VISUALITZACIÓ DE DADES

Dra. Vanessa Serrano, Universitat Rovira i Virgili
Dr. Jordi Cuadros, IQS Universitat Ramon Llull
Maig, 2023

#ProDigital

TEMARI

1. Principis bàsics de visualització

2. Visualitzacions efectives amb eines habituals

3. Conceptualització dels gràfics

4. Visualització de dades simples

VISUALITZACIÓ DE DADES

4. Visualització de dades simples

#ProDigital

CONJUNT DE DADES

Enquesta de pressupostos familiars a Espanya

(INE, 2016-2021)

CONJUNT DE DADES

CONJUNT DE DADES

https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176806&menu=ultiDatos&idp=1254735976608

188 variables

- 1. Información general (10 variables)
- 2. Características relativas al hogar (43 variables)
- 3. Datos del sustentador principal (35 variables)
- 4. Características de la vivienda principal (11 variables)
- 5. Otras viviendas a disposición del hogar (65 variables)
- 6. Gastos de consumo del hogar (7 variables)
- 7. Ingresos regulares mensuales del hogar (12 variables)
- 8. Número de comidas y cenas efectuadas durante la bisemana muestral (5 variables)

CONJUNT DE DADES

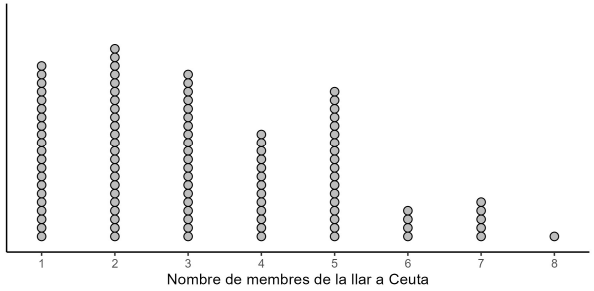
Nombre d’observacions per any

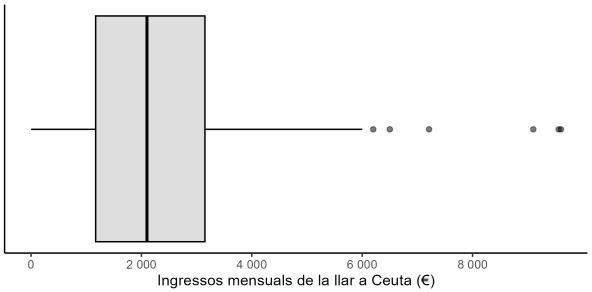
- 2016: 22011
- 2017: 22043
- 2018: 21395
- 2019: 20817
- 2020: 19170
- 2021: 19394

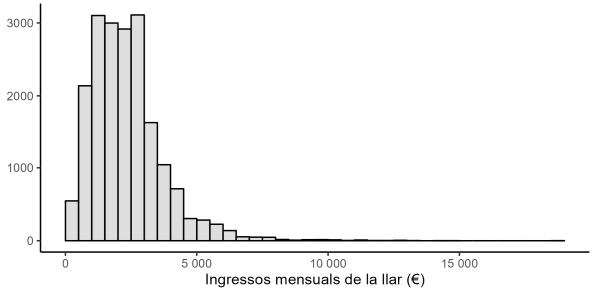
GRÀFICS DE DISTRIBUCIÓ/COMPOSICIÓ

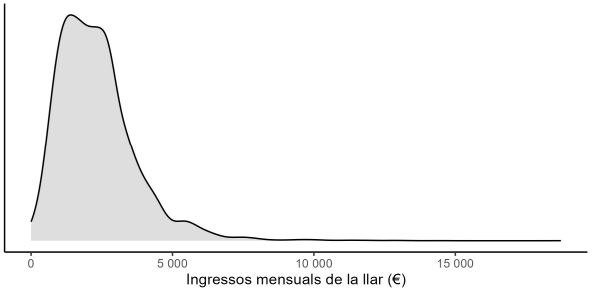
DISTRIBUCIÓ/COMPOSICIÓ

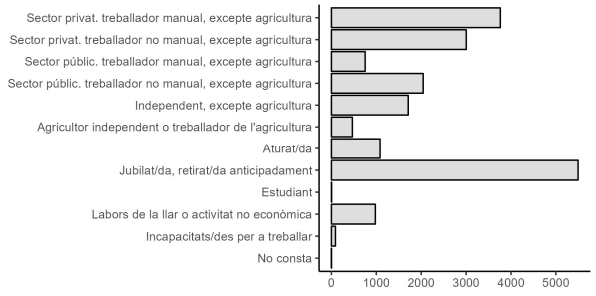
- Gràfic de punts
- Diagrama de caixa
- Histograma
- Corba de densitat
- Diagrama de barres
- Gràfic de sectors
- Waffle
- Mapa

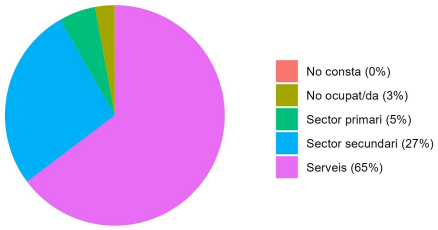


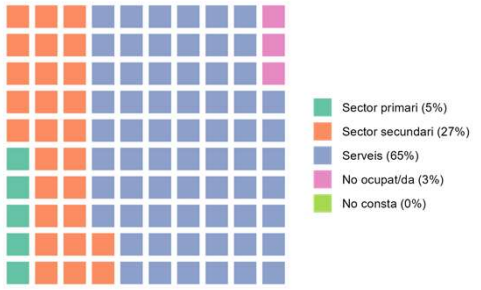










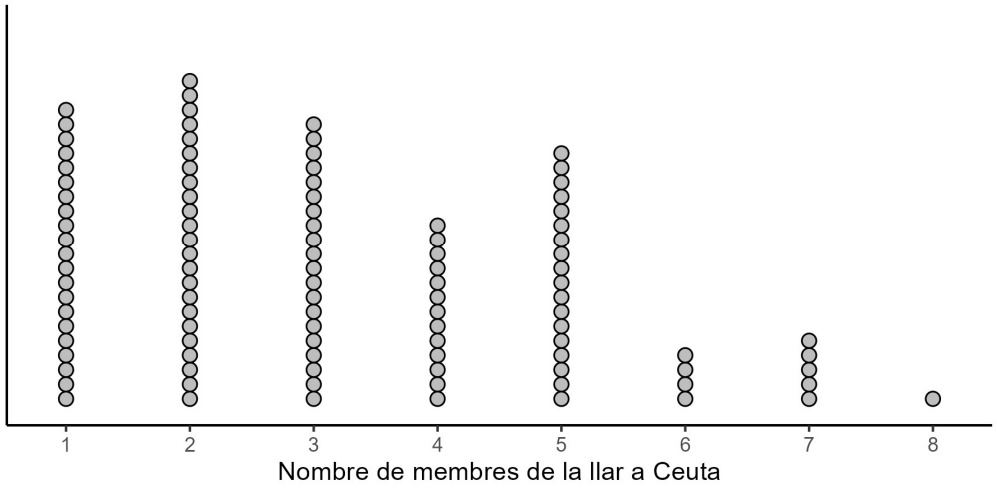


GRÀFIC DE PUNTS

Recomanat per visualitzar la distribució de conjunts de dades de mida petita (preferiblement no més de 100 dades) i referents a una única variable.

