ENNCIADO 2024-2025

Taller

Tabla de contenidos

Intrucciones para el taller

Se entrega en grupos que deben de estar constituidos en la actividad de grupos. Los grupos son de 2 o 3 ESTUDIANTES, loa caso especiales consultadlos con el profesor para que los autorice.

Enlaces y Bibliografía

- R for data science, Hadley Wickham, Garret Grolemund.
- Fundamentos de ciencia de datos con R.
- Tablas avanzadas: kable, KableExtra.
- Geocomputation with R, Robin Lovelace, Jakub Nowosad, Jannes Muenchow
- Apuntes de R-basico y tidyverse moodel MAT3.

Objetivo MALLORCA

Leeremos los siguientes datos de la zona de etiqueta `mallorca` con el código siguiente:

```
load("clean_data/mallorca/listing_common0_select.RData")
ls()
```

```
[1] "listings_common0_select"
```

```
str(listings_common0_select)
```

```
tibble [52,088 x 16] (S3: tbl_df/tbl/data.frame)
 $ date                   : Date[1:52088], format: "2023-12-17" "2023-12-17" ...
 $ id                     : chr [1:52088] "49752748" "935239498971961146" "24932587" "7825182
 $ price                  : num [1:52088] 2636 107 50 683 62 ...
 $ longitude              : num [1:52088] 2.71 3.12 2.62 3.21 3.24 ...
 $ latitude               : num [1:52088] 39.8 39.3 39.6 39.5 39.4 ...
 $ property_type          : chr [1:52088] "Entire home" "Entire home" "Entire rental unit" "En
 $ room_type              : chr [1:52088] "Entire home/apt" "Entire home/apt" "Entire home/apt
 $ accommodates           : num [1:52088] 14 5 2 10 4 8 5 2 6 10 ...
 $ bedrooms               : num [1:52088] NA NA NA NA NA NA NA NA NA NA ...
 $ beds                   : num [1:52088] 9 4 1 7 3 5 3 3 5 5 ...
 $ number_of_reviews      : num [1:52088] 0 0 124 0 18 0 0 73 0 0 ...
 $ review_scores_rating   : num [1:52088] NA NA 4.88 NA 4.89 NA NA 4.73 NA NA ...
 $ review_scores_value    : num [1:52088] NA NA 4.64 NA 4.83 NA NA 4.64 NA NA ...
 $ host_is_superhost      : logi [1:52088] FALSE FALSE TRUE FALSE FALSE FALSE ...
 $ host_name              : chr [1:52088] "Novasol" "Mallorca Villa Selection" "Juana" "Homer
 $ neighbourhood_cleansed: chr [1:52088] "Sóller" "Santanyí" "Palma de Mallorca" "Felanitx"
```

listings

Hemos cargado el objeto `listing_common0_select` que contiene los datos de las 4 muestras de apartamentos de inside Airbnb de Mallorca con unas 15 0 16 variables.

Notemos que cada apartamento:

- queda identificado por id y por date que ns da la muestra en la que apareció el dato.
- así que cada apartamento parece 4 veces ya que hemos elegido solo los apartamentos que aparecen en las 4 muestras.
- Las muestras son 2023-12-17, 2024-03-23, 2024-06-19, 2024-09-13,

```r
unique(listings_common0_select$date)
```

```
[1] "2023-12-17" "2024-03-23" "2024-06-19" "2024-09-13"
```

reviews

Estos datos necesitan leerse de forma adecuada, las columnas 1,2 y 4 deben ser de tipo `character` las otras son correctas

```r
reviews=read_csv("data/mallorca/2023-12-17/reviews.csv.gz")
str(reviews)
```

```
spc_tbl_ [344,651 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ listing_id   : num [1:344651] 69998 69998 69998 69998 69998 ...
 $ id           : num [1:344651] 881474 4007103 4170371 4408459 4485779 ...
 $ date         : Date[1:344651], format: "2012-01-24" "2013-04-02" ...
 $ reviewer_id  : num [1:344651] 1595616 3868130 5730759 5921885 810469 ...
 $ reviewer_name: chr [1:344651] "Jean-Pierre" "Jo And Mike" "Elizabeth" "Jone" ...
 $ comments     : chr [1:344651] "This place was charming! Lorenzo himself is a very warm and
 - attr(*, "spec")=
  .. cols(
  ..   listing_id = col_double(),
  ..   id = col_double(),
  ..   date = col_date(format = ""),
  ..   reviewer_id = col_double(),
  ..   reviewer_name = col_character(),
  ..   comments = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

