Examen Práctico
Semestre agosto-diciembre 2021

INSTRUCCIONES

a. Utiliza tu computadora con Spyder para resolver los ejercicios.
b. Puedes utilizar todo el material de consulta que necesites, apuntes, libros, hojas, códigos de ejercicios anteriores, etc.
c. No está permitido utilizar la conexión a internet con ningún dispositivo electrónico (reloj, laptop, celular, computadora, etc.) para consultas.
d. Solo puedes utilizar los módulos Pandas, Pyplot y Math. No puedes utilizar ningún otro módulo (**import**)
e. Cada ejercicio vale 5.7 puntos.
f. Guarda el código de cada ejercicio como **ejercicio_1.py**, **ejercicio_2.py**, **ejercicio_3.py**, **ejercicio_4.py**
g. Cuando termines de resolver los problemas, comprime los cinco archivos de los ejercicios en una carpeta **.zip** o **.rar** y nombra la carpeta comprimida de la siguiente manera: **apellido1_apellido2_nombre1_nombre2_nombre3.zip**, todo en **minúsculas y sin acentos**, por ejemplo: **gomez_carranza_juan_carlos.zip**
h. Sube el archivo comprimido a Teams.

EJERCICIOS

1. El archivo adjunto salaries.csv contiene datos de 5 variables de interés para puestos de trabajo en México. Las variables son (entre paréntesis está el nombre de la variable en el archivo), puesto (offer), salario (salary), horas diarias de trabajo (hours_worked), días semanales de trabajo (days_worked), y estado de la república (state). Escribe un programa en Python que lea el archivo con Pandas y calcule lo siguiente:
    a. El resumen de los cinco números, la media, la desviación estándar y dibuje el diagrama de caja para la variable salary para las respuestas del estado de Queretaro y solo para las personas que trabajan entre 20 y 40 horas a la semana (nota que la variable de horas de trabajo es por día).
    b. El resumen de los cinco números, la media, la desviación estándar y dibuje el diagrama de caja para la variable salary para las respuestas del estado de Guanajuato y solo para las personas que trabajan entre 20 y 40 horas a la semana (nota que la variable de horas de trabajo es por día).
    c. La correlación entre las variables salary y days_worked de forma independiente para cada estado que aparece en la variable state. Debes separar los datos de salary y days_worked por estado, y para cada estado calcular la correlación entre esas dos variables.

2. Con el mismo archivo del punto anterior escribe un programa en Python que:

a. Transforme la variable salary a una variable binaria, considerando sueldos altos o bajos. Un salario bajo es aquel menor o igual 10000, el alto es mayor a 10000.
b. Transforme la variable hours_worked en una variable binaria, considerando tiempo completo o tiempo parcial, el tiempo parcial es menor o igual a 6 horas, el tiempo completo es mayor a 6 horas
c. Genere una tabla de contingencia con estas dos variables como se muestra a continuación (los números son ficticios)

		Horas de Trabajo Diarias	
		Completo	Parcial
Salario	Alto	10	5
	Bajo	1	4

La cantidad en cada celda es el número de respuestas en el cruce de valores de ambas variables (en el ejemplo, hay 10 puestos de trabajo de tiempo completo y ofrecen un salario alto). La tabla se debe generar como una lista de listas, una lista para cada tipo de salario, con dos valores por lista.

d. Calcule el coeficiente Q de Yule, definido como

$$Q = \frac{ad - bc}{ad + bc}$$

En donde los valores a, b, c y d son las celdas de la tabla de contingencia como sigue.

		Horas de Trabajo Diarias	
		Completo	Parcial
Salario	Alto	a	b
	Bajo	c	d

En el ejemplo de la tabla superior, el valor del coeficiente Q sería:

$$Q = \frac{(10)(4) - (5)(1)}{(10)(4) + (5)(1)} = \frac{40 - 5}{40 + 5} = \frac{35}{45} = 0.777$$

3. Utilizando el mismo archivo de salarios, hacer un programa en Python que para cada puesto de trabajo tome el subconjunto de variables con salary, hours_worked y days_worked, de tal forma que cada puesto de trabajo quede representado como un vector de 3 valores. Este vector representa a su vez un punto en un espacio 3D. Después, debe calcular la distancia Euclideana entre cada par posible de puntos. La salida debe ser una lista de listas de *n*x*n* con las distancias (n = número de puestos de trabajo en la lista).

Ejemplo (datos ficticios):

Puestos de trabajo:

20000,10,5
5000,12,4
40000,8,7

Salida:

[[0.0, 30.4, 21.5],
[30.4, 0.0, 10.4],
[21.5, 10.4, 0.0]]

Explicación: Hay tres puestos de trabajo en el archivo, que se representan como un vector de las tres variables salary, hours_worked y days_worked. Cada vector representa un punto en el espacio 3D. La salida es una lista de listas 3x3 (3 listas con 3 elementos cada una), formando una matriz de distancias. La distancia de un punto consigo mismo es 0, por lo tanto, la diagonal contiene solo 0s.

La distancia Euclideana se calcula como:

$$d_{p,q} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Donde p y q son dos puntos (puestos de trabajo) y n es la dimensión de esos puntos, para nuestro caso n=3.

4. Utilizando el mismo archivo previo, escribe un programa en Python que calcule la asimetría (skewness) y la curtosis para las variables salary, hours_worked y days_worked, separadas por estado de la república. Primero debes separar los datos por estado y luego calcular las métricas indicadas.

   La asimetría se define como:

$$\mu_3 = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}}{\sigma^3}$$

   La curtosis se define como:

$$\mu_4 = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n}}{\sigma^4}$$

   Donde x_i es cada valor de la variable, $\bar{x}$ es la media de esa variable, σ es la desviación estándar de la variable y n es el número de observaciones (datos) de la variable.