Per interpretar-lo ens fixem en:

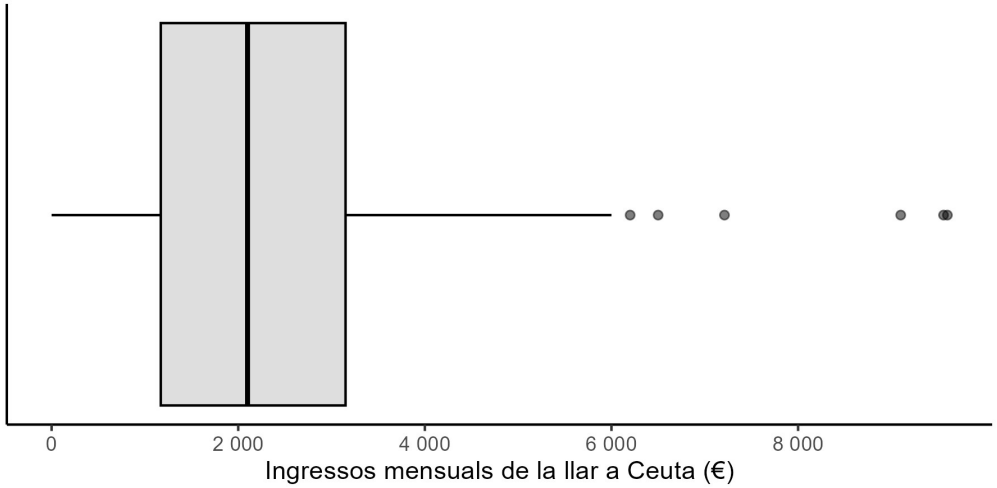
- on estan acumulats la majoria dels punts,
- el valor de dades extremes i
- la presència de dades allunyades




```
ggplot(df1CEUTA,aes(x=NMIEMB))+
  geom_dotplot(dotsize=0.5,col="black",fill="grey")+
  scale_y_continuous(NULL,breaks=NULL)+
  scale_x_continuous(breaks=1:max(df1CEUTA$NMIEMB))+
  xlab("Nombre de membres de la llar a Ceuta")+
  theme_classic()
```

DIAGRAMA DE CAIXA

Recomanat per visualitzar la distribució de conjunts de dades de mida mitjana (preferiblement entre 20 i 1000 dades) i referents a una única variable quantitativa.



Per interpretar-lo ens fixem en:

- la posició dels quartils,
- el rang interquartílic,
- la simetria de la distribució i
- la presència de dades allunyades

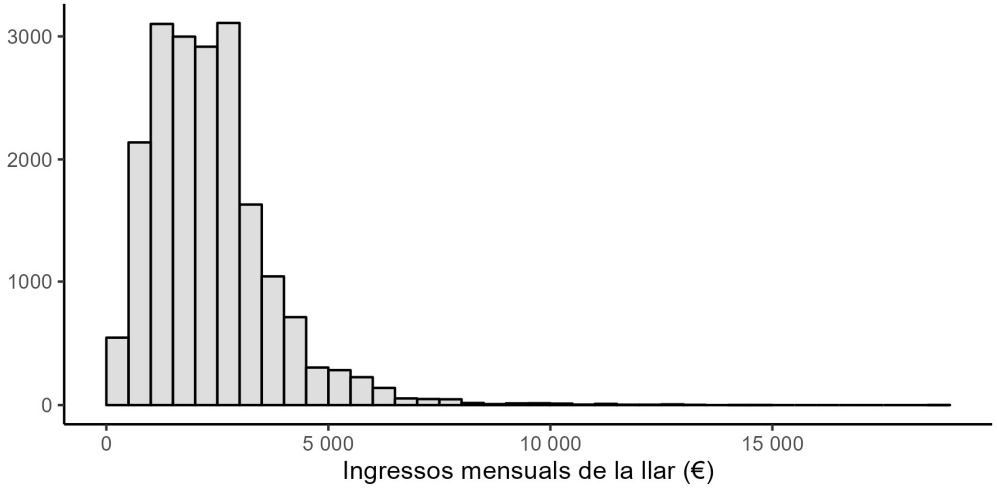
```
ggplot(df1CEUTA,aes(x=IMPEXAC))+
  geom_boxplot(col="black",fill="grey",alpha=0.5)+
  scale_y_continuous(NULL,breaks=NULL)+
  scale_x_continuous(
    labels=scales::number_format(accuracy=1),
    breaks=seq(0,max(df1CEUTA$IMPEXAC,na.rm=T),2000))+
  xlab("Ingressos mensuals de la llar a Ceuta (€)")+
  theme_classic()
```

HISTOGRAMA

Recomanat per visualitzar la distribució de conjunts de dades de mida gran (preferiblement més de 100 dades) i referents a una única variable quantitativa.

Per interpretar-lo ens fixem en:

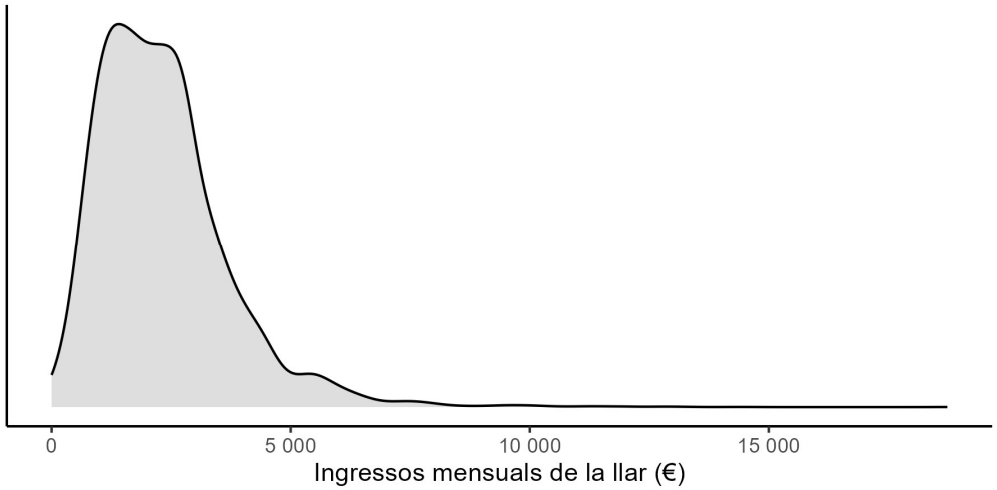
- el rang de valors,
- les regions on s'agrupen les dades,
- la simetria de la distribució i
- la presència de dades allunyades




```
ggplot(df1,aes(x=IMPEXAC))+
  geom_histogram(col="black",fill="grey",alpha=0.5,
    binwidth=500,boundary=0,closed="left")+
  scale_y_continuous()+
  scale_x_continuous(
    labels=scales::number_format(accuracy=1))+
  xlab("Ingressos mensuals de la llar (€)")+
  ylab(NULL)+
  theme_classic()
```

CORBA DE DENSITAT

Similar a l'histograma, però suavitzat. Mostra una corba que representa la distribució en un interval continu.



