**Descripción y respuestas de la Práctica a realizar**

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Las diferentes tareas a realizar (y **justificar**) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Es un dataset llamado tourisData.csv, recogido de Kaggle y en el podemos ver datos sobre el número de turistas internacionales que llegan a Brasil (a cada uno de los estados de Brasil) mes a mes entre los años 1989 y 2015. También podemos por qué medio han llegado (aéreo, por tierra, por rio o por mar).

Los datos que contempla cada registro son:

Continent: Continente de procedencia del turista

Country: País de procedencia del turista

State of arrival: estado brasileño por donde el turista ha llegado en Brasil

Way in: modo en el que ha entrado en Brasil

Year: año del dato

Month: mes del dato

Count: número de turistas internacionales que han cumplido las condiciones del registro.

Son datos interesantes ya que ayudan a ver eso, de donde son los turistas que llegan a Brasil, de qué forma llegan, en qué mes o meses del año hay más turistas y que estado visitan en primer o único lugar.

Puede ayudar a ver si hay algún problema con las conexiones desde un país a Brasil de forma que no lleguen clientes de ese país, o bien, si hay algún mes en el que no vienen clientes de un país. Se pueden realizar estudios para poder ver por qué causa puede ser y si se puede hacer algo para que mejore o cambie.

Lo mismo si hay países de los que no viajan a Brasil el ver el por qué y si puede mejorarse.

Este dataset es útil, principalmente**,** para estudiar acciones a realizar para conseguir que viajen más turistas extranjeros a Brasil.

1. Limpieza de los datos.

2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

En el Dataset escogido existe un campo, el Country, que puede llevarnos directamente a descubrir el Continent (que es otro de los campos), entonces el campo Continent no es imprescindible y podemos quitarlo.

En primer lugar estudiaremos la casuística ligada a ¿Cuántos turistas internacionales han visitado Brasil dividiéndolo por meses (acumulando todos los años )?, todo ello para poder ver la tendencia que hay entre los diferentes meses del año.

Para dar respuesta a mi problema serían suficientes los campos Month y Count. Acumulando el campo Count por cada mes.

En segundo lugar, vamos a realizar el estudio de año a año como se ha comportado la visita de Turistas a Brasil viéndolos país a país de procedencia.

Aunque vemos que en el dataset escogido se podrían realizar muchos otros estudios ya que dispone de más datos que podrían ser útiles según el estudio que se desee realizar. Según la pregunta que queramos resolver se pueden utilizar unos campos u otros.

2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

Los datos sí contienen ceros pero el cero es un valor posible. En cuanto a elementos vacíos, el campo Count lo que si contiene es valores NA que lo que haremos será trasformarlos en 0´s, En el caso de los campos Continent y Country, contienen valores “Not informed”. Para el Continent ya hemos dicho que nos olvidamos de él y lo eliminamos.

En cuanto a Country, para el primer estudio no vamos a utilizarlo, en cuanto al segundo estudio estos registros no suponen un problema ya que suponen un 1% de los registros y si vemos el sumatorio de los Turistas que suponen es muy pequeño, nos supondrá un error muy bajo. Tambien observamos que este error en el que el Atributo Country esta con valor “Not informed” esta distribuido en todos los años.

Para nuestro primer estudio recogeremos los campos Year, Month y Count, estos campos si tienen unidades y por ello los tendremos en cuenta.

El porcentaje de estos registros con el Count con valor NA es bajo: 1,16%.

Para el segundo estudio recogeremos los campos Country, Year, Month y Count. Son los que nos hacen falta para ver para cada año como se ha comportado la visita de Turistas desde los diferentes países, el país “Not informed” será un país adicional con su acumulado del año.

En cuanto a valores extremos No veo los haya para que resulten un problema para los cálculos que necesitamos realizar.

3. Análisis de los datos.

* 3.1.  Selección de los grupos de datos que se quieren analizar/comparar.

Primero, para los dos estudios, pondremos a 0 todos los valores de Count que tienen como valor NA.

Como hemos indicado, para el primer estudio, recogeremos únicamente los campos Month y Count, de forma que podamos ver cómo se ha comportado el número de turistas internacionales que han visitado Brasil acumulando los meses de todos los años. Todo ello para poder ver durante todos los años que meses se reciben más turistas.

