Ejercicios sesión 3

Alberto Torres Barrán

2020-01-29

diamonds

Con el conjunto de datos diamonds:

1. Ver el tipo de cada una de las variables.

2. Realizar un análisis estadístico de las variables numéricas: calcular la media, varianza, mediana y rangos ¿Tienen las distintas variables rangos muy diferentes?.

3. Hacer un gráfico de cajas de la variable price para cada uno de los distintos valores de color.

4. Hacer el mismo gráfico del punto anterior pero con un gráfico de cajas para cada uno de los valores de la variable cut.

5. Calcular la correlación de todas las variables numéricas con la variable price.

6. Crear un histograma de la variable carat para cada uno de los distintos valores de color. ¿Son muy diferentes las distribuciones?.

7. Realizar un gráfico de dispersión para las variables que tienen más y menos correlación con price y comentar los resultados. ¿Como seria el gráfico de dispersión entre dos vectores con correlación 1?.

8. Definimos los outliers como los elementos (filas) de los datos para los que cualquiera de las variables

(numéricas) está por encima o por debajo de la mediana más/menos 3 veces el MAD (Median Absolute Deviation). Identificar estos outliers y quitarlos.

9. Separar el conjunto de datos en dos subconjuntos disjuntos de forma aleatoria, el primero conteniendo un 70% de los datos y el segundo un 30%.

10. Escalar los datos para que tengan media 0 y varianza 1, es decir, restar a cada variable numérica su media y dividir por la desviación típica. Calcular la media y desviación en el conjunto de train, y utilizar esa misma media y desviación para escalar el conjunto de test.

gapminder

1. Instalar el paquete `gapminder`

2. Calcular cuantos niveles distintos tienen las columnas `country` y `continent`

3. ¿Cuántas filas del data.frame hay para cada una de las columnas anteriores?

4. Crea un data.frame con las continentes que empiezan por “A”

5. Realiza un gráfico de barras **ordenadas** de mayor a menos, donde el eje x representa los continentes y el eje y el número de paises (registros)

6. Representa en un gráfico de puntos la esperanza de vida (`lifeExp`) en el eje x para cada uno de los paises de Asia (eje y). Usa los datos de 2007 y ordena los puntos de menor a mayor.

7. Representa la evolución de la esperanza de vida para los paises “Egypt”, “Haiti”, “Romania”, “Thailand” y “Venezuela”. Realizar un único gráfico con una línea de un color distinto para cada país. Ordenar la leyenda para que el país con más esperanza de vida esté de primero (usar `fct_reorder2()`).

8. Cálcula la longitud media de los nombres de los países en los datos para cada continente.

9. Contar cuantas veces aparece cada vocal en los nombres de los paises (columna “country”)

Puerto Rico

El fichero `mortality_report.csv` contiene datos diarios de mortalidad en Puerto Rico desde el 1 de Enero de 2015 al 31 de Mayo de 2018 (ejercicio adaptado de Introduction to Data Science:

1. Leer los datos en R
2. Convertir la columna "month" a su valor numérico
3. Crear una columna con la fecha (tipo `date`) a partir de las columnas "day", "month" y "year"
4. Realizar un gráfico de líneas de las muertes (eje y) frente a la fechas (eje x)
5. Ignorar fechas posteriores al 1 de Mayo de 2018
6. Repetir un gráfico de barras pero ahora representar en el eje x el día del año.
7. Repetir el gráfico del ej 4 pero usar dos colores distintos para antes y después del 20 de septiembre de 2017 (no poner ninguna leyenda)
8. Repetir el mismo gráfico del ejercicio 6 pero representar ahora medias mensuales (pista: `round_date()`)