Universitat Oberta de Catalunya

Estudis d'Informàtica,
Multimedia i
Telecomunicacions

Practica 1: Selecció del conjunt de dades

Dario Cabrera Gurillo

02-12-2022

- 1 Introducció.
- 2 Explicacio, objectius i idees.
- 3 Manipulació i creació del dataset final
    - 3.1 Extracció del dataset final

1 Introducció.

Al llarg d'aquest document parlarem sobre el conjunt de dades elegit per a la creació de la visualització i lliurament del projecte, corresponent a la pràctica 2 de l'assignatura. Les dades elegides tracten el tema de l'atur present en els municipis d'Espanya, des de l'any 2006 fins al novembre del 2022. Per a realitzar aquest conjunt de dades hem d'agafar totes les dades que es troben en fent clic ací.

El motiu principal d'aquesta elecció és: disposar d'un lloc en comú on es representen les dades referents a l'atur de tot l'estat espanyol a diferents escales i períodes. Podríem representar l'atur en tots els municipis de l'estat, però seria molt difícil de llegir i d'interpretar, per aquest motiu, elegirem agafar les dades d'atur per província espanyola.

Aquestes dades poden servir a molts camps, des del ciutadà habitual per a veure l'evolució de l'atur en la seua província, fins a sociòlegs o econòmics per intentar donar explicació de les diferents fluctuacions del treball en Espanya. A més, aquestes dades tenen catalogades l'atur espanyol per 3 rangs d'edat i per sexe, més avant explicarem totes les característiques del conjunt.

En el nostre cas crearem un històric d'evolució de l'atur en Espanya des del 2006 fins a l'actualitat, a continuació veiem els atributs inclosos en un dataset inicial, ja que el nostre serà format per la unió de tots els datasets anuals aconseguits.

```
paro_2006 <- read.csv('inputs/paro_2006.csv', header = FALSE, sep = ";")
# Eliminem la primera fila, que es la cabecera
paro_2006 <- paro_2006[-1,]
# Convertim la nova primera fila en els atributs de les columnes
names(paro_2006) = paro_2006[1,]
# Eliminamos la repeticion de variables i posem els indexs be
paro_2006 <- paro_2006[-1,]
rownames(paro_2006) <- NULL
head(paro_2006)
```

```
##   Código mes             mes Código de CA Comunidad Autónoma Codigo Provincia
## 1     200601 Enero de 2006            1          Andalucía                4
## 2     200601 Enero de 2006            1          Andalucía                4
## 3     200601 Enero de 2006            1          Andalucía                4
## 4     200601 Enero de 2006            1          Andalucía                4
## 5     200601 Enero de 2006            1          Andalucía                4
## 6     200601 Enero de 2006            1          Andalucía                4
##   Provincia Codigo Municipio  Municipio total Paro Registrado
## 1   Almería             4001       Abla                    56
## 2   Almería             4002   Abrucena                    50
## 3   Almería             4003       Adra                   775
## 4   Almería             4004  Albánchez                    14
## 5   Almería             4005  Alboloduy                    24
## 6   Almería             4006      Albox                   360
##   Paro hombre edad < 25 Paro hombre edad 25 -45  Paro hombre edad >=45
## 1                     5                      11                     12
## 2                     3                      16                     15
## 3                    73                     168                     97
## 4                     0                       7                      3
## 5                     2                       6                      0
## 6                    14                      64                     65
##   Paro mujer edad < 25 Paro mujer edad 25 -45  Paro mujer edad >=45
## 1                    2                     23                    3
## 2                    6                     10                    0
## 3                   72                    280                   85
## 4                    0                      2                    2
## 5                    6                      9                    1
## 6                   36                    136                   45
##   Paro Agricultura Paro Industria Paro Construcción Paro Servicios
## 1               10              6                 9             27
## 2                7              6                 7             26
## 3              138             36               103            451
## 4                1              6                 2              4
## 5                1              1                 5             14
## 6               20             26                70            181
##   Paro Sin empleo Anterior
## 1                        4
## 2                        4
## 3                       47
## 4                        1
## 5                        3
## 6                       63
```

El nostre dataset conte les següents columnes:

- **Código mes**: Atribut numèric on tenim l'any més el mes, per exemple, 200601 correspon a gener del 2006, 200812 correspon a desembre del 2008.
- **mes**: Data actual, en el nostre dataset, podríem emprar l'atribut anterior.
- **Código de CA**: Aquest atribut correspon al codi de cada Comunitat Autonoma, tenim el registre d'aquest en el seguent enllaç.
- **Comunidad Autónoma**: Nom de la Comunitat Autonoma d'on extraiem les dades.
- **Codigo Provincia**: Codi postal corresponen a cada província de l'estat Espanyol, tenim el registre d'aquest en el seguent enllaç.
- **Municipio**: Dades corresponent al municipi elegit, en aquesta base de dades tenim els registres de tots els pobles d'Espanya.
- **total Paro Registrado**: Total d'atur registrat en el municipi elegit.
- **Paro hombre edad < 25**: Del total atur registrat, en aquesta columna tenim els hòmens menors de 25 anys.
- **Paro hombre edad 25 -45**: Del total atur registrat, en aquesta columna tenim els hòmens d'edat en el interval [25,45[.
- **Paro hombre edad >=45**: Del total atur registrat, en aquesta columna tenim els hòmens majors o iguals de 45 anys.
- **Paro mujer edad < 25**: Del total atur registrat, en aquesta columna tenim les dones menors de 25 anys.
- **Paro mujer edad 25 -45**: Del total atur registrat, en aquesta columna tenim les dones d'edat en el interval [25,45[.
- **Paro mujer edad >=45**: Del total atur registrat, en aquesta columna tenim les dones majors o iguals de 45 anys.
- **Paro Agricultura**: Del total atur registrat, quantes persones corresponen al camp de l'agricultura.
- **Paro Industria**: Del total atur registrat, quantes persones corresponen al camp de la indústria.
- **Paro Construcción**: Del total atur registrat, quantes persones corresponen al camp de la construcció.
- **Paro Servicios**: Del total atur registrat, quantes persones corresponen al camp de serveis (hostaleria, empleat públic, etc.).
- **Paro Sin empleo Anterior**: Del total atur registrat, aquelles persones que estan aturades per primera vegada en la seua vida.

Per acabar aquesta secció, el dataset ha sigut aportat per l'administració de l'estat, servei públic de treball estatal (SEPE), la llicencia és definit pel SEPE.

2 Explicacio, objectius i idees.

Després de mirar diferents tipus de datasets públics, per exemple, l'estudi actual del preu de la habitatge Base de dades, evolució de l'abandonament animal, on trobem algunes Gràfiques actuals, anàlisis de la violència de gènere i històric d'aquestes en l'estat espanyol aportades per la web del poder judicial, anàlisis dels taxis de Nova York trobats en el TLC Trip Record Data, fins a veure base de dates obertes en el camp de la medicina he acabat per mirar d'analitzar l'atur en Espanya per diferents motius:

- El conjunt de dades és immensament més gran que la resta de dades trobades, i més senzilles per a la manipulació i realització d'estudis. Com a apunt de les dades, tenim un conjunt enorme de dades, amb 20 columnes i al final més d'un milió de registres (sense aplicar el filtre de província), i en aquestes es mesclen algunes dades categòriques i moltes quantitatives.
- Aquestes dades són més fàcils de comprendre per a la població, comparat a la majoria de conjunts abans comentada.
- Els diferents conjunts de gràfiques obtingudes en la xarxa (com està o aquesta) donen una visió global de les dades, però es podria crear un històric, o una història, dels diferents impactes econòmics i polítics espanyols, ja siga veure l'evolució de la bombolla immobiliària, les diferents retallades imposades pels diferents partits polítics del moment, l'efecte del SARS-Cov-2, i més precedents. Així podríem veure clarament com van canviant les dades, siguen mensualment o anualment per província, i intentar trobar motius històrics, i no caure una altra vegada en la mateixa pedra.
- Gràcies a les dades recollides pel SEPE, tenim unes dades actualitzades i preparades, amb distinció de gènere i de rangs d'edat. A banda, el SEPE té dos conjunts de dades més que es podria afegir a aquest per tindre anàlisis més profunds, com Demandantes de Empleo por municipios i Contratos Registrados por municipios.

Per aquests motius hem decidit mirar de crear una visualització dinàmica amb dos formats. Un serà una visualització dinàmica històrica que posarem els diferents impactes que han pogut ocasionar l'augment o la disminució de l'atur, per a poder tenir un històric i intentar buscar un motiu d'aquesta diferència, i en altra instància, un grup de gràfiques i dades enfocades a cada província d'Espanya, per veure com aquestes varien al llarg dels anys (o inclòs com varia dins d'un mateix any). També podríem realitzar l'evolució de l'atur per cada municipi de l'estat espanyol, però segurament per la gran quantitat de dades serà difícil realitzar-ho.

Algunes preguntes que intentarem resoldre amb aquesta visualització seria:

- Hi ha molts moments històrics que provoquen l'augment o la disminució d'atur?, ha sigut provocat pels governs?, o la majoria és per la simple casualitat? Per a dur a terme aquesta pregunta es visualitzara els diferents governs del moment i les polítiques impulsades.
- Existeix en realitat un augment de treball significatiu en les zones de turisme? Trobem impacte negatiu o positiu en la feina segons l'època de l'any en concret?
- Adjuntant la quantitat de població que trobem en el INE, podríem traure percentatges d'atur per població, per preguntar-nos si hi ha diferencia notaria d'atur entre les grans ciutats i els petits pobles.

