**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?

Realmente son todos los campos de interés ya que según los que utilicemos nos servirá para realizar un estudio u otro. Es un dataset que podemos utilizar para muchos estudios diferentes.

Si decidimos, por ejemplo, estudiar 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 esta cuestión nos harán falta los campos: Month y Count.

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 “No Informado”, estos registros no suponen un problema ya que no nos resultan necesarios para nuestro estudio, solo recogeremos los campos Year, Month y Count, estos campos si tienen unidades y supondremos que aunque los otros algunos de los valores del registro sean “No informado”, el resto de datos este bien. De todas formas, esperemos que esto no suponga una variación muy alta en los resultados

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

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.

Una vez puestos a 0 todos los valores de Count que tienen como valor 0.

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 serán:

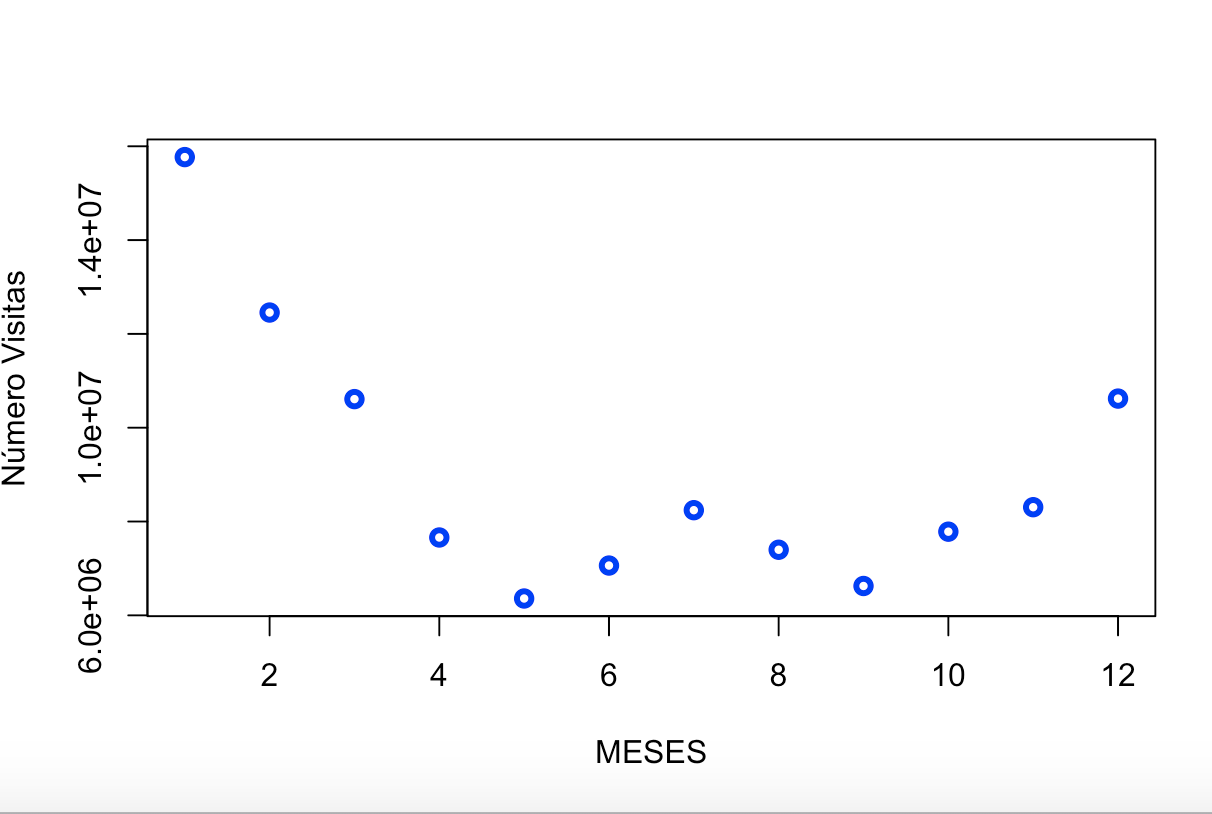
a. Leer los datos y cambiar los NA del campo Count por 0´s.

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.

Despues de realizar todo el proceso sobre los datos en programa con R vemos el gráfico:



Esto 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.

Esta gráfica 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.

* 3.2.  Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.
* 3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

1. Representación de los resultados a partir de tablas y gráficas.
2. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las  conclusiones? ¿Los resultados permiten responder al problema?
3. 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.