Los pasos que realizaremos para poder llegar a responder como se comporta la visita de turistas extranjeros a Brasil durante los diferentes meses del año serán:

a. Leer los datos y cambiar los NA del campo Count por 0´s.

b. Recoger del resultado solo los campos Month y Count.

c. Sumatorio del campo Count para cada Month de todos los Year.

d. Sustitución valores del campo Mes, de forma que sean numéricos, sustituiremos: janeiro, fevereiro, março, abril, maio, junho, julho, agosto, setembro, octubre, novembro, dezembro por : 1,2,3,4,5,6,7,8,9,10,11,12 respectivamente. De esta forma podremos aplicar un orden numérico y nos será mucho más fácil estudiar resultados.

Para el segundo estudio los pasos que realizaremos serán:

a. Leer los datos y cambiar los NA del campo Count por 0´s.

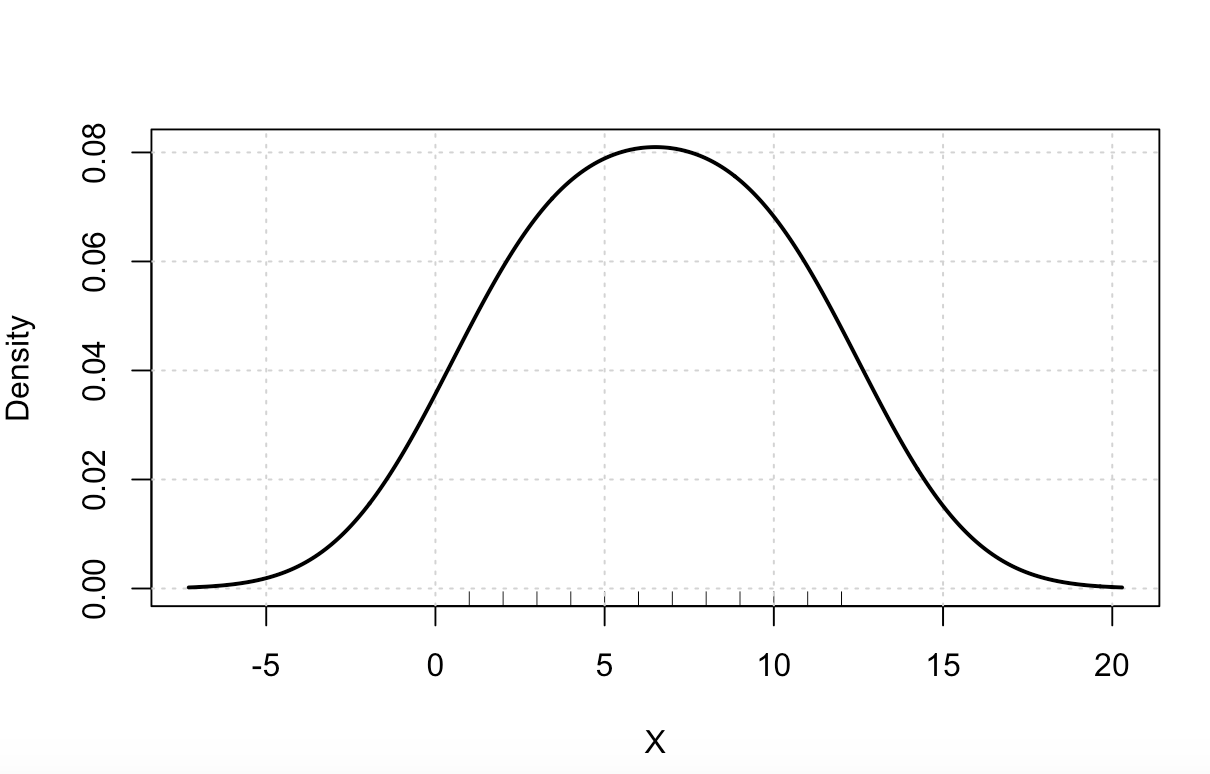
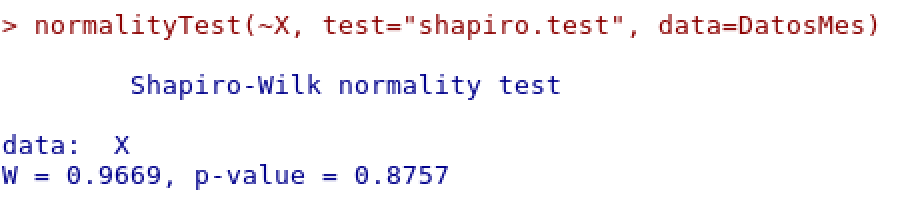
b. Trabajamos en una sentencia repetitiva utilizando los valores de Year, desde 1989(primero) a 2015(último) de forma que para cada vez recogemos solo los campos Country, Year, sumando los Count, agrupándolos por Country y Year.

