

Actividad - Estadística básica

- Nombre: Daniel Queijeiro Albo
- Matrícula: A01710441


**Entregar:** Archivo PDF de la actividad, así como el archivo .ipynb en tu repositorio. **Nota:** Recuerda habrá una penalización de 50 puntos si la actividad fue entregada fuera de la fecha límite.

Carga el conjunto de datos `insurance.csv` (se encuentra en el repositorio de la clase) y realiza un análisis estadístico de las variables.

```
# Carga las librerías necesarias.
import pandas as pd
import numpy as np
import random

# Carga el conjunto de datos al ambiente de Google Colab y muestra los primeros
# 6 renglones.
!curl -O https://raw.githubusercontent.com/Manchas2k4/tc1002S/main/datasets/insurance.csv

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

% Total			% Received		% Xferd		Average Speed		Time	Time	Time	Current
100	54289	100	54289	0	0	946k	0	--:--:--	--:--:--	--:--:--	946k	
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					

El conjunto de datos contiene información demográfica sobre los asegurados en una compañía de seguros:

- 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
1  29.0  26.8  1.0  1669.36
2  33.0  22.3  0.0  1686.47
3  33.0  31.1  3.0  1448.56
4  30.0  27.3  1.0  2615.69
5  33.0  30.7  3.0  2431.54
6  30.0  20.9  1.0  3861.64
7  33.0  33.4  4.0  4414.48
8  31.0  30.9  3.0  2831.07
9  31.0  22.4  1.0  866.35
10 30.0  29.8  3.0  2101.53

# ¿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
charges  0.299008  0.198341  0.067998  1.000000

# Determina si existe o no una correlación entre el índice de masa corporal
# (bmi) y el costo del seguro.
selected = df[['age', 'bmi', 'children', 'charges']]
print('Correlación pearson', selected['bmi'].corr(selected['charges'], method='pearson'))
print('Correlación spearman', selected['bmi'].corr(selected['charges'], method='spearman'))
print('Correlación kendall', selected['bmi'].corr(selected['charges'], method='kendall'))

Correlación pearson 0.19834096883362895
Correlación spearman 0.11939590358331145
Correlación kendall 0.08252397079981415

# ¿Cuántas personas aseguradas son hombre y cuántas son mujeres?
ndf = df.dropna(subset=['charges'])
h = ndf.groupby('sex').get_group('male')
m = ndf.groupby('sex').get_group('female')
print('Hay', len(h), 'hombres asegurados y ', len(m), 'mujeres aseguradas.')

Hay 676 hombres asegurados y 662 mujeres aseguradas.

# ¿Cuántos hombres y mujeres asegurados viven en cada región?
group = ndf.groupby('region')
count = group['sex'].value_counts()
resultado = pd.DataFrame(count)
resultado.columns = ['Cantidad']
print(resultado)

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

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

La cuota promedio de fumadores es de 32050.23183153285
La cuota promedio de no fumadores es de 8434.268297856202

# ¿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
	4	4561.18850	36580.28216
	5	4687.79700	19023.26000

```
# ¿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?
promedio_bmi = ndf.groupby(['sex', 'region', 'smoker'])['bmi', 'charges'].mean()
print("Promedio de bmi y tarifa por genero, región y fumadores")
print(promedio_bmi)
```

Promedio de bmi y tarifa por genero, región y fumadores			bmi	charges
sex	region	smoker		
female	northeast	no	29.777462	9640.426984
		yes	27.261724	28032.046398
	northwest	no	29.488704	8786.998679
		yes	28.296897	29670.824946
	southeast	no	32.780000	8440.205552
		yes	32.251389	33034.820716
male	southwest	no	30.050355	8234.091260
		yes	30.128571	31687.988430
	northeast	no	28.861760	8664.042222
		yes	29.560000	30926.252583
	northwest	no	28.930379	8320.689321
		yes	29.983966	30713.181419
	southeast	no	34.129552	7609.003587
		yes	33.650000	36029.839367
	southwest	no	31.019841	7778.905534
		yes	31.502703	32598.862854

```
<ipython-input-18-0a5db88234c3>:4: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecate
promedio_bmi = ndf.groupby(['sex', 'region', 'smoker'])['bmi', 'charges'].mean()
```

