Tipologia i cicle de vida de les dades: Pràctica 2

Diana Campillo Campbell

3 de gener de 2019

Table of Contents

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

En l'actualitat, malauradament, és comú en la comunitat científica la migració de investigadors per a poder-se desenvolupar professionalment. Molts països no tenen la cultura d'acollir aquest potencial que tenen i qui vol dedicar-se a la recerca sovint es veu amb la necessitat de marxar per poder continuar la seva carrera professional. Espanya és un clar exemple. Però ara tenim la oportunitat de comprovar-ho i poder veure quant de cert és aquesta afirmació.

Centrarem el nostre estudi a Espanya i EEUU. Les preguntes que volem respondre són:

- Com està Espanya en comparativa amb altres països de fuga de talents? Comparació respecte EEUU.
- Hi ha una relació de la migració amb la renta per càpita del país? (com origen o com destí)
- A quins països van els investigadors i de quins països venen a Espanya? I als EEUU?

Utilitzarem els següents datasets:

1. **Scientific Researcher Migrations**: obtingut de *Kaggle: https://www.kaggle.com/jboysen/scientist-migrations*. Són els registres corresponen a investigadors que tenen codi ORCID, on s'indica si tenen doctorat, país i any que l'han obtigut, així com informació de països que consten en el registre i si han migrat o no.
2. **gdp**: obtingut de *Github: https://github.com/datasets/gdp/blob/master/data/gdp.csv*. Conté la renta per càpita en dòlars, per països i anys.
3. **Country codes**: obtingut de *Github: https://github.com/datasets/country-codes/blob/master/data/country-codes.csv*. Conté la codificació de països amb les seves descripcions en diferents idiomes. Entre les codificacions hi ha les que utilitzen el fitxers anteriors (Alpha-2 i Alpha-3).

A continuació una breu descripció dels atributs de cada dataset:

```
#Obrim dataset "Scientific Researcher Migrations"
myfileResearcher <- "ORCID_migrations_2016_12_16_by_person.csv"
mydataResearcher <- read.csv(myfileResearcher,na.strings=c("", "NA"),
stringsAsFactors=FALSE)
#Obrim dataset "gdp"
myfileGDP <- "GPD.csv"
mydataGDP <- read.csv(myfileGDP,na.strings=c("", "NA"),
stringsAsFactors=FALSE)
#Obrim dataset "Country codes"
myfileCodes <- "CountryCodes.csv"
mydataCodes <- read.csv(myfileCodes,na.strings=c("", "NA"),
stringsAsFactors=FALSE)
```

1. **Scientific Researcher Migrations**:

741867 registres i 8 columes. Aquestes són: *orcid_id, phd_year, country_2016, earliest_year, earliest_country, has_phd, phd_country, has_migrated*. Concretament:

- **orcid_id**: identificador únic ORCID de la persona
- **phd_year**: *any* en el que va obtenir el doctorat
- **country_2016**: *codi Alpha-2* del país on vivia l'any 2016

- **earliest_year**: *any* anteror al 2016 que es té informació de la persona
- **earliest_country**: *codi Alpha-2* del país on vivía en l'any anterior al 2016 (any de l'atribut anterior)
- **has_phd**: *True/False* si té o no el doctorat
- **phd_country**: *codi Alpha-2* del país on ha obtingut el doctorat
- **has_migrated**: *True/False* si ha emigrat

```
res <- sapply(mydataResearcher,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

variables	clase
orcid_id	character
phd_year	numeric
country_2016	character
earliest_year	numeric
earliest_country	character
has_phd	character
phd_country	character
has_migrated	character

2. **gdp**:

11507 registres i 4 columes. Aquestes són: *Country.Name, Country.Code, Year, Value*. Concretament:

- **Country Name**: descripció del país
- **Country Code**: *codi Alpha-3* del país
- **Year**: *any* al que correspon el registre
- **Value**: quantitat en *USdòlars* de renta per càpita.

```
res <- sapply(mydataGDP,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

variables	clase
Country.Name	character
Country.Code	character
Year	integer
Value	numeric

3. **Country codes**:

250 registres i 56 columes. Aquestes són: *ï..FIFA, Dial, ISO3166.1.Alpha.3, MARC, is_independent, ISO3166.1.numeric, GAUL, FIPS, WMO, ISO3166.1.Alpha.2, ITU, IOC, DS, UNTERM.Spanish.Formal, Global.Code, Intermediate.Region.Code, official_name_fr, UNTERM.French.Short, ISO4217.currency_name, Developed...Developing.Countries, UNTERM.Russian.Formal, UNTERM.English.Short, ISO4217.currency_alphabetic_code, Small.Island.Developing.States..SIDS., UNTERM.Spanish.Short, ISO4217.currency_numeric_code, UNTERM.Chinese.Formal, UNTERM.French.Formal, UNTERM.Russian.Short, M49, Sub.region.Code, Region.Code, official_name_ar, ISO4217.currency_minor_unit, UNTERM.Arabic.Formal, UNTERM.Chinese.Short, Land.Locked.Developing.Countries..LLDC., Intermediate.Region.Name, official_name_es, UNTERM.English.Formal, official_name_cn, official_name_en, ISO4217.currency_country_name, Least.Developed.Countries..LDC., Region.Name, UNTERM.Arabic.Short, Sub.region.Name, official_name_ru, Global.Name, Capital, Continent, TLD, Languages, Geoname.ID, CLDR.display.name, EDGAR*. Només descriurem les que ens interessen per l'objectiu de descofidicar les dades dels fitxers anteriors:

- **ISO3166.1.Alpha.3**: codi de tres caràcters (utilitzat en el fitxer GPD).
- **ISO3166.1.Alpha.2**: codi de dos caràcters (utilitzat en el fitxer Researchers Migration).
- **official_name_en**: descripció en anglès del país

```
res <- sapply(mydataCodes,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
```

variables	clase
ï..FIFA	character
Dial	character
ISO3166.1.Alpha.3	character
MARC	character
is_independent	character
ISO3166.1.numeric	integer
GAUL	character
FIPS	character
WMO	character
ISO3166.1.Alpha.2	character
ITU	character
IOC	character
DS	character
UNTERM.Spanish.Formal	character
Global.Code	character
Intermediate.Region.Code	integer
official_name_fr	character

UNTERM.French.Short	character
ISO4217.currency_name	character
Developed...Developing.Countries	character
UNTERM.Russian.Formal	character
UNTERM.English.Short	character
ISO4217.currency_alphabetic_code	character
Small.Island.Developing.States..SIDS.	character
UNTERM.Spanish.Short	character
ISO4217.currency_numeric_code	character
UNTERM.Chinese.Formal	character
UNTERM.French.Formal	character
UNTERM.Russian.Short	character
M49	integer
Sub.region.Code	integer
Region.Code	integer
official_name_ar	character
ISO4217.currency_minor_unit	character
UNTERM.Arabic.Formal	character
UNTERM.Chinese.Short	character
Land.Locked.Developing.Countries..LLDC.	character
Intermediate.Region.Name	character
official_name_es	character
UNTERM.English.Formal	character
official_name_cn	character
official_name_en	character
ISO4217.currency_country_name	character
Least.Developed.Countries..LDC.	character
Region.Name	character
UNTERM.Arabic.Short	character
Sub.region.Name	character
official_name_ru	character
Global.Name	character
Capital	character
Continent	character
TLD	character
Languages	character

Geoname.ID	integer
CLDR.display.name	character
EDGAR	character

---

2. Integració i selecció de les dades d'interès a analitzar.

- **Scientific Researcher Migrations**

El fitxer *Scientific Researcher Migrations* conté diversos atributs relatius a països i anys. Per facilitar el tractament de dades, decidirem aquí quin país determinem com país Origen i quin país serà el Destí. També assignarem un any de referència.

ORIGEN

Serà el que, a partir de les dades, deduïm que deu ser la procedència de la persona. L'obtindrem seguin el següent criteri:

Com a atribut Origen posarem *earliest_country*. Si aquest està en blanc, *phd_country*. Si està en blanc, mirem si *has_migrated="False"*. Aleshores posarem *country_2016.* Si no, no podrem omplir el camp.

DESTI

Serà el que, a partir de les dades, deduïm que deu ser el lloc de residència actual de la persona. L'obtindrem seguin el següent criteri:

Com a atribut Origen posarem *country_2016* Si està en blanc, i *has_migrated="False"*, posarem el país *earliest_country.* Si també està en blanc, *phd_country* Si està en blanc, ho deixarem en blanc

ANY

Serà un punt de referència en el temps. El podem entendre quan l'investigador comença a formar part de la comunitat científica. Ja sigui perquè s'ha registrat a ORCID, o per fer el PhD. L'obtindrem seguin el següent criteri:

Com a atribut Any poseré el mínim any entre *earliest_year* i *phd_year* (si algun d'aquest no té valor, doncs el valor del atribut que tingui valor). Si està en blanc, si té informat el *country_2016* posarem l'any '2016' Si està en blanc, ho deixarem en blanc.

```
#ORIGEN
mydataResearcher$origen_country<-mydataResearcher$earliest_country

mydataResearcher$origen_country[which
(is.na(mydataResearcher$origen_country))]<-
mydataResearcher$phd_country[which
(is.na(mydataResearcher$origen_country))]
```

```r
mydataResearcher$origen_country[which
(is.na(mydataResearcher$origen_country) &
mydataResearcher$has_migrated=="False")]<-
mydataResearcher$country_2016[which
(is.na(mydataResearcher$origen_country) &
mydataResearcher$has_migrated=="False")]

#DESTÍ
mydataResearcher$desti_country<-mydataResearcher$country_2016

mydataResearcher$desti_country[which
(is.na(mydataResearcher$desti_country) &
mydataResearcher$has_migrated=="False")]<-
mydataResearcher$earliest_country[which
(is.na(mydataResearcher$desti_country) &
mydataResearcher$has_migrated=="False")]

mydataResearcher$desti_country[which
(is.na(mydataResearcher$desti_country) &
mydataResearcher$has_migrated=="False")]<-
mydataResearcher$phd_country[which (is.na(mydataResearcher$desti_country)
& mydataResearcher$has_migrated=="False")]


#ANY

mydataResearcher <- transform(mydataResearcher, min =
pmin(mydataResearcher$earliest_year,mydataResearcher$phd_year  ))
colnames(mydataResearcher)[colnames(mydataResearcher)=="min"] <- "any"


#Ara anem a introduir el valor pel que han quedat na:
mydataResearcher$any[which (is.na(mydataResearcher$any))]<-
mydataResearcher$earliest_year[which (is.na(mydataResearcher$any))]
mydataResearcher$any[which (is.na(mydataResearcher$any))]<-
mydataResearcher$phd_year[which (is.na(mydataResearcher$any))]

mydataResearcher$any[which (is.na(mydataResearcher$any) &
!is.na(mydataResearcher$country_2016))]<-2016

head(mydataResearcher)
```

```
##              orcid_id phd_year country_2016 earliest_year
earliest_country
## 1 0000-0001-5000-0138       NA           CO          2014
CO
## 2 0000-0001-5000-0736     2006         <NA>            NA
<NA>
## 3 0000-0001-5000-1018     2015           US          2005
```

```
US
## 4 0000-0001-5000-1181       NA           RU          1978
RU
## 5 0000-0001-5000-1923     2016           GB          2004
GB
## 6 0000-0001-5000-223X     1998           GB          1989
GB
##   has_phd phd_country has_migrated origen_country desti_country  any
## 1   False        <NA>        False             CO            CO 2014
## 2    True          PT        False             PT            PT 2006
## 3    True          US        False             US            US 2005
## 4   False        <NA>        False             RU            RU 1978
## 5    True          GB        False             GB            GB 2004
## 6    True          GB         True             GB            GB 1989
```

- **gdp**

El fitxer *gdp* afegirem la codificació de país Alpha-2, per poder creuar amb els registres del fitxer anterior.

```
#Creuem GDP amb fitxer de codis
#Reduïm les dades de CountryCodes i treiem els que no tenen codi o
descripció assignat:
mydataCodes<-subset(mydataCodes, select =
c(ISO3166.1.Alpha.3,ISO3166.1.Alpha.2,official_name_en) )
mydataCodes <- subset(mydataCodes,
((!is.na(mydataCodes$ISO3166.1.Alpha.3)&((!is.na(mydataCodes$ISO3166.1.Al
pha.2)&(!is.na(mydataCodes$official_name_en)))))))

#Reemplacem codi de país en fitxer GPD a Alpha-2
mydataGDP<-merge(x = mydataGDP, y = mydataCodes,
by.x=c("Country.Code"),by.y=c("ISO3166.1.Alpha.3"), all.x = TRUE)
#Eliminem registres que no tinguin info a Alpha-2
mydataGDP<-na.omit(mydataGDP, cols=c("ISO3166.1.Alpha.2"))
```

