Ejercicio diamonds

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.