3 Manipulació i creació del dataset final

En primer lloc, crearem tots els dataset i anirem incloent-ho al dataset final.

```
# Lectura de datasets
paro_2007 <- read.csv('inputs/paro_2007.csv', header = FALSE, sep = ";")
paro_2008 <- read.csv('inputs/paro_2008.csv', header = FALSE, sep = ";")
paro_2009 <- read.csv('inputs/paro_2009.csv', header = FALSE, sep = ";")
paro_2010 <- read.csv('inputs/paro_2010.csv', header = FALSE, sep = ";")
paro_2011 <- read.csv('inputs/paro_2011.csv', header = FALSE, sep = ";")
paro_2012 <- read.csv('inputs/paro_2012.csv', header = FALSE, sep = ";")
paro_2013 <- read.csv('inputs/paro_2013.csv', header = FALSE, sep = ";")
paro_2014 <- read.csv('inputs/paro_2014.csv', header = FALSE, sep = ";")
paro_2015 <- read.csv('inputs/paro_2015.csv', header = FALSE, sep = ";")
paro_2016 <- read.csv('inputs/paro_2016.csv', header = FALSE, sep = ";")
paro_2017 <- read.csv('inputs/paro_2017.csv', header = FALSE, sep = ";")
paro_2018 <- read.csv('inputs/paro_2018.csv', header = FALSE, sep = ";")
paro_2019 <- read.csv('inputs/paro_2019.csv', header = FALSE, sep = ";")
paro_2020 <- read.csv('inputs/paro_2020.csv', header = FALSE, sep = ";")
paro_2021 <- read.csv('inputs/paro_2021.csv', header = FALSE, sep = ";")
paro_2022 <- read.csv('inputs/paro_2022.csv', header = FALSE, sep = ";")


# Crearem la funcio de neteja
neteja <- function(x){
            # Eliminem la primera fila, que es la cabecera
            x <- x[-1,]
            head(x)
            # Convertim la nova primera fila en els atributs de les columnes
            names(x) = x[1,]
            # Eliminamos la repeticion de variables i posem els indexs be
            x <- x[-1,]
            rownames(x) <- NULL
            return(x)
}

# Arreglo
paro_2007 <- neteja(paro_2007)
paro_2008 <- neteja(paro_2008)
paro_2009 <- neteja(paro_2009)
paro_2010 <- neteja(paro_2010)
paro_2011 <- neteja(paro_2011)
paro_2012 <- neteja(paro_2012)
paro_2013 <- neteja(paro_2013)
paro_2014 <- neteja(paro_2014)
paro_2015 <- neteja(paro_2015)
paro_2016 <- neteja(paro_2016)
paro_2017 <- neteja(paro_2017)
paro_2018 <- neteja(paro_2018)
paro_2019 <- neteja(paro_2019)
paro_2020 <- neteja(paro_2020)
paro_2021 <- neteja(paro_2021)
paro_2022 <- neteja(paro_2022)

# Unio de totes les dades
paro_completo <- Reduce(function(x,y) merge(x,y,all = TRUE), list(paro_2006,  paro_2007,paro_2008,paro_20
09,paro_2010,paro_2011, paro_2012, paro_2013, paro_2014, paro_2015, paro_2016, paro_2017, paro_2018, paro
_2019, paro_2020, paro_2021, paro_2022))

# Tamany del conjunt de dates complet
dim(paro_completo)
```

```
## [1] 1640632      20
```