Ara procedirem a crear el nostre dataset amb el que farem els estudis, deixant només els atributs que ens interessen:

Migration

```
myvars <- c("origen_country", "desti_country", "any", "has_phd",
"has_migrated")

mydataMigration <- mydataResearcher[myvars]

#Ara afegim GDP origen i GDP destí per a l'any del registre:
mydataMigration<-merge(x = mydataMigration, y = mydataCodes,
by.x=c("origen_country"),by.y=c("ISO3166.1.Alpha.2"), all.x = TRUE)

colnames(mydataMigration)[colnames(mydataMigration)=="official_name_en"]
<- "origen_name"
```

```r
mydataMigration$Country.Code<-NULL
mydataMigration$Country.Name<-NULL
mydataMigration$Year<-NULL
mydataMigration$ISO3166.1.Alpha.3<-NULL

mydataMigration<-merge(x = mydataMigration, y = mydataGDP,
by.x=c("origen_country","any"),by.y=c("ISO3166.1.Alpha.2","Year"), all.x
= TRUE)


colnames(mydataMigration)[colnames(mydataMigration)=="Value"] <-
"origen_GDP"
mydataMigration$Country.Code<-NULL
mydataMigration$Country.Name<-NULL
mydataMigration$official_name_en<-NULL

#idem pel destí:

mydataMigration<-merge(x = mydataMigration, y = mydataCodes,
by.x=c("desti_country"),by.y=c("ISO3166.1.Alpha.2"), all.x = TRUE)

colnames(mydataMigration)[colnames(mydataMigration)=="official_name_en"]
<- "desti_name"
mydataMigration$Country.Code<-NULL
mydataMigration$Country.Name<-NULL
mydataMigration$Year<-NULL
mydataMigration$ISO3166.1.Alpha.3<-NULL

mydataMigration<-merge(x = mydataMigration, y = mydataGDP,
by.x=c("desti_country","any"),by.y=c("ISO3166.1.Alpha.2","Year"), all.x =
TRUE)
nrow(mydataMigration)
```

```
## [1] 741867
```

```r
colnames(mydataMigration)[colnames(mydataMigration)=="Value"] <-
"desti_GDP"
mydataMigration$Country.Code<-NULL
mydataMigration$Country.Name<-NULL
mydataMigration$official_name_en<-NULL

head(mydataMigration)
```

```
##   desti_country  any origen_country has_phd has_migrated origen_name
## 1            AD 1990             AD   False         True     Andorra
## 2            AD 1991             AD   False        False     Andorra
## 3            AD 1993             AD   False        False     Andorra
## 4            AD 1993             AD    True         True     Andorra
## 5            AD 1996             AD    True         True     Andorra
## 6            AD 1996             AD   False         True     Andorra
```

```
##   origen_GDP desti_name  desti_GDP
## 1 1029048482    Andorra 1029048482
## 2 1106928583    Andorra 1106928583
## 3 1007025755    Andorra 1007025755
## 4 1007025755    Andorra 1007025755
## 5 1223945357    Andorra 1223945357
## 6 1223945357    Andorra 1223945357
```

Per finalitzar factoritzarem les variables qualitatives i finalment, mostrarem un resum de les dades de les nostres dues taules:

```
mydataMigration$origen_country<-as.factor(mydataMigration$origen_country)
mydataMigration$origen_name<-as.factor(mydataMigration$origen_name)
mydataMigration$desti_country<-as.factor(mydataMigration$desti_country)
mydataMigration$desti_name<-as.factor(mydataMigration$desti_name)
mydataMigration$has_phd<-as.factor(mydataMigration$has_phd)
mydataMigration$has_migrated<-as.factor(mydataMigration$has_migrated)

summary(mydataMigration)
```

```
##  desti_country         any         origen_country    has_phd
##  US     :109268   Min.   :1913     US     :102521   False:412576
##  GB     : 40272   1st Qu.:1994     IN     : 41705   True :329291
##  BR     : 39842   Median :2003     BR     : 40393
##  IN     : 35744   Mean   :2000     GB     : 39756
##  CN     : 32668   3rd Qu.:2009     CN     : 39529
##  (Other):360032   Max.   :2017     (Other):368617
##  NA's   :124041   NA's   :130587   NA's   :109346
##  has_migrated
##  False:630718
##  True :111149
##
##
##
##
##
##                                                 origen_name
##  United States of America                             :102521
##  India                                                : 41705
##  Brazil                                               : 40393
##  United Kingdom of Great Britain and Northern Ireland: 39756
##  China                                                : 39529
##  (Other)                                              :363758
##  NA's                                                 :114205
##    origen_GDP
##  Min.   :8.985e+07
##  1st Qu.:2.350e+11
##  Median :8.796e+11
##  Mean   :2.689e+12
##  3rd Qu.:2.456e+12
```

```
##  Max.   :1.862e+13
##  NA's   :152340
##                                                    desti_name
##  United States of America                               :109268
##  United Kingdom of Great Britain and Northern Ireland: 40272
##  Brazil                                                 : 39842
##  India                                                  : 35744
##  China                                                  : 32668
##  (Other)                                                :354802
##  NA's                                                   :129271
##    desti_GDP
##  Min.   :4.910e+07
##  1st Qu.:2.390e+11
##  Median :8.916e+11
##  Mean   :2.850e+12
##  3rd Qu.:2.521e+12
##  Max.   :1.862e+13
##  NA's   :164434

#-------------------------afegir desviació estandar i mostrar comparant
origen amb destí---
#I indicar descripció completa de les dues taules amb les que
treballarem.
```

---

3. Neteja de les dades.

3.1 Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Revisem l'estat de les dades de la nostra taula d'estudi. Aquesta ja ha tingut un tractament respecte les dades originals (amb la creació de nous atributs Origen, Desti i Any), així que molts dels problemes ja els hem resolt prèviament. Revisem l'estat actual de les dades:

-**mydataMigration** Revisem si hi ha valors nulls:

```
sapply(mydataMigration,class)
```

```
##  desti_country            any origen_country        has_phd
has_migrated
##       "factor"      "numeric"       "factor"       "factor"
"factor"
##    origen_name     origen_GDP     desti_name      desti_GDP
##       "factor"      "numeric"       "factor"      "numeric"
```

```
#Com que al obrir el fitxer hem convertit tots els bancs en NA, només cal
buscar els NA:
sapply(mydataMigration, function(x) sum(is.na(x)))
```

```
##  desti_country            any origen_country        has_phd
has_migrated
##         124041         130587         109346              0
0
##    origen_name     origen_GDP     desti_name      desti_GDP
##         114205         152340         129271         164434
```

Anem a revisar els atributs que tenen valors nulls: 1. Ens interessen dades que tinguin origen i destí informats. Si algún dels dos atributs és buit, només tindríem una informació parcial. Pel nostre estudi optem per eliminar-los. També eliminarem els que no tenen any informat:

```
nrow(mydataMigration)
```

```
## [1] 741867
```

```
mydataMigration<-
mydataMigration[!is.na(mydataMigration$origen_country)&!is.na(mydataMigra
tion$desti_country)&!is.na(mydataMigration$any),]
nrow(mydataMigration)
```

```
## [1] 597787
```

```
sapply(mydataMigration, function(x) sum(is.na(x)))
```

```
##  desti_country            any origen_country        has_phd
has_migrated
##             0             0             0             0
0
##    origen_name     origen_GDP     desti_name      desti_GDP
##           4392          20964           4898          20380
```

2. Mirem quins països no tenim descripció. És una dada informativa, només és una codificació per ajudar-nos en la interpretació de les dades. Mirem de resoldre-ho:

```
unique(mydataMigration$origen_country[is.na(mydataMigration$origen_name)]
)
```

```
## [1] TW
## 224 Levels: AD AE AF AG AI AL AM AO AQ AR AS AT AU AW AX AZ BA BB BD
... ZW
```

```
unique(mydataMigration$desti_country[is.na(mydataMigration$desti_name)])
```

```
## [1] TW
## 227 Levels: AD AE AF AG AI AL AM AO AQ AR AS AT AU AW AX AZ BA BB BD
... ZW
```

Veiem que només és un codi que no ha trobat la descripció. Per les sigles sembla "Taiwan". Revisem la taula de codis i comprovem que no hi ha cap registre que correspongui amb la descripció de "Taiwan". Ho podem entendre al ser un país on no a tot arreu és reconegut. Procedim doncs a introduir manualment el nom corresponent:

```
#Comprovem que no tenim cap registre de descripció Taiwan:
mydataCodes$official_name_en[which
(mydataCodes$official_name_en=="Taiwan")]

## character(0)

#Al no retornar cap resultat, procedim doncs a introduir el nom
manualment:
mydataMigration$origen_name<-as.character(mydataMigration$origen_name)
mydataMigration$origen_name[which
(mydataMigration$origen_country=="TW")]<-"Taiwan"
mydataMigration$origen_name<-as.factor(mydataMigration$origen_name)

mydataMigration$desti_name<-as.character(mydataMigration$desti_name)
mydataMigration$desti_name[which (mydataMigration$desti_country=="TW")]<-
"Taiwan"
mydataMigration$desti_name<-as.factor(mydataMigration$desti_name)

sapply(mydataMigration, function(x) sum(is.na(x)))

##  desti_country            any origen_country        has_phd
has_migrated
##              0              0              0              0
0
##    origen_name     origen_GDP     desti_name      desti_GDP
##              0          20964              0          20380
```

4. Ara només queda revisar les dades de GPD. Aquests corresponen a països i anys dels que no disposem informació de renta per càpita. Com que centrarem l'estudi de renta per càpita a Espanya i EEUU, només revisarem aquests dos països:

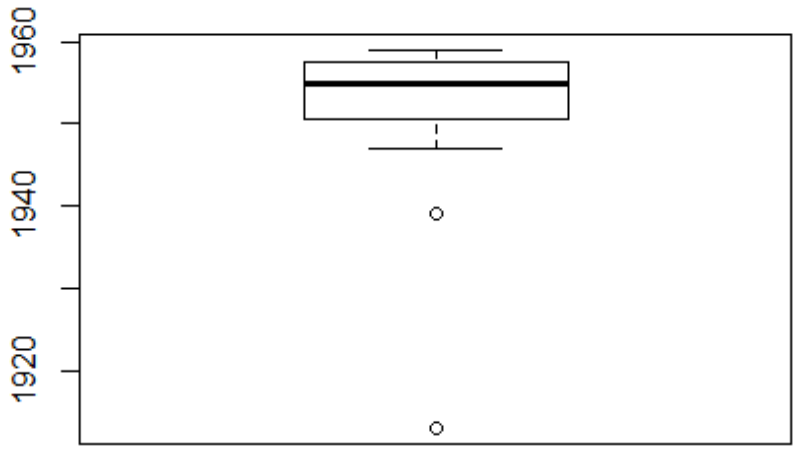
```
esp<-mydataMigration[which (mydataMigration$origen_country=="ES" &
is.na(mydataMigration$origen_GDP)),]
usa<-mydataMigration[which (mydataMigration$origen_country=="US" &
is.na(mydataMigration$origen_GDP)),]

min(esp$any)

## [1] 1913

min(usa$any)

## [1] 1913

max(esp$any)
```

```
## [1] 1959
```

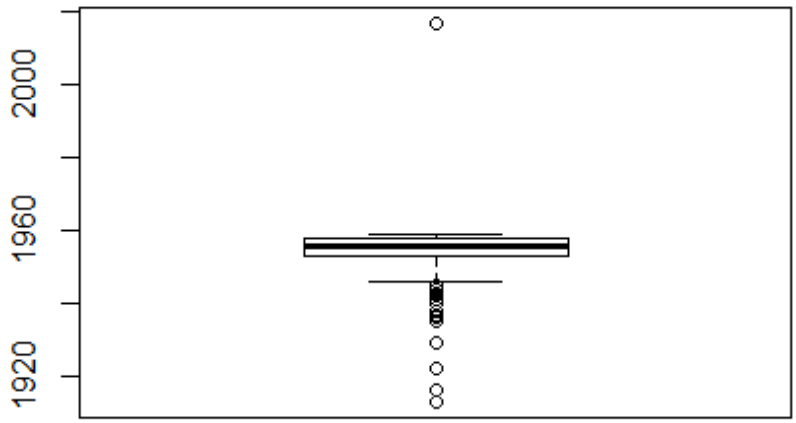
```
max(usa$any)
```

```
## [1] 2017
```

```
boxplot(esp$any)
```




```
boxplot(usa$any)
```