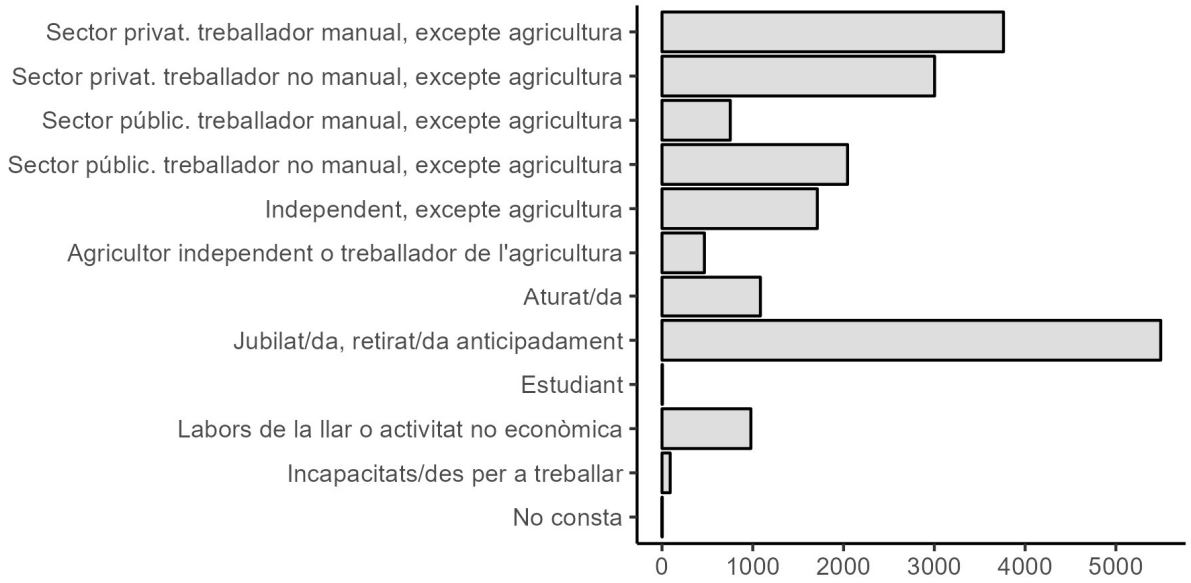

```
ggplot(df1,aes(x=IMPEXAC))+
  geom_density(col="black",fill="grey",
    alpha=0.5,adjust=2)+
  scale_y_continuous(breaks=NULL)+
  scale_x_continuous(
    labels=scales::number_format(accuracy=1))+
  xlab("Ingressos mensuals de la llar (€)")+
  ylab(NULL)+
  theme_classic()
```

DIAGRAMA DE BARRES

Recomanat per visualitzar la distribució de conjunts de dades referents a una única variable qualitativa o quantitativa amb pocs valors diferents.

Per interpretar-lo ens fixem en:

- les categories (o valors) més freqüents,
- les categories (o valors) que no apareixen o apareixen amb una freqüència menor i
- en el cas de dades ordinals, la simetria de la distribució




```
ggplot(df1m,aes(y=SITSOCI))+
  geom_bar(col="black",fill="grey",alpha=0.5)+
  scale_y_discrete(limits=rev)+
  scale_x_continuous(breaks=(0:5)*1000)+
  ylab(NULL)+
  xlab(NULL)+
  theme_classic()
```

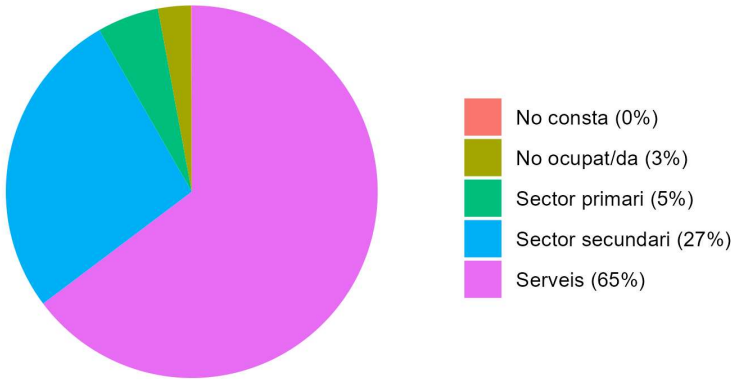
GRÀFIC DE SECTORS

S'utilitza per visualitzar la distribució relativa de conjunts de dades referents a una única variable qualitativa o quantitativa amb pocs valors diferents.

Recomanat exclusivament quan el que volem és destacar una categoria (o valor) més freqüent.

Per interpretar-lo ens fixem en:

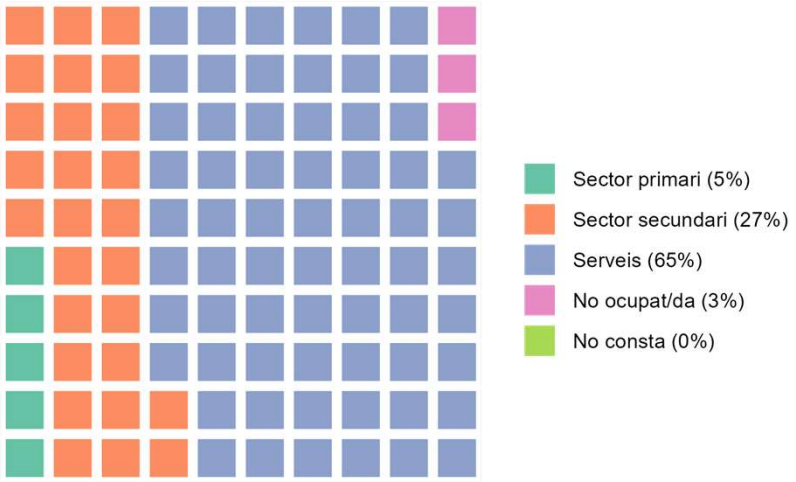
- les categories (o valors) amb major freqüència.




```
ggplot(dfPlt,aes(x="",y=Prop,fill=Var1))+
  geom_bar(stat="identity")+
  coord_polar(theta="y",start=0)+
  labs(x=NULL,y=NULL)+
  scale_y_continuous(breaks=NULL)+
  theme_classic()+
  theme(legend.title=element_blank(),
    axis.line=element_blank(),
    axis.ticks=element_blank())
```

WAFFLE

Similar al gràfic de sectors, però facilita la lectura de la freqüència de totes les categories (o valors) que apareixen.



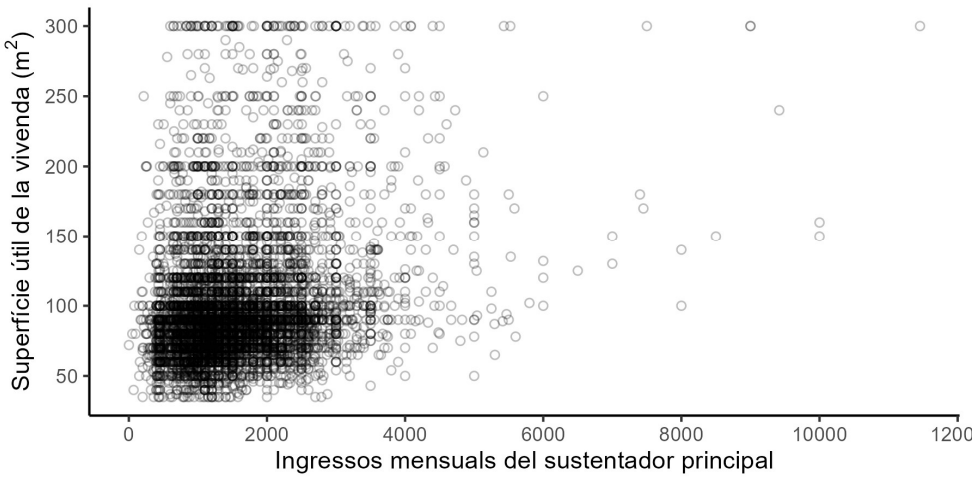

```
v<-as.integer(round(df1PIE$Prop*100,0))
names(v)<-df1PIE$Var1
waffle(v, rows=10)
```

