REPRESENTACIÓ GRÀFICA DE LES DADES: SOLUCIÓ

Una vegada hem carregat correctament la base de dades i hem importat els mòduls necessaris, podem començar a resoldre les qüestions.

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


file_path = ('insurance.csv')

insurance_data = pd.read_csv(file_path)
```

Per tal de resoldre el primer paquet de qüestions, només ens cal cridar la funció *head()* d'aquesta manera:

```
insurance_data.head()
```

Així, obtenim les cinc primeres files de la taula. Ara podem llegir les dades de les files que ens demanen.

	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

Podem veure que l'observació 4 pertany a la regió northwest, l'observació 2 té un *bmi* de 33.000, l'observació 3 és del sexe femení i l'observació 1 té un cost de 1.725,55.

Per resoldre les qüestions del segon paquet, hem de cridar la funció *describe()*. Ho fem així:

insurance_data.describe()

La taula que obtenim és aquesta:

	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

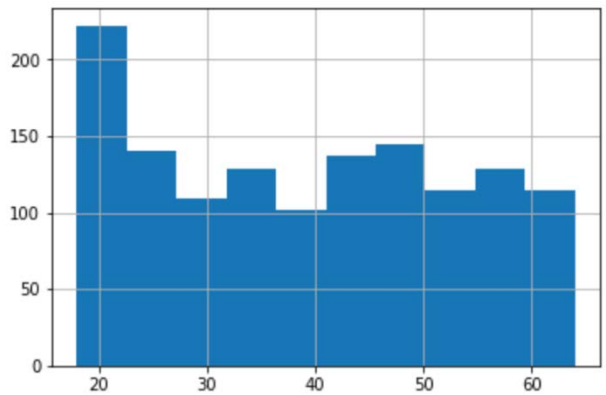
Així que el nombre d'observacions (*count*) és de 1.338. La desviació (*std*) dels valors *bmi* és de 6.09, el cost (*charge*) mínim que apareix a la taula és 1.121,87 i el nombre màxim de fills o filles que es pot veure en una observació és 5.

Per fer front al tercer grup, hem de de fer servir la funció *hist()*, que és l'encarregada de crear els histogrames. Aquesta eina ens permet veure com estan distribuïts els valors pel que fa al nombre d'observacions.

Per veure la distribució de les edats, podem escriure:

insurance_data['age'].hist()

El gràfic que obtenim és:

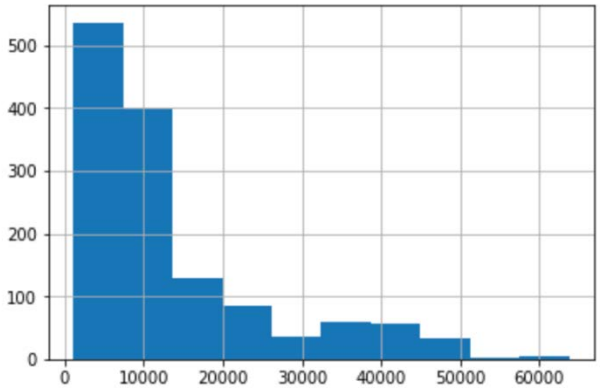


Aquí veiem, doncs, que la distribució de la població està força equilibrada, tret de l'edat que està al voltant dels 20 anys, que mostra una població significativament més gran que la resta.

Per veure la distribució del cost, podem escriure:

insurance_data['charge'].hist()

El gràfic que obtenim és:

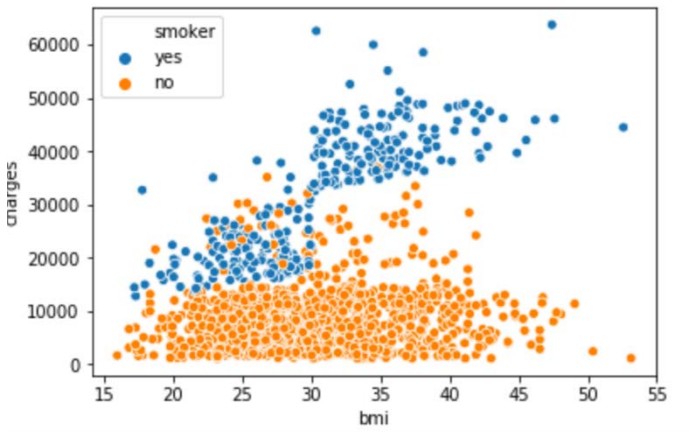


Podem veure que la gran part de les assegurances tenen un cost per sota del valor de 10.000.

Per tal de veure el gràfic de dispersió que es presenta al respecte, farem servir la funció *scatterplot()*. Per construir el primer gràfic corresponent a *charges* davant de *bmi*, escriurem aquest codi:

sns.scatterplot(x=insurance_data['bmi'],y=insurance_data['charges'], hue=insurance_data['smoker'])

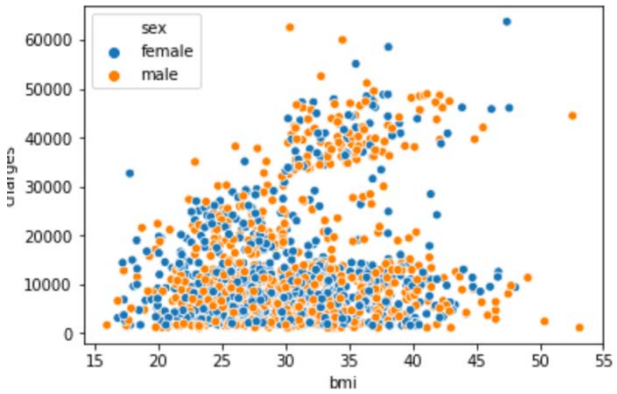
El gràfic que obtenim és:



Els dos colors diferents identifiquen si el punt pertany a una persona fumadora o no. Per tal de fer el segon gràfic, aquell que ha de mostrar els punts de diferents colors depenent del sexe, he de canviar el valor del paràmetre *hue*:

sns.scatterplot(x=insurance_data['bmi'],y=insurance_data['charges'], hue=insurance_data['sex'])

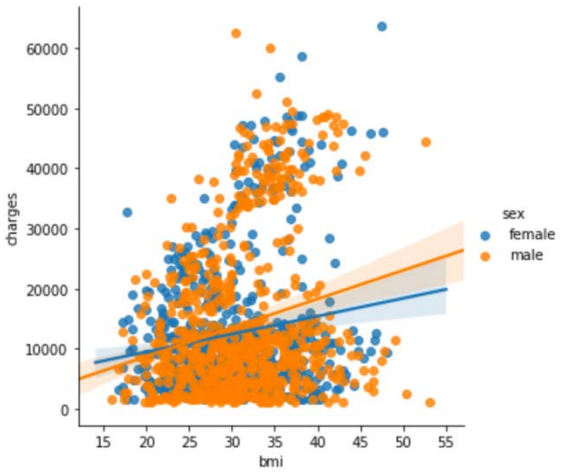
El gràfic que s'obté és:



En el primer gràfic, es veu clarament que, per a un valor de *bmi,* el cost és superior quan la persona és fumadora. Aquesta relació no es veu en el cas del sexe. Sembla que el sexe, doncs, no afecti al cost per a un *bmi* donat. En qualsevol cas, però, podem projectar una recta de regressió per veure si realment el sexe pot ocasionar cap rellevància. Recorda que hem de fer servir la funció *lmplot()* del mòdul *Seaborn*. Escrivim el codi:

Sns.lmplot(x="bmi", y="charges", hue="sex", data=4nsurance_data)

El gràfic que obtenim és aquest:

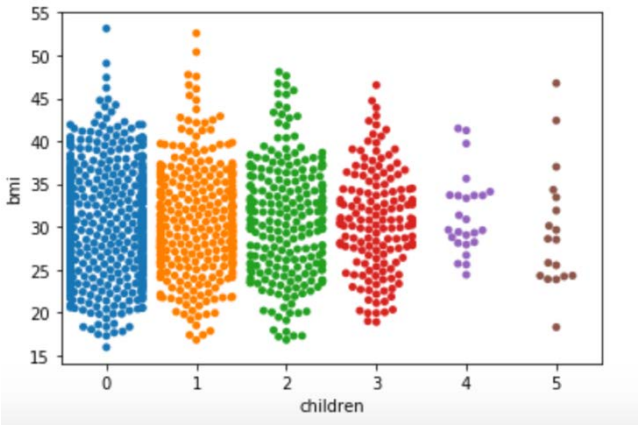


Aquest gràfic ens mostra que el sexe femení pot ser més sensible que el sexe masculí a l'augment del valor *bmi*. Això vol dir que, per a un increment donat del valor de *bmi*, l'augment del cost de l'assegurança creixerà una mica més en el sexe femení que en el masculí.

Per últim, anem a crear gràfics de dispersió per veure els valors de *bmi* davant de variables categòriques. Comencem per la variable del nombre de fills o filles (*children*). Escrivim aquest codi:

```
sns.swarmplot(x=insurance_data['children'],
              y=insurance_data['bmi'])
```

El gràfic que obtenim és:

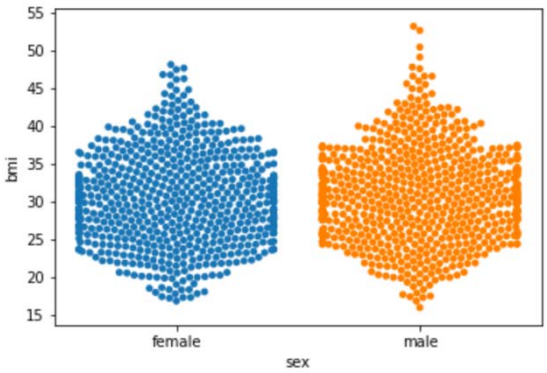


Aquí veiem, per exemple que, des de 2 fins a 4 fills o filles, el rang de *bmi* es fa cada vegada més estret.

Seguim amb el mateix gràfic de dispersió, però ara els valors de *bmi* els posarem en relació amb el sexe. Aquest és el codi:

```
sns.swarmplot(x=insurance_data['sex'],
              y=insurance_data['bmi'])
```

El gràfic que obtenim és aquest:

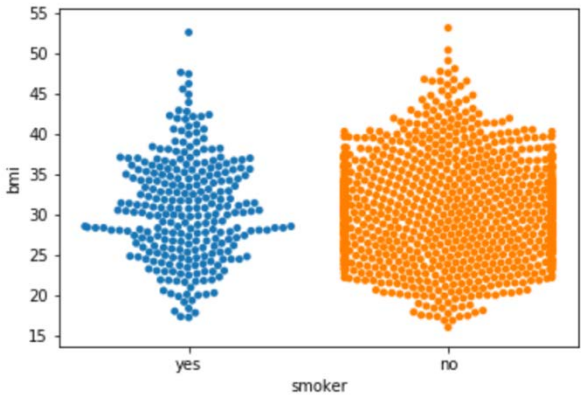


Bé, el gràfic ens ensenya que els valors més grans de *bmi* són del sexe masculí. També podem veure que els extrems dels valors de *bmi*, tan màxim com mínim, tenen forma de fletxa. Això vol dir que, a mesura que ens apropem als extrems, el nombre d'observacions disminueix. En canvi, el sexe femení no mostra aquesta forma de fletxa tan pronunciada. Això vol dir que, a mesura que ens apropem als extrems, el nombre d'observacions, tot i que disminueix, no té una forma tan marcada.

Pel que fa al valor *bmi* en relació amb si la persona és fumadora o no, només cal tornar a canviar la variable que composarà l'eix x. El codi serà:

```
sns.swarmplot(x=insurance_data['smoker'],
        y=insurance_data['bmi'])
```

El gràfic que veurem és:



El primer que veiem clarament és que el nombre d'observacions que corresponen a persones no fumadores és força més gran que el de les fumadores. A part d'això, no es veu, en un primer moment, que el sexe afecti als valors de *bmi*.

Descobreix tot el que Barcelona Activa pot fer per a tu

Acompanyament durant tot el procés de recerca de feina

barcelonactiva.cat/treball

Suport per posar en marxa la teva idea de negoci

barcelonactiva.cat/emprenedoria

Serveis a les empreses i iniciatives socioempresarials

barcelonactiva.cat/empreses

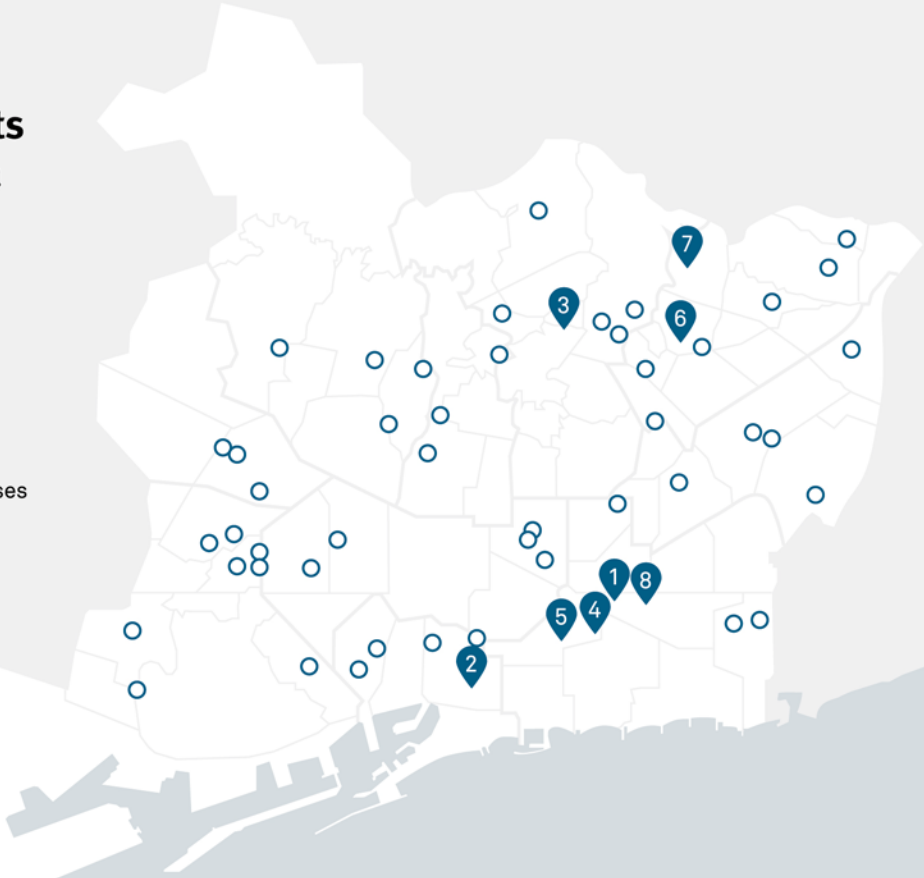
Formació tecnològica i gratuïta per a la ciutadania

barcelonactiva.cat/cibernarium

Xarxa d'equipaments de Barcelona Activa

1. Seu Central Barcelona Activa
   Porta 22
   Centre per a la Iniciativa Emprenedora Glòries
   Incubadora Glòries
2. Convent de Sant Agustí
3. Ca n'Andalet
4. Oficina d'Atenció a les Empreses
   Cibernàrium
   Incubadora MediaTIC
5. Incubadora Almogàvers
6. Parc Tecnològic
7. Nou Barris Activa
8. innoBA

○ Punts d'atenció a la ciutat



Barcelona Activa

Ajuntament de Barcelona

Darrera actualització 2019

Cofinançat per:

UNIÓ EUROPEA
Fons Europeu de Desenvolupament Regional

Segueix-nos a les xarxes socials:

barcelonactiva.cat/cibernarium
barcelonactiva
barcelonactiva
company/barcelona-activa