

1rParcial20232Soluciones.pdf



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 Primer Parcial – 11 Abril 2023
ENUNCIATS MODELS 1,2

Nom i Cognom: _____

NIU: _____ Grup de Matrícula: _____

PARTE 1 (5 pt)

Dataset: *PIB_population_lifeExpectancy.csv*

1.1. (0.5 pt) Abre el fichero. ¿Qué tipo de atributo son: Country, Region, RegionCode, LifeExpectancy_2018, PIBperCapita_2018, y Population_2018?

RESPOSTA:

Country-Categórica; Region-Categórica; RegionCode-Ordinal; LifeExpectancy-Cuantitativa; PIBperCapita- Cuantitativa; Population-Cuantitativa

1.2. Haz un bubble chart con las variables LifeExpectancy_2018, PIBperCapita_2018, y Population_2018 (puedes controlar el tamaño de los puntos con `aes(..., size)`). Cualquiera de las 3 variables se puede codificar en cualquiera de los canales disponibles: X, Y, y tamaño.

A) (1 pt) ¿Cuál es el mapeo óptimo de esos 3 canales para explorar la correlación entre variables? Justifica brevemente tu respuesta.

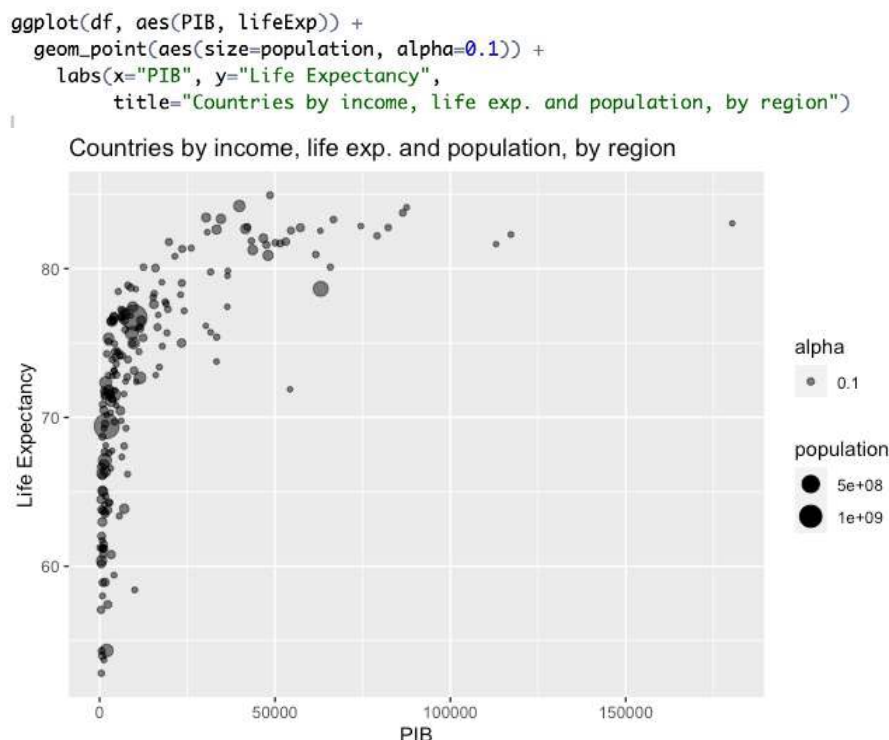
RESPOSTA:

Life Expectancy y PIB son dos variables con una correlación alta, expresarlas como posición en los ejes X e Y del scatterplot potencia una de las fortalezas de esta gráfica, que es mostrar correlaciones.

Por otro lado, Population tiene outliers. Expresarlos como posición resulta en una concentración de puntos en los rangos más bajos y más altos, siendo difícil distinguir entre los bajos y dejando mucho espacio sin usar en el medio. En cambio, esta variable codificada como tamaño también resalta la existencia de los outliers pero nos proporciona una capa más de información al no interferir significativamente con la correlación observada entre las otras dos.

B) (1 pt) Sube el código y la gráfica.

RESPOSTA:



1.3. Mejora la gráfica añadiendo transparencia, color, bordes, etc., y las leyendas necesarias para que la gráfica sea comprensible.

A) (1 pt) Haz una gráfica en la que el color represente la variable *Region*. Explica brevemente la ventaja de usar el color para esta variable en relación a las demás, y el tipo de paleta que utilizas.

