

# 1rParcial20233Soluciones.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

pierdo espacio



Necesito concentración

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

WUOLAH

Visualització de Dades (Enginyeria de Dades – EE - UAB)  
Examen Recuperació Primer Parcial – 3 Juliol 2023  
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 (1 pt)**

Dataset: `filmdeathcounts.csv`

Haz un bubble Chart que permita explorar la relación entre *Year*, *Body\_Count* y *IMDB\_Rating*.

1.1. (0.5 pt) ¿Cuál es el mapeo óptimo de variables a canales visuales en este caso (qué variable va en el tamaño, en el eje X o en el Y)? Explica el porqué de tu elección.

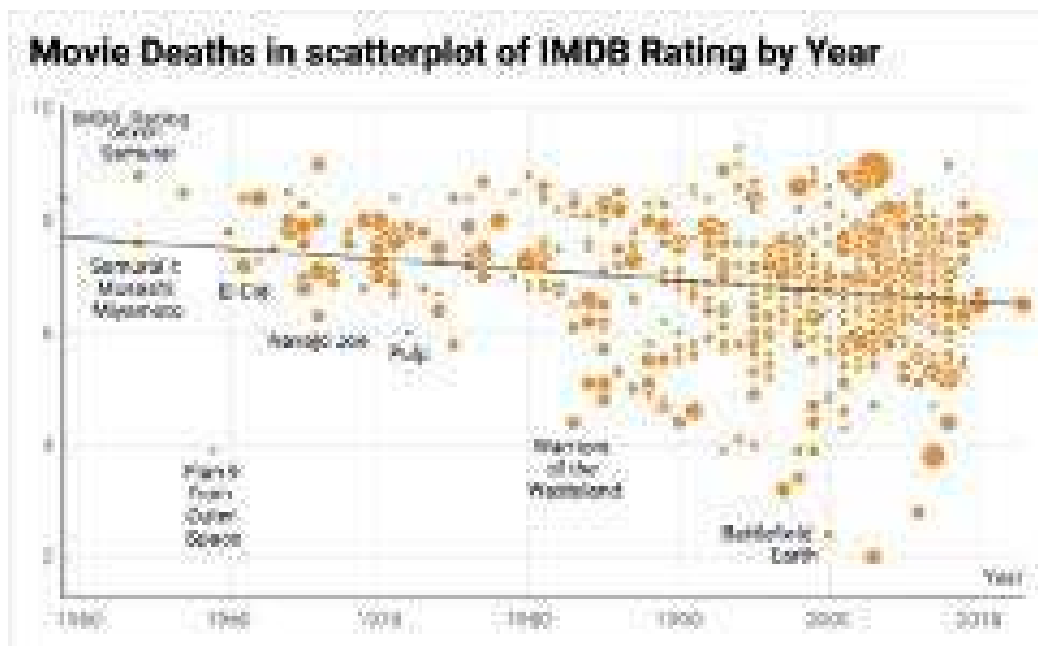
**RESPOSTA:**

El tiempo es mejor en el eje X porque lo asociaremos fácilmente a la progresión temporal. BodyCount tiene mucha diferencia entre valores exxtremos que hacen difícil ver patrones en zonas de la gráfica con alta densidad de puntos, funciona mejor en el tamaño de los círculos.

1.2. (0.5 pt)

Haz la gráfica con las leyendas y títulos necesarios para que sea comprensible. Sube las versiones que creas convenientes para acompañar la respuesta anterior, más el código usado para hacer la versión final.

**RESPOSTA:**



## PARTE 2 (3 pt)

Dataset: `titanic.csv`

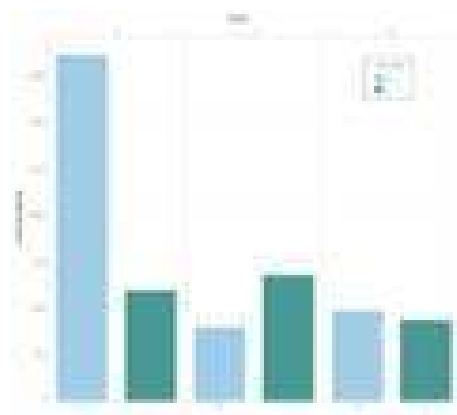
2.1. (0.5 pt) Abre el fichero. ¿Qué tipo de atributo son: *Survived*, *Pclass*, *Sex*, *Age*, y *Fare*? ¿Qué atributo es el key (clave primaria) del dataset?

