativa entre A13 y Gasto Medio (-0.462 de C.V vs -0.506 de las Islas Baleares). Con lo que podemos concretar que la variable A13 tiene un mayor peso en la turistas que visitan las islas que en los que visitan la C.V.
Terminando con la correlación entre Gasto total y medio vemos que es mayor en la Comunidad Valenciana que en las Baleares.

8.4 Gasto medio en Agosto y Diciembre

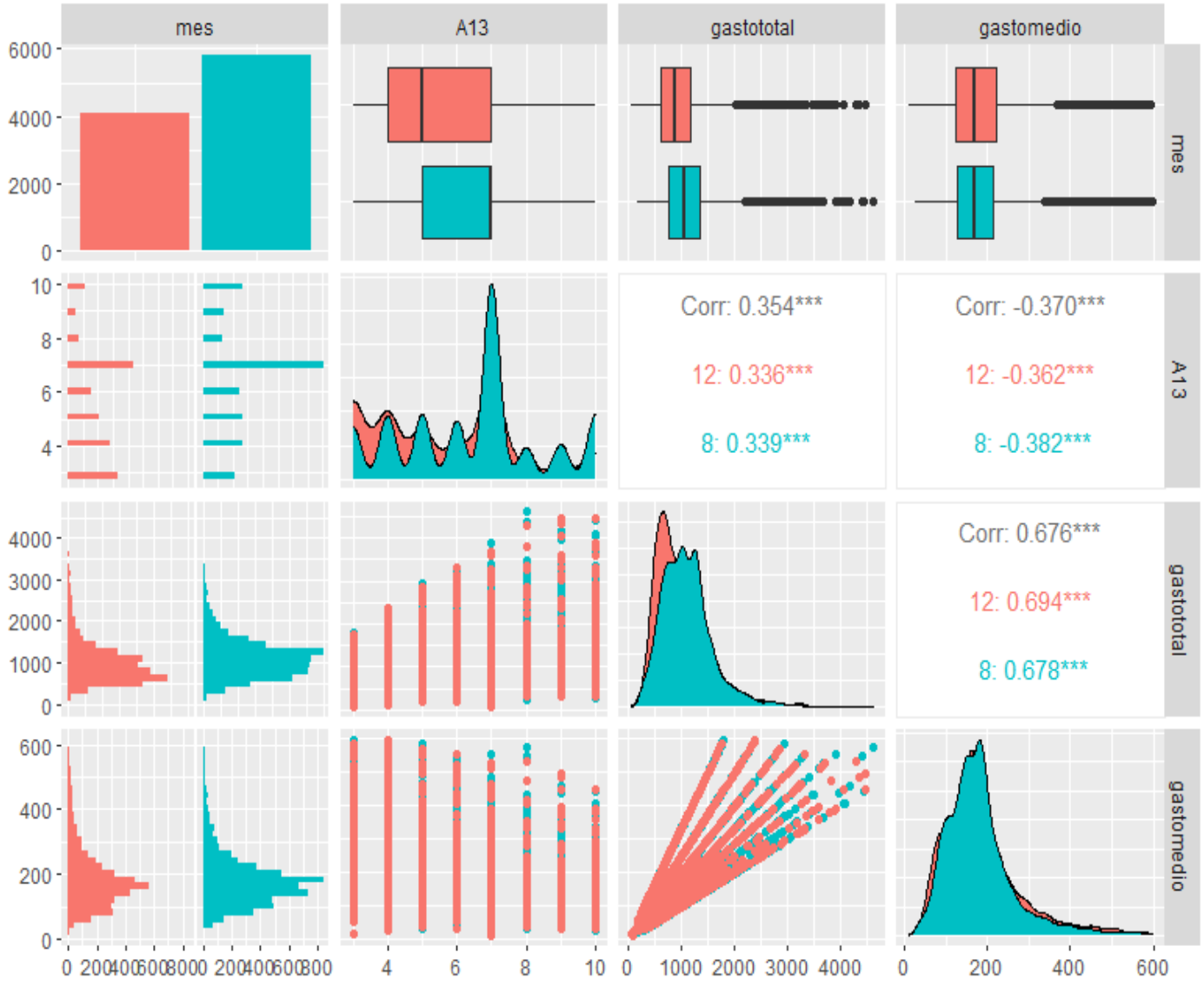


Gráfico 8.5 Agosto vs Diciembre

En el gráfico 8.5 tenemos las correlaciones de las variables mes, A13, gastototal y gastomedio con respecto a dos meses, Agosto y Diciembre.
Aparte de la mayor dispersión de la variable A13 de diciembre con respecto a agosto y la mayor afluencia de turistas en Agosto, por lo demás los dos meses se comportan bastante parecidos.
Incluso sus covarianzas son prácticamente iguales.

8.5 Resumen de Conclusiones

Haciendo un resumen de todas las preguntas planteadas tenemos:

- El gasto medio del turista que visitó España en 2019 es menos de 200 euros al día.
- Los turistas Alemandes gastan de media 3 euros más al día que los Británicos.
- El gasto medio en Agosto es menor que el gasto medio en Diciembre.
- El gasto medio es mayor en la Comunidad de Madrid que en Cataluña.
- El gasto total es mayor en las Islas Baleares que en la Comunidad Valenciana.

Entonces "cogido con pinzas" podemos decir que para el PIB Español es mejor un turista Aleman que venga de vacaciones en Diciembre a la Comunidad de Madrid o a las islas Baleares.

Este ha sido un proyecto muy interesante que ha despertado en mí las ganas de seguir investigando y de utilizar la inferencia estadística para la comprobación de hipótesis. Ya tengo en mente próximos proyectos con los que seguir aprendiendo y perfeccionando estas herramientos y mis conocimientos.

