

PujarNota-Examen-1-Parcial-Soluc...



alucero



Visualització de Dades



3º Grado en Ingeniería de Datos



Escuela de Ingeniería
Universidad Autónoma de Barcelona

antes



**Descarga sin publi
con 1 coin**



Después

WUOLAH



Importante

Puedo eliminar la publi de este documento con 1 coin

¿Cómo consigo coins? → Plan Turbo: barato
→ Planes pro: más coins

perdo
espacio



Necesito
concentración

ali ali ooh
esto con 1 coin me
lo quito yo...

WUOLAH

Visualització de Dades (Enginyeria de Dades – EE - UAB)
Examen PUJAR NOTA Primer Parcial – 1 Juliol 2024
SOLUCIONS

Nom i Cognoms: _____

NIU: _____ Grup de Matrícula: _____

Només es permet l'ús d'internet per l'accés al campus virtual en el moment de descarregar el full d'enunciats y d'entregar l'examen.

Sólo se permite el uso de internet para el acceso al campus virtual en el momento de descargar la hoja de enunciados y de entregar el examen.

PARTE 1 (2 pt)

Preguntas de teoría, no hay dataset.

1.1. (1 pt) Supón que tenemos un dataset que contiene la edad media en España desde 1990 a 2020. Tiene 2 columnas, "hombres" y "mujeres" y cada fila es el valor de 1 año.

Año	hombre	mujeres
1990	45	47
1991	46	49
...		
2020	48	48

¿Cuál sería la gráfica más adecuada para visualizar la **diferencia** entre hombres y mujeres **a lo largo del tiempo**? La gráfica debería mostrar la serie temporal completa.

Explica qué gráfica es más adecuada y razona tu respuesta. **No hace falta hacer la gráfica.**

RESPOSTA:

La gráfica más adecuada es un difference chart, una gráfica de líneas en la que el espacio entre las dos líneas (la diferencia) se representa como un área de color. Alternativamente, una gráfica de líneas serviría porque se trata de una serie temporal de un atributo cuantitativo continuo, aunque sería menos efectiva que la anterior.

1.2. (0.25 pt) Marca el tipo de datos que es mejor codificar con cada uno de los componentes del color:

- | | | | |
|----|------------|-------------|---------------|
| a) | Tono | categoricos | cuantitativos |
| b) | Saturación | categoricos | cuantitativos |
| c) | Luminancia | categoricos | cuantitativos |

RESPOSTA:

a) categoricos, b) cuantitativos, c) cuantitativos

1.3. (0.25 pt) ¿Cuál es el número máximo recomendado de colores distintos en una paleta categórica?

RESPOSTA:

12

1.4. (0.5 pt) Nombra 2 problemas derivados de usar la escala de color arcoíris para codificar datos cuantitativos.



RESPOSTA:

- No tiene un orden inherente
- Hay colores que percibimos más intensamente que otros
- “Bandas” perceptuales
- Pérdida de detalle

PARTE 2 (3 pt)

Dataset: 2017_accidents_vehicles_gu_bcn.csv

2.1. (0.5 pt) Inspecciona el fichero. ¿Qué tipo de atributo son: “Mes de any”, “Descripció torn”, “Longitud” y “Número de víctimes”? ¿Qué atributo sería el key (clave primaria) del dataset?

RESPOSTA:

Mes de any- Ordinal, Descripció torn- Ordinal, Longitud- Cuantitativo (espacial), Número de víctimes- Cuantitativo.

El key es Número d'expedient

2.2. (1 pt) Visualiza la distribución de “Número de vehicles implicats”. Sube la gráfica debidamente anotada y el código (0.5 pt).

¿Cual es el número de vehículos implicados más frecuente? ¿Cual es la gráfica más adecuada para visualizar la distribución? Justifica brevemente tu respuesta. (0.5 pt)

RESPOSTA:

Lo más adecuado es un histograma con bin 1 o un bar chart porque permite comparar los tamaños de las barras entre si y todas a la vez. El número más frecuente es 2.