```r
head(reviews)
```

```
# A tibble: 6 x 6
  listing_id      id date       reviewer_id reviewer_name comments
       <dbl>   <dbl> <date>           <dbl> <chr>         <chr>
1      69998  881474 2012-01-24     1595616 Jean-Pierre   "This place was charm~
2      69998 4007103 2013-04-02     3868130 Jo And Mike   "We had a four night ~
3      69998 4170371 2013-04-15     5730759 Elizabeth     "Lor's apartment look~
4      69998 4408459 2013-05-03     5921885 Jone          "Wonderful place! 10/~
5      69998 4485779 2013-05-07      810469 Andrea        "My boyfriend and I, ~
6      69998 4619699 2013-05-15     3318059 Devii         "We had a very last m~
```

neighbourhoods.csv

Son dos columnas y la primera es una agrupación de municipios (están NA) y la segunda es el nombre del municipio

```
municipios=read_csv("data/mallorca/2023-12-17/neighbourhoods.csv")
str(municipios)
```

```
spc_tbl_ [53 x 2] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ neighbourhood_group: logi [1:53] NA NA NA NA NA NA ...
 $ neighbourhood      : chr [1:53] "Alaró" "Alcúdia" "Algaida" "Andratx" ...
 - attr(*, "spec")=
  .. cols(
  ..   neighbourhood_group = col_logical(),
  ..   neighbourhood = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
head(municipios)
```

```
# A tibble: 6 x 2
  neighbourhood_group neighbourhood
  <lgl>               <chr>
1 NA                  Alaró
2 NA                  Alcúdia
3 NA                  Algaida
4 NA                  Andratx
5 NA                  Ariany
6 NA                  Artà
```

neighbourhoods.geojson

Es el mapa de Mallorca, o podemos leer así:

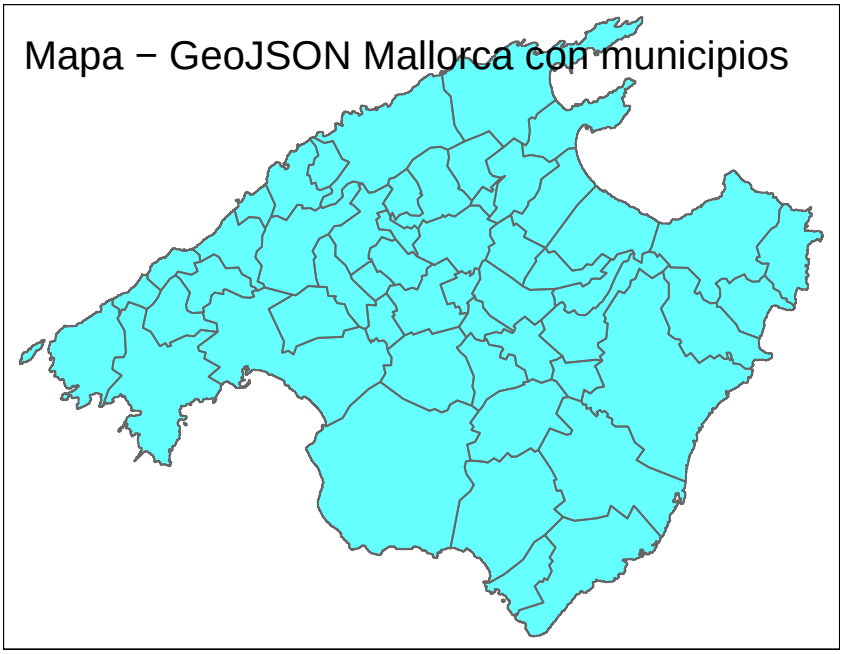
```
library(sf)
library(tmap)

