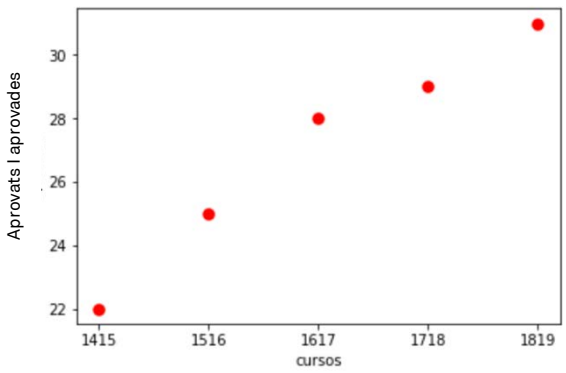
SCATTER PLOTS

Un gràfic de dispersió (*scatter plot*, en anglès) mostra els punts que relacionen els valors de l'eix *x* amb els valors de l'eix *y*. Si, per exemple, agafem les dades d'alumnes aprovats i aprovades respecte als cursos, com podem generar un gràfic de dispersió? Si canviem la funció *plot()* per *scatter()*, ja ho tindrem. Escrivim aquest codi:

```
cursos = ['1415', '1516', '1617', '1718', '1819']

aprovats = [22,25,28,29,31]

dades = plt.scatter(cursos,aprovats)

plt.xlabel('cursos')

plt.ylabel('aprovats i aprovades')

plt.setp(dades, color='red', linewidth=2.5)

plt.show()
```

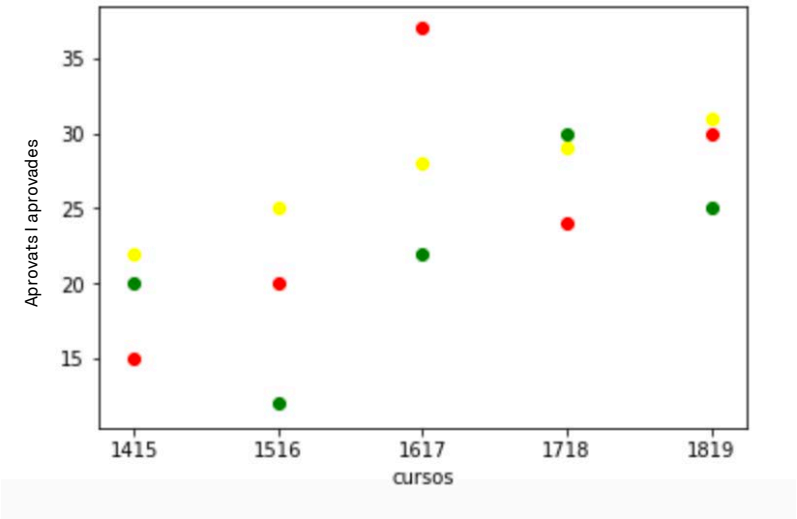
La funció plt.scatter rep els paràmetres que formen els valor de l'eix *x* (cursos) i *y* (persones aprovades). El gràfic que obtenim, quan executem el codi, és:



Si suposem que tenim més d'un grup fent la mateixa assignatura, per exemple, tres grups, podem visualitzar el nombre de persones aprovades per any d'aquests tres grups en el mateix gràfic. Ho farem així:

```
cursos = ['1415', '1516', '1617', '1718', '1819']

aprovats1 = [22,25,28,29,31]

aprovats2 = [15,20,37,24,30]

aprovats3 = [20,12,22,30,25]

plt.scatter(cursos,aprovats1, color='yellow')

plt.scatter(cursos,aprovats2, color='red')

plt.scatter(cursos,aprovats3, color='green')

plt.xlabel('cursos')

plt.ylabel('aprovats i aprovades')

plt.show()
```

Els i les alumnes aprovades del grup 1 (aprovats1) es representaran de color groc, el del grup 2 de color vermell i el del grup 3 de color verd.



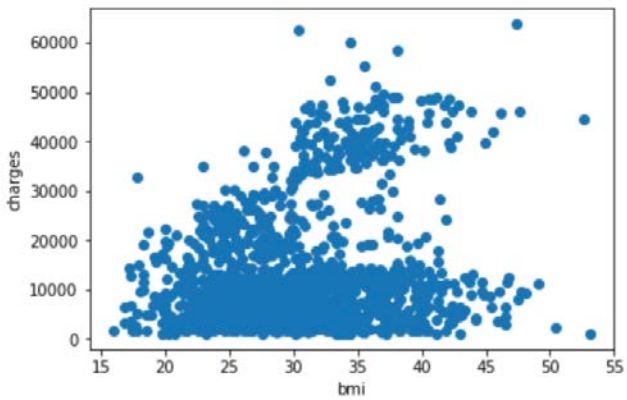
Anem, ara, a treballar amb una taula de dades més gran. Fes clic aquí per descarregar-te el fitxer en format CSV. Aquesta taula és un exemple d'una empreses d'assegurances. Després, escriu aquest codi:

```
file_path = ('insurance.csv')

insurance_data = pd.read_csv(file_path)

print(insurance_data.head(10))
```

