

ACTIVIDAD 06 - ESTADÍSTICA BÁSICA

```
[ ] # Carga las librerías necesarias.
```

```
import pandas as pd
import numpy as np
import random
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])))

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

Elegir archivos. Sin archivos seleccionados Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving insurance.csv to insurance.csv

user uploaded file "insurance.csv" with length 54289 bytes

	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

```
[ ] # Crea una tabla resumen con los estadísticas generales de las variables
# numéricas.
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ---
 0   age         1338 non-null   int64
 1   sex         1338 non-null   object
 2   bmi         1338 non-null   float64
 3   children    1338 non-null   int64
 4   smoker      1338 non-null   object
 5   region      1338 non-null   object
 6   charges     1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

```
[ ] # ¿Cómo se correlacionan las variables 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

```
[ ] # Determina si existe o no una correlación entre el índice de masa corporal
# (bmi) y el costo del seguro.
```

```
select = df[['bmi', 'charges']]
select.corr()
```

```
print('Correlación Pearson: ',
      select['bmi'].corr(select['charges'], method = 'pearson'))
```

Correlación Pearson: 0.19834096883362895

```
[ ] # ¿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']]
```

charges	
smoker	
no	8434.268298
yes	32050.231832

```
[ ] # ¿Cuáles son las cuotas mínimas y máximas que las personas 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

```
[ ] # ¿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?

impact = df.groupby(['sex', 'smoker', 'region']).mean()[['bmi', 'charges']]
impact
```

			bmi	charges
sex	smoker	region		
female	no	northeast	29.777462	9640.426984
		northwest	29.488704	8786.998679
		southeast	32.780000	8440.205552
		southwest	30.050355	8234.091260
	yes	northeast	27.261724	28032.046398
		northwest	28.296897	29670.824946
		southeast	32.251389	33034.820716
		southwest	30.128571	31687.988430
male	no	northeast	28.861760	8664.042222
		northwest	28.930379	8320.689321
		southeast	34.129552	7609.003587
		southwest	31.019841	7778.905534