

Exemple d’aplicació de tècniques d’anàlisi multivariant a l’estudi de l’economia familiar filipina

**Data d’entrega:** 21/02/2019

**Components del grup:**

Miquel de Jover Boira (m.dejover@gmail.com)

Laura Julià Melis (laaurajulia@gmail.com)

Ferran Lacasta Roig (ferranlacasta@gmail.com)

Víctor Navarro Garcés ([vng1997@gmail.com](mailto:vng1997@gmail.com))

Guiu Puigcercós Vilar (96guiu@gmail.com)

Guillem Querol Llaveria (guillem.querolet@gmail.com)

Carles Requena Sánchez (carlesrequenasanchez@gmail.com)

# Definició del projecte i assignació.

Aquest treball es desenvoluparà amb l’objectiu de conèixer l’activitat econòmica familiar a la República de Filipines a partir dels ingressos i les despeses així com també altres característiques rellevants. Per a assolir-ho, s’utilitzaran una sèrie de tècniques d’anàlisi multivariant.

* **Font i informació sobre la base de dades.**

La base de dades s’ha obtingut de la pàgina web *kaggle*, una comunitat online de científics de dades i *Machine learners* que permet als usuaris obtenir dades públiques de manera gratuïta. L’enllaç on es troba la base de dades a *kaggle* és el següent: <https://www.kaggle.com/grosvenpaul/family-income-and-expenditure/home>, malgrat que les dades són propietat de la *Philippine Statistics Authority* i es poden trobar també a la seva pàgina, <http://openstat.psa.gov.ph/search>.

El conjunt de dades va ser recollit realitzant enquestes d’ingressos i despeses familiars (Family Income and Expenditure Survey, FIES) cada 3 anys, amb l’objectiu de predir els ingressos a les llars de Filipines en funció de determinades característiques i conèixer les principals fonts d’ingressos de les famílies.

A la pàgina es pot trobar la base de dades amb el nom de “Family Income and Expediture.csv”.

* **Estructura de la base de dades.**

La base de dades original conté més de 40.000 observacions i 60 variables, 45 de les quals són numèriques (nombres enters i continus, dates i codis identificadors) i 15 categòriques (dicotòmiques i politòmiques).

Per tal d’ajustar-nos a la dimensionalitat de les dades proposada a les bases del treball, s’han seleccionat aleatòriament un subconjunt de 5000 observacions fent servir R i s’han eliminat de la base de dades aquelles variables que es consideren poc rellevants en relació als objectius del treball. A més, s’han agrupat algunes columnes de manera que abans hi havia unes 10 variables en relació a les despeses en menjar (fruites, verdures, carn, etc.) i ara només n’hi ha una. D’aquesta forma, assegurem que tots els procediments requerits podran ser implementats de manera eficient i satisfactòria amb les nostres dades.

Així doncs, finalment la base de dades que utilitzarem d’ara en endavant té 5000 observacions i 34 variables. A continuació es mostra una taula amb totes les variables i els valors missing que té cada una (hi ha XX variables amb missings). Per acabar, s’adjunta el codi d’R utilitzat.

* Taula de valors missing.

|  |  |  |
| --- | --- | --- |
| **Nom de la variable** | **Nre missings** | **Freq missings** |
| **id** | 0 | 0,0000 |
| **host\_id** | 0 | 0,0000 |
| **host\_since** | 0 | 0,0000 |
| **host\_response\_time** | 0 | 0,0000 |
| **host\_response\_rate** | 0 | 0,0000 |
| **host\_identity\_verified** | 0 | 0,0000 |
| **neighbourhood\_cleansed** | 0 | 0,0000 |
| **property\_type** | 0 | 0,0000 |
| **room\_type** | 0 | 0,0000 |
| **accommodates** | 0 | 0,0000 |
| **bathrooms** | 3 | 0,0001 |
| **bedrooms** | 2 | 0,0001 |
| **beds** | 5 | 0,0002 |
| **bed\_type** | 0 | 0,0000 |
| **square\_feet** | 4945 | 0,1919 |
| **price** | 0 | 0,0000 |
| **weekly\_price** | 4474 | 0,1736 |
| **monthly\_price** | 4584 | 0,1779 |
| **security\_deposit** | 1641 | 0,0637 |
| **cleaning\_fee** | 1215 | 0,0472 |
| **guests\_included** | 0 | 0,0000 |
| **extra\_people** | 0 | 0,0000 |
| **number\_of\_reviews** | 0 | 0,0000 |
| **review\_scores\_rating** | 1268 | 0,0492 |
| **review\_scores\_accuracy** | 1269 | 0,0493 |
| **review\_scores\_cleanliness** | 1268 | 0,0492 |
| **review\_scores\_checkin** | 1273 | 0,0494 |
| **review\_scores\_communication** | 1269 | 0,0493 |
| **review\_scores\_location** | 1274 | 0,0494 |
| **review\_scores\_value** | 1275 | 0,0495 |
| **cancellation\_policy** | 0 | 0,0000 |
| **TOTAL** | 25765 | 1,0000 |

* Codi d’R utilitzat.

**#1.1 FILTRATGE DADES**

listing.melbourne <- read.csv("C:/Users/carles.requena/Downloads/

listing melbourne.csv")

View(listing.melbourne)

nrow(listing.melbourne)

set.seed(2019)

p<-sample(x=nrow(listing.melbourne),size = 5000)

noudataframe <- listing.melbourne[p,]

summary(noudataframe)

write.csv(noudataframe, file="datanueva.csv")

**# 1.2 ESTRUCTURA DE LES DADES**

d <-read.csv("E:/multi/datanueva.csv",header=TRUE)

summary(d)

str(d)

dim(d)

# borrem variable X, que és un index per la mostra

d <- d[,-1]

length(d)

# missings

sum(is.na(d))

colSums(is.na(d))

colSums(is.na(d))/5000

n\_var\_miss(d)

library(naniar)

vis\_miss(d)