Nombre: Aldo Tena García

Matrícula: A01275222

```
# Carga las librerías necesarias.
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import pearsonr
```

```
# Carga el conjunto de datos al ambiente de Google Colab y muestra los primeros
# 6 renglones.
from google.colab import files

uploaded = files.upload()

for fn in uploaded.keys():
  print('User uploaded file "{name}" with length {length} bytes'.format(
      name=fn, length=len(uploaded[fn])))
```

Elegir archivos insurance.csv

- **insurance.csv**(text/csv) - 54289 bytes, last modified: 8/5/2022 - 100% done

```
Saving insurance.csv to insurance (2).csv
User uploaded file "insurance.csv" with length 54289 bytes
```

```
df = pd.read_csv('insurance.csv')
df.head(6)
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160

age: Edad del asegurado principal

sex: Género del asegurado. female o male

bmi: Índice de masa corporal

children: Número de hijos que estan cubiertos con la poliza.

smoke: ¿El beneficiario fuma? (yes/no)

region: ¿Dónde vive el beneficiario? Estos datos son de Estados Unidos. Regiones disponibles: northeast, southeast, southwest, northwest

charges: Costo del seguro.

```
# Crea una tabla resumen con los estadísticas generales de las variables
# numéricas.
df.describe()
```

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

```
# ¿Cómo se correlacionan las varaibles numéricas entre sí?
df.corr()
```

	age	bmi	children	charges
age	1.000000	0.109272	0.042469	0.299008
bmi	0.109272	1.000000	0.012759	0.198341
children	0.042469	0.012759	1.000000	0.067998
charges	0.299008	0.198341	0.067998	1.000000

```
selected = df[['bmi', 'charges']]
```

```
selected.head(5)
```

	bmi	charges
0	27.900	16884.92400
1	33.770	1725.55230
2	33.000	4449.46200
3	22.705	21984.47061
4	28.880	3866.85520

```
# Determina si existe o no una correlación entre el índice de masa corporal
# (bmi) y el costo del seguro.
r, p = stats.pearsonr(selected['bmi'], selected['charges'])
print(f"Correlación Pearson: r={r}, p-value={p}")

r, p = stats.spearmanr(selected['bmi'], selected['charges'])
print(f"Correlación Spearman: r={r}, p-value={p}")

r, p = stats.kendalltau(selected['bmi'], selected['charges'])
print(f"Correlación Pearson: r={r}, p-value={p}")
```

```
Correlación Pearson: r=0.1983409688336288, p-value=2.459085535117846e-13
Correlación Spearman: r=0.11939590358331145, p-value=1.1926059544526874e-05
Correlación Pearson: r=0.08252397079981415, p-value=6.25690064095591e-06
```

```
# ¿Cuántas personas aseguradas son hombre y cuántas son mujeres?
df['sex'].value_counts()
```

```
male      676
female    662
Name: sex, dtype: int64
```

```
# ¿Cuántos hombres y mujeres asegurados viven en cada región?
pd.crosstab(df['sex'], df['region'])
```

region	northeast	northwest	southeast	southwest
sex				
female	161	164	175	162
male	163	161	189	163

```
# En promedio, ¿quién paga más de cuota de seguro? ¿Los fumadores o los no
# fumadores? Muéstralo con los datos.
df.groupby(['smoker']).mean()[['charges']]
```

smoker	charges
no	8434.268298
yes	32050.231832

```
# ¿Cuáles son las cuotas mínimas y máximas que las personan pagan dependiendo
# del género y del número de hijos?
df.groupby(['sex', 'children']).agg(['min', 'max'])[['charges']]
```

		charges	
		min	max
sex	children		
female	0	1607.51010	63770.42801
	1	2201.09710	58571.07448
	2	2801.25880	47305.30500
	3	4234.92700	46661.44240
	4	4561.18850	36580.28216
	5	4687.79700	19023.26000
male	0	1121.87390	62592.87309
	1	1711.02680	51194.55914
	2	2304.00220	49577.66240
	3	3443.06400	60021.39897
	4	4504.66240	40182.24600
	5	4915.05985	14478.33015

```
# ¿Cuál es el índice de masa corporal promedio para hombre y mujeres dependiendo
# región en la que viven y si son fumadores? ¿Impacta eso en la tarifa del
# seguro?
df.groupby(['sex','region','smoker']).mean()[['bmi']]
```

			bmi
sex	region	smoker	
female	northeast	no	29.777462
		yes	27.261724
	northwest	no	29.488704
		yes	28.296897
	southeast	no	32.780000
		yes	32.251389
	southwest	no	30.050355
		yes	30.128571
male	northeast	no	28.861760
		yes	29.560000
	northwest	no	28.930379
		yes	29.983966
	southeast	no	34.129552
		yes	33.650000
	southwest	no	31.019841
		yes	31.502703

✓ 0 s completado a las 10:50