Veuràs les deu primeres observacions de la taula. Anem a veure la relació entre les dades *bmi* i *charges*:

```
plt.scatter(insurance_data['bmi'], insurance_data['charges'],)

plt.xlabel('bmi')

plt.ylabel('charges')

plt.show()
```

Aquest és el gràfic que obtenim:



Nota: *bmi* és el valor de *Body Mass Index* i *charges és* l'import de risc de l'assegurança.

Fixa't que, en el codi, la llista de valors de l'eix *x* i l'eix *y* estan dins d'*insurance_data['bmi'], insurance_data['charges].*

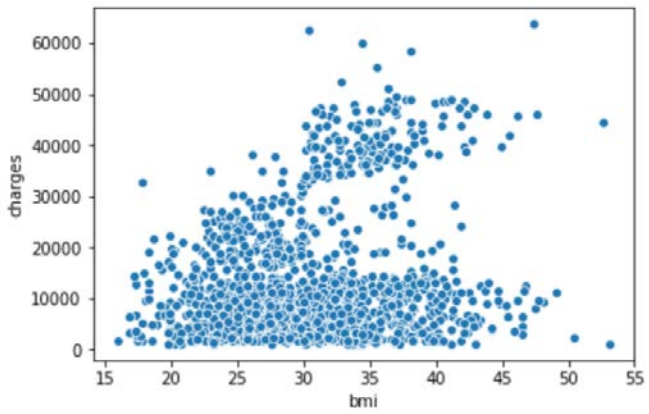
A continuació, utilitzarem una altra llibreria basada en *Matplotlib* que es diu *Seaborn*. Per tal de carregar aquesta llibreria i poder fer-la servir, hem d'importar-la.

```
import seaborn as sns
```

Podem generar un gràfic de dispersió amb les mateixes dades que en el cas anterior:

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

El gràfic que obtenim és el següent:

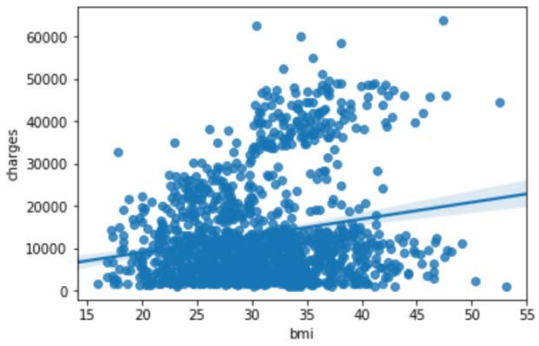


Fixa't que els punts tenen més definició i que no hem hagut de declarar el nom dels eixos. La mateixa funció *scatterplot()* ja els ha generat.

Anem a veure, ara, si hi ha una correlació entre els valors *bmi* i el preu de la pòlissa. Per fer això, hem de fer servir una altra funció anomenada *regplot()*:

```
sns.regplot(x=insurance_data['bmi'], y=insurance_data['charges'])
```

El gràfic que obtenim és el següent:

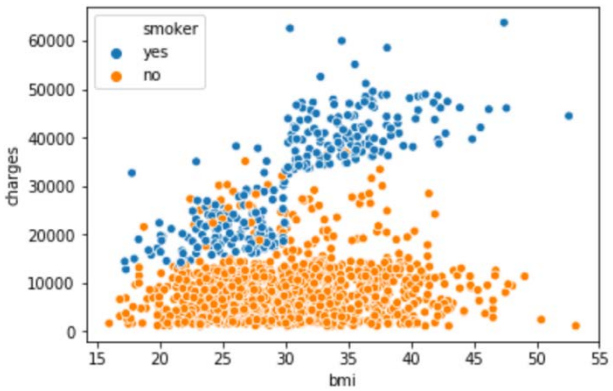


Segons el gràfic, efectivament hi ha una correlació positiva. A mesura que el valor *bmi* d'una p ersona augmenta, també creix el preu.