Anexo I

Nombre	Tipo	Longitud	Valores	Literal
mm_aaaa	string	6.0		Mes y año de referencia
A0	string	1.0	2: Egatur	Encuesta de procedencia
A0_1	string	14.0		Identificador cuestionario
A0_7	string	1.0	2: Turista no residente (no tránsito), 8: Turista no residente en tránsito	Código de cuestionario (TEN)
A1	string	1.0	1: carretera, 2: aeropuerto, 3: puerto, 4: tren	Vía de salida
pais	string	2.0	01:Alemania. 02:Bélgica. 03:Francia. 04: Irlanda. 05: Italia. 06: Países Bajos. 07: Portugal. 08: Reino Unido. 09: Suiza. 10:Rusia. 11: Países Nórdicos (Dinamarca, Finlandia, Noruega, Suecia). 12: Resto de Europa. 13: EEUU. 14: Resto de América. 15:Resto del mundo	Pais de residencia habitual
ccaa	string	2.0	01: Andalucía. 02: Aragón. 03: Principado de Asturias. 04: Illes Balears. 05: Canarias. 06: Cantabria. 07: Castilla y León. 08: Castilla-La Mancha. 09: Cataluña. 10: Comunitat Valenciana. 11: Extremadura. 12: Galicia. 13: Comunidad de Madrid. 14: Región de Murcia. 15: Comunidad Foral de Navarra. 16: País Vasco. 17: La Rioja. 18: Ceuta. 19: Melilla	Comunidad Autónoma de destino principal del viaje
A13	positiveinteger	3.0		Total pernoctaciones
aloja	string	1.0	1: Hoteles y similares, 2: Resto de mercado, 3: Alojamiento no de mercado	Alojamiento principal
motivo	string	1.0	1: Ocio/vacaciones, 2: Negocios, 3: Resto	Motivo principal del viaje
A16	string	1.0	1: Sí, 6: No	Paquete turístico
gastototal	decimal	12		Gasto total del viaje/excursión
factoregatur	decimal	12		Factor de elevación de Egatur