Actividad3

March 18, 2021

1 Actividad 3. Mapas de calor y boxplots

1.0.1 Equipo 6:

A01706095 - Naomi Estefanía Nieto Vega

A01706189 - Alejandro Angel Calderon Berges

A01706596 - Carlos Soria de la Cabada

Paso 1. Cargar los datos usando tu lector de csv o con pandas. Para esta sección cambiamos nuestra base de datos por una con registros de infartos en personas y lo actualizamos para la actividad anterior. A continuación se presentan algunas de las variables o columnas existentes en los datos y su signficado.

- age - edad en años
- sex - (1 = hombre; 0 = mujer)
- cp - tipo de dolor en el pecho previo al infarto
- trestbps - presión sanguínea en mmHg
- chol - colesterol en sangre en mg/dl
- fbs - azucar en sangre mayor a 120 mg/dl (1 = si; 0 = no)
- restecg - resultados de electrocardiograma
- thalach - máxima frecuencia cardiaca
- oldpeak - depresión del segmento ST
- slope - pendiente del segmento ST
- target - tuvo infarto o no (1=si, 0=no)

Pero al tener datos booleanos en algunas, esto nos genera conflictos al momento de hacer algunas gráficas por lo que omitiremos estos datos. Asimismo realizamos la carga de datos usando pandas en la misma sección de código como podemos ver a continuación.

[52]:
```
import pandas as pd

data = pd.read_csv('heart.csv')
data.head()
```

[52]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	\
0	63	1	3	145	233	1	0	150	0	2.3	0	
1	37	1	2	130	250	0	1	187	0	3.5	0	
2	41	0	1	130	204	0	0	172	0	1.4	2	

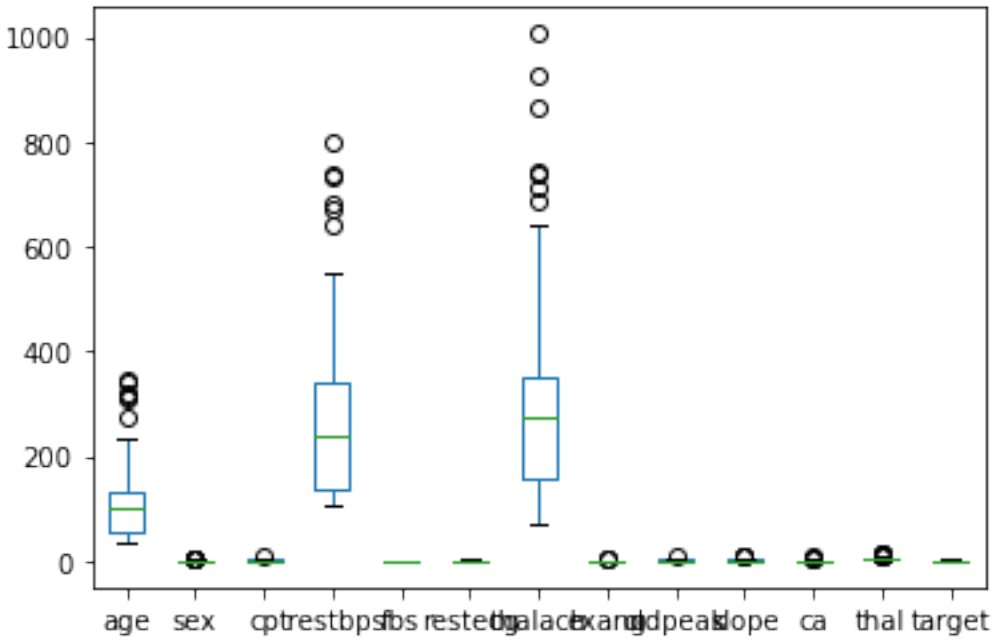
```
3   56    1   1        120   236    0         1       178       0       0.8       2
4   57    0   0        120   354    0         1       163       1       0.6       2

    ca  thal  target
0   0     1       1
1   0     2       1
2   0     2       1
3   0     2       1
4   0     2       1
```

a) Diagrama de cajas y bigotes

[23]:
```
import matplotlib.pyplot as plit
import seaborn as sns
df.groupby('chol').sum().plot(kind='box',legend='Reverse')
```

[23]: <AxesSubplot:>



En esta gráfica podemos ver el diagrama un poco disperso únicamente en 3 variables porque son las que cuentan con números más grandes o mayor variedad que se puede observar al graficarse a diferencia de las variables que tienen solo datos booleanos no hay mucho que observar. De acuerdo con esto las variables más significativas son la edad, la presión sanguínea y la frecuencia cardiaca.