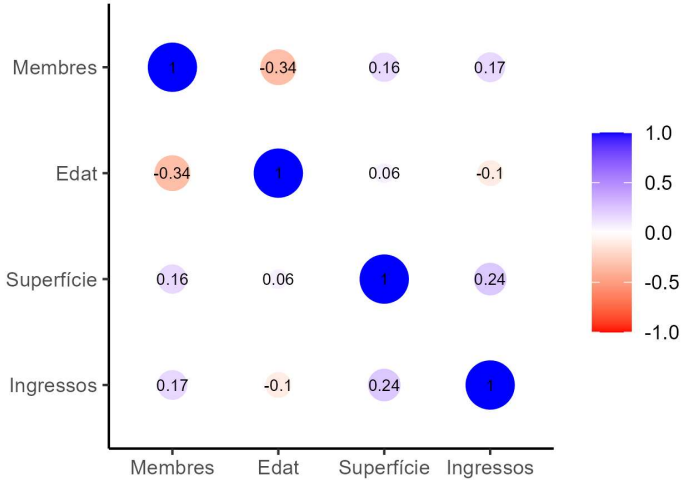
GRÀFICS DE RELACIÓ

RELACIÓ

- Gràfic de dispersió
- Gràfic d'intensitat de colors
- Correlograma

Alternativament a partir de comparacions entre valors diferents





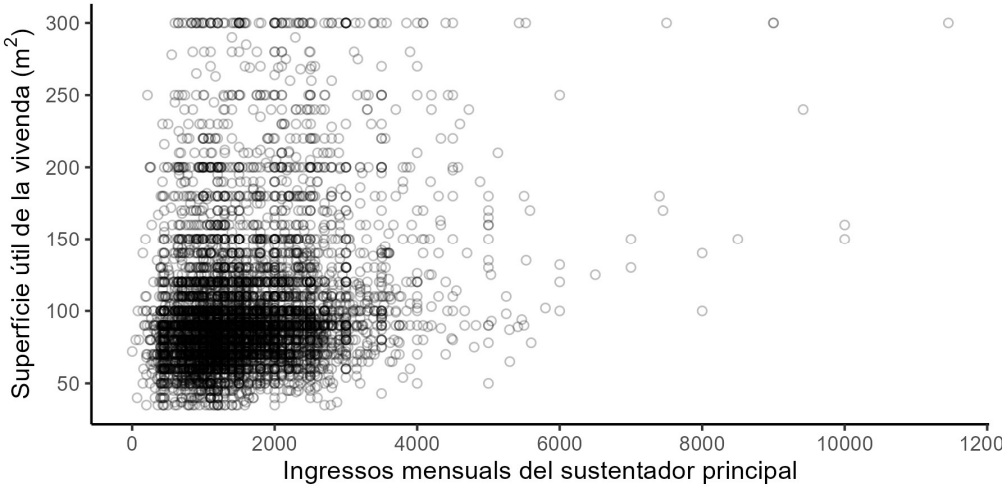
GRÀFIC DE DISPERSIÓ

Recomanat per visualitzar la relació entre dues variables quantitatives.

Quan el volum de dades és elevat, pot haver-hi moltes superposicions de punts. En aquest cas és recomanable afegir transparència.

Per interpretar-lo ens fixem en:

- si hi ha algun tipus de relació funcional entre les dues variables.
- la direcció d'aquesta relació, en cas d'existir,
- la concentració de punts en determinades regions i
- la presència de dades allunyades per a qualsevol de les dues variables.




```
ggplot(df1[df1$IMPEXACPSP>=0 & df1$SUPERF>0,],
    aes(x=IMPEXACPSP,y=SUPERF))+
  geom_point(shape=21,alpha=0.25)+
  scale_x_continuous(breaks=seq(0,12000,2000))+
  scale_y_continuous(breaks=seq(0,300,50))+
  labs(x="Ingressos mensuals del sostenidor principal“,
    y=expression(
      "Superfície útil de la vivenda (m"^2*")"))+
  theme_classic()
```

GRÀFIC D'INTENSITAT DE COLORS

Similar al gràfic de dispersió però per a dues variables qualitatives o quantitatives amb pocs valors diferents.

El color indica la freqüència absoluta o relativa de les observacions corresponents a l'encreuament de les categories (o valors) de cada variable.

Per interpretar-lo ens fixem en:

- les regions amb elevada freqüència de dades,
- Les regions amb una freqüència especialment baixa i
- en el cas de dades ordinals, la direcció de la relació entre les dues variables, en cas d'existir.

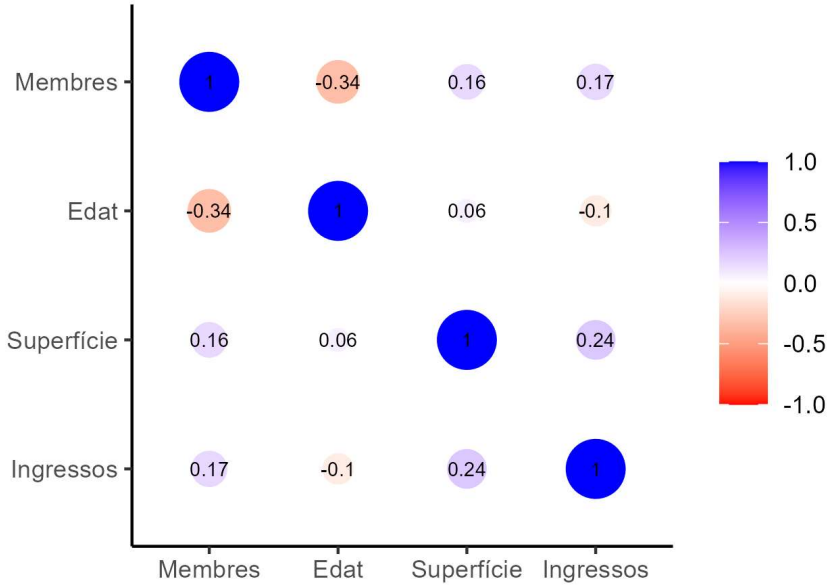

```
ggplot(df1Heat,aes(x=Var1,y=Var2,fill=Freq,label=Freq))+
  geom_tile(col="white")+
  geom_text(
    color=ifelse(df1Heat$Freq>3000,"white","black"))+
  scale_fill_viridis_c(direction=-1)+
  scale_y_discrete(limits=rev)+
  labs(x=NULL, y=NULL)+
  guides(fill=guide_colorbar(title="Nombre \nde llars"))+
  theme_classic()
```

CORRELOGRAMA

Recomanat per visualitzar les correlacions (habitualment lineals) entre parells de variables quantitatives.

Per interpretar-lo ens fixem en:

- els parells de variables amb alta correlació, valors propers a 1 (positiva) o -1 (negativa) i
- els parells de variables no correlacionats, valors propers a 0.