c. Para cada resultado de la repetición realizamos estudio de los datos obtenidos. Hemos guardado 2 archivos para cada Year, estos los genera el mismo proceso en R. El primero con el dibujo de la grafica de distribución de puntos con el Count del año por país y el segundo con el fichero .csv de los datos que se han tratado con los acumulados del año.

* 3.2.  Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

Para los dos casos como comprobación de la normalidad he aplicado el test de Saphiro.

Vemos que la gráfica sobre el Count en los datos preparados sigue una distribución normal.

* + - * 1. 
        2. Y comprobándolo con el test de Shapiro y me ha dado los datos siguientes:
      1. 

Para el segundo estudio hemos aplicado también el test de Shapiro. Para cada uno de los años. Hemos visto las gráficas que parece siguen una distribución Normal pero al aplicar el Test el valor de p que nos ha salido ha sido:

Desde 1989 a 2003: p=0,084

Desde 2004 a 2006: p=0,058

Desde 2007 a 2015: p=0499 y podemos rechazar que la distribución sea de tipo normal.

* 3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

Para el primer estudio como el p-value es mayor que 0,05 (0,8757>0,05) aceptaremos la hipótesis nula, podemos afirmar que nuestros datos se distribuyen siguiendo una normal.

Para el segundo estudio, tal como hemos comentado en el punto anterior, hemos realizado el Test de Shapiro y hemos visto que:

Desde 1989 a 2003: p=0,084 podemos afirmar que sigue una distribución Normal.

Desde 2004 a 2006: p=0,058 podemos afirmar que sigue una distribución normal

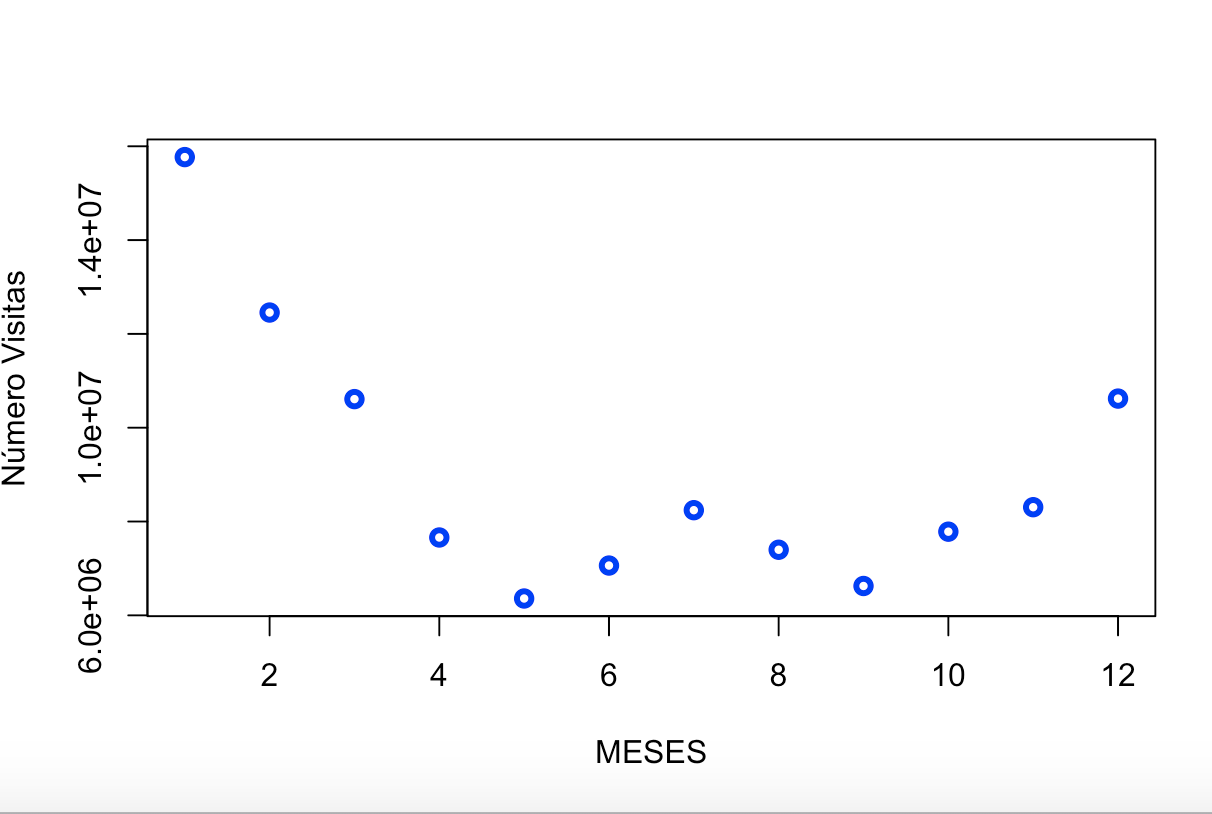
Desde 2007 a 2015: p=0499 y para estos años podemos rechazar que la distribución sea de tipo normal Se deberá estudiar más a fondo que está ocurriendo en estos años con las visitas de Turistas a Brasil.