RESPOSTA:

Colorear por región nos permite añadir una variable más a la gráfica: una capa más de información que, por las características propias de este dataset (solo 5 categorías y una distribución en el scatter que encaja con los bloques por regiones), no solo no interfiere con la legibilidad de la gráfica, sino que además aporta una lectura más rica. La paleta o escala de color debe ser categórica porque la variable también lo es.

B) (1 pt) Haz una gráfica en la que el color sea una de las 3 variables que ya has usado en X, Y, o tamaño. Explica brevemente la ventaja de usar el color para esa variable en relación a las demás, y el tipo de paleta que utilizas.

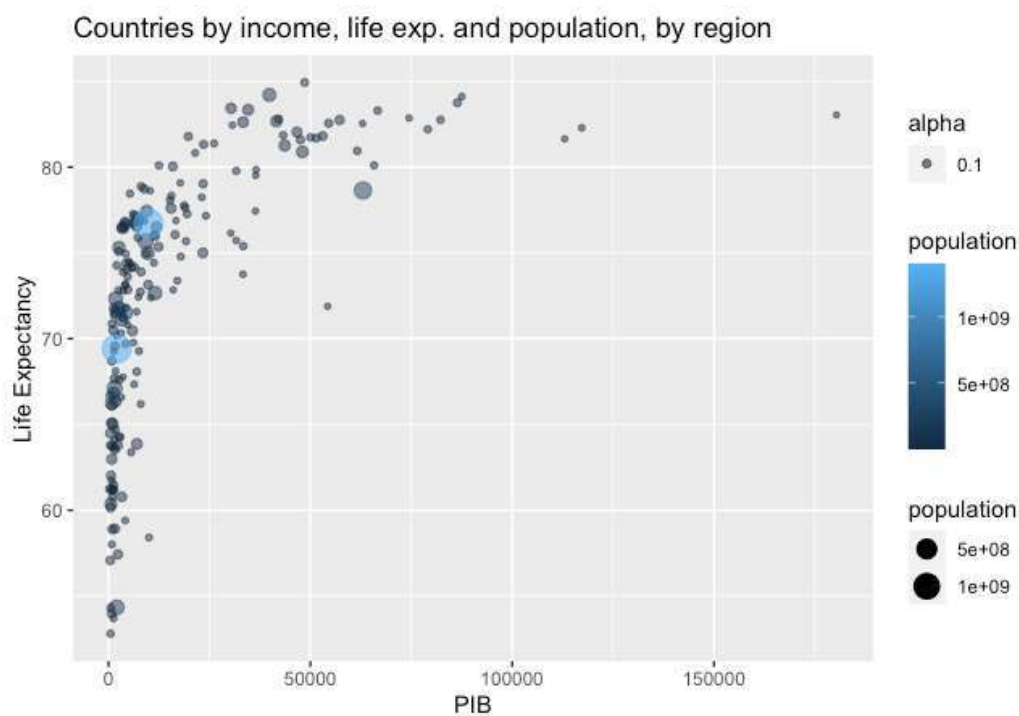
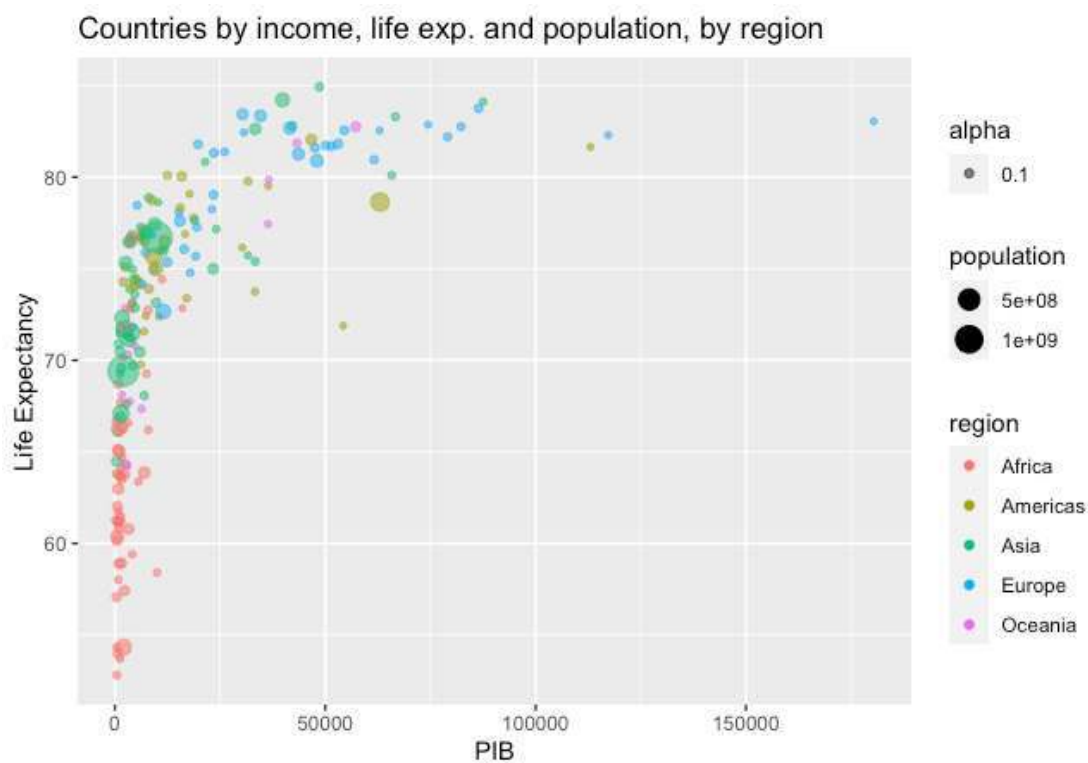
RESPOSTA:

En cualquiera de las 3 variables la ventaja consiste en resaltar por redundancia un aspecto de los datos, con tal de que se vea más claramente o de darle más importancia respecto a las otras. En el caso de Population se resaltan los outliers y las diferencias de población a lo largo de toda la distribución; en el PIB se resaltan los valores extremos. La paleta o escala debe ser continua para cualquiera de las tres.

C) (0.5 pt) Sube el código y la gráfica de una de las dos opciones, A o B.

RESPOSTA:

```
ggplot(df, aes(PIB, lifeExp)) +  
  geom_point(aes(size=population, color=region, stroke=0.5, alpha=0.1)) +  
  labs(x="PIB", y="Life Expectancy",  
       title="Countries by income, life exp. and population, by region")
```



Importante

Puedo eliminar la publi de este documento con 1 coin

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

pierdo espacio



PARTE 2 (5 pt)

Dataframe: iris de R



NOTA: En los ejercicios de esta parte, hacer uso de las *pipes*

2.1. (0.75 pt) NOTA: Para este ejercicio no se piden hacer las visualizaciones, solo responder las preguntas

- a) ¿Qué tipo de variables tiene este dataframe? (0.25 pt)
- b) ¿Qué gráfico de los siguientes podemos hacer que nos permita comparar las variables, conectándolas con líneas que permitan visualizar patrones y relaciones entre ellas? ¿Por qué? (0.5 pt)
- b.1. Un *parallel coordinate plot* que nos permita comparar las cinco variables de este dataframe juntas.
- b.2. Un *parallel coordinate plot* que nos permita comparar algunas variables de este dataframe juntas. ¿Qué variables?
- b.3. Ninguna de las anteriores. De hecho, no haríamos un *parallel coordinate plot* con este dataframe, haríamos un *parallel set plot*.

RESPOSTA:

a) Tiene 5 variables:

- 4 variables de métricas "Sepal.Length, Sepal.Width, Petal.Length, Petal.Width" – variables numéricas cuantitativas
- 1 variable que indica la especie de la flor ("Species") – variable categórica nominal

La respuesta es: podemos conectar las variables correspondientes a las métricas (datos numéricos multivariados cuantitativos). Esto corresponde a b.2

Vimos en clase que el *parallel coordinate plot* es una técnica de visualización que se usa para analizar datos numéricos multivariados. En este dataframe tenemos cuatro variables numéricas (las métricas) y una variable categórica (las especies). Así que podríamos usar el gráfico de coordenadas paralelas conectando con líneas dichas métricas para buscar patrones y relaciones entre ellas. Sin embargo, las especies, no podríamos conectarlas con líneas a otras variables (ya que *Species* es una variable categórica) – podríamos introducir dicha variable como una variable de agrupamiento poniendo un color, por ejemplo.

