

Hoja de Ejercicios

DESCRIPCIÓN DE LOS DATOS

Contiene tres tipos de entidades:

- La especificación de un carro en términos de varias características
- La tasa de riesgo:
 - Corresponde al grado en el cual el auto es más riesgoso de lo que su precio indica. El valor de 3 en este indicador dice que el auto es riesgoso mientras el valor -3 indica que el auto es bastante seguro
- La frecuencia de pérdida comparada con otros carros:
 - El promedio relativo de pagos de pérdida por vehículo asegurado. Este valor representa el promedio de pérdidas por carro por año.

Número de filas: 205

Número de atributos o columnas: 26

Información de los atributos:

Atributo:	Rango del atributo:
1. symboling:	-3, -2, -1, 0, 1, 2, 3.
2. normalized-losses:	numérico 65 hasta 256.
3. make:	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, hastayota, volkswagen, volvo
4. fuel-type:	diesel, gas.
5. aspiration:	std, turbo.
6. num-of-doors:	four, two.
7. body-style:	hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels:	4wd, fwd, rwd.
9. engine-location:	front, rear.
10. wheel-base:	numérico desde 86.6 hasta 120.9.
11. length:	numérico desde 141.1 hasta 208.1.
12. width:	numérico desde 60.3 hasta 72.3.
13. height:	numérico desde 47.8 hasta 59.8.
14. curb-weight:	numérico desde 1488 hasta 4066.
15. engine-type:	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.

Los datos que contiene fueron extraídos de UCI Machine Learning Repository.

(<https://archive.ics.uci.edu/datasets.html>)

16. num-of-cylinders:	eight, five, four, six, three, twelve, two.
17. engine-size:	numérico desde 61 to 326.
18. fuel-system:	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore:	numérico desde 2.54 hasta 3.94.
20. stroke:	numérico desde 2.07 hasta 4.17.
21. compression-ratio:	numérico desde 7 hasta 23.
22. horsepower:	numérico desde 48 hasta 288.
23. peak-rpm:	numérico desde 4150 hasta 6600.
24. city-mpg:	numérico desde 13 hasta 49.
25. highway-mpg:	numérico desde 16 hasta 54.
26. price:	numérico desde 5118 hasta 45400.

Los datos que contiene fueron extraídos de UCI Machine Learning Repository.
(<https://archive.ics.uci.edu/datasets.html>)

EJERCICIOS

Ejercicio 1.

- Cargue los datos del archivo data.csv a una variable llamada carros
- Cuando lo haya hecho, muestre las primeras 5 filas.
- Elaboren una tabla con cada una de las variables del dataset y clasifiquenla en el tipo que tiene cada uno. Para cada variable definan el tipo y subtipo (ej. Cuantitativa continua)

Ejercicio 2

Responda las siguientes preguntas:

1. ¿Cuántos carros tipo hatchback tiene la muestra?
2. ¿Cuál es el promedio de pérdidas de los carros convertibles? Obvie los carros que tienen valores NA en el atributo "normalized_losses"

Ejercicio 3

La siguiente instrucción crea un nuevo data frame con el promedio del indicador de riesgo y de las pérdidas por marca de carro. Ejecútela.

Puede usar la función aggregate de R o groupby de python

Luego de ejecutar la instrucción anterior al solicitar las primeras 6 filas debe mostrarse lo siguiente:

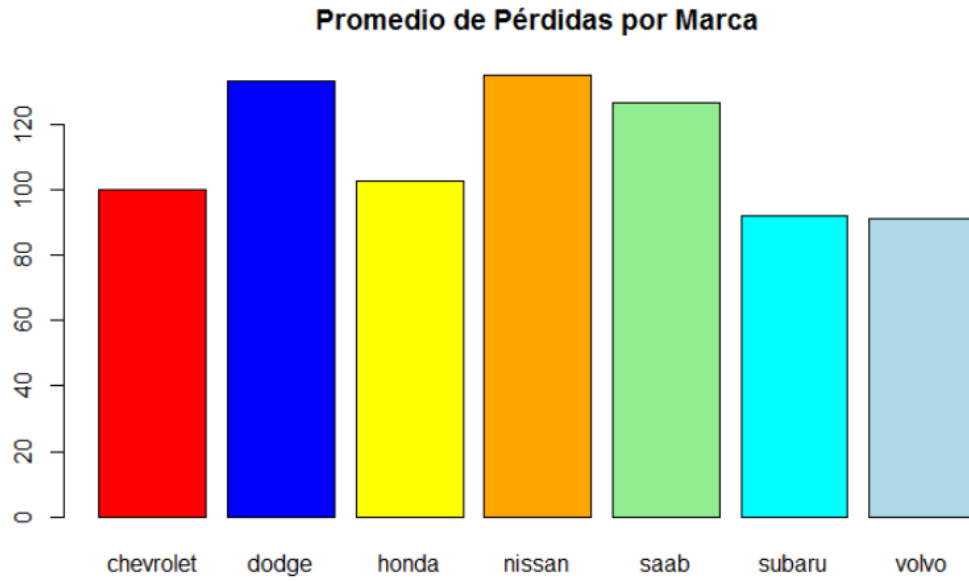
```
> head(peligrosos)
  marca indicadorPeligro promedioPerdidas
1 alfa-romero      2.333333              NA
2      audi      1.285714              NA
3      bmw      0.375000              NA
4 chevrolet      1.000000      100.0000
5      dodge      1.000000      133.4444
6      honda      0.615384      103.0000
```

- a) Cámbiele los nombres a las columnas del data frame peligrosos
- b) Guarde en una nueva variable los datos que no tienen NA en el promedio de las pérdidas. Esta variable será la que usará para los siguientes ejercicios.
- c) Haga un vector con el promedio de pérdidas normalizadas de cada una de las marcas de carros.
- d) Utilice el vector creado en el inciso anterior para hacer un gráfico de barras que permita visualizar las marcas que más pérdidas han tenido.

Nota: El gráfico debe tener un título, un color diferente por cada barra, además de la marca correspondiente a cada barra. Se espera que obtenga un gráfico parecido al siguiente:

Los datos que contiene fueron extraídos de UCI Machine Learning Repository.

(<https://archive.ics.uci.edu/datasets.html>)



- e) Basado en el gráfico que generó responda las siguientes preguntas:
- ¿Cuáles podrían decirse que eran las marcas de carros de menor riesgo en la época en que fueron obtenidos los datos? ¿Por qué?
 - ¿Cuáles podrían decirse que eran las marcas de carros de mayor riesgo en la época en que fueron obtenidos los datos? ¿Por qué?
- f) Investigue un poco más el conjunto de datos:
- Explique cómo se distribuyen los datos en las variables cuantitativas. Saque medidas de tendencia central y dispersión y orden.
 - ¿Se puede decir que las variables cuantitativas siguen una distribución normal? Compruébelo tanto gráficamente como usando pruebas de normalidad. Explique los resultados.
 - ¿Cuál es la distribución de los datos de las variables cualitativas? Haga, tablas de frecuencias de cada una. Explique los resultados.
 - ¿Hay correlaciones entre algunas de las variables? ¿Cuáles? Explique.