**RESPOSTA:**

*Survived*=Categórico, *Pclass*=Ordinal, *Sex*=categórico, *Age*=Cuantitativo, *Fare*=Cuantitativo. El key es *PassengerID*

2.2. (0.5 pt) Queremos saber si la clase en la que viajaban los pasajeros (*Pclass*) influye en la probabilidad de sobrevivir (*Survived*). Haz una gráfica que permita visualizar la relación entre esas dos variables, teniendo en cuenta su tipo.

Sube la gráfica debidamente anotada y el código.



### 2.3. (1 pt)

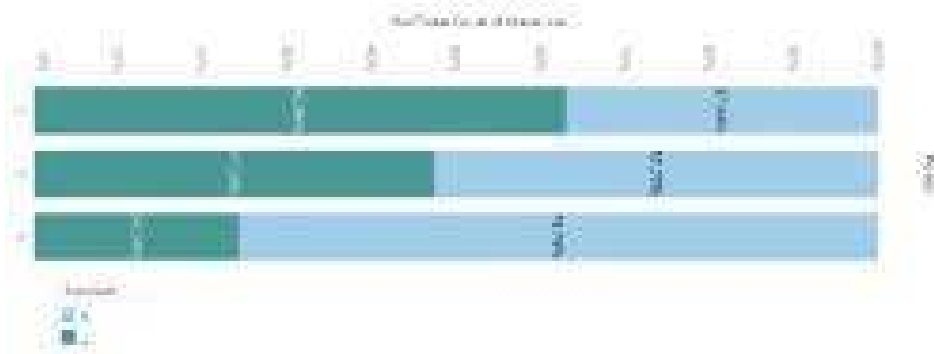
Explica por qué la gráfica que has hecho sirve para visualizar la relación entre esas dos variables de acuerdo con el framework Datos/Tareas/Codificación (Cuántos atributos usas, de qué tipo son; qué tarea ayuda a llevar a cabo la gráfica; marcas y canales empleados).

**RESPOSTA:**

Survived y Pclass son de tipo categórico y categórico ordinal. Se puede visualizar como una gráfica de barras agrupadas o apiladas, donde un categórico se utiliza para determinar la posición en el eje horizontal y el otro para separar cada barra en dos secciones de colores distintos. Si se utiliza el Count de pasajeros como atributo cuantitativo son barras normales. ( En R vimos en seminario 2 cómo hacer estos gráficos en R usando: `geom_bar(position="dodge")`)

**2.4. (1 pt)** Haz una variación de esa gráfica usando porcentajes en lugar de cantidades y súbela junto al código. ¿Ofrece ventajas respecto a la anterior? Explica cuales.

Si se utilizan porcentajes son barras normalizadas. Las barras normalizadas son mejores porque permiten comparar con más precisión entre categorías a pesar de las diferencias en número de pasajeros.



## PARTE 3 (4 pt)

**Dataframe:** white-wine.csv

Llegu el **dataframe** amb la següent comanda

```
➤ white_wine<-read.csv("./white-wine.csv", sep=";")
```

Aquest **dataframe** té unes dades basades en proves fisicoquímiques de vins blancs :

- fixed acidity
- volatile acidity
- citric acid
- residual sugar

Importante

Puedo eliminar la publi de este documento con 1 coin

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

perdo  
espacio



- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

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 (2.5 pt)

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

b) Fes un gràfic que mostri la distribució de l'alcohol. Afegiu etiquetes als eixos (0.5 pts)

c) Fes un multipanel que permeti comparar les distribucions del *volatile.acidity* i *citric.acid* (1.5 pts). *Nota: En cas de no poder realitzar un multipanel i fer dues figures separades, es comptarà només 0.75 pts.*

d) Fes un gràfic que mostri la distribució del pH versus qualitat (*quality*). Afegiu etiquetes als eixos (0.5 pts)

**RESPOSTA:**

a) Llegim el dataframe

```
>white_wine<-read.csv("./white-wine.csv",sep=";")  
>str(white_wine)
```

El *dataframe* té 4898 observacions i 12 variables.