Fins al 1960 aprox no tenim informació de renta per càpita tant per EEUU com per Espanya. Al ser dates semblants, i per no perdre informació de migracions, mantenim els registres i deixem les dades nul·les a 0. No mirem de fer estimació de la renta per càpita perquè no seria un valor fiable, i tenim prous dades per als anys posteriors al 1960.

```
mydataMigration$origen_GDP[is.na(mydataMigration$origen_GDP)]<-0
mydataMigration$desti_GDP[is.na(mydataMigration$desti_GDP)]<-0
sapply(mydataMigration, function(x) sum(is.na(x)))
```

```
##  desti_country            any origen_country        has_phd
has_migrated
##              0              0              0              0
0
##    origen_name     origen_GDP     desti_name      desti_GDP
##              0              0              0              0
```

Resum dataset:

```
summary(mydataMigration)
```

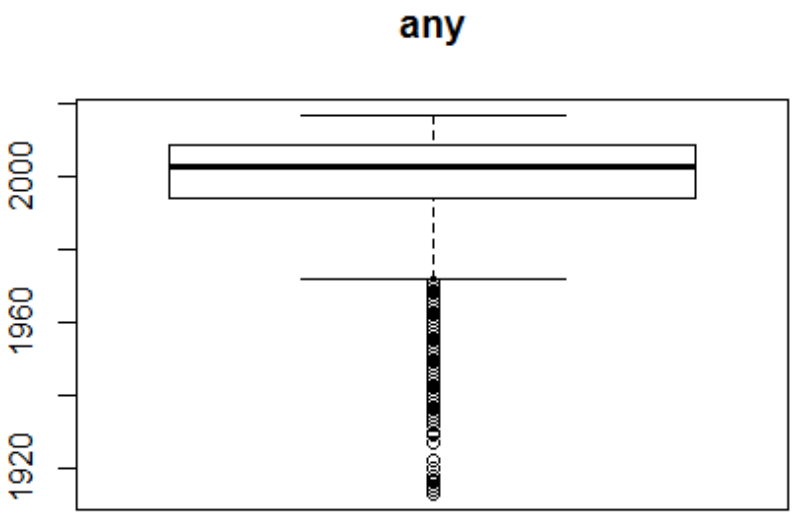
```
##  desti_country         any       origen_country    has_phd
##  US     :105995   Min.   :1913   US     : 97820   False:301472
##  BR     : 39424   1st Qu.:1994   BR     : 39630   True :296315
##  GB     : 39321   Median :2003   IN     : 39471
##  IN     : 34331   Mean   :2000   GB     : 37636
##  CN     : 31253   3rd Qu.:2009   CN     : 36887
```

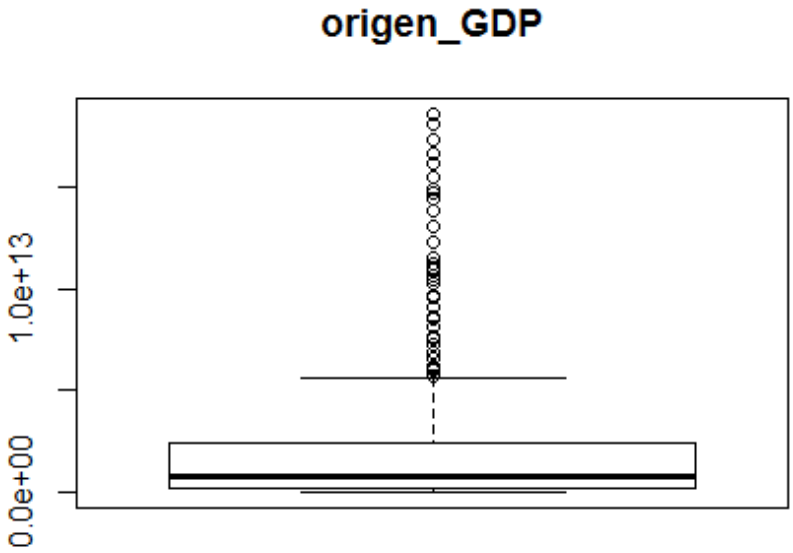
```
##  ES     : 30640   Max.   :2017   ES     : 30880
##  (Other):316823                  (Other):315463
##  has_migrated
##  False:502668
##  True : 95119
##
##
##
##
##
##                                                 origen_name
##  United States of America                              : 97820
##  Brazil                                                : 39630
##  India                                                 : 39471
##  United Kingdom of Great Britain and Northern Ireland: 37636
##  China                                                 : 36887
##  Spain                                                 : 30880
##  (Other)                                               :315463
##    origen_GDP
##  Min.   :0.000e+00
##  1st Qu.:2.044e+11
##  Median :8.089e+11
##  Mean   :2.624e+12
##  3rd Qu.:2.399e+12
##  Max.   :1.862e+13
##
##                                                  desti_name
##  United States of America                              :105995
##  Brazil                                                : 39424
##  United Kingdom of Great Britain and Northern Ireland: 39321
##  India                                                 : 34331
##  China                                                 : 31253
##  Spain                                                 : 30640
##  (Other)                                               :316823
##    desti_GDP
##  Min.   :0.000e+00
##  1st Qu.:2.064e+11
##  Median :8.089e+11
##  Mean   :2.752e+12
##  3rd Qu.:2.465e+12
##  Max.   :1.862e+13
##
```

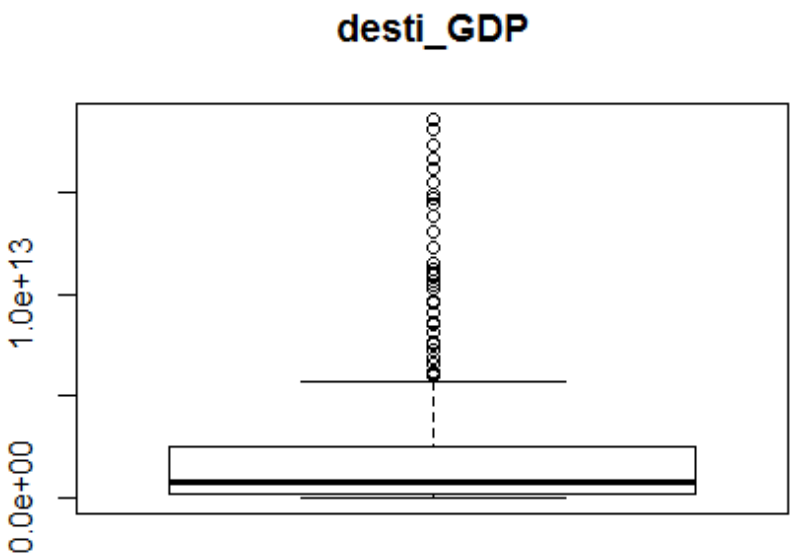
3.2 Identificació i tractament de valors extrems.

Revisem dades d'origen US i ES Revisem els *outliers* de les dades qualitatives mitjançant boxplots:

```
for(i in 1:ncol(mydataMigration)) {
  if (is.numeric(mydataMigration[,i])){
    boxplot(mydataMigration[,i], main = colnames(mydataMigration)[i],
width = 100)
  }
}
```

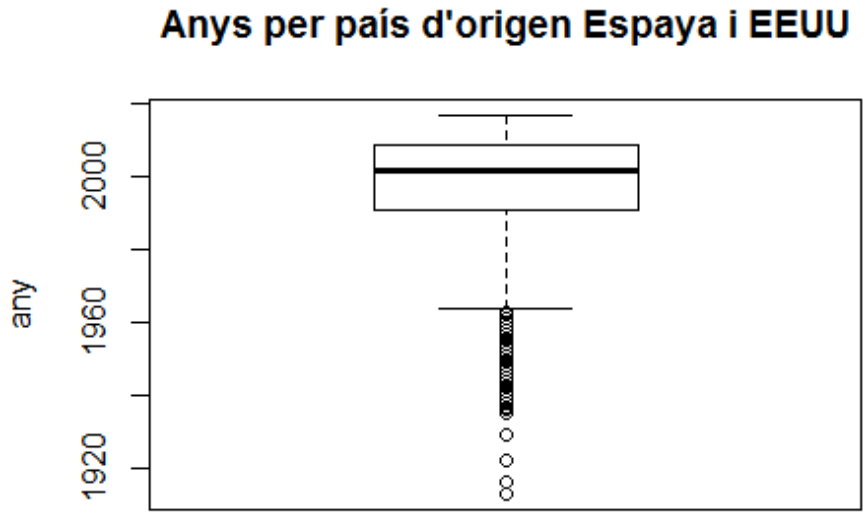

any
2000
1960
1920


origen_GDP
1.0e+13
0.0e+00

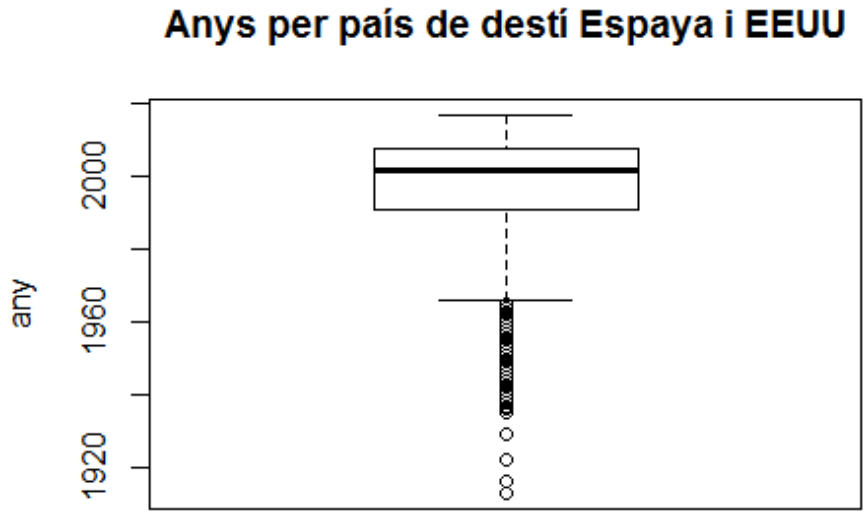


Referent a **GDP** (com país d'origen o destí), ja ens està bé tenir outliers. És ben sabut la desigualtat en el repartiment de la riquesa entre països en el món. D'altra banda, respecte els **anys** trobem outliers per sota del 1971. Anem a veure'n els outliers pels nostres països d'estudi (Espanya i EEUU tant com país d'origen com de destí).