2.3. (1.5 pt)

Haz una gráfica que te permita visualizar el número de **accidentes por hora** del día, para los **7 días de la semana a la vez**. Puedes usar “facet” para hacer las gráficas de los 7 días. **Sube las gráficas y el código completo** (1 pt).

¿Qué tipo de visualización es la más adecuada? Justifica brevemente tu respuesta (0.5 pt).

RESPOSTA:

La visualización más adecuada son gráficas yuxtapuestas de conteo de accidentes por hora del día, una gráfica por cada día de la semana. Así se puede comparar valores entre horas para cada día y también entre días. Cada gráfica sería un area chart, aunque un bar chart serviría también.

```
day <- df$'Dia setmana'
hour <- df$'Hora de dia'
# Small multiples de accidentes por hora y por dia de la semana
ggplot(df, aes(hour)) + geom_bar() + labs(title="accidentes", x="Hora", y="count") +
  facet_wrap(day, ncol=1)
```

PARTE 3 (5 pt)

Dataframe: beers.csv

En aquesta part de l'examen cal incloure, a més del que demani l'enunciat:

- *Les llibreries necessàries*
- *les comandes R*
- *una captura de pantalla de la gràfica.*

Aquest dataframe té les següents variables:

- **Brewery:** Nom de la cerveseria
- **Beer:** Nom de la cervesa
- **Description:** Descripció
- **Style:** Estil de la cervesa
- **ABV:** El percentatge d'alcohol contingut per volum
- **IBU :** Unitats Internacionals d'Amargor - una mesura aproximada de l'amargor en una cervesa a causa de la quantitat de llúpols que conté
- **Rating:** Puntuació que ha rebut de 1 a 100. Sent 100 la més puntuada

I unes altres dades basades en dades sensorials

- **quality** (amb una puntuació entre 0 i 10)

NOTA: Fer ús de pipes quan sigui possible

3.1 (3 pt)

a) Quin tipus de variables són *Style* i *ABV*? Quantes observacions i variables té el dataframe? (0.25 pts)

b) Fes un gràfic que mostri en una sola figura d'un sol panel, la distribució de la puntuació aconseguida per cada estil de cervesa. Afegiu etiquetes als eixos i doneu una conclusió (0.5 pts)

Importante

Puedo eliminar la publi de este documento con 1 coin

¿Cómo consigo coins? → Plan Turbo: barato
→ Planes pro: más coins

perdo
espacio



c) Fes un gràfic que mostri la distribució de l'ABV. Posa un títol a la gràfica i dóna una conclusió. (0.5 pts)

c) Mostra la distribució de la variable **IBU** (0.25 pts). Després, fes un multipanel que permeti comparar les distribucions del **ABV** i **IBU**. Extreu conclusions veient ambdós gràfics. Pots veure bé la distribució de la variable **IBU** en el multipanel? Per què creus que és degut? (1.5 pts). Nota: En cas de no poder realitzar un multipanel i fer dues figures separades, es comptarà només 0.75 pts.

RESPOSTA:

Càrrega de les llibreries i dataset:

```
> library(tidyverse)
> library(dplyr)

> setwd("C:/Users/...")
> beers <- read.csv('./beers.csv')
> str(beers) ...o...>view(beers)
```