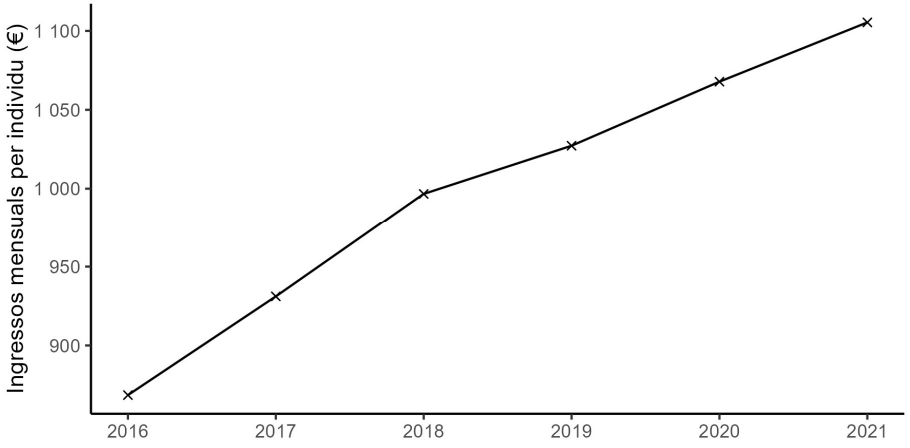

```
ggcorrplot(M[,4:1],method="circle",
    lab=T,lab_size=2.5,
    colors=c("red","white","blue"),
    outline.color="white")+
  labs(x=NULL,y=NULL)+
  theme_classic()+
  theme(legend.title=element_blank())
```

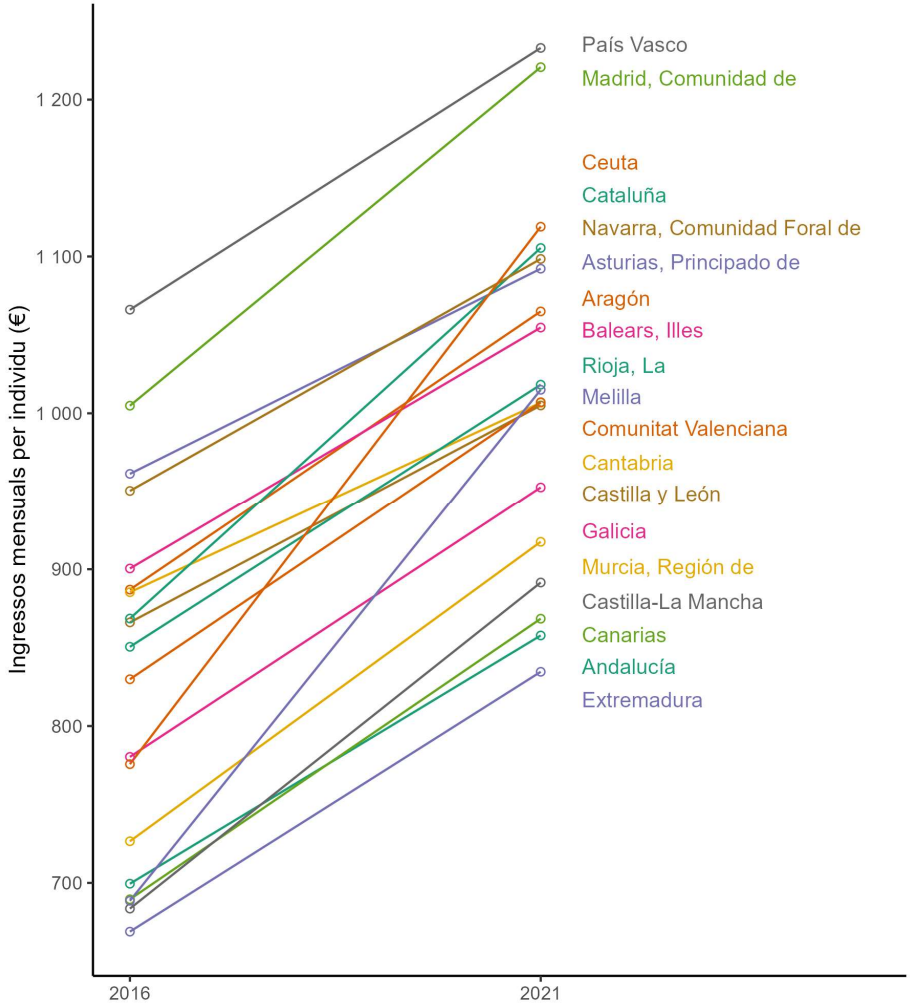
GRÀFICS D'EVOLUCIÓ TEMPORAL

EVOLUCIÓ TEMPORAL

- Gràfic de línies
- Gràfic de pendents

Alternativament a partir de comparacions entre moments diferents



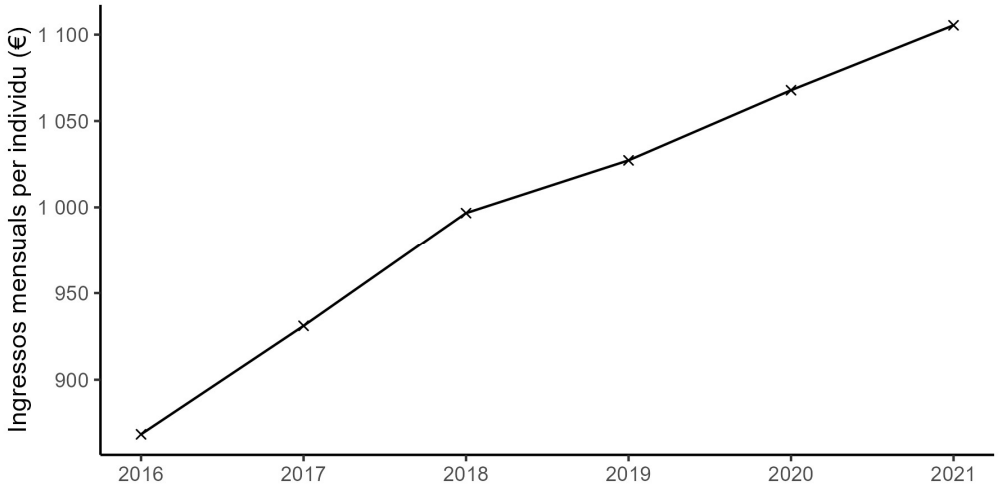


GRÀFIC DE LÍNIES

Recomanat per visualitzar l'evolució d'una variable quantitativa en funció d'una variable temporal.

Per interpretar-lo ens fixem en:

- la tendència de les dades,
- la presència de comportaments periòdics i
- la presència de valors que s'allunyen de la tendència general.