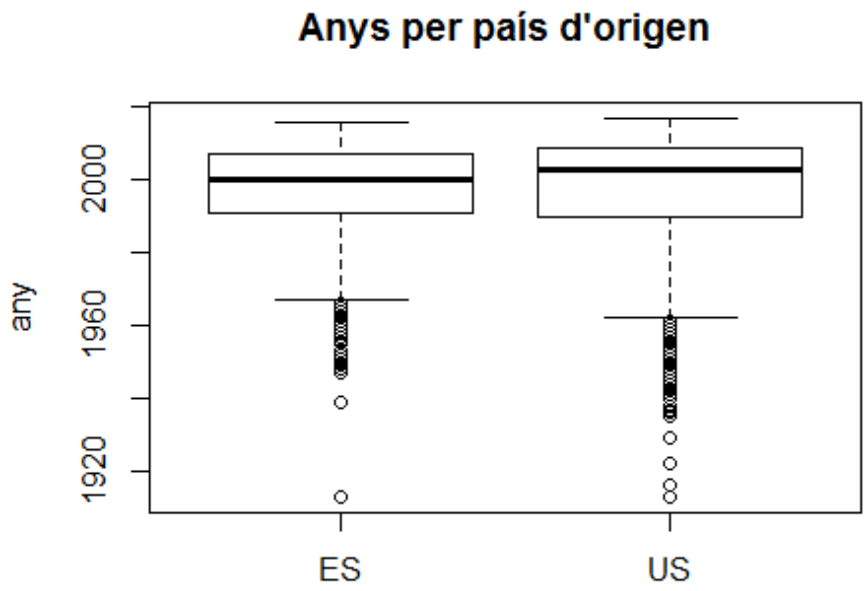
```
OrigenUSAESP<-
mydataMigration[(mydataMigration$origen_country=="US")|(mydataMigration$o
rigen_country=="ES"),]
OrigenUSAESP$origen_country<-as.character(OrigenUSAESP$origen_country)
OrigenUSAESP$origen_country<-as.factor(OrigenUSAESP$origen_country)

#1.Boxplot per ORIGEN=US, mirar per any.

DestiUSAESP<-
mydataMigration[(mydataMigration$desti_country=="US")|(mydataMigration$de
sti_country=="ES"),]
DestiUSAESP$desti_country<-as.character(DestiUSAESP$desti_country)
DestiUSAESP$desti_country<-as.factor(DestiUSAESP$desti_country)

#1.Boxplot per ORIGEN=US, mirar per any.
#par(mfrow=c(2,2))
out1<-boxplot(OrigenUSAESP$any, main="Anys per país d'origen Espaya i
EEUU",ylab="any")$out
```



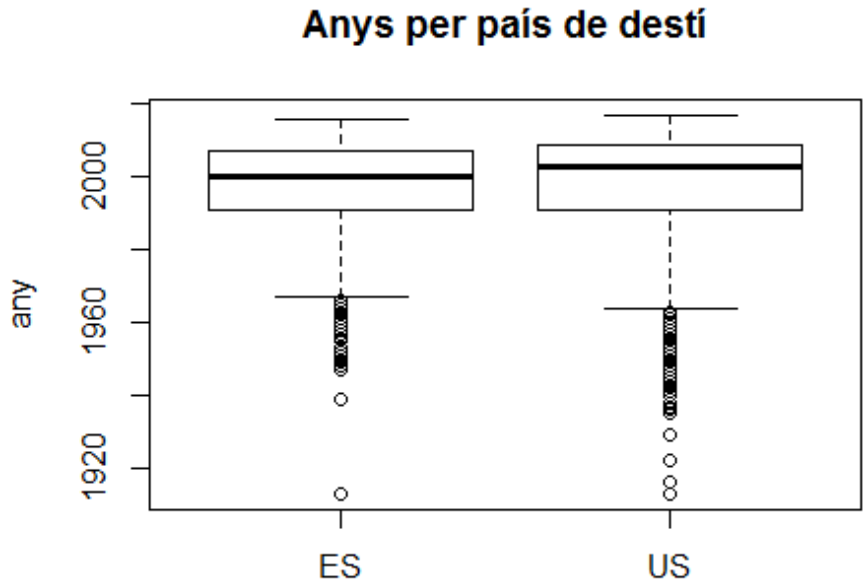

```
out2<-boxplot(DestiUSAESP$any, main="Anys per país de destí Espaya i
EEUU",ylab="any")$out
```



```
out3<-boxplot(OrigenUSAESP$any~OrigenUSAESP$origen_country, main="Anys
per país d'origen",ylab="any")$out
```




```
out4<-boxplot(DestiUSAESP$any~DestiUSAESP$desti_country, main="Anys per
país de destí",ylab="any")$out
```




```
max(out1)
```

```
## [1] 1963
```

```
max(out2)
```

```
## [1] 1965
```

```
max(out3)
```

```
## [1] 1966
```

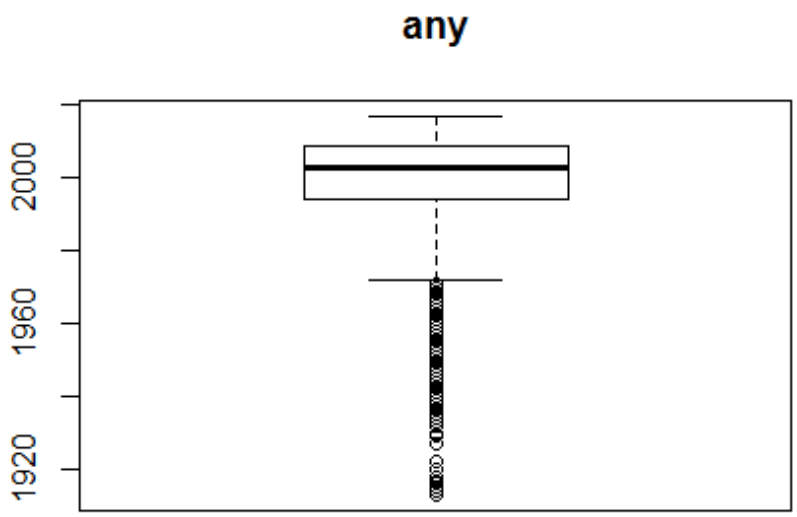
```
max(out4)
```

```
## [1] 1966
```

Veiem que EEUU té dades més antigues que Espanya. Com a curiositat, detecto que en les dades d'Espanya, el valor mínim es troba aïllat i allunyat respecte la resta. La resta de outliers tenen com a punt mínim entorn 1940, que curiosament coincideix amb la fi de la guerra civil.

Hem de tenir en compte que el registre ORCID no es va crear fins l'any 2012, així que les dades *outliers* serà informació no tant completa com les dades actuals. Per aquest motiu optem per eliminar els outliers del nostre estudi, establint així una data de tall. Veient que l'any màxim d'outlier dels nostres països d'estudi és 1966 quan entre tot el dataset se situa a 1971, entenem que 1971 ja és un bon punt de tall, així tractem totes les dades a partir del mateix punt independentment dels països d'estudi:

```
library(ddpcr)

length(boxplot(mydataMigration$any, main = "any")$out)
```




```
## [1] 16147

mydataMigration$any<-as.numeric(mydataMigration$any)
mydataMigration<-mydataMigration[!((mydataMigration$any) %in% outs),]


nrow(mydataMigration)

## [1] 581640
```

Exportem el dataset un cop finalitzada la neteja de dades:

```
my.newfile <- "Migration.csv"
write.csv(mydataMigration, file=my.newfile, row.names = FALSE)
```

4. Anàlisi de les dades.

4.1 Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

1. Comparar fuga de talents a Espanya amb EEUU. Per això necessitem la taula de migrants agrupada per data i país origen i per data i país destí. La construirem a partir de *mydataMigration*

2. Anem a veure si les migracions és a països amb més renta per càpita, comparant les migracions de Espanya i EEUU. Utilitzarem la taula *mydataMigration.*

3. Veure cap a quins països van desde EEUU o Espanya i d'on venen. Hem de tenir taula origen-desti, *mydataMigration.*

A continuació construïm els dos datasets d'informació de migrants acumulats per origen i any:

```
# mydataOrigen:

#Taula agrupada per origen i any, calculant els que han migrat
myvars<-c("origen_country", "any", "has_migrated")
data1<-mydataMigration[myvars]

grouped_data1 <- aggregate(data1, by=list(data1$origen_country,
data1$any, data1$has_migrated), FUN=length);   #notem que ignora
registres NA en origen o any. Que ja ens va bé per el nostre estudi.

grouped_data11<-grouped_data1[which (grouped_data1$Group.3=="True"),]

colnames(grouped_data11)[colnames(grouped_data11)=="origen_country"] <-
"has_migrated_True"
grouped_data11$any<-NULL
grouped_data11$has_migrated<-NULL
grouped_data11$Group.3<-NULL
colnames(grouped_data11)[colnames(grouped_data11)=="Group.1"] <-
"origen_country"
colnames(grouped_data11)[colnames(grouped_data11)=="Group.2"] <- "any"

#Taula agrupada per origen i any, calculant el total
grouped_data2 <- aggregate(data1, by=list(data1$origen_country,
data1$any), FUN=length);

colnames(grouped_data2)[colnames(grouped_data2)=="origen_country"] <-
"TOTAL"
grouped_data2$any<-NULL
grouped_data2$has_migrated<-NULL
colnames(grouped_data2)[colnames(grouped_data2)=="Group.1"] <-
"origen_country"
```

```r
colnames(grouped_data2)[colnames(grouped_data2)=="Group.2"] <- "any"

#Unim les dues taules
grouped_data3<-
merge(grouped_data11,grouped_data2,by=c("origen_country","any"),all.y=TRU
E)
head(grouped_data3)
```

```
##   origen_country  any has_migrated_True TOTAL
## 1             AD 1972                 1     1
## 2             AD 1990                 1     1
## 3             AD 1991                NA     1
## 4             AD 1993                 1     2
## 5             AD 1996                 2     2
## 6             AD 2001                 1     1
```

```r
#i li afegim les dades de phd:
myvars<-c("origen_country", "any", "has_phd")
data1<-mydataMigration[myvars]

grouped_data4 <- aggregate(data1, by=list(data1$origen_country,
data1$any, data1$has_phd), FUN=length);

grouped_data44<-grouped_data4[which (grouped_data4$Group.3=="True"),]

colnames(grouped_data44)[colnames(grouped_data44)=="origen_country"] <-
"has_phd_True"
grouped_data44$any<-NULL
grouped_data44$has_phd<-NULL
grouped_data44$Group.3<-NULL
colnames(grouped_data44)[colnames(grouped_data44)=="Group.1"] <-
"origen_country"
colnames(grouped_data44)[colnames(grouped_data44)=="Group.2"] <- "any"

#I ho unim de nou a la taula anterior
grouped_data5<-
merge(grouped_data3,grouped_data44,by=c("origen_country","any"),all.x=TRU
E)
mydataOrigen<-grouped_data5

#Finalment els valors nulls de has_migrated_True i has_phd_True els
asignem 0:
mydataOrigen$has_migrated_True[is.na(mydataOrigen$has_migrated_True)]<-0
mydataOrigen$has_phd_True[is.na(mydataOrigen$has_phd_True)]<-0
head(mydataOrigen)
```

```
##   origen_country  any has_migrated_True TOTAL has_phd_True
## 1             AD 1972                 1     1            1
## 2             AD 1990                 1     1            0
## 3             AD 1991                 0     1            0
```

```
## 4             AD 1993                1     2            1
## 5             AD 1996                2     2            1
## 6             AD 2001                1     1            1
```

```
#Taula agrupada per destí i any, calculant els que han migrat (anàleg al
anterior)
myvars<-c("desti_country", "any", "has_migrated")
data1<-mydataMigration[myvars]

grouped_data1 <- aggregate(data1, by=list(data1$desti_country, data1$any,
data1$has_migrated), FUN=length);

grouped_data11<-grouped_data1[which (grouped_data1$Group.3=="True"),]

colnames(grouped_data11)[colnames(grouped_data11)=="desti_country"] <-
"has_migrated_True"
grouped_data11$any<-NULL
grouped_data11$has_migrated<-NULL
grouped_data11$Group.3<-NULL
colnames(grouped_data11)[colnames(grouped_data11)=="Group.1"] <-
"desti_country"
colnames(grouped_data11)[colnames(grouped_data11)=="Group.2"] <- "any"

#Taula agrupada per origen i any, calculant el total
grouped_data2 <- aggregate(data1, by=list(data1$desti_country,
data1$any), FUN=length);

colnames(grouped_data2)[colnames(grouped_data2)=="desti_country"] <-
"TOTAL"
grouped_data2$any<-NULL
grouped_data2$has_migrated<-NULL
colnames(grouped_data2)[colnames(grouped_data2)=="Group.1"] <-
"desti_country"
colnames(grouped_data2)[colnames(grouped_data2)=="Group.2"] <- "any"

#Unim les dues taules
grouped_data3<-
merge(grouped_data11,grouped_data2,by=c("desti_country","any"),all.y=TRUE
)
head(grouped_data3)
```

```
##   desti_country  any has_migrated_True TOTAL
## 1            AD 1990                1     1
## 2            AD 1991               NA     1
## 3            AD 1993                1     2
## 4            AD 1996                2     2
## 5            AD 1997               NA     1
## 6            AD 2001                1     1
```

```r
#i li afegim les dades de phd:
myvars<-c("desti_country", "any", "has_phd")
data1<-mydataMigration[myvars]

grouped_data4 <- aggregate(data1, by=list(data1$desti_country, data1$any,
data1$has_phd), FUN=length);

grouped_data44<-grouped_data4[which (grouped_data4$Group.3=="True"),]

colnames(grouped_data44)[colnames(grouped_data44)=="desti_country"] <-
"has_phd_True"
grouped_data44$any<-NULL
grouped_data44$has_phd<-NULL
grouped_data44$Group.3<-NULL
colnames(grouped_data44)[colnames(grouped_data44)=="Group.1"] <-
"desti_country"
colnames(grouped_data44)[colnames(grouped_data44)=="Group.2"] <- "any"

#I ho unim de nou a la taula anterior
grouped_data5<-
merge(grouped_data3,grouped_data44,by=c("desti_country","any"),all.x=TRUE
)
mydataDesti<-grouped_data5

#Finalment els valors nulls de has_migrated_True i has_phd_True els
asignem 0:
mydataDesti$has_migrated_True[is.na(mydataDesti$has_migrated_True)]<-0
mydataDesti$has_phd_True[is.na(mydataDesti$has_phd_True)]<-0
head(mydataDesti)

##   desti_country  any has_migrated_True TOTAL has_phd_True
## 1            AD 1990                1     1            0
## 2            AD 1991                0     1            0
## 3            AD 1993                1     2            1
## 4            AD 1996                2     2            1
## 5            AD 1997                0     1            1
## 6            AD 2001                1     1            1

summary(mydataOrigen)

##  origen_country      any       has_migrated_True     TOTAL
##  AU     :  46   Min.   :1972   Min.   :  0.00    Min.   :   1.00
##  BR     :  46   1st Qu.:1986   1st Qu.:  1.00    1st Qu.:   3.00
##  CA     :  46   Median :1997   Median :  3.00    Median :  11.00
##  CH     :  46   Mean   :1996   Mean   : 15.57    Mean   :  98.02
##  CN     :  46   3rd Qu.:2007   3rd Qu.: 12.00    3rd Qu.:  58.75
##  DE     :  46   Max.   :2017   Max.   :614.00    Max.   :4482.00
##  (Other):5658
##   has_phd_True
##  Min.   :   0.00
```

```
##  1st Qu.:   1.00
##  Median :   5.00
##  Mean   :  48.58
##  3rd Qu.:  29.00
##  Max.   :2569.00
##
```

```
summary(mydataDesti)
```

```
##  desti_country       any       has_migrated_True     TOTAL
##  AU     :  46   Min.   :1972   Min.   :   0.00   Min.   :   1.00
##  BR     :  46   1st Qu.:1986   1st Qu.:   1.00   1st Qu.:   2.00
##  CA     :  46   Median :1997   Median :   2.00   Median :  10.00
##  CH     :  46   Mean   :1996   Mean   :  15.16   Mean   :  95.44
##  CN     :  46   3rd Qu.:2007   3rd Qu.:  10.75   3rd Qu.:  58.00
##  DE     :  46   Max.   :2017   Max.   :1011.00   Max.   :4903.00
##  (Other):5818
##   has_phd_True
##  Min.   :   0.00
##  1st Qu.:   1.00
##  Median :   4.00
##  Mean   :  47.31
##  3rd Qu.:  27.00
##  Max.   :2987.00
##
```