# Leer el archivo GeoJSON
geojson_sf <- sf::st_read("data/mallorca/2024-09-13/neighbourhoods.geojson")
```

```
Reading layer `neighbourhoods' from data source
  `C:\Users\ricuib\Documents\Docencia_24_25\MatGIN\Taller_evaluable_24_25\tallerMat3_24_25\da
  using driver `GeoJSON'
Simple feature collection with 53 features and 2 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 2.303195 ymin: 39.26403 xmax: 3.479028 ymax: 39.96236
Geodetic CRS:  WGS 84
```

```r
# Crear un mapa

# interactivo
tmap_mode("plot") # Cambiar a modo  view/plot   que es interactivo/estático
tm_shape(geojson_sf) +
  tm_polygons(col = "cyan", alpha = 0.6) +
  tm_layout(title = "Mapa - GeoJSON Mallorca con municipios")
```



Tenéis que consultar en la documentación de inside Airbnb para saber que significad cada variable. Os puede ser útil leer los ficheros DATA_ABB_modelo_de_datos.html y DATA_ABB_modelo_de_datos.pdf en los que se explica el modelo de datos de Airbnb y como se cargan en el espacio de trabajo.

Responder las siguientes preguntas como formato Rmarkdown (.Rmd) o quarto (.qmd) y entregad la fuente un fichero en formato html como salida del informe. Se puntúa la claridad de

la respuesta, la calidad de la redacción y la corrección de la respuesta.

Pregunta 1 (1punto)

Del fichero con los datos de listings da los estadísticos descriptivos de las variable `price` y de la variable `number_of_reviews` agrupados por municipio y por año.

Presenta los resultados con una tabla de kableExtra.

Pregunta 2 (1punto)

Consideremos las variables `price` y `number_of_reviews` de Pollença y Palma del perido "2024-09-13", del fichero `listings.csv.gz`. Estudiad si estos datos se aproximan a una distribución normal gráficamente . Para ello, dibujad el histrograma la función kernel que aproxima la desidad y la densidad de la normal de media y varianza las de las muestras de las variables `price` y `number_of_reviews` agrupadas por municipio y por año.

Pregunta 3 (1punto)

Contrastar si las media del precio en Pollença es igual a la de Palma contra que mayor que en Palma para los precios >50 euros y menores de 400$. Construid la hipótesis nula y alternativa, calculad el p-valor y elintervalo de confiaza asociado al contraste. Justifica tecnicamente la conclusión del contraste.

Pregunta 4 (1punto)

Contrastar si las medias de los precios en Palma entre los periodos "2023-12-17" y "2024-03-23" son iguales. Construid la hipótesis nula y alternativa, calculad el p-valor y el intervalo de confianza asociado al contraste. Haced un boxplot

Pregunta 5 (1punto)

Calcular la proporción de apartamentos de la muestra "2024-03-23" con media de valoración `review_scores_rating` mayor que 4 en Palma y en Pollença son igules contra que son distintas. Construid un intervalo de confianza para la diferencia de proporciones.

Pregunta 6 (1punto)

Calcular la proporción de apartamentos de los periodos "2023-12-17" y "2024-03-23" con media de valoración `review_scores_rating` mayor que 4 en Palma y en Pollença son iguales contra que son distintas. Construid un intervalo de confianza para la diferencia de proporciones.

Pregunta 7 (1punto)

La Zipf's law es una ley empírica que dice que la frecuencia de las palabras en un texto es inversamente proporcional a su rango. Decidid si la ley se ajusta a los datos de la longitud de los comentarios de los apartamentos de la muestra "2023-12-17" de Palma. Para ello, haced un análisis de regresión lineal de la frecuencia de las longitudes de los comentarios de los apartamentos de Palma y el rango de las longitudes de los comentarios. Justificad la respuesta.

Como ayuda estudiar el siguiente código, ultilizarlo y comentarlo.