Si et fixes en les dades de la taula, veuràs que hi ha una columna anomenada *smoke*, que ind ica si aquella persona és fumadora o no. Ara farem que els punts del gràfic de dalt tinguin un color o un altre, depenent de si la persona és fumadora o no. Afegim un nou paràmetre a la fu nció *scatterplot()*:

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

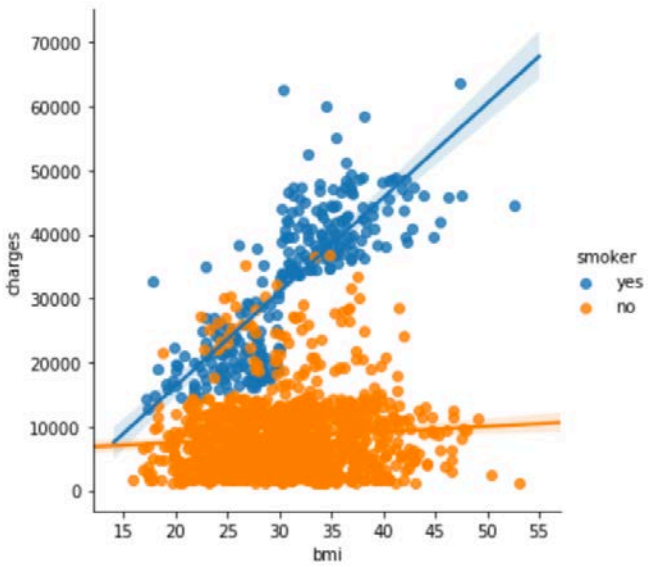
El gràfic que obtenim és el següent:



Aquí veiem que, clarament, si la persona és fumadora, el preu de la seva assegurança serà més alt. I, si volem veure si la relació entre *bmi* i *charges* és diferent, en funció de si una persona és fumadora o no, hem d'executar aquest codi:

```
sns.lmplot(x="bmi", y="charges", hue="smoker", data=insurance_data)
```

El gràfic generat per la funció *lmplot()* és el següent:



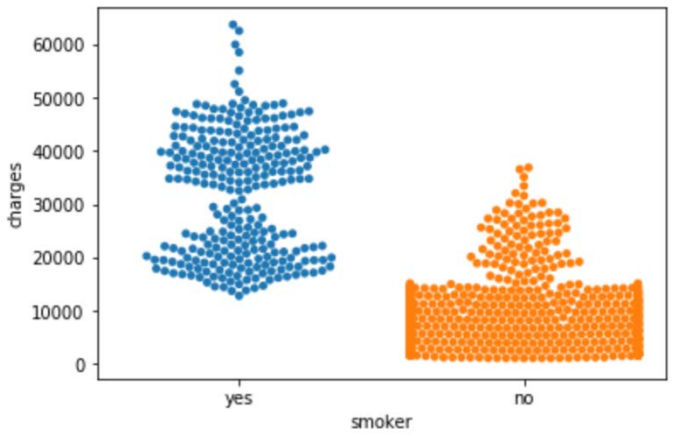
Efectivament, veiem que, per a una persona fumadora, el preu de la seva pòlissa augmenta considerablement quan té un *bmi* més alt, mentre que les persones no fumadores tenen un augment molt més petit.

Fins ara, hem fet gràfics de dispersió entre dues variables numèriques, però també podem fer que una variable sigui categòrica (no numèrica). Per exemple, podem veure la dispersió de preus de les pòlisses, en funció de si la persona és fumadora o no.

En aquest cas, farem servir la funció *swarmplot().* Escriurem el codi següent:

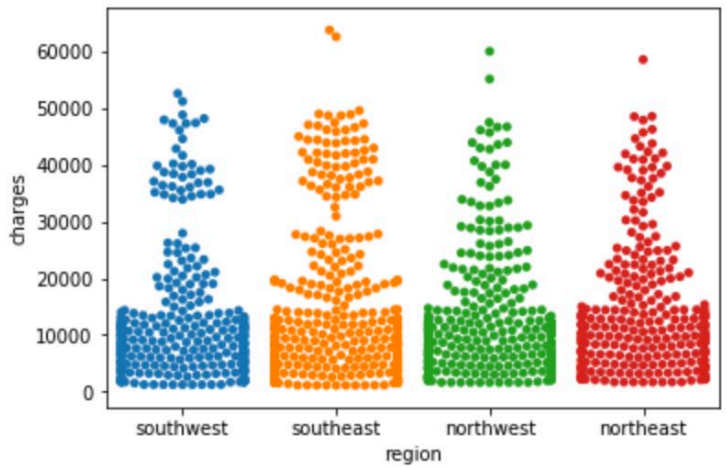
```
sns.swarmplot(x=insurance_data['smoker'], y=insurance_data['charges'])
```

El gràfic que genera la funció és aquest:



Aquest gràfic ens ensenya que la gran majoria de persones no fumadores tenen un preu inferior al de les persones fumadores.