Les variables són totes numèriques quantitatives, excepte *quality* que és una variable numèrica també però categòrica.

b) Com alcohol 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



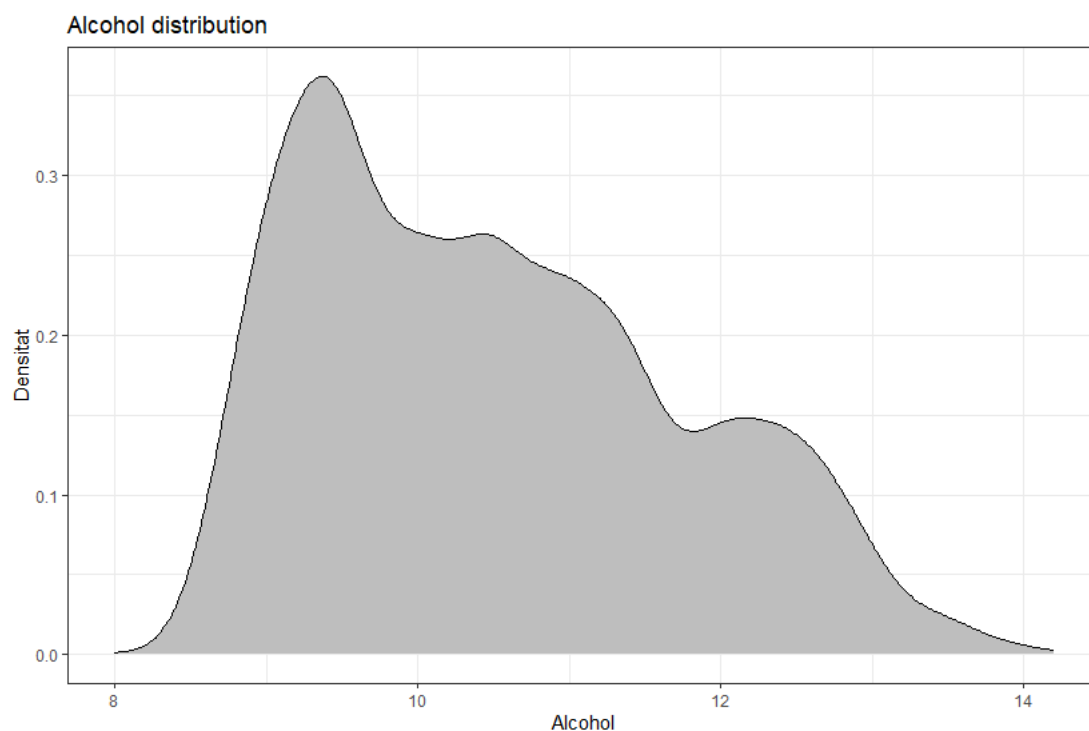
**c + geom\_freqpoly()**  
x, y, alpha, color, group, linetype, size  
a + geom\_freqpoly(aes(y = ..density..))



**c + geom\_histogram(binwidth = 5)**  
x, y, alpha, color, fill, linetype, size, weight  
a + geom\_histogram(aes(y = ..density..))

*Una posible soluci3n ser3a:*

```
>ggplot(white_wine)+aes(alccohol)+geom_density(fill='grey')+theme_bw(
)+ggtitle("Alcohol distribution")+xlab('Alcohol')+ylab('Densitat')
```



- c) Tenim quatre distribucions variables num3riques cont3nues i s'haur3 de fer un histograma/densitats per cada variable. Podeu fer un facet (dues columnes i dues files) amb un histograma/gr3fic de densitats per cada variable ( en cada casella).

Abans per3 hem de construir un amb les m3triques que necessitem. Primer per simplificar, fem un dataframe que contingui cadascuna d'elles en una columna:



# 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 <span>Anual <input type="checkbox"/></span> | 0,99 € / mes   | 3,99 € / mes  | 7,99 € / mes  |