En cuanto a la pregunta (c) es falsa porque como vimos en clase, los *parallel set plots* (PSPs) se utilizan con datos multivariados categóricos. De hecho, en los PSPs, las barras horizontales en la visualización muestran la frecuencia absoluta de la frecuencia con la que ocurrió cada categoría. En este dataframe solo tenemos una variable categórica (species).

2.2. (2.25 pt)

a) Crea dos nuevas variables “Petal.shape” y “Sepal.shape” que correspondan en cada caso a la ratio entre su anchura y su longitud. (0.25 pt)

RESPOSTA:

```
> iris2<-iris %>% mutate(Petal.Shape = Petal.Width / Petal.Length,  
Sepal.Shape = Sepal.Width / Sepal.Length)
```

b) Compara la distribución de las nuevas variables en un gráfico multi-panel (1.25pt)
Nota: En el caso de no saber hacer el multi-panel y realizar dos figuras distintas se contará solo un máximo de 0.75 pt.

RESPOSTA:




























Petal.shape y Sepal.shape son variables numéricas continuas. Como nos piden un multi-panel, indirectamente nos piden un panel para mostrar la distribución de la primera variable y otro para la segunda. Podemos pues hacer un histograma, por ejemplo, para cada variable.

Si queremos hacer un multi-panel, lo primero es usar gather como hicimos en la parte 3 del seminario 4, que nos construya un dataframe con las métricas que necesitamos

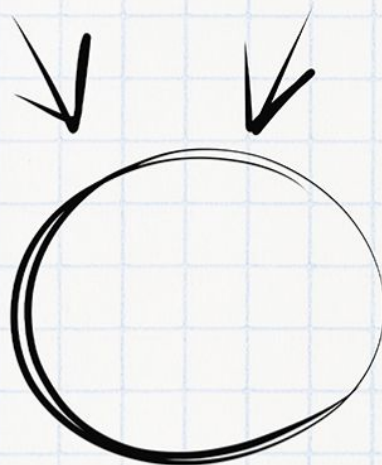
```
>iris_long <-iris2%>%gather(Petal.Shape, Sepal.Shape, key='metric',  
value='value')  
>ggplot(iris_long)+aes(value)+geom_histogram(binwidth =  
0.01)+facet_wrap(~ metric, ncol=1)+xlab('valor')+ylab('quantitat')
```