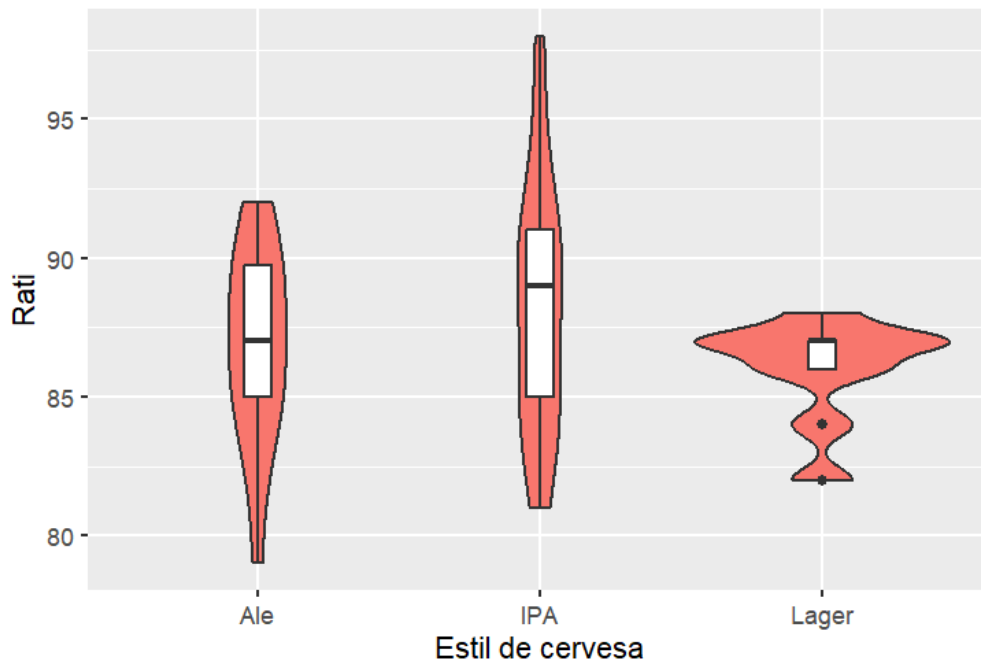
a) Style és una variable categòrica i ABV és una variable quantitativa
El dataframe té 44 observacions i 7 variables.

The screenshot shows the RStudio interface. At the top, the 'Environment' pane displays the 'beers' dataset with 44 rows and 7 columns. Below it, the 'Console' pane shows the output of the `str(beers)` command, which identifies the data as a tibble with 44 rows and 7 columns. The columns are: `Brewery` (character), `Beer` (character), `Description` (character), `Style` (character), `ABV` (numeric), `IBU` (numeric), and `Rating` (numeric).

Brewery	Beer	Description	Style	ABV	IBU	Rating
Urban Growler	Cowbell Cream Ale	Cream Ale	Ale	5.2	20	82
Urban Growler	Midwest IPA	English IPA	IPA	6.2	60	83
Urban Growler	De-Lovely Porter	Porter	Ale	5.6	33	86
Urban Growler	Kentucky Uncommon	Kentucky Common Beer	Ale	5.5	40	84
Urban Growler	Big Boot Rye IPA	Rye IPA	IPA	6.5	66	84
Surly	Bender	Oatmeal Brown Ale	Ale	5.5	45	92
Surly	Coffee Bender	Brown Ale with Coffee	Ale	5.5	45	92
Surly	CynicAle	Belgian Style Pale Ale	Ale	6.5	33	90
Surly	Furious	India Pale Ale	IPA	6.6	99	95
Surly	Hell	German Style Munich Helles Lager	Lager	5.3	20	87

b) Tenim una variable ABV contínua de la que hem de mostrar la seva distribució per les tres categories d'*style* (discreta), per tant, el més adequat é fer un diagrama de violí i/o boxplot (no cal fer els dos, tot i que també es poden adherir si voleu)

```
> beers%>%ggplot(aes(x=Style, y=Rating))+geom_violin(aes(fill="red"))
+geom_boxplot(width = 0.1)+xlab("Estil de cervesa")+ylab("Rati")+the
me(legend.position = "none")
```

En tots els casos els valors es mouen per sobre de 75 punts. En el cas d'ALE la distribució de la puntuació es mou entre 75-92 punts amb una mediana d'uns 87 punts. En el cas de Lager en canvi, veiem dos outliers, però fora d'aquests outliers podem dir que tenim una distribució molt més focalitzada entre els 85 i 87,5 punts, amb una mediana una mica superior a 86 punts.

- c) Com ABV és una variable contínua, farem ús de un *geom_density* o *geom_histogram*.

