# Actividad - Estadística básica

- **Nombre:** Ramona Nájera Fuentes
- **Matrícula:** A01423596

**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.

```
import pandas as pd
import numpy as np


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)
```

```
Sélect...fichiers...  Aucun fichier choisi  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 55628 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 |
| **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 |
|---|---|---|---|---|
| **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 variables numéricas entre sí?
df.corr()

'''
0.299008 (age - charges)
0.198341 (bmi - charges)
0.109272 (age - bmi)
0.067998 (children - charges)
0.042469 (age - children)
0.012759 (bmi - children)

Rápidamente, podemos concluir lo siguiente:
  - Todas las variables están directamente relacionadas
  - En general, las relaciones observadas son débiles
'''
```

| | 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.

'''
r (coeficiente de correlación [-1, 1])
  -1: Correlación "perfecta" negativa (inversamente relacionados)
   0: Relación lineal débil
   1: Correlación "perfecta" positiva (directamente relacionados)

p (significancia estadística [útil en pruebas de hipótesis])
  Probabilidad de un r != 0 cuando la hipótesis nula es verdadera

Hipótesis
  Nula: Relación producto del azar (Rechazada: p < 0,05)
  Alternativa: Relación presente en los datos
'''

from scipy import stats
from scipy.stats import pearsonr

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

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

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

'''
CONCLUSIÓN
El factor de relación no es muy fuerte; sin embargo, existe una relación causa-efecto entre el bmi y el costo del seguro
'''
```

```
Correlación Pearson: r=0.1983409688336288, p-value=2.459085535116766e-13
Correlación Spearman: r=0.11939590358331145, p-value=1.1926059544526874e-05
Correlación Kendall: r=0.08252397079981415, p-value=6.256900640955888e-06
'\nCONCLUSIÓN\nEl factor de relación no es muy fuerte; sin embargo, existe una relac
ión causa-efecto entre el bmi y el costo del seguro\n'
```

```
# ¿Cuántas personas aseguradas son hombres 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['region'], df['sex'])
```

| sex | female | male |
|---|---|---|
| region | | |

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

| | charges |
|---|---|
| smoker | |
| 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(['region', 'sex', 'smoker']).mean()[['bmi', 'charges']]

'''
HALLAZGOS
1. El bmi por sexo en cada región no varía mucho entre fumadores y no fumadores
2. La tarifa del seguro es mucho más baja para los no fumadores
'''
```

| | | | bmi | charges |
|---|---|---|---|---|
| region | sex | smoker | | |
| northeast | female | no | 29.777462 | 9640.426984 |
| | | yes | 27.261724 | 28032.046398 |
| | male | no | 28.861760 | 8664.042222 |
| | | yes | 29.560000 | 30926.252583 |
| northwest | female | no | 29.488704 | 8786.998679 |
| | | yes | 28.296897 | 29670.824946 |
| | male | no | 28.930379 | 8320.689321 |
| | | yes | 29.983966 | 30713.181419 |
| southeast | female | no | 32.780000 | 8440.205552 |
| | | yes | 32.251389 | 33034.820716 |
| | male | no | 34.129552 | 7609.003587 |
| | | yes | 33.650000 | 36029.839367 |
| southwest | female | no | 30.050355 | 8234.091260 |