Imagínate aprobando el examen

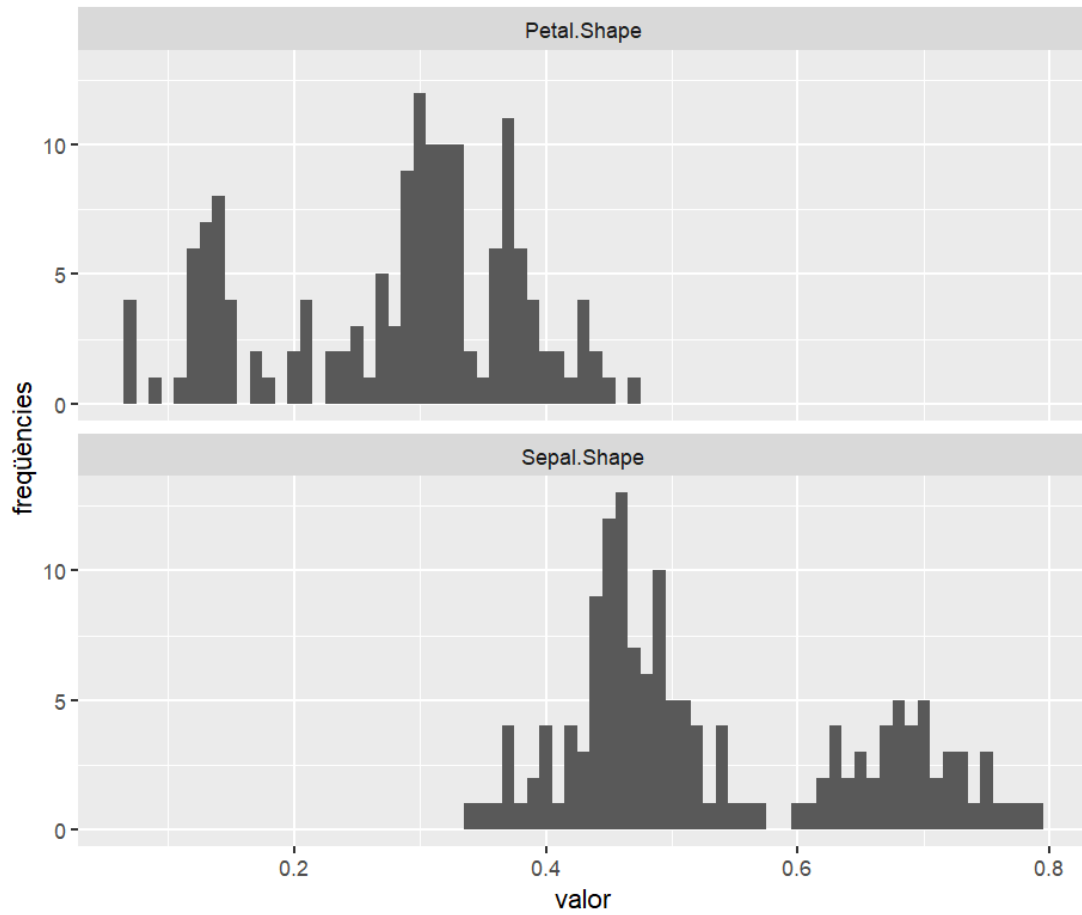
Necesitas tiempo y concentración

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



Como conclusiones podemos decir, por ejemplo, que, todo y que sabemos que hay tres especies de flores, hay claramente dos grupos separados de flores (totales) según la distribución de la forma del sépalo, el grupo con más flores tiene una moda de 0.45 y el grupo con menos flores tiene una moda alrededor de 0.7.

En la forma del petal, podemos intuir tres grupos. Uno con moda inferior a 0.2 y los otros tres superior a 0.3. Pero no podemos realmente diferenciar entre grupos de forma clara.

c) Ahora compara en una sola visualización la distribución de la nueva variable Petal.Shape para cada especie (0.75pt)

RESPOSTA:

Petal.Shape es una variable numérica cuantitativa continua y especies una variable categórica (discreta), podemos hacer pues un boxplot o violín plot para comparar distribuciones

```
>ggplot(iris2)+aes(Species,Petal.Shape)                                     +
geom_violin()+xlab('Especies')+ylab('Forma del pètal')
```


Importante

Puedo eliminar la publi de este documento con 1 coin

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

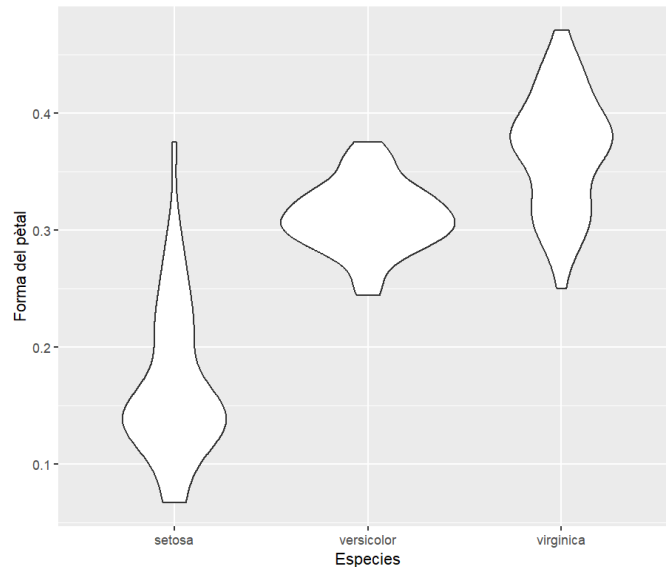
perdo
espacio



Necesito
concentración

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

WUOLAH



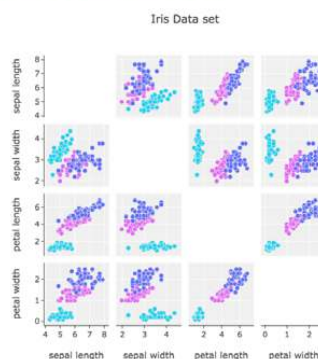
Aquí en cambio ya vemos que en el primer histograma, el grupo con moda inferior a 0.2, debe corresponder a las flores de la especie setosa.

2.3. (1.5pt) Haz un gráfico que permita ver claramente cuáles de las métricas del dataframe tendríamos en cuenta para tener una máxima correlación. ¿Y para tener una mínima correlación? Razona tu respuesta

RESPOSTA:

Una manera que hemos visto en teoría es hacer un SPLOM, vimos que hasta (3 o 4 variables) era adecuado, con lo cual nos serviría. De hecho vimos en teoría una visualización con este dataframe concreto

We saw: **Scatterplot matrix (SPLOM)** uses multiple scatterplots to determine the correlation (if any) between a series of variables.