One Variable

Continuous

```
c <- ggplot(mpg, aes(hwy))
```




























- 
c + geom_area(stat = "bin")
 x, y, alpha, color, fill, linetype, size
 a + geom_area(aes(y = ..density..), stat = "bin")
- 
c + geom_density(kernel = "gaussian")
 x, y, alpha, color, fill, group, linetype, size, weight
- 
c + geom_dotplot()
 x, y, alpha, color, fill
- 
c + geom_freqpoly()
 x, y, alpha, color, group, linetype, size
 a + geom_freqpoly(aes(y = ..density..))
- 
c + geom_histogram(binwidth = 5)
 x, y, alpha, color, fill, linetype, size, weight
 a + geom_histogram(aes(y = ..density..))

Una possible solució seria:

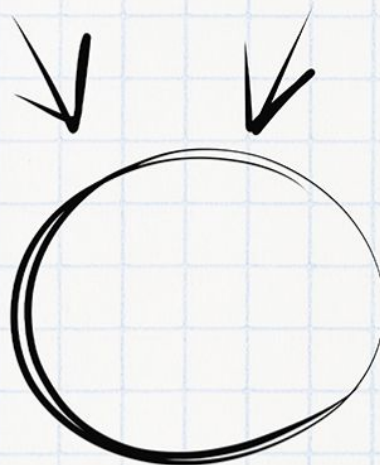
```
>ggplot(beers)+aes(ABV)+geom_density(fill='grey')+theme_bw()+ggtitle
("Distribució del percentatge d'alcohol contingut per
volum")+xlab('ABV')+ylab('Densitat')
```

Imagínate aprobando el examen

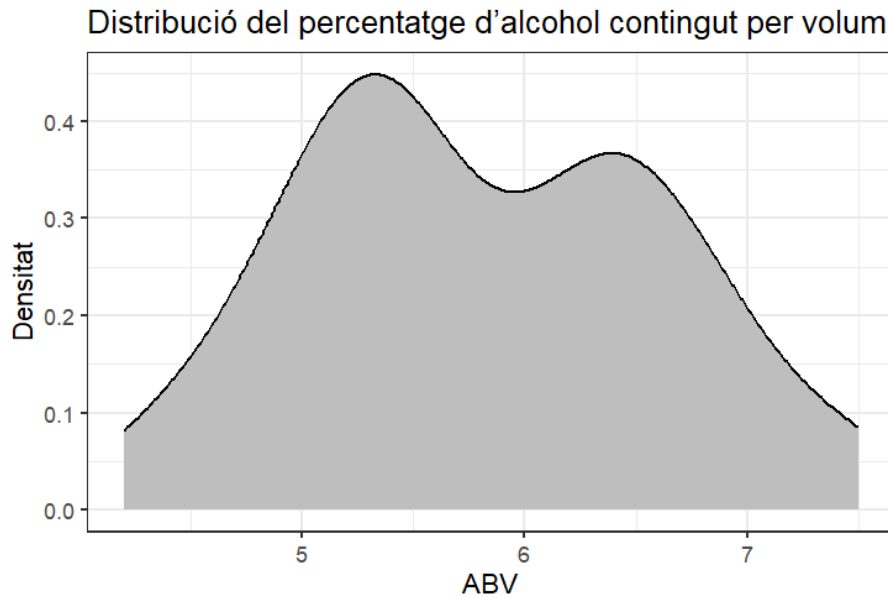
Necesitas tiempo y concentración

Planes	 PLAN TURBO	 PLAN PRO	 PLAN PRO+
 Descargas sin publi al mes	10 	40 	80 
 Elimina el video entre descargas			
 Descarga carpetas			
 Descarga archivos grandes			
 Visualiza apuntes online sin publi			
 Elimina toda la publi web			
 Precios Anual <input type="checkbox"/>	0,99 € / mes	3,99 € / mes	7,99 € / mes

Ahora que puedes conseguirlo,
¿Qué nota vas a sacar?