Si ara volem veure com són els preus, depenent de la regió, hem d'escriure el codi següent:

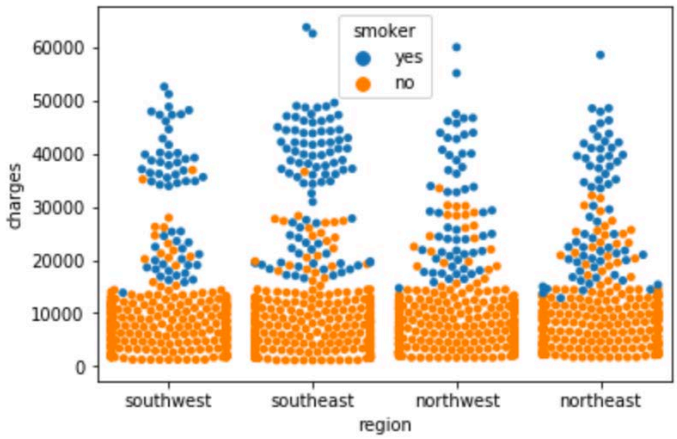


sns.swarmplot(x=insurance_data['region'], y=insurance_data['charges'])

Amb el gràfic generat, veiem la dispersió de preus agrupats en les quatre regions que té la taula. En principi, la majoria de la població no presenta diferents preus, segons la seva regió. I què passa, si afegim a l'anàlisi la variable *smoke*? Per saber-ho, haurem d'afegir a la funció *swarmplot* el paràmetre *hue* d'aquesta manera:

```
sns.swarmplot(x=insurance_data['region'],y=insurance_data['charges'],
hue=insurance_data['smoker'])
```

I visualitzem aquest gràfic:



Aquí, una altra vegada, comprovem que el factor més determinant per veure com varia el preu d'una assegurança és saber si la persona és fumadora o no.

Seguim!

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

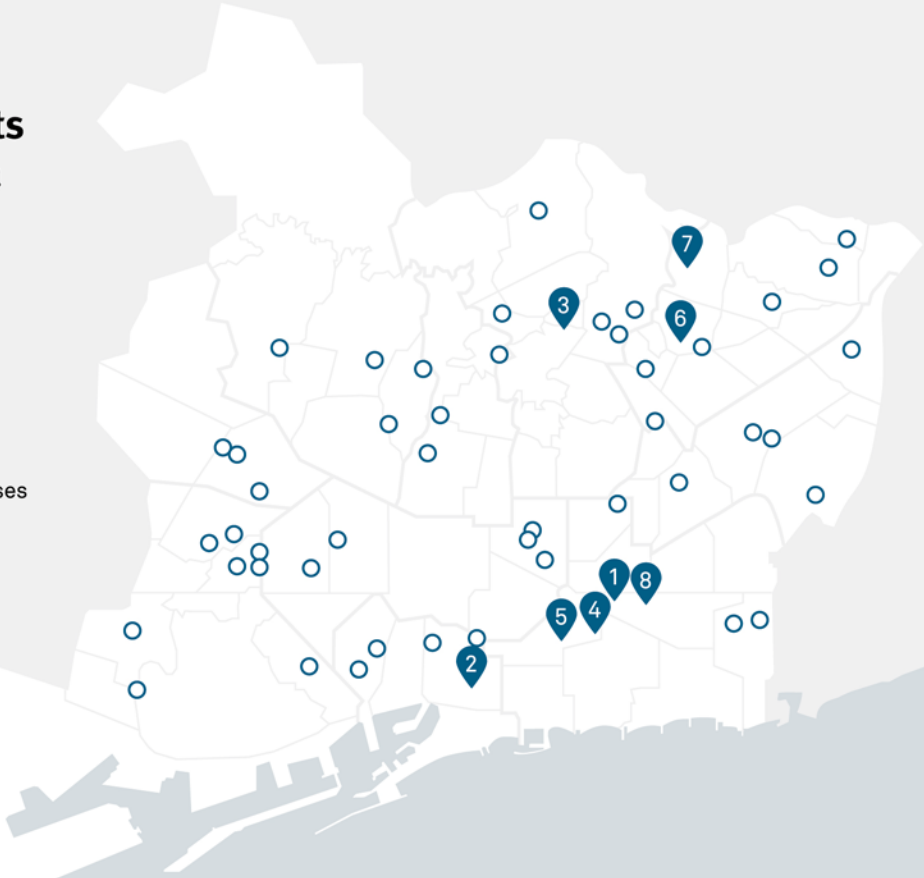
Formació tecnològica i gratuïta per a la ciutadania

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