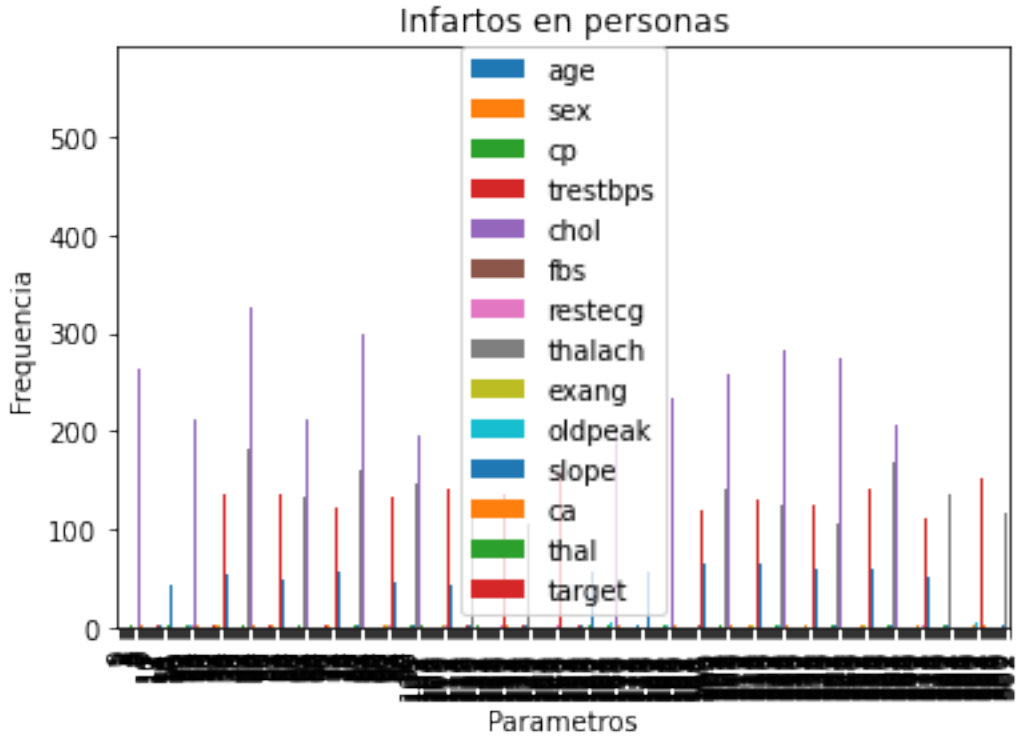
b) Histogramas

[50]:
```
import matplotlib.pyplot as plt
import pandas as pd

data = pd.read_csv('heart.csv')

data.plot(kind = 'bar')
plt.ylabel('Frequencia')
plt.xlabel('Parametros')
plt.title('Infartos en personas')

plt.show()
```



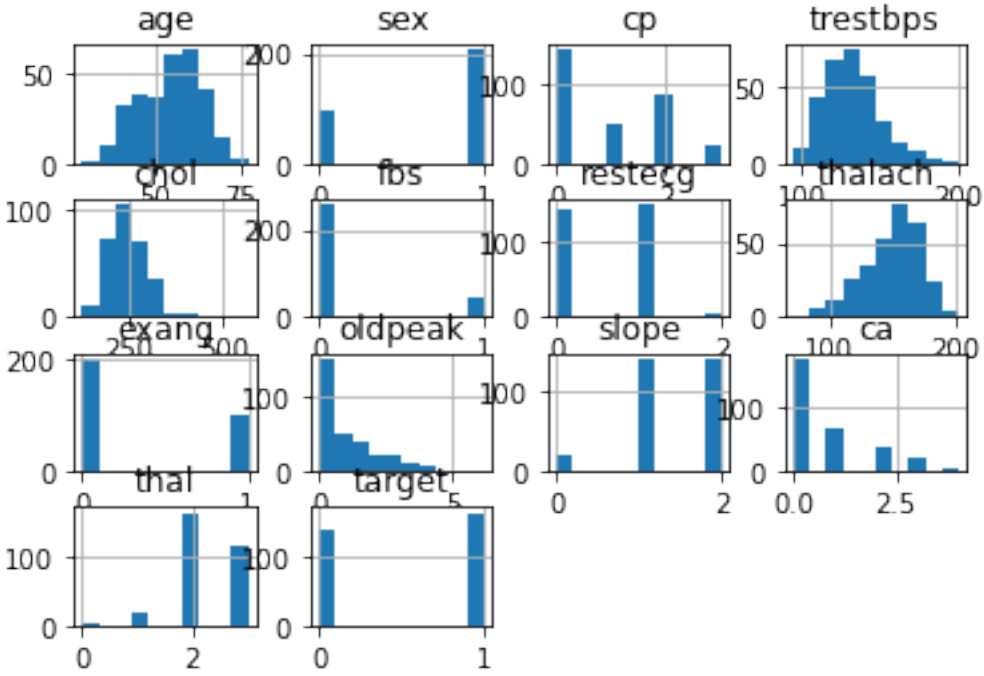
En esta sección podemos ver el histograma, que nos muestra la frecuencia con la que ciertos hechos sucedem en este caso la frecuencia de las variables relevantes para los infartos.