Com podem observar, el nostre conjunt de dates al complet té una mida de 20 columnes i 1.640.632 de files, és molt gran per a poder treballar en ell, el reduirem per províncies per a tindre cada província les seues dades corresponents, a més eliminarem algunes columnes.

```
# Columnes a eliminar
paro_completo$`total Paro Registrado` <- as.integer(paro_completo$`total Paro Registrado`)
paro_completo$`Paro hombre edad < 25` <- as.integer(paro_completo$`Paro hombre edad < 25`)
paro_completo$`Paro hombre edad 25 -45 ` <- as.integer(paro_completo$`Paro hombre edad 25 -45 `)
paro_completo$`Paro hombre edad >=45` <- as.integer(paro_completo$`Paro hombre edad >=45`)
paro_completo$`Paro mujer edad < 25` <- as.integer(paro_completo$`Paro mujer edad < 25`)
paro_completo$`Paro mujer edad 25 -45 ` <- as.integer(paro_completo$`Paro mujer edad 25 -45 `)
paro_completo$`Paro mujer edad >=45` <- as.integer(paro_completo$`Paro mujer edad >=45`)
paro_completo$`Paro Agricultura` <- as.integer(paro_completo$`Paro Agricultura`)
paro_completo$`Paro Industria` <- as.integer(paro_completo$`Paro Industria`)
paro_completo$`Paro Construcción` <- as.integer(paro_completo$`Paro Construcción`)
paro_completo$`Paro Servicios` <- as.integer(paro_completo$`Paro Servicios`)
paro_completo$`Paro Sin empleo Anterior` <- as.integer(paro_completo$`Paro Sin empleo Anterior`)

# Valors nuls per <5 canviats a 2
paro_completo <- mutate_if(paro_completo, is.numeric, ~replace(.,is.na(.),2))

# Agrupacio de dades
paro_reduit <- paro_completo %>% group_by(`Código mes `, `Comunidad Autónoma`) %>%
  summarise( "total Paro Registrado" = sum(`total Paro Registrado`),
             "Paro hombre edad < 25" = sum(`Paro hombre edad < 25`),
             "Paro hombre edad 25 -45" = sum(`Paro hombre edad 25 -45 `),
             "Paro hombre edad >=45" = sum(`Paro hombre edad >=45`),
             "Paro mujer edad < 25" = sum(`Paro mujer edad < 25`),
             "Paro mujer edad 25 -45" = sum(`Paro mujer edad 25 -45 `),
             "Paro mujer edad >=45" = sum(`Paro mujer edad >=45`),
             "Paro Agricultura" = sum(`Paro Agricultura`),
             "Paro Industria" = sum(`Paro Industria`),
             "Paro Construcción" = sum(`Paro Construcción`),
             "Paro Servicios" = sum(`Paro Servicios`),
             "Paro Sin empleo Anterior" = sum(`Paro Sin empleo Anterior`))
```

```
## `summarise()` has grouped output by 'Código mes '. You can override using the `.groups` argument.
```

```
# Eliminem les columnes sobtrants
paro_reduit$`Código mes `<- sub("$", "01", paro_reduit$`Código mes `)
paro_reduit$`Código mes ` <- as.Date(paro_reduit$`Código mes `, format("%Y%m%d"))
paro_reduit$`Código mes ` <- format(paro_reduit$`Código mes `, "%m-%Y")

# Dimensio Final
head(paro_reduit)
```

```
## # A tibble: 6 x 14
## # Groups:   Código mes [1]
##   `Código mes ` `Comunidad Autónoma`    `total Paro Registrado` `Paro hombre ed~
##   <chr>         <chr>                                     <dbl>            <dbl>
## 1 01-2006       Andalucía                                490580            33676
## 2 01-2006       Aragón                                    43166             2910
## 3 01-2006       Asturias, Principado de                   59953             4065
## 4 01-2006       Balears, Illes                            48036             3594
## 5 01-2006       Canarias                                 133647             7745
## 6 01-2006       Cantabria                                 25727             1592
## # ... with 10 more variables: Paro hombre edad 25 -45 <dbl>,
## #   Paro hombre edad >=45 <dbl>, Paro mujer edad < 25 <dbl>,
## #   Paro mujer edad 25 -45 <dbl>, Paro mujer edad >=45 <dbl>,
## #   Paro Agricultura <dbl>, Paro Industria <dbl>, Paro Construcción <dbl>,
## #   Paro Servicios <dbl>, Paro Sin empleo Anterior <dbl>
```

```
dim(paro_reduit)
```

```
## [1] 3838   14
```

Al llarg de l'elaboració del dataset final, hem vist que hi ha uns valors que no poden convertir-se a numèric (més exactament als últims anys), ja que en vegada de tenir un valor té el símbol menys de 5 (<5), per aquest motiu, hem realitzat el canvi d'aquests valors (que es convertiran a nuls una vegada canviats a numèric) a un valor de 2.

El dataset final és compon de 14 columnes i 3838 registres.

3.1 Extracció del dataset final

Per acabar, extraurem el conjunt de dades que hem creat en format *.csv* i en format *.xslx*, per a la Pràctica 2: Creació de la visualització i lliurament del projecte.

```
write.csv2(paro_reduit, "outputs/csv/paro_provincies.csv")
write.csv2(paro_completo, "outputs/csv/paro_municipis.csv")
write.xlsx(paro_reduit, "outputs/excel/paro_provincies.xlsx")
write.xlsx(paro_completo, "outputs/excel/paro_municipis.xlsx")
```

Podeu visualitzar el conjunt de dades obtingut en el meu GitHub personal, sera obert el dia seguent a la finalitzacio d'aquesta tasca, el enllaç és troba fent clic ací