Visualización de la Información - Tarea 1

Andrés Urbano Guillermo Gerardo

8 de Febrero del 2022

[2]:
```
# Bibliotecas para el analisís
import pandas as pd
import numpy as np
# Bibliotecas de visualizacion
import cufflinks as cf
from IPython.display import display, HTML
cf.set_config_file(sharing='public',theme='white',offline=True)
```

Primer paso para nuestro analisis será cargar nuesto conjunto de datos:

[3]:
```
df_heart = pd.read_csv('./heart.csv')
df_heart.head()
```

[3]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Verificamos si existen datos faltantes:

[4]:
```
df_heart.isnull().sum()
```

[4]:
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
```

```
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Vemos que nuestro conjunto de datos esta completo. Procedemos a generar estadistica básica a nuestro conjunto para conocer más sobre el.

[5]:
```
print(f'Shape: {df_heart.shape}')
df_heart.describe()
```

```
Shape: (303, 14)
```

[5]:

	age	sex	cp	trestbps	chol	fbs	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	

	restecg	thalach	exang	oldpeak	slope	ca	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	
std	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	
min	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	
50%	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	
75%	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	
max	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	

	thal	target
count	303.000000	303.000000
mean	2.313531	0.544554
std	0.612277	0.498835
min	0.000000	0.000000
25%	2.000000	0.000000
50%	2.000000	1.000000
75%	3.000000	1.000000
max	3.000000	1.000000

Determinar el porcentaje de personas que han sido diagnosticados con una enfermedad cardíaca

(target, donde 0=no, 1=sí).

[6]:
```
people_disease, healthy_people = np.bincount(df_heart.target)
total_people = df_heart.shape[0]
print('Cantidad de muestras por clase:\n {}'
      .format({k: v for k, v in zip(['Diagnosticado', 'No diagnosticado'],␣
 ↪[people_disease, healthy_people])}))
print(f'Personas con una enfermedad cardiaca: {people_disease / total_people:.
 ↪2%}')
print(f'Personas saludable: {healthy_people / total_people:.2%}')
```

```
Cantidad de muestras por clase:
 {'Diagnosticado': 138, 'No diagnosticado': 165}
Personas con una enfermedad cardiaca: 45.54%
Personas saludable: 54.46%
```

[7]:
```
df_diagnosed = pd.DataFrame({'Clase': ['Diagnosticado', 'No diagnosticado']␣
 ↪,'Total': [people_disease, healthy_people]})
df_diagnosed.iplot(kind='pie', labels='Clase', values='Total',
                   title='Cantidad de pacientes con enfermedad')
```

0.1 Actividad 1

Elabore un histograma o grafica de barras que permita visualizar la edad (age) comparada con el porcentaje de personas si diagnosticadas con una enfermedad cardiaca y las que no han sido diagnosticadas (target, donde 0=no, 1=sí).

[8]:
```
df_heart.tail()
```

[8]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	\
298	57	0	0	140	241	0	1	123	1	0.2	
299	45	1	3	110	264	0	1	132	0	1.2	
300	68	1	0	144	193	1	1	141	0	3.4	
301	57	1	0	130	131	0	1	115	1	1.2	
302	57	0	1	130	236	0	0	174	0	0.0	

	slope	ca	thal	target
298	1	0	3	0
299	1	0	3	0
300	1	2	3	0
301	1	1	3	0
302	1	1	2	0

Contamos el número de personas de la misma edad que tienen una enfermedad cardiaca:

[9]:
```
ages = {}
for age, target in df_heart[['age', 'target']].values:
    if age not in ages:
        ages[age] = 0
```

```
    if target == 1:
        ages[age] += 1
# Convertimos nuestro dic a una estructura de datos de pandas (Serie)
serie_ages = pd.Series(ages)
#serie_ages
```

[10]:
```
serie_ages.iplot(kind='bar', title='Número de personas de la misma edad con enfermedad cardiaca',
                 xTitle='Edades', yTitle='Cantidad de personas')
```

En este histograma podemos ver que el mayor número de personas diagnosticada con problemas del corazon es con edad de 54 años y tambien podemos detectar que la edad más temprana es de 29 años.

0.2 Actividad 2

Elabore un histograma que permita visualizar la presión arterial (trestbps) comparada con el porcentaje de personas si diagnosticadas con una enfermedad cardiaca y las que no han sido diagnosticadas.

[11]:
```
blood_pressure = {}
for trestbp, target in df_heart[['trestbps', 'target']].values:
    if trestbp not in blood_pressure:
        blood_pressure[trestbp] = 0
    if target == 1:
        blood_pressure[trestbp] += 1
# Convertimos nuestro dic a una estructura de datos de pandas (Serie)
blood_pressure
serie_trestbp = pd.Series(blood_pressure)
#serie_trestbp
```

[12]:
```
serie_trestbp.iplot(kind='bar', color='blue', xTitle='Presión arterial',
yTitle='Número de personas',
                  title='Personas diagnosticas')
```

En este histograma podemos ver que en nuestro conjunto de datos no hay tantas personas con alta presión alterial, la mayoria se concentra entre 110 y 150.

Por último, haremos un histograma de los vasos sanguineos de nuestro conjunto de datos:

[21]:
```
df_heart['ca'].value_counts().iplot(kind='bar', color='red',
                                    xTitle='Vasos sanguineos', yTitle='Número de vasos sanguineos')
```

Vemos que hay mayor cantidad de personas que que tiene 0 vasos sanguineos principales.

0.3 Resources

- https://plotly.com/python/pandas-backend/
- https://plotly.com/python/ipython-notebook-tutorial/

- https://stackoverflow.com/questions/49880314/what-is-difference-between-plot-and-iplot-in-pandas
- https://www.kaggle.com/ronitf/heart-disease-uci