1. Construya una función con dos argumentos, un dataframe/matriz de entrada y un dataframe/matriz de salida. La función deberá devolver un objeto (dataframe/matriz) que contenga en cada columna la cantidad de valores NA que existen en cada columna homónima del dataframe original. Ejemplo:

Df\_original

|  |  |  |
| --- | --- | --- |
| x1 | x2 | x3 |
| 1 | 5 | 26 |
| NA | 15 | NA |
| 5 | 68 | 8 |

Resultado

|  |  |  |
| --- | --- | --- |
| x1 | x2 | x3 |
| 2 | 0 | 1 |

Input: construya un dataframe con lo siguiente y utilícelo como prueba

SMP <- c(10,30, 033, 966, NA)

CALLENOMBRE <- c(4519, NA, NA, 966,966)

CALLEALTURA <- c(4519, NA, 5640, 1418, 966)

1. a- Haga una función en la que tenga cuatro argumentos, un dataframe de entrada, un dataframe de salida y dos numeros “N” y “n”. La función debe extraer “N” muestreos aleatorios de tamaño “n”. De cada muestra aleatoria calcule la media, la varianza y la moda e incorpore cada uno de estos resultados en el dataframe de salida que contenga columnas llamadas igual que la base original y que tenga una fila para las medias obtenidas por el sampleo.

b- Del resultado del punto a) haga un resumen estadístico y un histograma para cada variable (construya un grafico multiple con los cuatro histogramas)

dos filas y dos columnas

c- Modifique la función del punto a) para que el tamaño de la muestra sea aleatorio.

Input de prueba: irisnum=iris[,-5]

Varios promedios de varias muestras aleatorias 🡪 Teorema central del limite

Cuantas mas muestras tomo mas normal se parece

1. Matrices:

Installe el paquete “wooldridge” y utilice la base de datos Wage1 corriendo la línea

data("wage1")

Estos son datos de la Encuesta de Población Actual de 1976, recopilada por Henry Farber cuando él y Wooldridge fueron colegas en el MIT en 1988.

Haga un help sobre la base y describa de que se trata esta base, que tipo de variables tiene, cuantas variables tiene? Use no mas de tres renglones

Signo de interrogación

Help(iris) 🡪 Da información sobre la base de datos

Construya una función que realice la siguiente operación sobre un dataframe, siendo “Y”, la primera columna del dataframe (trate que la selección de la variable Y sea uno de los argumentos de la función) y “X” las columnas 2 a 6 del dataframe.

Separar una columna cualquiera de un df, con una columna hacer una cosa y con el resto hacer otra cosa. Si y=3, las columnas correspondientes a x son las otras.

Resultado=[inv(transpuesta(X)\*X)]\*[transuesta(X)\*Y]

Input

Utilice este segmento de la base como input

wage=wage1[-c(520:526),]

1. Haga una función que tome las primeras cuatro columnas s de la base wage (del punto 3 como input) y cree una matriz nueva en la cual, a la matriz original (wage), adicione el cuadrado de cada variable, y las variables centralizadas (variable menos la media) y cada variable estandarizada (variable menos la media y dividido por el desvío estándar)

Agregarle columnas a la base de datos

5 - Cargue en memoria el archivo "netflix.csv" y asignelo a una variable  
  
a) Cree una columna que indique la duracion total en minutos (hours\*60+minutes)  
  
b) ¿Cual es el año con mayor duracion promedio? [obtener duracion total promedio con la columna anterior, agrupar por año, calcular la duracion promedio, y ordenar de mayor a menor] ¿cual es el año con mayor variabilidad? [sd(x) calcula el desvio estandar del vector x]

Sobre la duración   
  
c) De las peliculas para mayores de 13 anios (maturity U/A 13+), el año de la duracion promedio mas  
alta, es igual a la poblacion general? [operacion anterior + filtro]  
  
6 - GGPLOT  
a) Utilizando GGPLOT, realice un gráfico que permita visualizar la distribución   
de la duracion calculada en el punto (2.b) por cada "maturity\_rating" (puede ser boxplot, histograma, densidad, o el que sea relevante)