[47]:
```
import pandas as pd
datos = pd.read_csv('heart.csv')
df = pd.DataFrame(datos)

data.drop ([0,1]).hist()
```

[47]:
```
array([[<AxesSubplot:title={'center':'age'}>,
        <AxesSubplot:title={'center':'sex'}>,
```

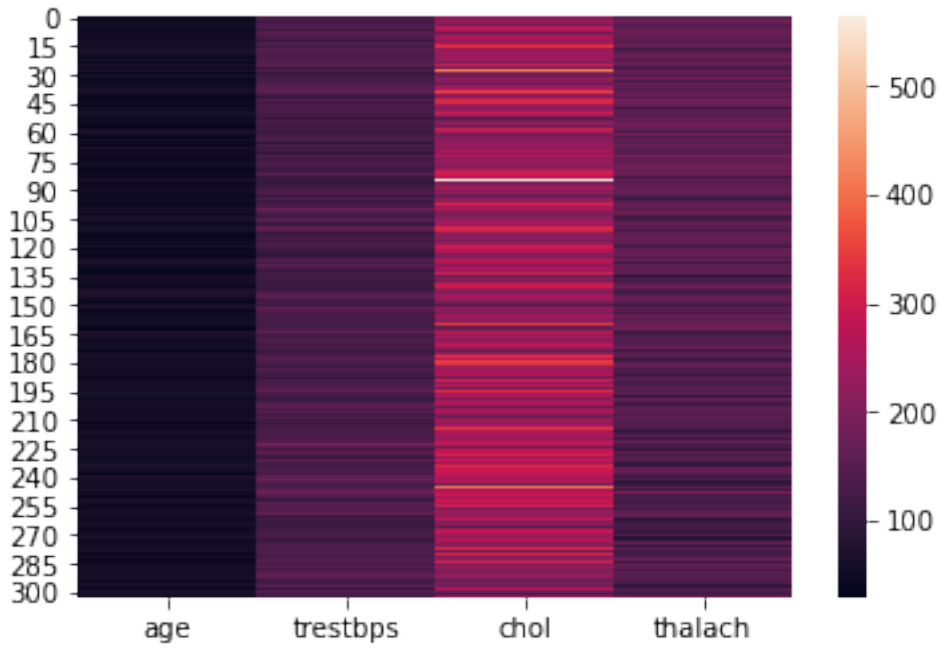
```
 <AxesSubplot:title={'center':'cp'}>,
 <AxesSubplot:title={'center':'trestbps'}>],
[<AxesSubplot:title={'center':'chol'}>,
 <AxesSubplot:title={'center':'fbs'}>,
 <AxesSubplot:title={'center':'restecg'}>,
 <AxesSubplot:title={'center':'thalach'}>],
[<AxesSubplot:title={'center':'exang'}>,
 <AxesSubplot:title={'center':'oldpeak'}>,
 <AxesSubplot:title={'center':'slope'}>,
 <AxesSubplot:title={'center':'ca'}>],
[<AxesSubplot:title={'center':'thal'}>,
 <AxesSubplot:title={'center':'target'}>, <AxesSubplot:>,
 <AxesSubplot:>]], dtype=object)
```



3.- Mapas de calor

[43]:
```
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
data = pd.read_csv('heart.csv')

uniform_data = data.
 ↪drop(['sex','cp','fbs','restecg','exang','oldpeak','slope','ca','thal','target'],␣
 ↪axis=1)
ax = sb.heatmap(uniform_data)
```



El mapa de color es una ayuda visual, que nos permite ver en conjunto los valores que estan asociados con colores para que sea más perceptible notar patrones. De este se eliminaron los datos con valores mínimos o booleanos ya que no mostraban un cambio relevante en la gráfica final por lo que en la línea 6 se excluyen.

Conclusiones

1. ¿Hay alguna variable que no aporta información? Si, existen variables que no se consideran muy relevantes por el tipo de dato que tienen en sus valores, esto nos impide analizarlo de esta forma y lo que debemos hacer para considerarlo es el proceso de normalización de la información para tener una base de datos limpia, sin valores extraños o valores NULL.

2. Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué? Quitaría las variables de resultados de electrocardiograma, depresión y pendiente del segmento ST del ciclo cardiaco de la persona. Las borraría porque considero que los datos que aportan son muy mínimos y no tan necesarios para el alcance de análisis que tendrá nuestro proyecto en este momento.

3. ¿Existen variables que tengan datos extraños? No existen valores con caracteres extraños, unicamente los booleanos que nos complican un poco el análisis pero fuera de eso al momento de buscar la base de datos intentamos que tuviera mayormente valores númericos. Las variables con valores bool son fbs, target y el sexo.

4. Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte? No todas están en rangos similares pero hay algunas que si comparten similitudes como la presión sanguínea o el colesterol en sangre ya que estos son indicios principales que nos dan una posible

alerta de que esa persona puede sufrir un infarto. El hecho de que estén en rangos similares si afecta un poco ya que al estar los datos tan parecidos y no tener variedad es posible que se convierta en un patrón repetitivo de datos y esto nos limita en cuanto a predicciones.

5. ¿Puedes encontrar grupos qué se parezcan? ¿Qué grupos son estos? Sí, como ya mencioné antes la presión sanguínea en milimetros de mercurio de las personas y el el colesterol en sangre.