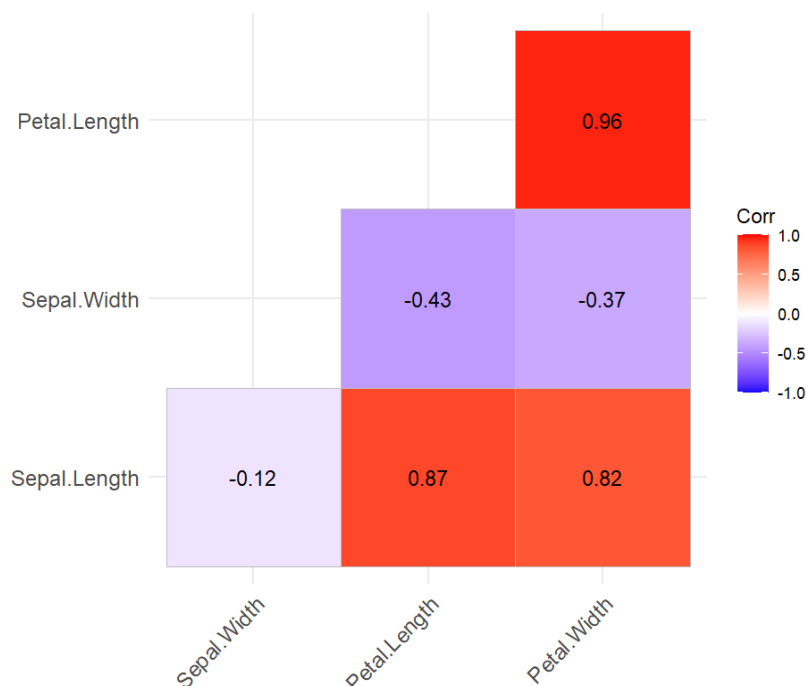


!! When we have >3 or 4 quantitative variables – scatterplot matrices quickly become unwieldy

Otra forma es con el correlograma y esto sí lo hemos visto en teoría y seminarios. Además, este nos permitirá “cuantificar” y mostrar la correlación máxima y mínima.

Para hacerlo seguimos de forma análoga a como vimos en el seminario 5. Una opción sería:

```
> library(ggcorrplot)
> cormat <- cor(iris[1:4]) #el enunciado ya nos indica que le demos 'correlaciones entre métricas'. Pero además vimos que los correlogramas solo aceptan variables numéricas (por lo que no aceptaría species, tal y como es-categoría nominal- aunque quisiéramos)
> ggcorrplot(cormat, lab=TRUE, type = "lower") #mostramos las etiquetas para facilitar ver max y min, y además, muy importante: como sabemos que la matriz de correlación es simétrica, mostramos sólo la parte inferior o superior de la matriz (evitando dar información que solo complica la visualización, ya que ya está)
```



La mínima correlación es entre sepal.length y sepal.width (con una correlación negativa de -0.12). La máxima correlación en cambio, es entre petal.length i petal.width (con una correlación positiva de 0.96).

2.4. (0.5 pt) Data massage

a) Calcula el valor mínimo de Petal.Length por Species

b) Crea un dataframe que para todas las flores con una longitud de sépalo menor que 5 cm, contenga como columnas la anchura del sépalo, la longitud del sépalo y las especies. ¿Cuántas observaciones tiene el nuevo dataframe?

RESPOSTA:

a)

```
> iris %>% group_by(Species) %>% summarise(Min.Petal.Length = min(Petal.Length))
```

```
# A tibble: 3 × 2
  Species    Min.Petal.Length
  <fct>          <dbl>
1 setosa            1
2 versicolor        3
3 virginica         4.5
```

b)

```
> ex3_4<-iris %>% filter(Sepal.Length < 5) %>% select(Sepal.Width,
Sepal.Length, Species)
```

Con `str(ex3_4)` Podemos ver que tiene 22 observaciones.

Versión examen modelo 2 (57 observaciones):

```
> ex3_4<-iris %>% filter(Sepal.Width < 3) %>% select(Sepal.Width,
Sepal.Length, Species)
```

```
> str(ex3_4)
```