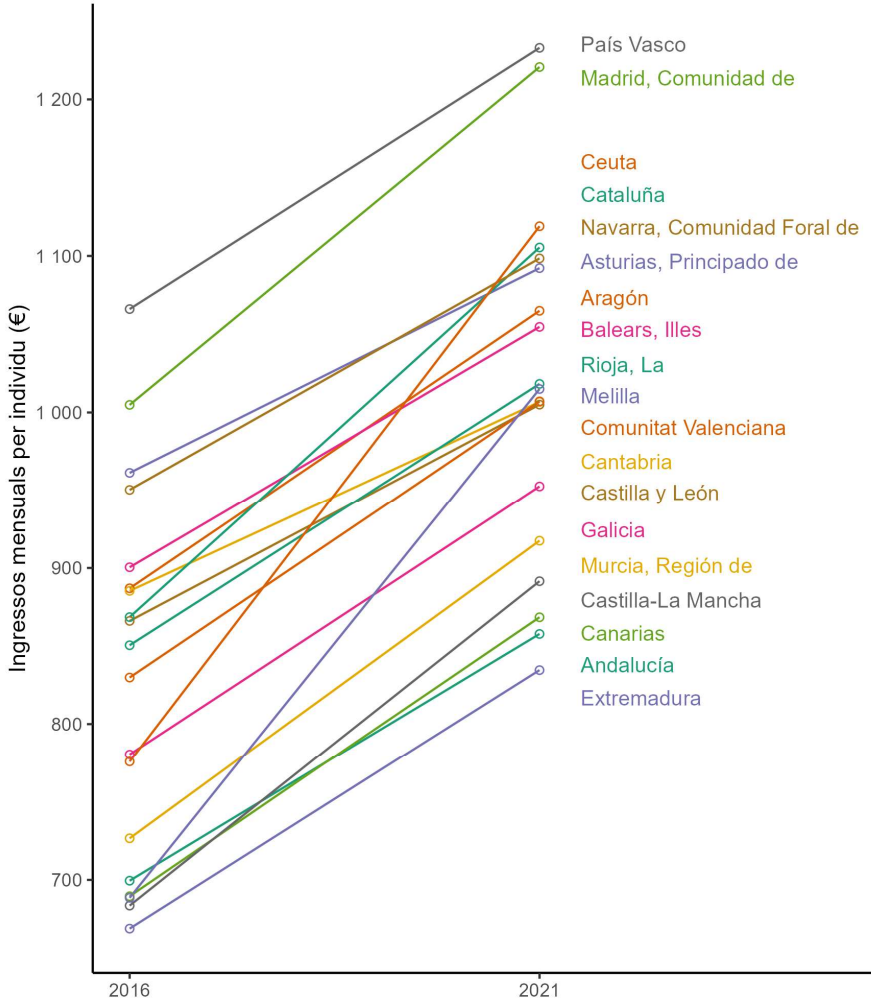

```
ggplot(dfTCATxany,aes(x=ANOENC,y=IMPEXACxpm))+
  geom_line()+
  geom_point(shape=4)+
  scale_y_continuous(
    labels=scales::number_format(accuracy=1))+
  labs(x=NULL,y="Ingressos mensuals per individu (€)")+
  theme_classic()
```

GRÀFIC DE PENDENTS

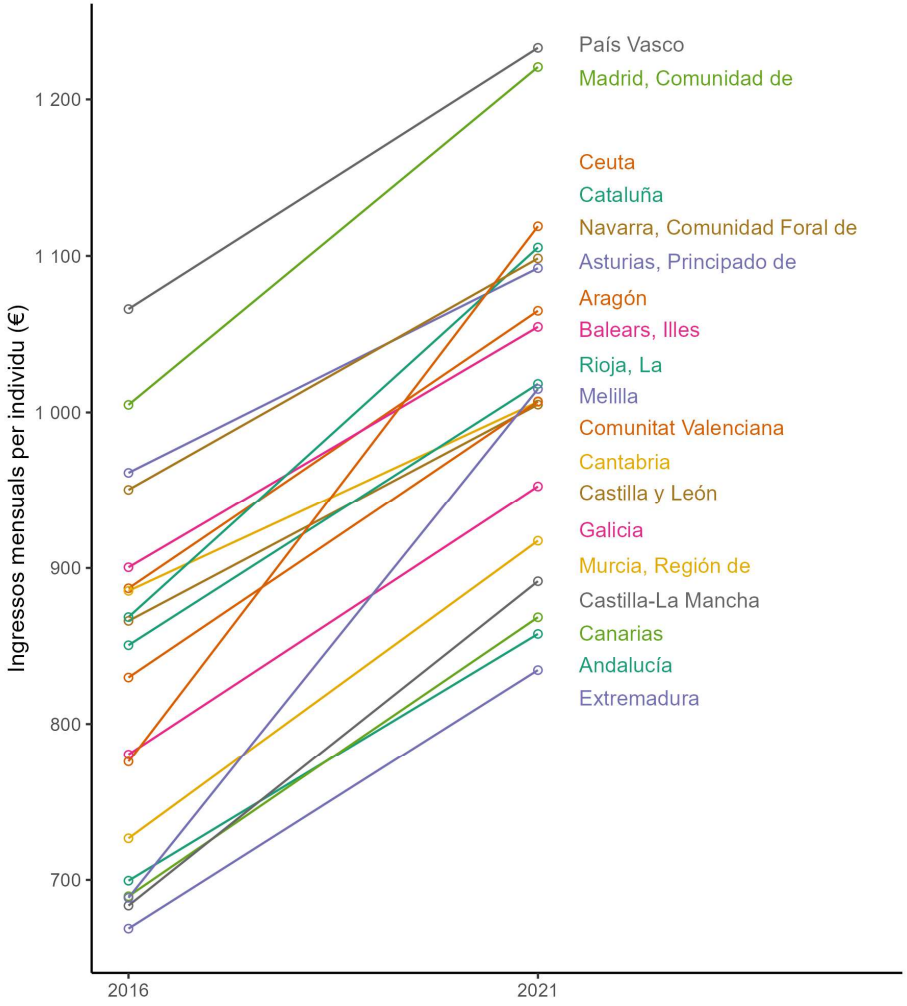
Resulta adequat per visualitzar l'evolució o canvi d'una variable quantitativa entre dos moments temporals. S'usa sobretot quan es vol mostrar simultàniament l'evolució de més d'un subjecte o entitat.

Per interpretar-lo ens fixem en:

- el pendent i la variació de la variable per a cada entitat,
- les entitats amb pendents màxims i mínims,
- les entitats amb valors extrems de la variable en qualsevol dels dos moments.