Exportem els dos datasets un cop acumulada la informació:

```
my.newfile <- "PerOrigen.csv"
write.csv(mydataOrigen, file=my.newfile, row.names = FALSE)

my.newfile <- "PerDesti.csv"
write.csv(mydataDesti, file=my.newfile, row.names = FALSE)
```

4.2 Comprovació de la normalitat i homogeneïtat de la variància.

Mitjançant la prova de normalitat *Anderson-Darling* comprovem la normalitat de les dades quantitatives. Si el p-valor és superior a 0,05 ho considerarem com distribució normal:

```
library(nortest)

alpha = 0.05
col.names = colnames(mydataMigration)

for (i in 1:ncol(mydataMigration)) {
  if (i == 1) cat("Variables que no segueixen una distribució normal:\n")
  if (is.integer(mydataMigration[,i]) | is.numeric(mydataMigration[,i]))
{
    p_val = ad.test(mydataMigration[,i])$p.value
    if (p_val < alpha) {
```

```
      cat(col.names[i])
      # Format output
      if (i < ncol(mydataMigration) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}

## Variables que no segueixen una distribució normal:
## any, origen_GDP, desti_GDP
```

Vejem-ho ara acotant les dades per origen EEUU i origen Espanya i anàleg com a destí:

```
US<-mydataOrigen[which(mydataOrigen$origen_country=="US"),]
ES<-mydataOrigen[which(mydataOrigen$origen_country=="ES"),]

alpha = 0.05
col.names = colnames(mydataOrigen)

print('--ORIGEN US---')

## [1] "--ORIGEN US---"

for (i in 1:ncol(US)) {
  if (i == 1) cat("Variables que no segueixen una distribució normal:\n")
  if (is.integer(US[,i]) | is.numeric(US[,i])) {
    p_val = ad.test(US[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(US) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}

## Variables que no segueixen una distribució normal:
## TOTALhas_phd_True

cat("\n")

cat("\n")

print('--ORIGEN ES---')

## [1] "--ORIGEN ES---"

for (i in 1:ncol(ES)) {4
  if (i == 1) cat("Variables que no segueixen una distribució normal:\n")
  if (is.integer(ES[,i]) | is.numeric(ES[,i])) {
    p_val = ad.test(ES[,i])$p.value
    if (p_val < alpha) {
```

```
      cat(col.names[i])
      # Format output
      if (i < ncol(ES) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}

## Variables que no segueixen una distribució normal:
## has_migrated_True,
## has_phd_True

cat("\n")

cat("\n")

print('--DESTI US---')

## [1] "--DESTI US---"

for (i in 1:ncol(US)) {
  if (i == 1) cat("Variables que no segueixen una distribució normal:\n")
  if (is.integer(US[,i]) | is.numeric(US[,i])) {
    p_val = ad.test(US[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(US) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}

## Variables que no segueixen una distribució normal:
## TOTALhas_phd_True

cat("\n")

cat("\n")

print('--DESTI ES---')

## [1] "--DESTI ES---"

for (i in 1:ncol(ES)) {4
  if (i == 1) cat("Variables que no segueixen una distribució normal:\n")
  if (is.integer(ES[,i]) | is.numeric(ES[,i])) {
    p_val = ad.test(ES[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(ES) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
```

```
    }
  }
}

## Variables que no segueixen una distribució normal:
## has_migrated_True,
## has_phd_True
```

Per la homgenietat, apliquem el *test Fligner-Killeen*. Ho aplicarem sobre les variables de migració i phd filtrant per EEUU i Espanya:

```
dataUS<-mydataMigration[mydataMigration$origen_country=="US",]
dataES<-mydataMigration[mydataMigration$origen_country=="ES",]

fligner.test( has_migrated_True~ has_phd_True, data = US)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  has_migrated_True by has_phd_True
## Fligner-Killeen:med chi-squared = 45, df = 44, p-value = 0.4298

fligner.test( has_migrated_True~ has_phd_True, data = ES)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  has_migrated_True by has_phd_True
## Fligner-Killeen:med chi-squared = 44, df = 43, p-value = 0.429
```

p-valor superior a 0,05 en tots dos casos. Acceptem la hipótesis de que les variances de les dues mostres són homogènies.

4.3 Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis,correlacions, regressions, etc.

Mirem si hi ha correlació entre la renta per càpita origen i destí. A més ho mirarem restringint origen US i origen ES.

```
res <- cor(mydataMigration[,c("origen_GDP","desti_GDP")])
round(res, 2)

##            origen_GDP desti_GDP
## origen_GDP       1.00      0.89
## desti_GDP        0.89      1.00

res <-
cor(mydataMigration[mydataMigration$origen_country=="US",c("origen_GDP","
desti_GDP")])
round(res, 2)
```

```
##             origen_GDP desti_GDP
## origen_GDP        1.00      0.87
## desti_GDP         0.87      1.00
```

```
res <-
cor(mydataMigration[mydataMigration$origen_country=="ES",c("origen_GDP","
desti_GDP")])
round(res, 2)
```

```
##             origen_GDP desti_GDP
## origen_GDP        1.00      0.41
## desti_GDP         0.41      1.00
```

Observem que no hi ha correlació. De fet hi ha més relació a EEUU que a Espanya.

---

5. Representació dels resultats a partir de taules i gràfiques.

1. Gràfics on es visualitzi informació de Espanya i EEUU soperposades amb dades de freq.relatives per poder-les comparar, per anys (migració per anys)

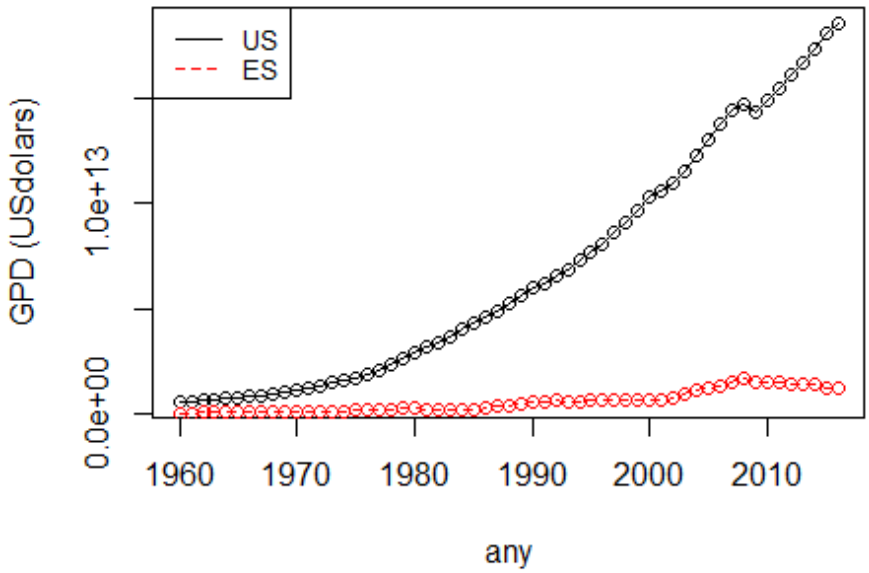
2. Gràfics on es superposi renta per càpita amb migració per país i any.

3. Mapes de destinacions EEUU i destinacions Espanya

4. Mapes d'origen del que tenen com destí EEUU i Espanya

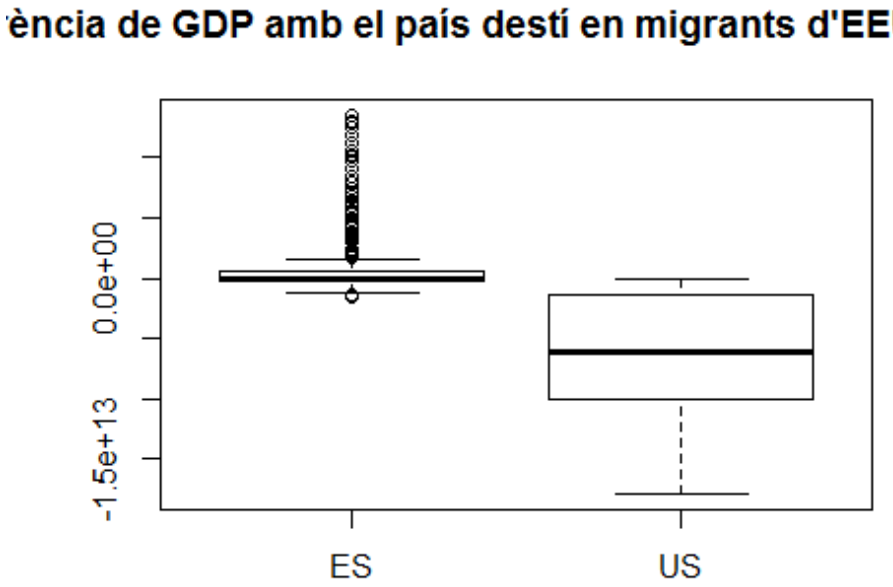
Vegem l'evolució de la renta per càpita de EEUU i Espanya

```
USf0<-mydataGDP$Value[which(mydataGDP$ISO3166.1.Alpha.2=="US")]
USg0<-mydataGDP$Year[which(mydataGDP$ISO3166.1.Alpha.2=="US")]
ESf1<-mydataGDP$Value[which(mydataGDP$ISO3166.1.Alpha.2=="ES")]
ESg1<-mydataGDP$Year[which(mydataGDP$ISO3166.1.Alpha.2=="ES")]

plot(USg0, USf0,xlab="any",ylab="GPD (USdolars)", type = "o")
lines(ESg1, ESf1, type = "o", lty = 2, col = "red")
legend("topleft",legend=c("US", "ES"),
       col=c("black", "red"), lty=1:2, cex=0.8)
```