1. Representación de los resultados a partir de tablas y gráficas.

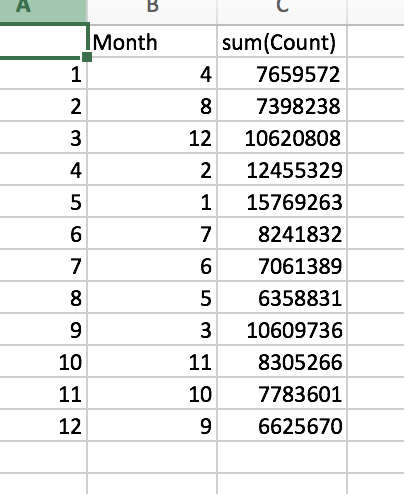
Después de trabajarlo con el programa en R copiado en la pregunta 5.

Para el primer estudio en la gráfica vemos como se ha comportado el número de turistas internacionales acumulados desde el año 1989 al 2015 en los diferentes meses del año

Gráfico 1

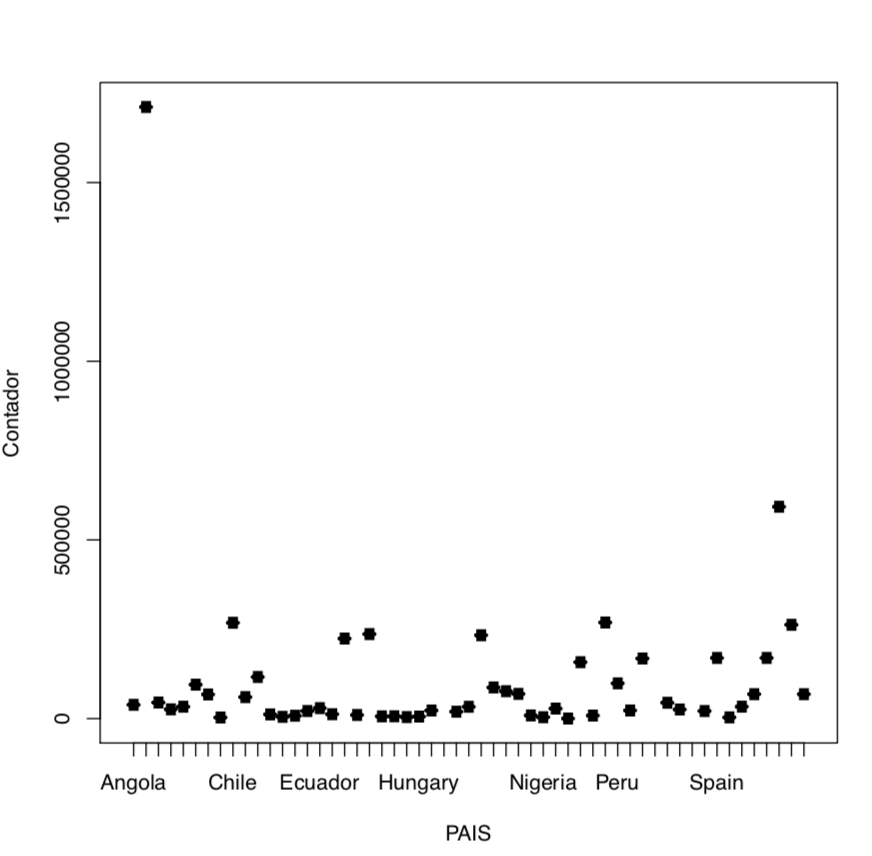


El fichero de donde salen los datos:



Para el segundo estudio si vemos para el año 2013 tenemos el gráfico de distribución de puntos y los datos:

Grafico 2:



Y a nivel del conjunto de datos:



Disponemos de un fichero con el gráfico y el conjunto de datos para cada año, esto nos ayudará a ver que esta ocurriendo con los países y nos ayudará a ver en el fichero de los puntos que país es el que no va bien ya que están codificados desde el 1 al 53.

1. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las  conclusiones? ¿Los resultados permiten responder al problema?

Para el primero de los estudios:

Analizando el gráfico 1 donde vemos la progresión de Count mes a mes, acumulando todos los años y países vemos que:

Nos supone pensar que los meses fuertes en Brasil son principalmente Enero y Febrero. Vemos que en Diciembre va creciendo la ocupación, Enero está en el máximo y Febrero, es el segundo mes más fuerte pero va bajando. Hay algo que me resulta extraño ya que Febrero es el mes de los carnavales y parecería que debería ser el mes con más ocupación extranjera.

No podemos perder de vista que hemos puesto a 0 todos los datos con valor NA en el campo Count. Estos datos estaban distribuidos en los años 1996, 1999, 2004 y 2007 pero a nivel de meses tocaban todos los meses. Lo que no hemos estudiado es de que países eran principalmente los turistas de los que nos faltaba el número ya que dependiendo de los países podemos pensar que la cantidad que se nos ha escapado es más o menos importante. De todas formas, el margen de error es pequeño ya que hemos dicho que eran un 1,6% de los datos.

El Grafico 1 supone el que veamos que ocurre en los meses más bajos de ocupación y podamos planificar acciones para que esta ocupación crezca.

Podemos ver sobre qué países podemos realizar acciones y si hay que mejorar algún tipo de canal de acceso a Brasil pero esto ya sería a consecuencia de más estudios a realizar con los datos que en este punto no hemos utilizado.

La forma de trabajar con ellos podría ser parecida a la utilizada en este caso.

El clima es uno de los factores importantes cuando hablamos de viajar, aunque en Brasil el clima es bueno todo el año y no debería ser un factor importante.

Para el segundo de los casos en estudio:

Vemos que año a año van creciendo el número de turistas internacionales que van a Brasil, nada de extrañar, lo que si llama la atención es que los turistas de EEUU bajan y eso no parece normal.

Año a año aparecen mas países como procedentes de turistas que viajan a Brasil, países de todo el mundo. En el 1989 teniamos 45 y en el 2015 tenemos 52. Tambien tenemos un país denominado “Other Countries” y el “Not Informed”.

El valor del contador de Turistas para “Other Countries” va creciendo y se debería ver a que es debido ya que para el año 2015 ya tiene un valor de 173.682 visitantes.

1. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís,  también podéis trabajar en Python.
2. library(stats)
3. library(readr)
4. ###install.packages("dplyr", dependencies=TRUE)
5. ###install.packages("sqldf", dependencies=TRUE)
6. ###install.packages("https://cran.r-project.org/src/contrib/nortest\_1.0-4.tar.gz", repos=NULL)
7. library(proto)
8. library(dplyr)
9. library(RSQLite)
10. library(gsubfn)
11. library(sqldf)
12. library(nortest)
13. library(splines)
14. library(RcmdrMisc)
15. library(car)
16. library(sandwich)
17. library(effects)
18. library(carData)
19. setwd("/Users/mariaantoniallabresespina/Documents/Tipologia y ciclo de vida de los datos/Practica 2")
20. touristData <- read.csv("touristData.csv")
21. View(touristData)
22. ### Queremos poner 0 en todos los campos Count que contienen NA
23. touristData$Count[ is.na(touristData$Count)] <- 0
24. ### Escogemos solo las columnas Month, Count sumadas por cada mes
25. touristSumaMes <- sqldf("select Month, sum(Count) from touristData group by Month")
26. View (touristSumaMes)
27. ###Cambiamos de nombre de mes a numero de mes en el fichero resultante
28. touristSumaMes$Month[touristSumaMes$Month == "janeiro"] <- 1
29. touristSumaMes$Month[touristSumaMes$Month == "fevereiro"] <- 2
30. touristSumaMes$Month[touristSumaMes$Month == "abril"] <- 4
31. touristSumaMes$Month[touristSumaMes$Month == "maio"] <- 5
32. touristSumaMes$Month[touristSumaMes$Month == "junho"] <- 6
33. touristSumaMes$Month[touristSumaMes$Month == "julho"] <- 7
34. touristSumaMes$Month[touristSumaMes$Month == "agosto"] <- 8
35. touristSumaMes$Month[touristSumaMes$Month == "setembro"] <- 9
36. touristSumaMes$Month[touristSumaMes$Month == "outubro"] <- 10
37. touristSumaMes$Month[touristSumaMes$Month == "novembro"] <- 11
38. touristSumaMes$Month[touristSumaMes$Month == "dezembro"] <- 12
39. touristSumaMes$Month[touristSumaMes$Month >= "mar"] <- 3
40. ### Dibujamos resultado
41. plot(touristSumaMes[,1],touristSumaMes[,2],col="blue", lwd=3, xlab="MESES",ylab="Número Visitas")
42. ### Guardamos fichero .csv
43. write.csv(touristSumaMes,"Turistas en Brasil por mes.csv")
44. DatosMes <- read.table("Turistas en Brasil por mes.csv",
45. header=TRUE, sep="", na.strings="NA", dec=",", strip.white=TRUE)
46. ### Dibujamos la Normal de los datos
47. densityPlot( ~ X, data=DatosMes, bw="SJ", adjust=1, kernel="gaussian")
48. ### Aplicamos test de Shaoiro
49. normalityTest(~X, test="shapiro.test", data=DatosMes)
50. ### Desde aqui realizamos nuevos estudios viendo para cada año como se ha distribuido la visita de Turistas a Brasil según pais de procedencia
51. touristData <- read.csv("touristData.csv")
52. View(touristData)
53. ### Queremos poner 0 en todos los campos Count que contienen NA
54. touristData$Count[ is.na(touristData$Count)] <- 0
55. ### Quitamos columna Continent que es la primera
56. touristData <- touristData[,-1]
57. View(touristData)
58. ### Escogemos solo las columnas Pais, Año y Count acumulado(numero de turistas que visitan Brasil)
59. ### de cada año agrupando por Pais y Año para ver su evolución. Hacemos una repetitiva para cada año
60. a<- min(touristData$Year)
61. b<- max(touristData$Year)
62. i<-a
63. for (i in a:b)
64. {seleccion<-paste ("select Country, Year, sum(Count) from touristData where Year=",i," group by Country, Year")
65. touristporpaisSumaMes <- sqldf(seleccion)
66. plot(touristporpaisSumaMes[,1],touristporpaisSumaMes[,3],col="blue", lwd=3, xlab="PAIS",ylab="Contador")
67. Resultado<- paste("Grafica turistas a Brasil anyo",i)
68. Resultadopdf<- paste(Resultado,".pdf")
69. Resultadocsv<- paste(Resultado,".csv")
70. ### Guardamos el resultado de la grafica por cada año y todos los paises en un pdf, es un pdf donde salen todas las cantidades de clientes acumuladas
71. ### por pais, el nombre del pdf que se guarda incluye el año del resultado.
72. dev.print(pdf, file=Resultadopdf, width=7, height=7,pointsize=12)
73. ### Guardamos el resultado del fichero csv resultante por cada año y todos los paises en un csv, es un csv donde salen todas las cantidades de clientes acumuladas
74. ### por pais, el nombre del csv que se guarda tiene incluye el año del resultado.
75. write.csv(touristporpaisSumaMes,Resultadocsv)
76. DatosAnyo <- read.table(Resultadocsv,header=TRUE, sep=",", na.strings="NA", strip.white=TRUE)
77. ### Dibujamos la Normal de los datos según el csv resultante
78. densityPlot( ~ X, data=DatosAnyo, bw="SJ", adjust=1, kernel="gaussian")
79. ### Aplicamos test de Shaoiro para cada csv resultante
80. normalityTest(~X, test="shapiro.test", data=DatosAnyo)}
81. ###Con los test de Shapiro vemos que durante los años 1998-2003: p=0,084, del 2004-2006: p=0,058, de 2007-2015: p=0,0499(p<0,05)