GRÀFIC DE PENDENTS

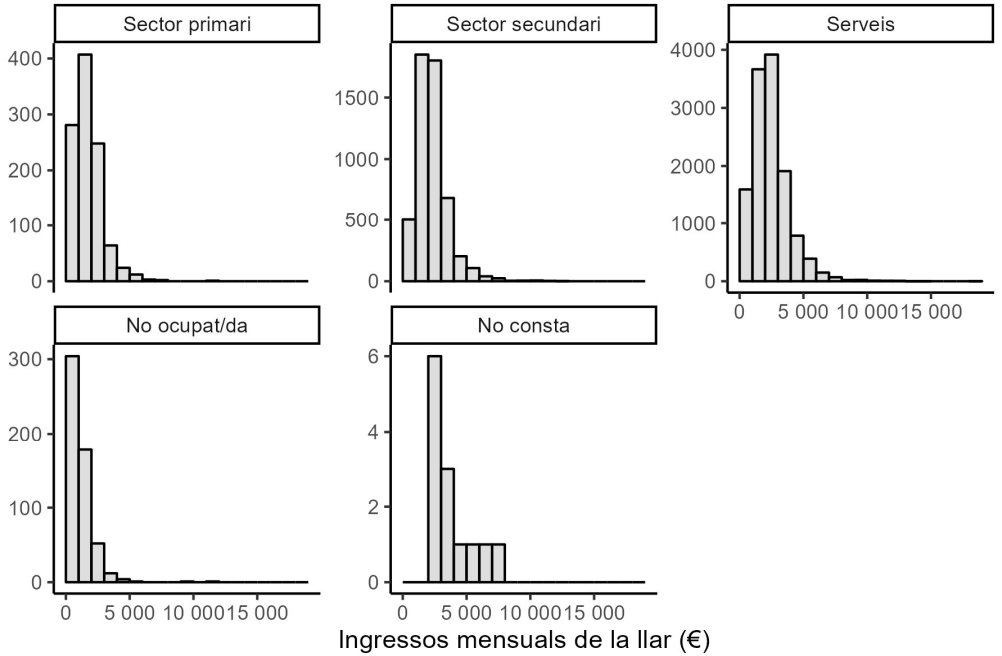


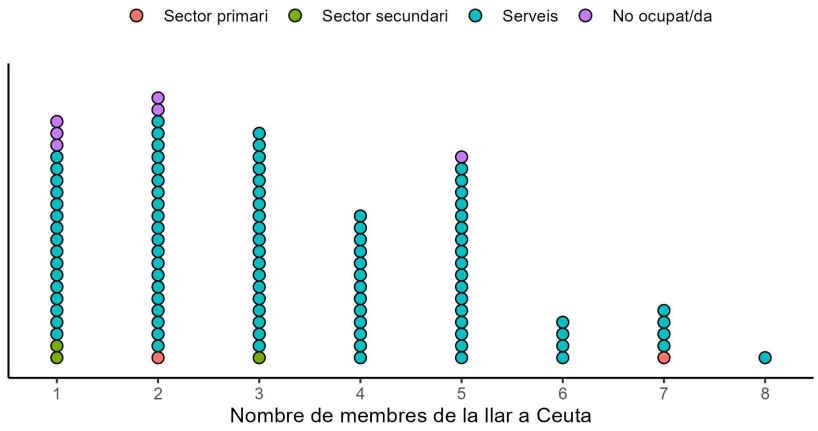

```
ggplot(dfTxanyxccaa[
    dfTxanyxccaa$ANOENC %in% c(2016,2021),],
    aes(x=ANOENC,y=IMPEXACxpm,group=CCAA,col=CCAA))+
  geom_line()+geom_point(shape=21)+
  geom_text_repel(aes(label=CCAA),
    data=dfTxanyxccaa[dfTxanyxccaa$ANOENC==2021,],
    nudge_x=0.5,direction="y",
    min.segment.length=10,size=3.5,max.overlaps = 10,
    force = 2, hjust=0)+
  scale_x_continuous(breaks=c(2016,2021),
    limits = c(2016,2025))+
  scale_y_continuous(
    labels=scales::number_format(accuracy=2),
    breaks=seq(0,2000,100))+
  scale_color_manual(
    values=rep(brewer_pal(palette="Dark2")(8),3)[1:19])+
  labs(x=NULL,y="Ingressos mensuals per individu (€)")+
  theme_classic()+ theme(legend.position="none")
```

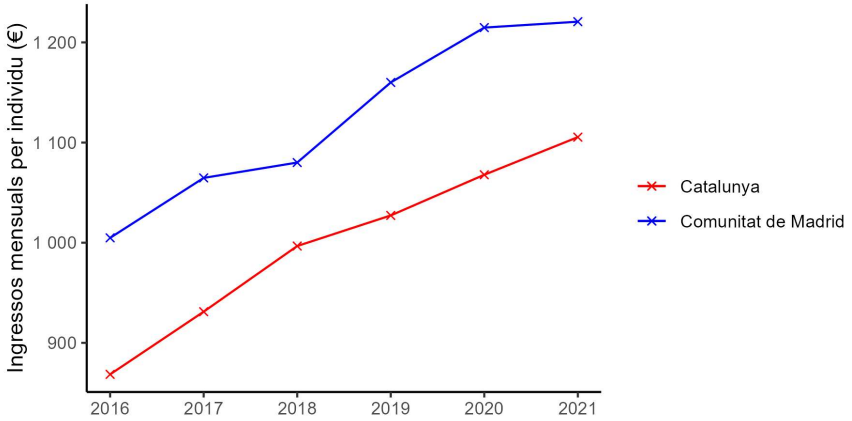
GRÀFICS DE COMPARACIÓ

COMPARACIÓ

>> Variables addicionals
>> Sèries de dades addicionals
>> Múltiples gràfics





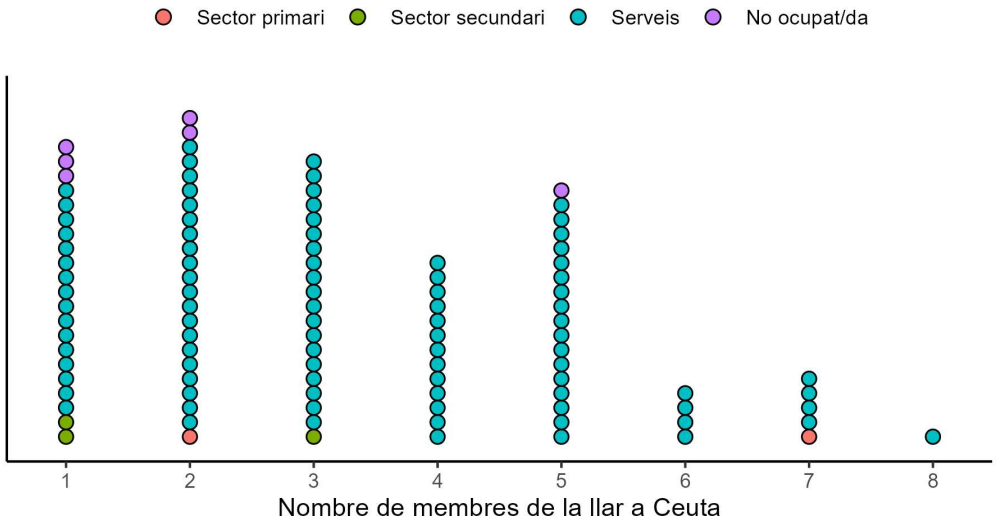


VARIABLES ADDICIONALS

Utilitzem aquesta estratègia quan volem afegir una variable addicional a un gràfic en forma d'element gràfic: mida, color, forma, etc.

Per interpretar-lo ens fixem en:

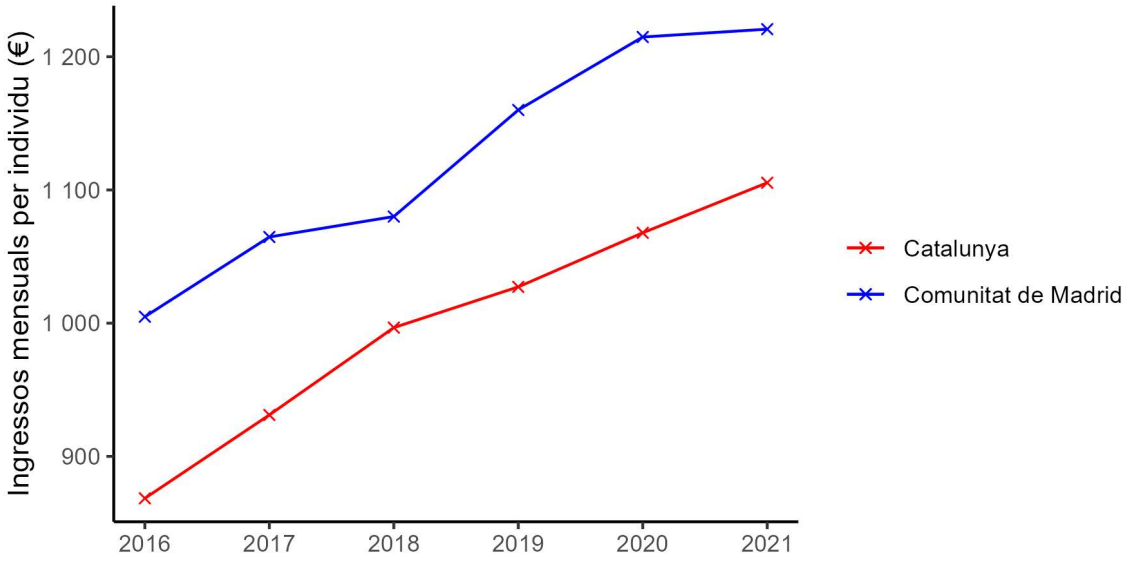
- si les categories (o valors) de la nova variable donen lloc a patrons diferenciables i
- la freqüència de les categories (o valors) de la nova variable.




```
ggplot(df1CEUTA,aes(x=NMIEMB,fill=ACTESTBRED))+
  geom_dotplot(dotsize=0.5,col="black",
    stackgroups=T)+
  scale_y_continuous(NULL,breaks=NULL)+
  scale_x_continuous(breaks=1:max(df1CEUTA$NMIEMB))+
  xlab("Nombre de membres de la llar a Ceuta")+
  theme_classic()+
  theme(legend.position="top",
    legend.title=element_blank())
```