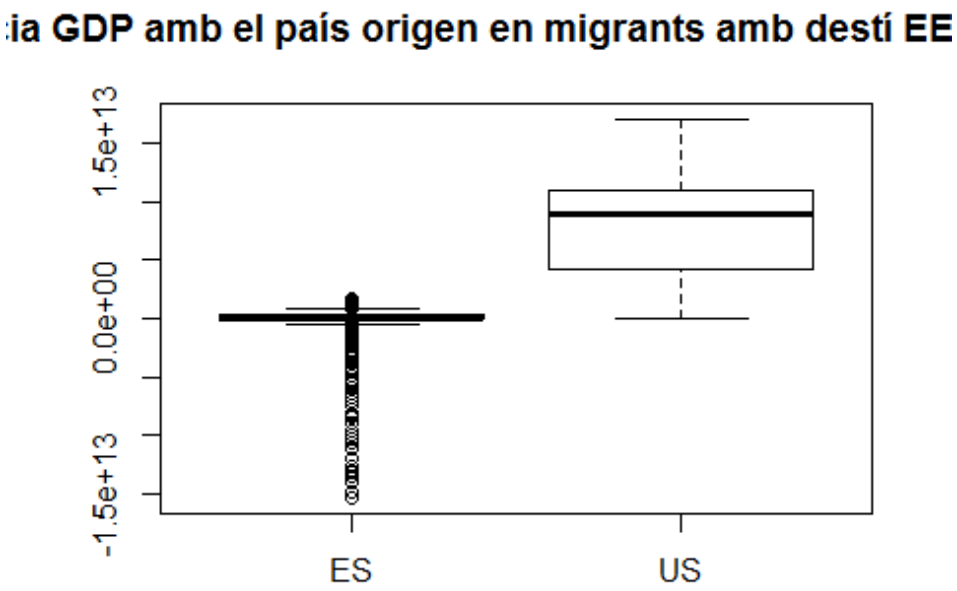


El següent boxplot ens mostra com migrants d'origen Espanya, el seu destí tendeix a tenir una renta per càpita similar que Espanya. En canvi per EEUU, la meitat dels casos el destí té una renta per càpita inferior. Si el que mirem és l'origen dels migrants amb destí Espanya o EEUU, a Espanya venen de rentes per càpita similars a Espanya, en canvi a EEUU en vénen d'inferiors:

```
mydataMigration$has_migrated<-
as.character((mydataMigration$has_migrated))
mydataMigration$has_migrated[mydataMigration$has_migrated==1]<-"True"
mydataMigration$has_migrated[mydataMigration$has_migrated==0]<-"False"
#Seleccionem els migrants amb origen EEUU o Espanya:
Migrate<-mydataMigration[mydataMigration$has_migrated=="True",]
MigrateOrigESUS<-
Migrate[Migrate$origen_country=="US"|Migrate$origen_country=="ES",]
MigrateOrigESUS$origen_country<-
as.character(MigrateOrigESUS$origen_country)
MigrateOrigESUS$origen_country<-as.factor(MigrateOrigESUS$origen_country)

#Vegem si GDP destí és superior que el GDP d'origen
boxplot((MigrateOrigESUS$desti_GDP-
MigrateOrigESUS$origen_GDP)~MigrateOrigESUS$origen_country,
main="Diferència de GDP amb el país destí en migrants d'EEUU/Espanya")
```

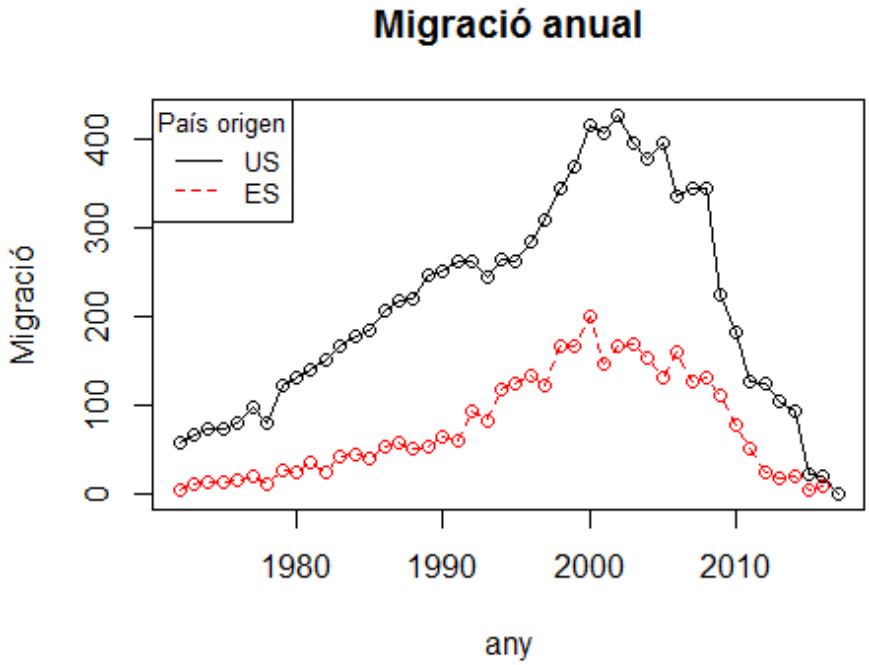



```
#Anàleg, per migrants amb destí EEUU o Espanya:
MigrateDestiESUS<-
Migrate[Migrate$desti_country=="US"|Migrate$desti_country=="ES",]
MigrateDestiESUS$desti_country<-
as.character(MigrateDestiESUS$desti_country)
MigrateDestiESUS$desti_country<-as.factor(MigrateDestiESUS$desti_country)

#Vegem si GDP destí és superior que el GDP d'origen
boxplot((MigrateDestiESUS$desti_GDP-
MigrateDestiESUS$origen_GDP)~MigrateDestiESUS$desti_country,
main="Diferència GDP amb el país origen en migrants amb destí
EEUU/Espanya")
```

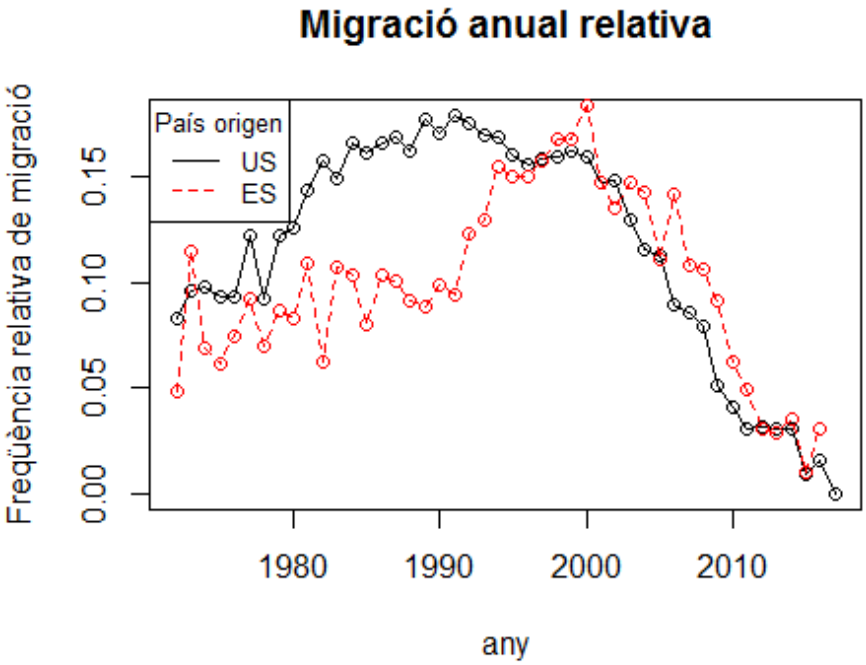


Anem a veure quantitat de migració entre els dos països d'estudi, per anys:

```
OrigenMigrated_US<-
mydataOrigen$has_migrated_True[which(mydataOrigen$origen_country=="US")]
OrigenMigratedAny_US<-
mydataOrigen$any[which(mydataOrigen$origen_country=="US")]
OrigenMigrated_ES<-
mydataOrigen$has_migrated_True[which(mydataOrigen$origen_country=="ES")]
OrigenMigratedAny_ES<-
mydataOrigen$any[which(mydataOrigen$origen_country=="ES")]

plot(OrigenMigratedAny_US, OrigenMigrated_US,xlab="any",ylab="Migració",
type = "o",main="Migració anual")
lines(OrigenMigratedAny_ES, OrigenMigrated_ES, type = "o", lty = 2, col =
"red")
legend("topleft",legend=c("US", "ES"),
       col=c("black", "red"), lty=1:2, cex=0.8,title="País origen")
```



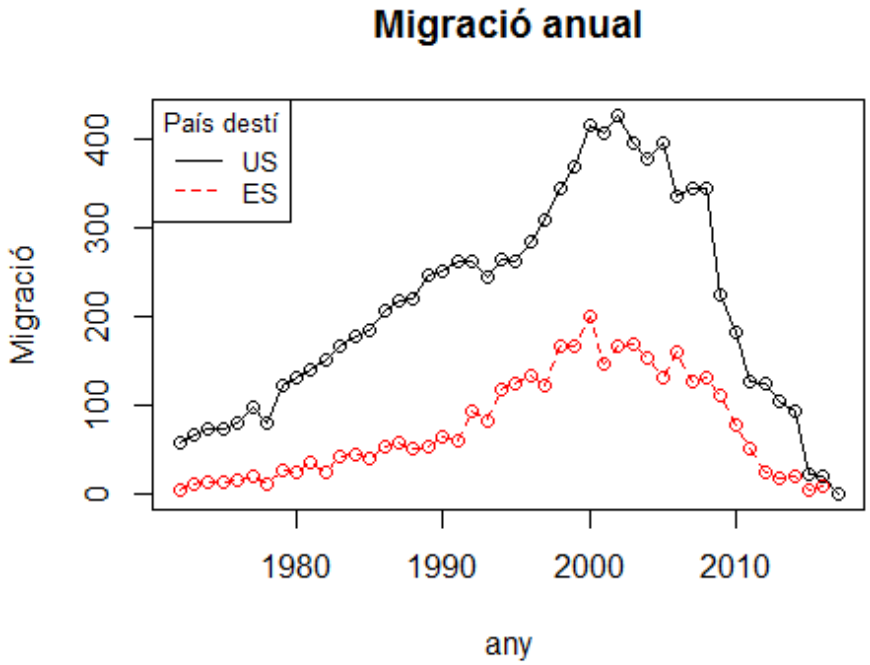

```
OrigenMigrated_US2<-
OrigenMigrated_US/mydataOrigen$TOTAL[which(mydataOrigen$origen_country=="
US")]
OrigenMigrated_ES2<-
OrigenMigrated_ES/mydataOrigen$TOTAL[which(mydataOrigen$origen_country=="
ES")]

plot(OrigenMigratedAny_US, OrigenMigrated_US2,xlab="any",ylab="Freqüència
relativa de migració", type = "o",main="Migració anual relativa")
lines(OrigenMigratedAny_ES, OrigenMigrated_ES2, type = "o", lty = 2, col
= "red")
legend("topleft",legend=c("US", "ES"),
       col=c("black", "red"), lty=1:2, cex=0.8,title="País origen")
```



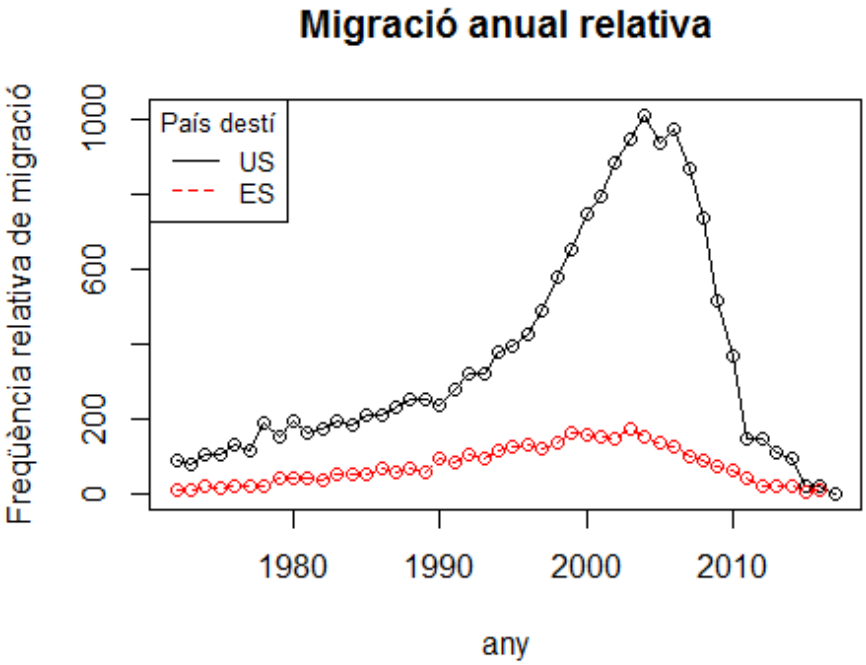
Observem en el primer gràfic com la migració segueix una corba semblant entre els dos països. El segon gràfic mostra la freqüència relativa de migrants sobre el total d'investigadors registrats. Observem com fins finals dels 90, hi havia més migració a EEUU. A finals dels 90 Espanya té més migració però és molt semblant i amb la mateixa fluctuació que a EEUU.