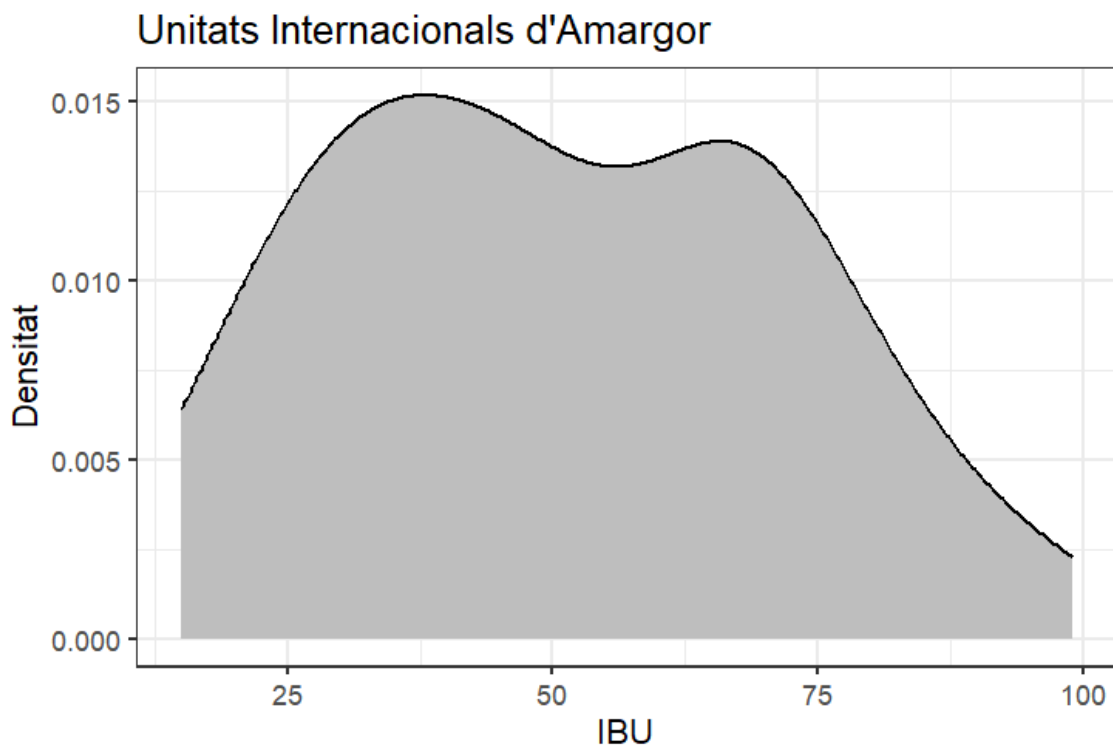


WUOLAH



L'ABV es mou entre 5-7 %. La majoria de cerveses però tenen un ABV entre 5-5,5%. Podem dir inclús que l'ABV segueix una distribució bimodal amb una moda al voltant del 5,25% i una altra al voltant del 6,5%.

d) `ggplot(beers)+aes(IBU)+geom_density(fill='grey')+theme_bw()+ggtitle("Unitats Internacionals d'Amargor")+xlab('IBU')+ylab('Densitat')`



Importante

Puedo eliminar la publi de este documento con 1 coin

¿Cómo consigo coins? → Plan Turbo: barato
→ Planes pro: más coins

pierdo espacio



Necesito concentración

ali ali ooh
esto con 1 coin me
lo quito yo...

WUOLAH

Com abans tenim una variable bimodal, però distribuïda entre molts valors d'IBU, el que fa que la densitat màxima sigui de 0,015 en la primera moda (IBU al voltant de 40) i de poc més de 0,0125 (per un IBU al voltant de 73).