SÈRIES DE DADES ADDICIONALS

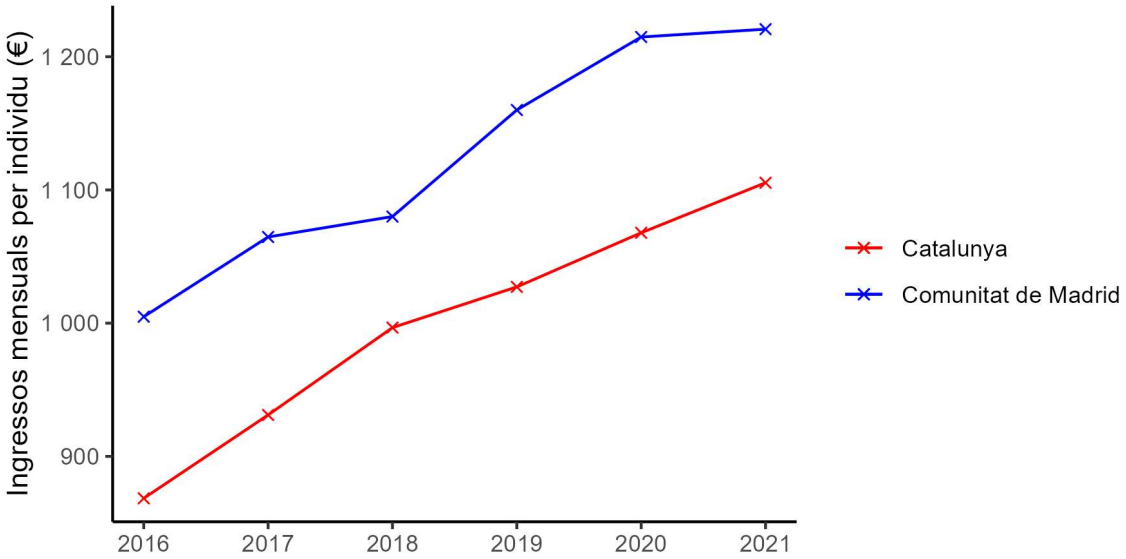
Utilitzem sèries o capes addicionals quan volem comparar els comportaments d'un mateix conjunt de variables per conjunt de dades diferents.



Per interpretar-lo ens fixem en:

- si els diferents conjunts de dades donen lloc a patrons diferenciables i
- si hi ha diferències en els descriptors de cada conjunt de dades.

SÈRIES DE DADES ADDICIONALS



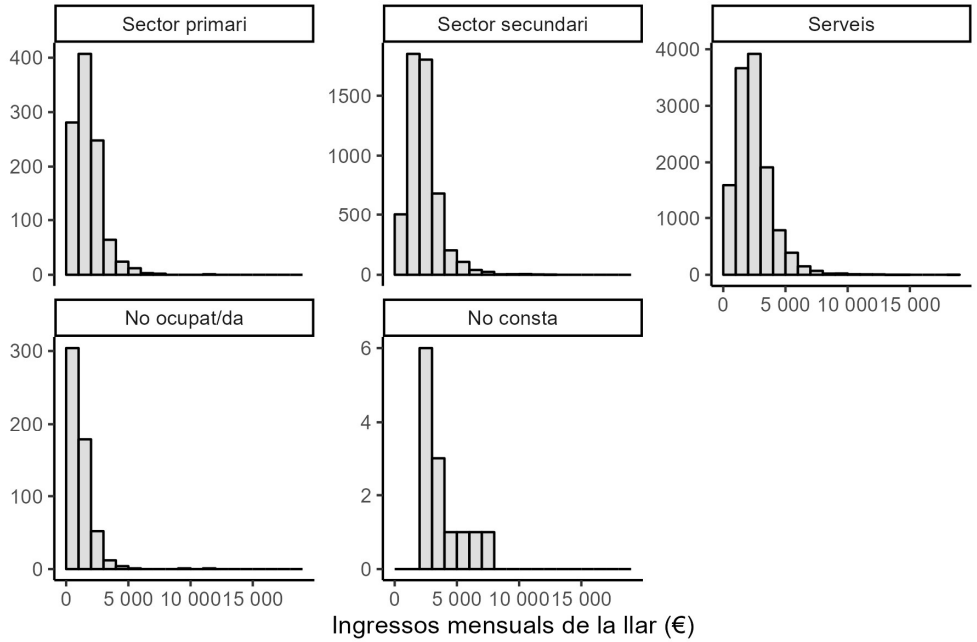

```
ggplot()+
 geom_line(aes(x=ANOENC,y=IMPEXACxpm,
    color="Catalunya"),dfTCATxany)+
 geom_point(aes(x=ANOENC,
    y=IMPEXACxpm,color="Catalunya"),
    dfTCATxany,shape=4)+
 geom_line(aes(x=ANOENC,y=IMPEXACxpm,
    color="Comunitat de Madrid"),
    dfTMADxany)+
 geom_point(aes(x=ANOENC,y=IMPEXACxpm,
    color="Comunitat de Madrid"),
    dfTMADxany,shape=4)+
  scale_y_continuous(
    labels=scales::number_format(accuracy=1))+
  labs(x=NULL,
    y="Ingressos mensuals per individu (€)")+
  scale_color_manual(name=NULL,
    breaks=c("Catalunya","Comunitat de Madrid"),
    values=c("red","blue"))+
  theme_classic()
```

MÚLTIPLES GRÀFICS

Utilitzem múltiples gràfics per comparar representacions completes per diferents categories (o valors) d'una o més variables addicionals.

Per interpretar-lo ens fixem en:

- si les categories (o valors) de les noves variables donen lloc a gràfics diferenciables i
- si hi ha diferències en els descriptors corresponents a cada categoria (o valor) de les variables addicionals.




```
ggplot(df1m,aes(x=IMPEXAC))+
  geom_histogram(col="black",fill="grey",alpha=0.5,
    binwidth=1000,boundary=0,closed="left")+
  scale_y_continuous()+
  scale_x_continuous(
    labels=scales::number_format(accuracy=1))+
  xlab("Ingressos mensuals de la llar (€)")+
  ylab(NULL)+
  facet_wrap(~ACTESTBRED,nrow=2,scales="free_y")+
  theme_classic()
```

Dra. Vanessa Serrano, Dr. Jordi Cuadros
vanessa.serrano@urv.cat, jordi.cuadros@iqs.url.edu

UNIVERSITAT JAUME I

#ProDigital