Vegem com es comporta com a països destins:

```
DestiMigrated_US<-
mydataDesti$has_migrated_True[which(mydataDesti$desti_country=="US")]
DestiMigratedAny_US<-
mydataDesti$any[which(mydataDesti$desti_country=="US")]
DestiMigrated_ES<-
mydataDesti$has_migrated_True[which(mydataDesti$desti_country=="ES")]
DestiMigratedAny_ES<-
mydataDesti$any[which(mydataDesti$desti_country=="ES")]

plot(OrigenMigratedAny_US, OrigenMigrated_US,xlab="any",ylab="Migració",
type = "o",main="Migració anual")
lines(OrigenMigratedAny_ES, OrigenMigrated_ES, type = "o", lty = 2, col =
"red")
legend("topleft",legend=c("US", "ES"),
       col=c("black", "red"), lty=1:2, cex=0.8,title="País destí")
```



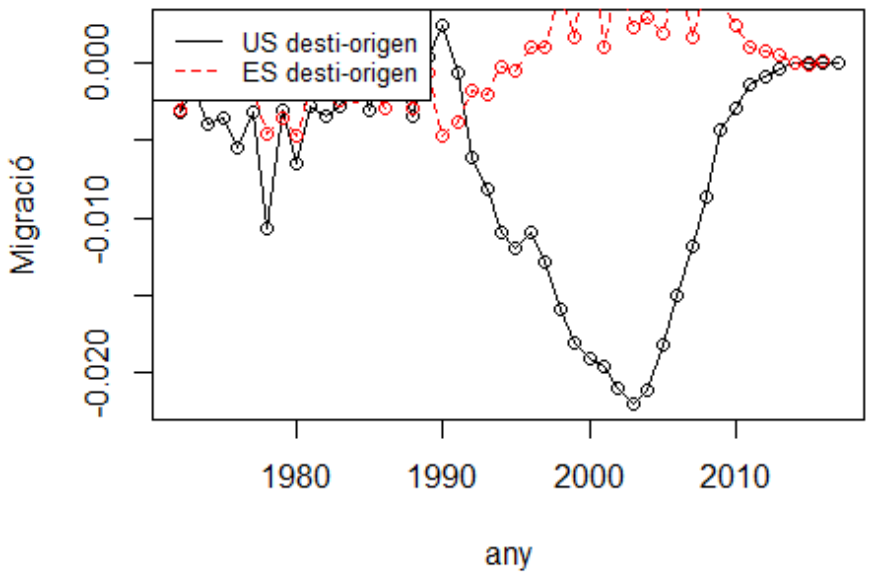

```
DestiMigrated_US2<-
OrigenMigrated_US/mydataOrigen$TOTAL[which(mydataOrigen$desti_country=="U
S")]
DestiMigrated_ES2<-
OrigenMigrated_ES/mydataOrigen$TOTAL[which(mydataOrigen$desti_country=="E
S")]

plot(DestiMigratedAny_US, DestiMigrated_US,xlab="any",ylab="Freqüència
relativa de migració", type = "o",main="Migració anual relativa")
lines(DestiMigratedAny_ES, DestiMigrated_ES, type = "o", lty = 2, col =
"red")
legend("topleft",legend=c("US", "ES"),
       col=c("black", "red"), lty=1:2, cex=0.8,title="País destí")
```



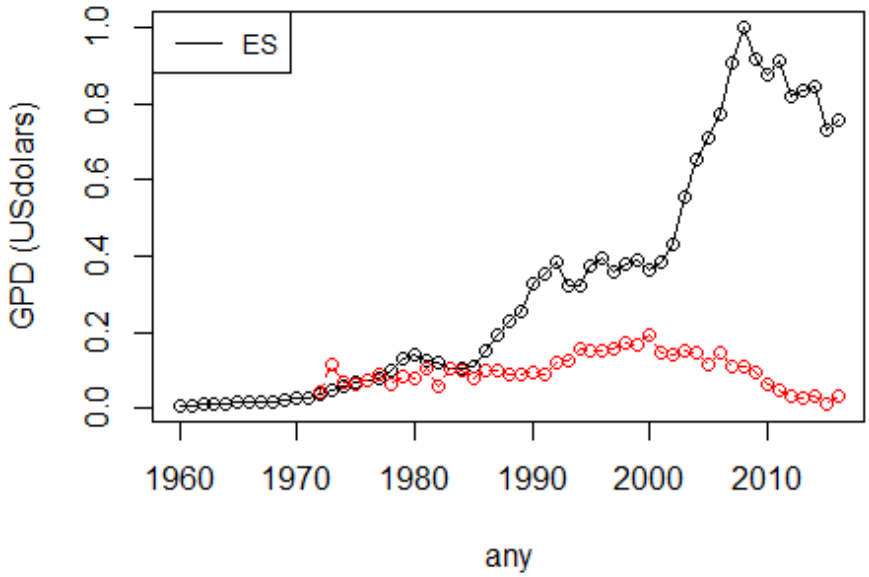
També veiem una corba semblant entre països, però mirant el segon gràfic, veiem com EEUU és un destí clarament preferit, tot i que en els últims anys hi ha hagut una davallada important que ha fet igualar amb les dades d'Espanya.

També podem donar un cop d'ull a la diferència entre la emmigració i immigració:

```
USf4<-
OrigenMigrated_US/mydataOrigen$TOTAL[which(mydataOrigen$origen_country=="
US")]
USg4<-mydataOrigen$any[which(mydataOrigen$origen_country=="US")]
ESf5<-
OrigenMigrated_ES/mydataOrigen$TOTAL[which(mydataOrigen$origen_country=="
ES")]
ESg5<-mydataOrigen$any[which(mydataOrigen$origen_country=="ES")]
USff4<-
OrigenMigrated_US/mydataDesti$TOTAL[which(mydataDesti$desti_country=="US"
)]
USgg4<-mydataDesti$any[which(mydataDesti$desti_country=="US")]
ESff5<-
OrigenMigrated_ES/mydataDesti$TOTAL[which(mydataDesti$desti_country=="ES"
)]
ESgg5<-mydataDesti$any[which(mydataDesti$desti_country=="ES")]

plot(USg4, (USff4-USf4),xlab="any",ylab="Migració", type = "o")
lines(ESg5, (ESff5-ESf5), type = "o", lty = 2, col = "red")
legend("topleft",legend=c("US desti-origen", "ES desti-origen"),
       col=c("black", "red"), lty=1:2, cex=0.8)
```




```
plot(ESg1, ESf1/max(ESf1),xlab="any",ylab="GPD (USdolars)", type = "o")
lines(ESgg5, ESff5, type = "o", lty = 2, col = "red")
legend("topleft",legend=c("ES"),
       col=c("black"), lty=1:2, cex=0.8)
```

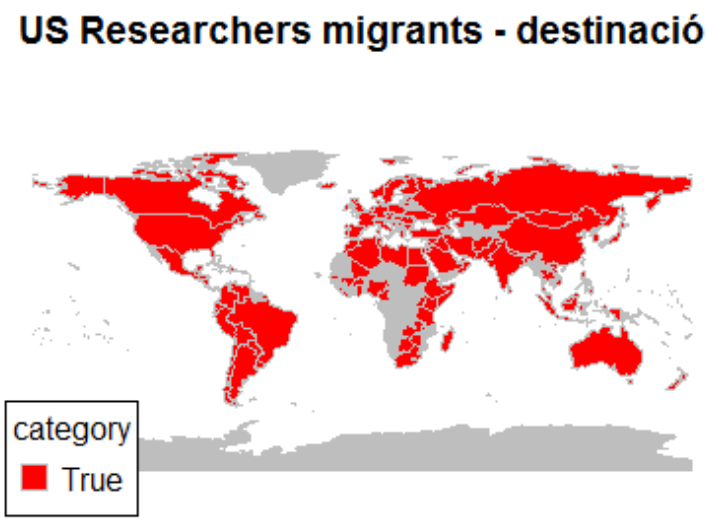


Veiem que a EEUU hi havia una època on marxava més gent que no pas es quedava. En canvi a Espanya en aquella època inclús va augmentar una mica la immigració. Això és entre els anys 90 i 2010 aprox (abans de la crisi). La tendència sembla ser a equilibrar-se de nou, amb indexos semblants entre EEUU i Espanya. Sembla que a Espanya hi ha més gent que ve que no pas que marxa. Però la gent que ve prové d'una renta per càpita inferior, mentres que la que marxa va a països de renta per càpita superior. Pot ser que aquí ens considerem mal pagats i per això marxen, tot i no superar la quantitat dels que venen?

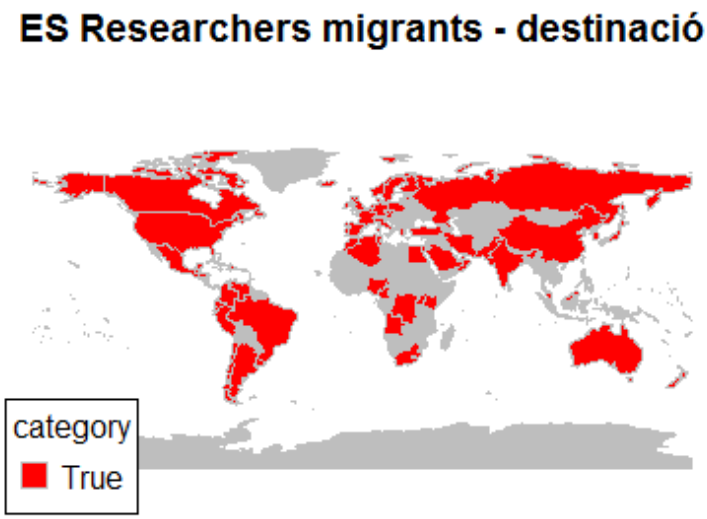
Per últim vegem les destinacions dels investigadors. Per visualitzar-ho ens ajudarem d'un mapa mundi:

```
library(rworldmap)


origen<-mydataMigration[mydataMigration$origen_country=="US" &
mydataMigration$has_migrated=="True",]

mapa=joinCountryData2Map(origen, joinCode="ISO2",
nameJoinColumn="desti_country")

## 9686 codes from your data successfully matched countries in the map
## 1 codes from your data failed to match with a country code in the map
## 96 codes from the map weren't represented in your data

mapCountryData(mapa,catMethod="fixedWidth",nameColumnToPlot="has_migrated
",
```

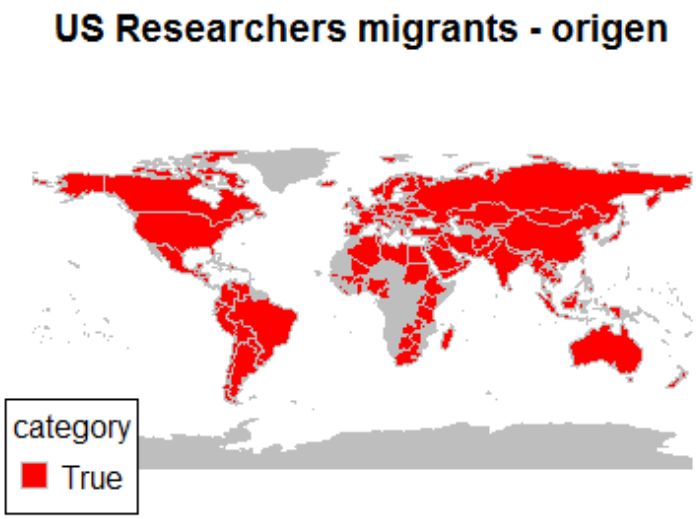
```
            mapTitle="US Researchers migrants -
destinació",missingCountryCol='grey', numCats=9)

## using catMethod='categorical' for non numeric data in mapCountryData
```