```
library(stringr)
head(reviews)
```

```
# A tibble: 6 x 6
  listing_id      id date       reviewer_id reviewer_name comments
       <dbl>   <dbl> <date>           <dbl> <chr>         <chr>
1      69998  881474 2012-01-24     1595616 Jean-Pierre   "This place was charm~
2      69998 4007103 2013-04-02     3868130 Jo And Mike   "We had a four night ~
3      69998 4170371 2013-04-15     5730759 Elizabeth     "Lor's apartment look~
4      69998 4408459 2013-05-03     5921885 Jone          "Wonderful place! 10/~
5      69998 4485779 2013-05-07      810469 Andrea        "My boyfriend and I, ~
6      69998 4619699 2013-05-15     3318059 Devii         "We had a very last m~
```

```
length_rewiews=stringr::str_length(reviews$comments)
head(table(length_rewiews))
```

```
length_rewiews
   1    2    3    4    5    6
1056  262  249  302  299  281
```

```
aux=table(length_rewiews)
head(aux)
```

```
length_rewiews
   1    2    3    4    5    6
1056  262  249  302  299  281
```

```
head(names(aux))
```

```
[1] "1" "2" "3" "4" "5" "6"
```

```
tbl=tibble( L=as.numeric(names(aux)),Freq=as.numeric(aux),
            Rank=rank(L),Log_Freq=log(Freq),Log_Rank=log(Rank))
str(tbl)
```

```
tibble [2,624 x 5] (S3: tbl_df/tbl/data.frame)
 $ L       : num [1:2624] 1 2 3 4 5 6 7 8 9 10 ...
 $ Freq    : num [1:2624] 1056 262 249 302 299 ...
 $ Rank    : num [1:2624] 1 2 3 4 5 6 7 8 9 10 ...
 $ Log_Freq: num [1:2624] 6.96 5.57 5.52 5.71 5.7 ...
 $ Log_Rank: num [1:2624] 0 0.693 1.099 1.386 1.609 ...
```

```
tbl2=tbl %>% filter(Rank>10) %>% filter(Rank<1000)
sol1=lm(tbl2$Freq~tbl2$Rank)
summary(sol1)
```

```
Call:
lm(formula = tbl2$Freq ~ tbl2$Rank)

Residuals:
    Min      1Q  Median      3Q     Max
-222.53  -91.93  -13.81   85.88  291.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 811.34692    6.42350  126.31   <2e-16 ***
tbl2$Rank    -0.94612    0.01107  -85.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 99.42 on 987 degrees of freedom
Multiple R-squared:  0.8809,    Adjusted R-squared:  0.8808
F-statistic:  7301 on 1 and 987 DF,  p-value: < 2.2e-16
```

```r
sol2=lm(tbl2$Freq~tbl2$Log_Rank)
summary(sol2)
```

```
Call:
lm(formula = tbl2$Freq ~ tbl2$Log_Rank)

Residuals:
    Min      1Q  Median      3Q     Max
-736.30  -52.21  -28.62   70.28  253.16

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2102.273     25.969   80.95   <2e-16 ***
tbl2$Log_Rank -296.983      4.313  -68.85   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 119.6 on 987 degrees of freedom
Multiple R-squared:  0.8277,    Adjusted R-squared:  0.8275
F-statistic:  4741 on 1 and 987 DF,  p-value: < 2.2e-16
```

```r
sol3=lm(tbl2$Log_Freq~tbl2$Log_Rank)
summary(sol3)
```

```
Call:
lm(formula = tbl2$Log_Freq ~ tbl2$Log_Rank)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3672 -0.3816  0.1029  0.4701  0.7228

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.28947    0.11542   97.81   <2e-16 ***
tbl2$Log_Rank -0.99792    0.01917  -52.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5315 on 987 degrees of freedom
```

```
Multiple R-squared:  0.733, Adjusted R-squared:  0.7327
F-statistic:  2710 on 1 and 987 DF,  p-value: < 2.2e-16
```