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

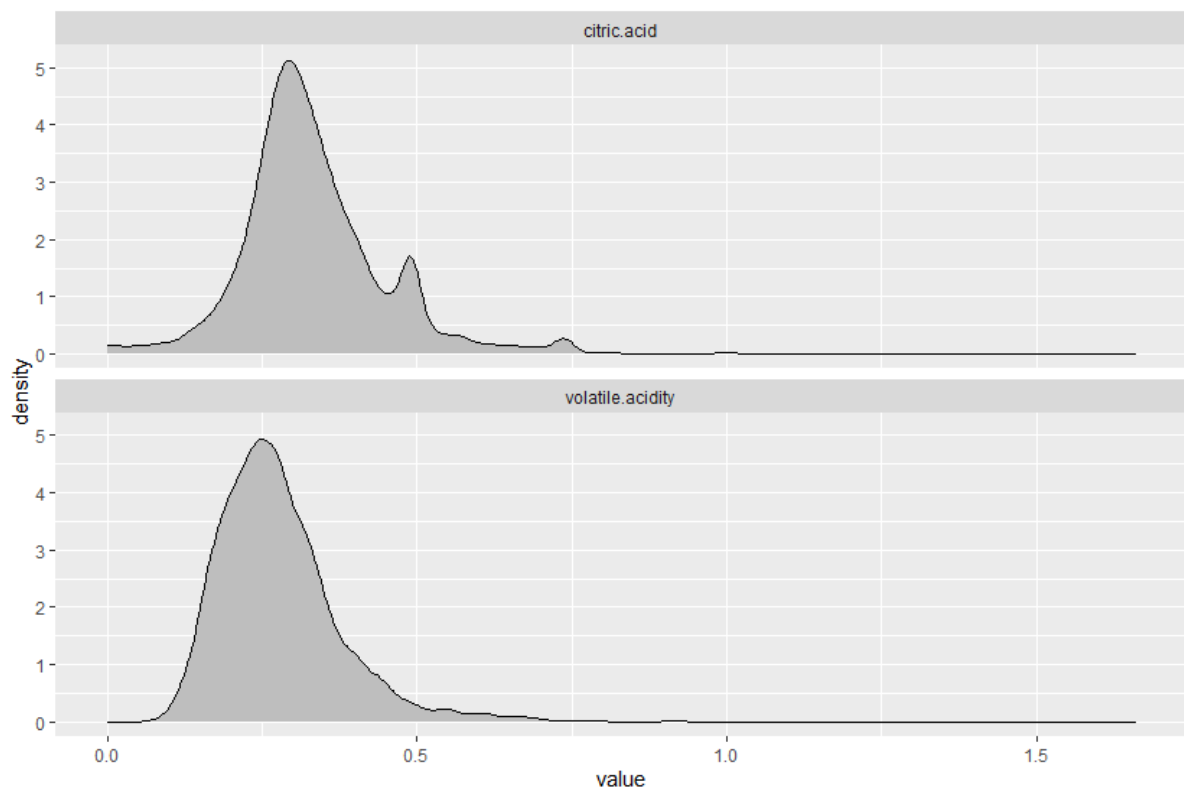


# WUOLAH

```
> df<-white_wine%>%select(c("volatile.acidity", "citric.acid"))
```

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(volatile.acidity, citric.acid, key='metric',  
value='value')  
ggplot(df_long)+aes(value)+geom_density(fill="grey")+facet_wrap(~ me  
tric, ncol=1)
```



Podem dir per exemple que ambdues variables tenen molts dels seus valors d'acidesa entre 0,25 i 0,30.



Importante

Puedo eliminar la publi de este documento con 1 coin

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

perdo  
espacio



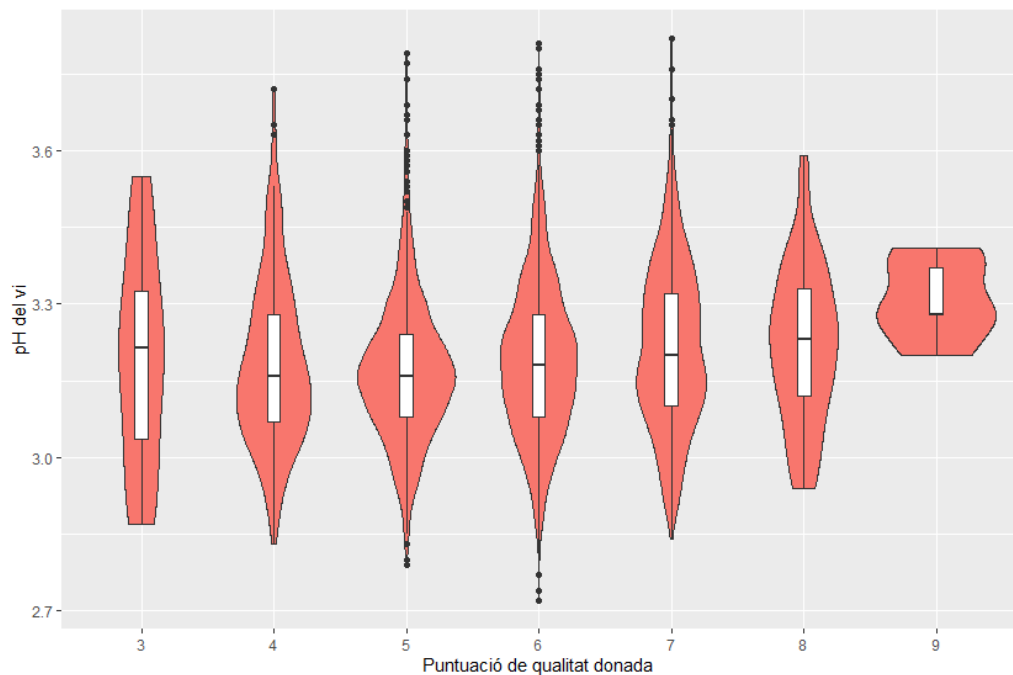
Necesito  
concentración

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

WUOLAH

d) Hem dit ja abans que *quality* actuava com categòrica, fem-la amb *as.factor*. pH és numèrica i contínua podem mostrar la distribució amb un *violin\_plot* o *box\_plot*, o inclús combinant ambdós