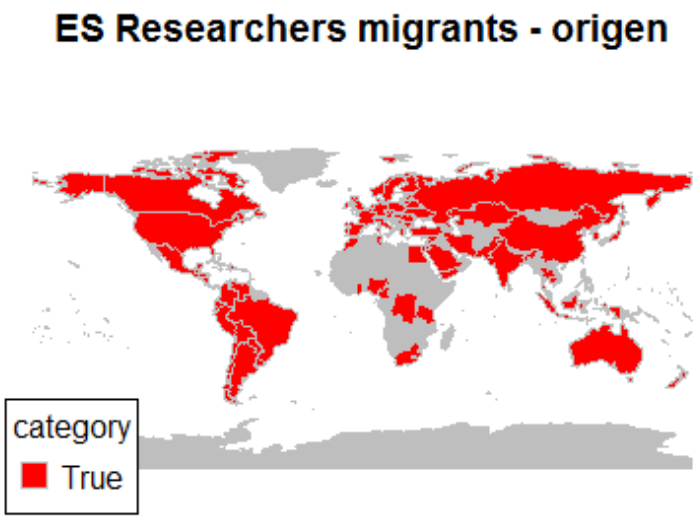

```
origen<-mydataMigration[mydataMigration$origen_country=="ES" &
mydataMigration$has_migrated=="True",]

mapa=joinCountryData2Map(origen, joinCode="ISO2",
nameJoinColumn="desti_country")

## 3367 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 163 codes from the map weren't represented in your data

mapCountryData(mapa,catMethod="fixedWidth",nameColumnToPlot="has_migrated
",
            mapTitle="ES Researchers migrants -
destinació",missingCountryCol='grey', numCats=9)

## using catMethod='categorical' for non numeric data in mapCountryData
```




```
desti<-mydataMigration[mydataMigration$desti_country=="US" &
mydataMigration$has_migrated=="True",]

mapa=joinCountryData2Map(desti, joinCode="ISO2",
nameJoinColumn="origen_country")

## 16400 codes from your data successfully matched countries in the map
## 3 codes from your data failed to match with a country code in the map
## 96 codes from the map weren't represented in your data

mapCountryData(mapa,catMethod="fixedWidth",nameColumnToPlot="has_migrated
",
            mapTitle="US Researchers migrants -
origen",missingCountryCol='grey', numCats=9)

## using catMethod='categorical' for non numeric data in mapCountryData
```

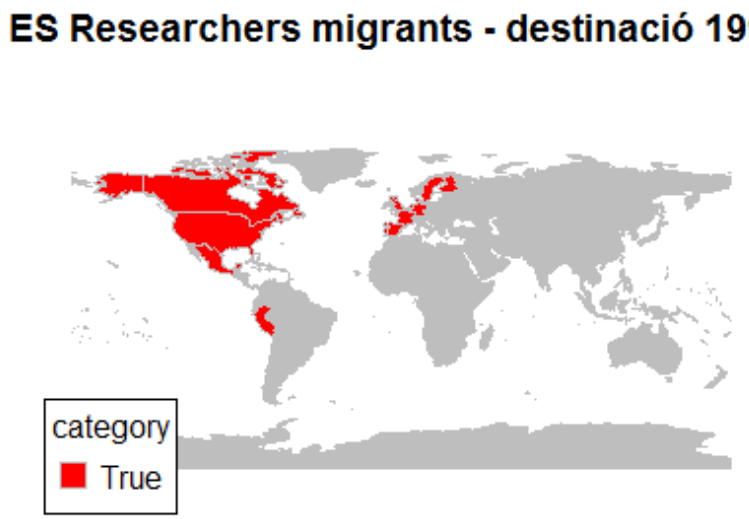



```
desti<-mydataMigration[mydataMigration$desti_country=="ES" &
mydataMigration$has_migrated=="True",]

mapa=joinCountryData2Map(desti, joinCode="ISO2",
nameJoinColumn="origen_country")

## 3334 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 149 codes from the map weren't represented in your data

mapCountryData(mapa,catMethod="fixedWidth",nameColumnToPlot="has_migrated
",
            mapTitle="ES Researchers migrants -
origen",missingCountryCol='grey', numCats=9)

## using catMethod='categorical' for non numeric data in mapCountryData
```

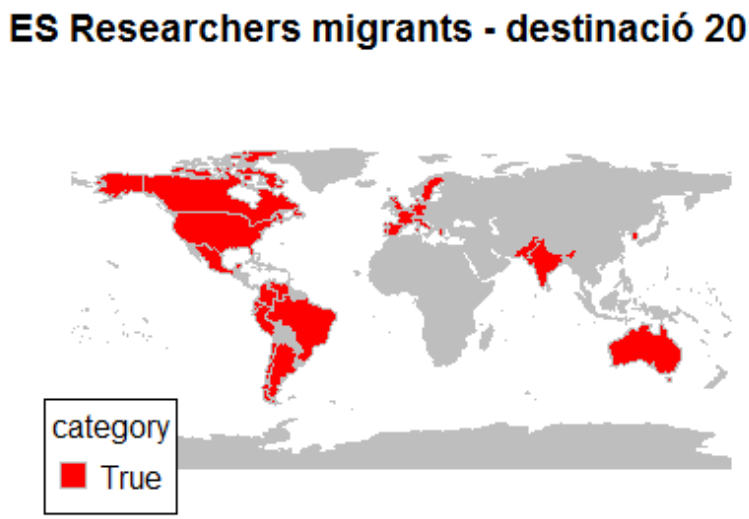


Ara ens centrarem a Espanya, compararem dos anys (1990 i 2010) per veure les destinacions dels espanyols i la procedència de immigrants investigadors:

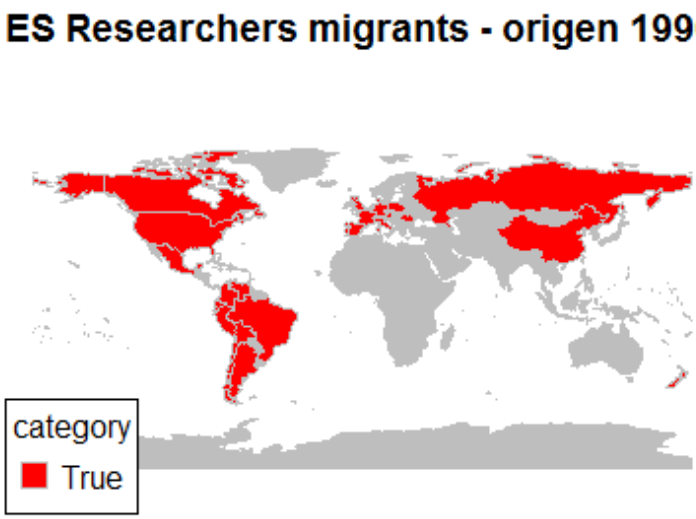
```
origen<-mydataMigration[mydataMigration$origen_country=="ES" &
mydataMigration$has_migrated=="True"&(mydataMigration$any=="1990"),]

mapa=joinCountryData2Map(origen, joinCode="ISO2",
nameJoinColumn="desti_country")

## 64 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 229 codes from the map weren't represented in your data

mapCountryData(mapa,catMethod="fixedWidth",nameColumnToPlot="has_migrated
",
            mapTitle="ES Researchers migrants - destinació
1990",missingCountryCol='grey', numCats=9)

## using catMethod='categorical' for non numeric data in mapCountryData
```



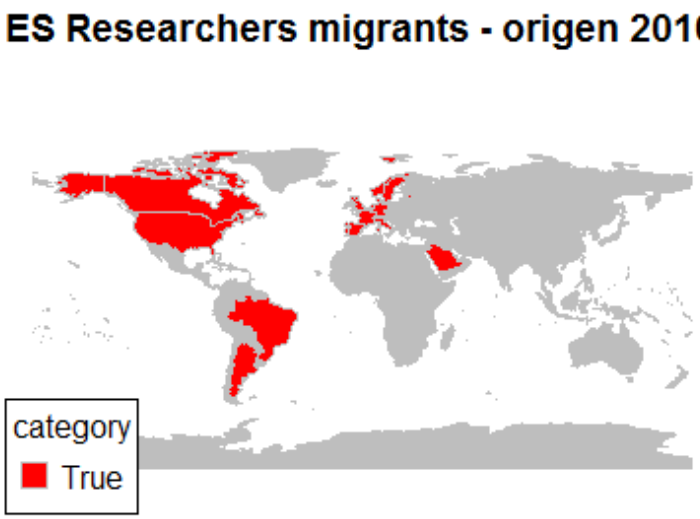

```
origen<-mydataMigration[mydataMigration$origen_country=="ES" &
mydataMigration$has_migrated=="True"&(mydataMigration$any=="2010"),]

mapa=joinCountryData2Map(origen, joinCode="ISO2",
nameJoinColumn="desti_country")

## 78 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 216 codes from the map weren't represented in your data

mapCountryData(mapa,catMethod="fixedWidth",nameColumnToPlot="has_migrated
",
            mapTitle="ES Researchers migrants - destinació
2010",missingCountryCol='grey', numCats=9)

## using catMethod='categorical' for non numeric data in mapCountryData
```




```
desti<-mydataMigration[mydataMigration$desti_country=="ES" &
mydataMigration$has_migrated=="True"&(mydataMigration$any=="1990"),]

mapa=joinCountryData2Map(desti, joinCode="ISO2",
nameJoinColumn="origen_country")

## 92 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 217 codes from the map weren't represented in your data

mapCountryData(mapa,catMethod="fixedWidth",nameColumnToPlot="has_migrated
",
            mapTitle="ES Researchers migrants - origen
1990",missingCountryCol='grey', numCats=9)

## using catMethod='categorical' for non numeric data in mapCountryData
```




```
desti<-mydataMigration[mydataMigration$desti_country=="ES" &
mydataMigration$has_migrated=="True"&(mydataMigration$any=="2010"),]

mapa=joinCountryData2Map(desti, joinCode="ISO2",
nameJoinColumn="origen_country")

## 61 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 226 codes from the map weren't represented in your data

mapCountryData(mapa,catMethod="fixedWidth",nameColumnToPlot="has_migrated
",
            mapTitle="ES Researchers migrants - origen
2010",missingCountryCol='grey', numCats=9)

## using catMethod='categorical' for non numeric data in mapCountryData
```



---

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Després d'aquest estudi mirem de respondre les qüestions plantejades a l'inici

- Com està Espanya en comparativa amb altres països de fuga de talents? Comparació respecte EEUU.

Amb Espanya o EEUU com país d'origen, en nombres relatius podem dir que tots dos països es comporten amb un index similar al llarg dels anys. Als anys 90 hi havia més migració a l'exterior a EEUU de forma diferenciada amb Espanya, però per l'any 2000 fins ara la migració és molt similar i segueix la mateixa tendència entre ells al llarg d'aquest periode. Així que el mite que a Espanya marxen els talents, podem confirmar que marxa amb la mateixa relació que ho fan d'EEUU. En canvi, pel que fa la immigració el comportament tot i seguir una mateixa corba, a EEUU els índexos són molt més elevats que a Espanya. Al 2005 és on la diferència és més gran però cap al 2010 sorprenentment s'igualen els índexos dels dos països.

Es podria potser torbar les causes en els moments que viu cada país. Però aquests resultats no podem saber la fiabilitat degut a que no coneixem el context exacte de l'origen de les dades. Ens manca informació de la fiabilitat de les dades, de si la

informació dels anys és més completa passats uns anys que els recents (per exemple, podria ser que d'aquí 5 anys, la informació de l'any 2016 sigui superior a les dades que hi ha ara, perquè s'introdueixen durant els 5 anys següents). Això podria explicar la corba que observem en el gràfic i com disminueixen les dades més recents.

- Hi ha una relació de la migració amb la renta per càpita del país? (com origen o com destí)

Els migrants d'origen Espanya, el seu destí tendeix a tenir una renta per càpita similar que Espanya o superior. En canvi per EEUU, la meitat dels casos el destí té una renta per càpita inferior i hi ha una variació més gran de GDP.

Si el que mirem és l'origen dels migrants amb destí Espanya o EEUU, a Espanya venen de rentes per càpita similars a Espanya, en canvi a EEUU en vénen d'inferiors.

- A quins països van els investigadors i de quins països venen a Espanya? I als EEUU?

Amb els mapes podem veure les diferents procedències i destinacions. Els últims mapes on visualitzem les destinacions pels anys 1990 i 2010 es veu com hi ha nous països que s'afegeixen i d'altres que deixen de tenir fluxe de professionals, però n'hi ha que es mantenen en els dos mapes. El que sí que es visualitza en els mapes que comprèn les dades de tots els anys (a partir del 1971) que el fluxe d'investigadors és per quasi tots els països del món. Hi ha una gran àrea de movilització entre científics.

---

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

En el repositori adjunto el codi en R (dcampilloca_TCD_PRAC2.Rmd)