Ejercicios sesión 2

Alberto Torres Barrán

2019-12-14

NBA shots

Descargar los datos que se encuentran en Kaggle sobre tiros de la NBA en la temporada 14-15:

1. Cargar los datos en R
2. Identificar que representa cada una de las filas y ver las variables (columnas) disponibles.
3. Vamos a comparar las estadísticas de tiro de dos jugadores, por ejemplo Kobe Bryant y James Harden. Por tanto, lo primero es seleccionar las filas que contienen la información de dichos jugadores.
4. A continuación vamos a crear una nueva variable, (`total_touch_time`), que acumula el tiempo total que cada jugador ha tocado el balón antes de cada tiro (`TOUCH_TIME`) en cada partido. Pista: función `cumsum()`.
5. ¿Cuántas filas tiene el `data.frame` resultado de la operación anterior?
6. Vamos a resumir ahora el `data.frame` anterior calculando una nueva variable, `points_per_tt` dividiendo la suma total de puntos de cada jugador/partido (variable `PTS`) por el tiempo total que ha tocado la pelota (máximo de la variable anterior, `total_touch_time`).

7. Ahora podemos volver a resumir el `data.frame` anterior para obtener un valor por cada jugador, calculando la media de la variable `points_per_tt` para todos los partidos.
8. Por último, hacer un histograma de la distancia de tiro (`SHOT_DIST`), ¿qué se puede observar en la distribución?
9. Comprobar la hipótesis anterior coloreando el histograma por la variable `PTS_TYPE`.

pew

1. Cargar el paquete `tidyr`.
2. Importar el fichero `pew.txt`, que contiene datos del Pew Research Center sobre el número de personas de diferentes religiones y rangos salariales.
3. ¿Cuáles son las variables en esos datos? ¿Se corresponden las columnas del data frame con las variables?
4. Convertir el data frame a otro “ordenado”
5. Hacer un gráfico de barras para cada religión, donde el eje x representa el rango salarial y el eje y el número de personas.
6. Calcular el número total de personas por religión
7. Agrupar todas las religiones en una única categoria “Others”, excepto las 5 con más personas
8. Repetir el gráfico del ejercicio con estos datos pero hacer las barras horizontales

weather

1. Cargar el paquete `tidyr`.
2. Leer el conjunto de datos `weather.txt` en R.
3. Identificar cuales son las variables en los datos.
4. Agrupar las variables d1–d31 en dos variables día y temperatura.
5. Convertir las columnas element y temperatura en dos variables `TMAX` y `TMIN`.
6. Separar la columna id en dos variables, país e id.
7. Eliminar la “d” de la columna dia
8. Juntar las columnas dia, mes y año en otra de tipo fecha
9. Calcular la temperatura máxima y mínima media para cada semana

tb

1. Cargar el conjunto de datos tb.csv, que contiene casos de tuberculosis por año, país, edad, sexo y método de diagnóstico.
2. Identificar si se trata de datos “ordenados”, donde cada columna representa una variable.
3. Eliminar las columnas `new_sp`, `new_sp_m04`, `new_sp_m514`, `new_sp_f04`, `new_sp_f514`.
4. Convertir el data.frame a formato “ancho”, creando dos nuevas variables “Edad-Sexo” y “NCasos”.
5. Eliminar la cadena `new_sp_` de la columna “Edad-Sexo”
6. Separar la variable “Edad-Sexo” en dos.
7. Representar en un gráfico de lineas con la evolución anual de los casos totales para los paises España (ES), Francia (FR), Italia (IT) y Reino Unido (GB)
8. Crear distintos gráficos como el anterior para los diferentes grupos de edad
9. Mejorar el gráfico anterior para que la escala del eje y sea independiente en cada sub-gráfico

ventas

1. Cargar el conjunto de datos `ventas.csv` en R.
2. Ver que columnas tiene y su tipo. ¿Ha identificado readr bien el tipo de todas ellas?. ¿Por qué?
3. Convertir la columna `Codigo` a factor.
4. Calcular la diferencia media en valor absoluto entre las ventas y su previsión.
5. Eliminar la variable `Prevision`.
6. Calcular la matriz de correlación de las Ventas para todos los distintos productos (identificados con su código). Pista: la matriz con tantas filas y columnas como códigos de productos distintos.
7. Transformar la matriz de correlación anterior en un data.frame que esté en formato largo. Pista: identificar que variables deberían ir en las columnas.
8. Hacer un heatmap que represente la matriz de correlación anterior. Pista: `geom_tile()`.
9. Con los datos iniciales, convertir la columna `Fecha` a tipo “date”
10. Filtrar los 5 productos con más ventas (ignorar ventas negativas)
11. Calcular las ventas mensuales para cada uno de los productos anteriores y representar su evolución con un

gráfico de barras (ignorar ventas negativas)

12. Mejorar el gráfico anterior rotando las etiquetas 90 grados

Plantas

Datos: https://github.com/gavinsimpson/plant-phys/raw/master/f18ph.xls

1. Importar los datos originales:

- Las columnas `treatment`, `cultivar` y `plantid` contienen el tratamiento aplicado, tipo de la planta e ID

- El resto de columnas tienen el formato `variable:dias`, es decir, la medición de distintas variables a lo largo del crecimiento de la planta (dias)

2. Convertir en un dataframe que tenga las variables: `treatment`, `cultivar`, `plantid`, `day`, `height`, `internodes`, `freshwt`