Pel facet:

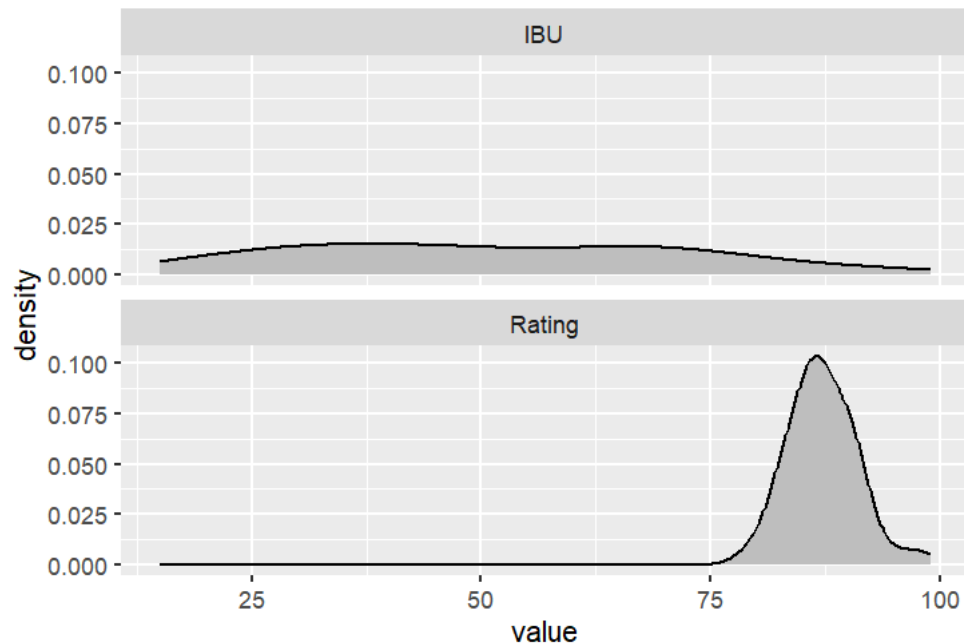
Tenim dues distribucions de variables numèriques contínues i s'haurà de fer un histograma/densitats per cada variable. Podeu fer un facet (dues files) amb un histograma/gràfic de densitats per cada variable (en cada casella).

Abans però hem de construir un amb les mètriques que necessitem. Primer per simplificar, fem un dataframe que contingui cadascuna d'elles en una columna:

```
> df<-beers%>%select(c("IBU", "Rating"))
```

Un cop el tenim, usem *gather*, com vam fer en el seminari 4 part 3, per construir un dataframe que ens construeixi les mètriques que necessitem i ja podem fer el facet

```
> df_long <-df%>%gather(IBU, Rating, key='metric', value='value')  
> ggplot(df_long)+aes(value)+geom_density(fill="grey")+facet_wrap(~  
metric, ncol=1)
```



Podem dir que el rating segueix una distribució bastant normalitzada entre 75 i 100 punts, amb una moda al voltant de 87 punts. Ara bé, al fer el gràfic multipanel ens podem adonar que perdem una mica la informació referent a la distribució de la IBU, això és degut a que quan fem un multipanel per facilitar la comparació entre els panels, vam veure que havíem de posar els mateixos eixos. Ara bé, els valors que prenen aquestes dues variables es distribueixen entre dos rangs força diferents, el que fa que no podem comparar-les com voldríem en un multipanel. Ens seria útil si volem justament mostrar aquesta diferència de distribució de valors, però si volem saber quelcom més sobre com és la distribució perdem informació.

3.2. (0,75 pt)

a) Volem mostrar en una figura la distribució d'una variable contínua per diferents categories. Pots utilitzar un violin plot, un boxplot, o ambdues? Si la resposta és ambdues quina és la gràfica més adient i per què?

b) Digues dos gràfics dels que hem vist que serveixin per reduir la dimensionalitat i explica alguna de les avantatges d'utilitzar-los