```
>ggplot(white_wine,aes(as.factor(quality),pH))  
+geom_violin(aes(fill="red"))+geom_boxplot(width = 0.1)+xlab("Puntuació  
de qualitat donada")+ylab("pH del vi")+theme(legend.position = "none")  
(0 una de les dues geometries)
```

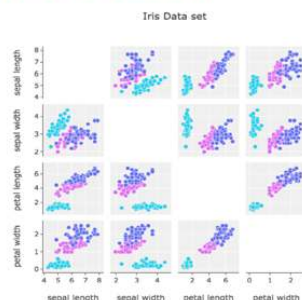


3.2. (1.5 pt) Fes un gràfic que permeti veure clarament quines de les mètriques del dataframe tindriem en compte per tenir una màxima correlació. I per tenir una mínima correlació? Raona la teva resposta

**RESPOSTA:**

Una manera que hem vist és el SPLOM

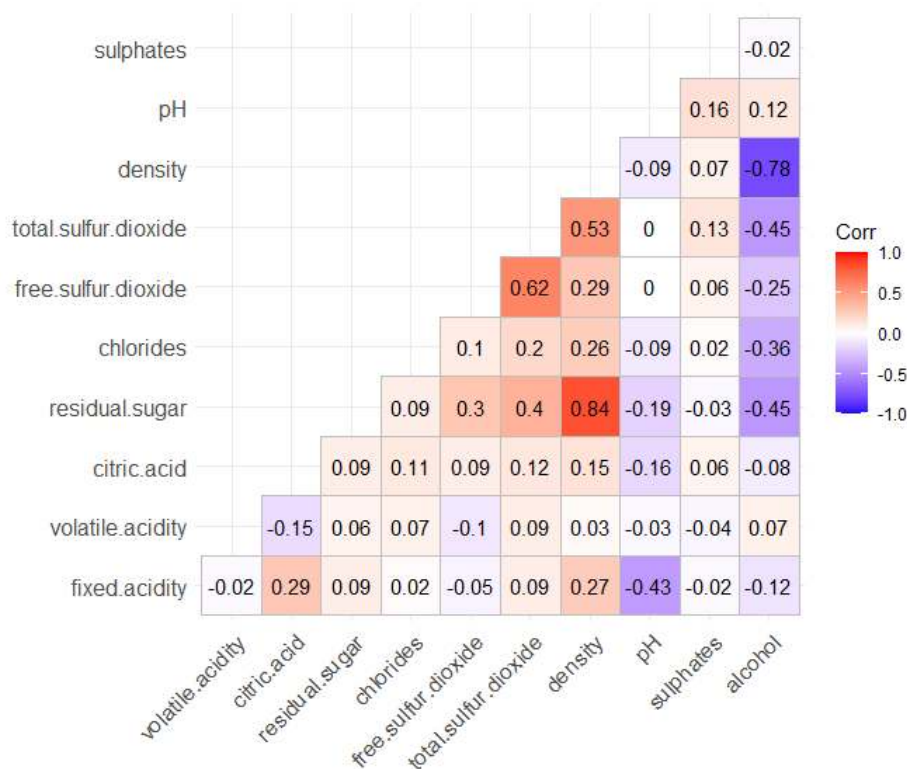
We saw: **Scatterplot matrix (SPLOM)** uses multiple scatterplots to determine the correlation (if any) between a series of variables.



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

Una altra forma és un correlogram. Per fer-lo fem com fèiem en el seminari 5. Una opció seria

```
> library(ggcorrplot)
> cormat <- cor(white_wine[1:11]) #no posaríem la variable "quality" ja que actua com
categòrica
> ggcorrplot(cormat, lab=TRUE, type = "lower") #mostrem las etiquetes per facilitar
veure max i min, a més, per simetria agafem sol la diagonal inferior o superior de la matriu, aquí hem triat
la primera
```



La mínima correlació és entre el pH i el total.sulfur.dioxide i free.sulfur.dioxide respectivament. La màxima (positiva) és entre la densitat i el sucre residual (i negativa la de densitat amb alcohol).

La majoria de variables estan però poc correlacionades.