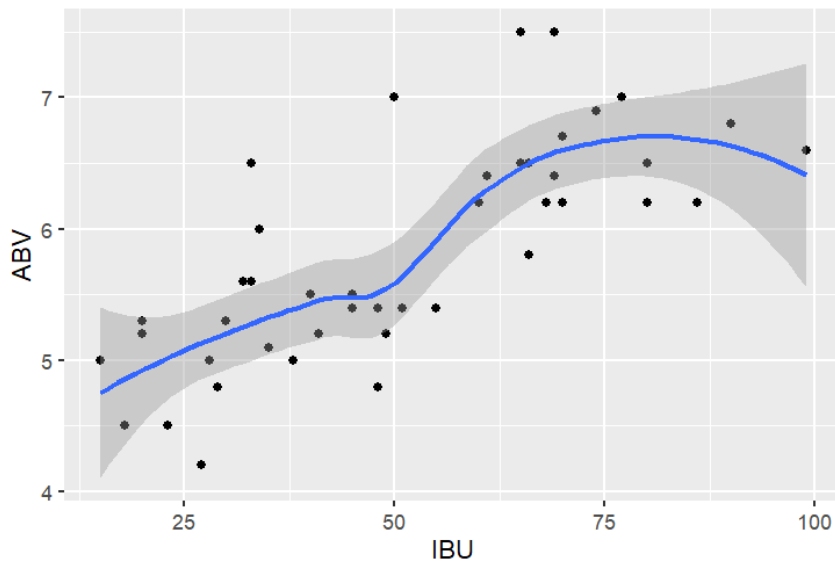
c) Dona un exemple on un bubble plot seria millor que un scatter plot

3.3 (1,25 pt) Voleu conèixer la relació entre les unitats internacionals d'Amargor (IBU) i el percentatge d'alcohol contingut per volum (ABV). Pots ajustar aquesta relació a algun patró? Afegeix un gràfic i títol a la gràfica. Què ajusta millor, un 'Locally Estimated Scatterplot Smoothing (loess)' amb un interval del 95 % de confiança o un model lineal (lm) amb un 85% de confiança? Presenta el que més s'ajusti amb etiquetes als eixos.

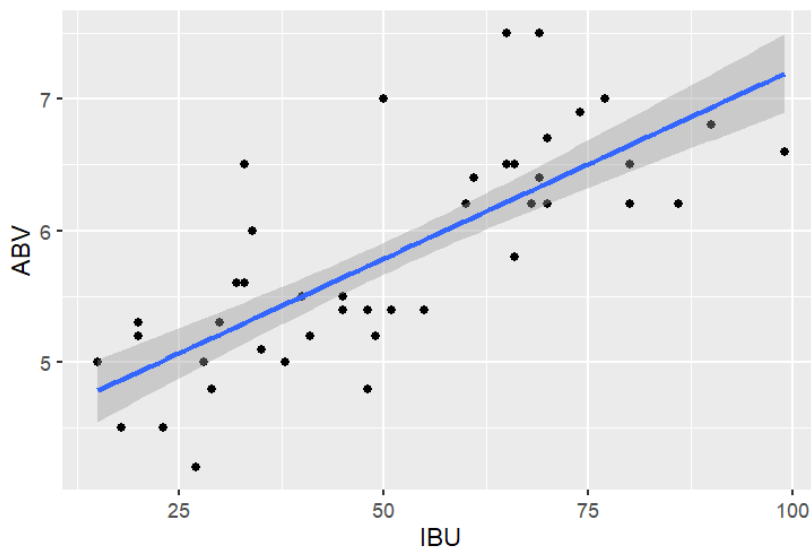
Fem la gràfica de punts i ajustem pels dos patrons que ens diuen

```
>beers%>%ggplot(aes(IBU,ABV))+geom_point()+geom_smooth(method='loess', level=0.95)
0:
```

```
>beers%>%ggplot(aes(IBU,ABV))+geom_point()+geom_smooth(method='loess')
'
```



```
>beers%>%ggplot(aes(IBU,ABV))+geom_point()+geom_smooth(method='lm',
level=0.85)
```



S'ajusta millor el loess amb un interval de confiança del 95%. Posem títol i etiquetes als eixos

```
>beers%>%ggplot(aes(IBU,ABV))+geom_point()+geom_smooth(method='loess')
)+xlab('Unitats Internacionals Amargor')+ylab('Percentatge d'alcohol
contingut per volum')
```

Importante

Puedo eliminar la publi de este documento con 1 coin

¿Cómo consigo coins? → Plan Turbo: barato
→ Planes pro: más coins

perdo
espacio



Necesito
concentración

ali ali ooh
esto con 1 coin me
lo quito yo...

WUOLAH

