PRA2: Com realitzar la neteja i l'anàlisi de dades?

Àlex Franco Granell; Roger Esteban Fabró

Gener 2023

1. Descripció del dataset

Hem escollit investigar quins factors sociodemogràfics tenen un major impacte sobre la incidència de càncer als Estats Units. Per fer aquest anàlisi s'ha escollit tres datasets amb dades dels Estats Units que mostren la mortalitat per comtat, i diversos paràmetres demogràfics. Concretament utilitzem els datasets de "Cancer Mortality & Incidence Rates: (Country LVL)", les dades dels EUA de "Demographics & observation for pandemic escalation", i un dataset d'usafacts.org que recull la població dels diversos comtats del país. Els dos primers es troben disponibles a kaggle i contenen les dades bàsiques per als anàlisis fets posteriorment.

2. Integració i selecció de dades

A continuació, carreguem les dades originals i mostrem els primers registres per pantalla.

```
dpoblacio <- read.csv('./CSVs Originals/covid_county_population_usafacts.csv', sep=',')
head(dpoblacio)
```

```
##   ï..countyFIPS           County.Name State population
## 1             0 Statewide Unallocated    AL          0
## 2          1001        Autauga County    AL      55869
## 3          1003        Baldwin County    AL     223234
## 4          1005        Barbour County    AL      24686
## 5          1007           Bibb County    AL      22394
## 6          1009         Blount County    AL      57826
```

```
dcancerdeath <- read.csv('./CSVs Originals/death.csv', sep=',')
head(dcancerdeath)
```

```
##   index                     County  FIPS Met.Objective.of.45.5...1.
## 1     0              United States     0                         No
## 2     1      Perry County, Kentucky 21193                         No
## 3     2     Powell County, Kentucky 21197                         No
## 4     3 North Slope Borough, Alaska  2185                         No
## 5     4     Owsley County, Kentucky 21189                         No
## 6     5       Union County, Florida 12125                         No
##   Age.Adjusted.Death.Rate Lower.95..Confidence.Interval.for.Death.Rate
## 1                      46                                         45.9
## 2                   125.6                                        108.9
```

```
## 3                   125.3                                       100.2
## 4                   124.9                                          73
## 5                   118.5                                        83.1
## 6                   113.5                                        89.9
##   Upper.95..Confidence.Interval.for.Death.Rate Average.Deaths.per.Year
## 1                                         46.1                157,376
## 2                                        144.2                     43
## 3                                        155.1                     18
## 4                                        194.7                      5
## 5                                        165.5                      8
## 6                                        141.4                     19
##   Recent.Trend..2. Recent.5.Year.Trend..2..in.Death.Rates
## 1          falling                                  -2.4
## 2           stable                                  -0.6
## 3           stable                                   1.7
## 4               **                                    **
## 5           stable                                   2.2
## 6          falling                                  -2.2
##   Lower.95..Confidence.Interval.for.Trend
## 1                                   -2.6
## 2                                   -2.7
## 3                                      0
## 4                                     **
## 5                                   -0.4
## 6                                   -4.3
##   Upper.95..Confidence.Interval.for.Trend
## 1                                   -2.2
## 2                                    1.6
## 3                                    3.4
## 4                                     **
## 5                                    4.8
## 6                                      0
```

```
dusparam <- read.csv('./CSVs Originals/us-county.csv', sep=',')
head(dusparam)
```

```
##   fips   state  county Confirmed Deaths  Smokers Obesity Food.Environment.index
## 1 1001 Alabama Autauga        19      1 18.08156    33.3                    7.2
## 2 1003 Alabama Baldwin        78      1 17.48903    31.0                    8.0
## 3 1005 Alabama Barbour        10      0 21.99998    41.7                    5.6
## 4 1007 Alabama    Bibb        17      0 19.11420    37.6                    7.8
## 5 1009 Alabama  Blount        15      0 19.20867    33.8                    8.4
## 6 1011 Alabama Bullock         6      0 22.89466    37.2                    4.3
##    Exercise overcrowding Diabetics Insufficient.Sleep Traffic.Volume
## 1 69.130124    1.2019231      11.1           35.90541      88.457040
## 2 73.713549    1.2707918      10.7           33.30587      86.997430
## 3 53.166770    1.6885965      17.6           38.56317     102.291762
## 4 16.251364    0.2553191      14.5           38.14887      29.335580
## 5 15.634486    1.8913676      17.0           35.94501      33.411782
## 6  2.501374    0.1125176      23.7           45.02064       4.066538
##   X65..Above.Population Rural.Population
## 1              15.56267         42.00216
## 2              20.44335         42.27910
## 3              19.42044         67.78963
```

```
## 4              16.47321         68.35261
## 5              18.23651         89.95150
## 6              16.38390         51.37438
```

A continuació modifiquem individualment els datasets per a la seva posterior integració.

```
d1 <- dcancerdeath %>%
  # Eliminem les dades a nivell estatal
  filter(index != 0) %>%
  # Generem les columnes County i State
  separate(County, c("county","state"), sep=", ") %>%
  # Seleccionem i modifiquem les variables d'interès
  transmute(FIPS,
            county = str_remove_all(county, " County"),
            state,
            met_obj_reduction = Met.Objective.of.45.5...1.,
            age_adj_deathrate = Age.Adjusted.Death.Rate,
            avg_deaths_year = Average.Deaths.per.Year,
            trend_recent_deaths = Recent.Trend..2.,
            trend_5y_deaths = Recent.5.Year.Trend..2..in.Death.Rates #,
            )

d2 <- dpoblacio %>%
  # Seleccionem les columnes de l'ID del county i la població
  transmute(FIPS = ï..countyFIPS,
            population) %>%
  # Eliminem els registres a nivell estatal
  filter(FIPS != 0)

d3 <- dusparam %>%
  transmute(FIPS = fips,
            smokers = Smokers,
            obesity = Obesity,
            food_env_index = Food.Environment.index,
            exercise = Exercise,
            overcrowding = overcrowding,
            diabetics = Diabetics,
            insuf_sleep = Insufficient.Sleep,
            traffic_vol = Traffic.Volume,
            above_65 = X65..Above.Population,
            rural_pop = Rural.Population)

uscancer <- left_join(d1, d2, by="FIPS") %>%
  left_join(d3, by="FIPS")

head(uscancer)
```

```
##    FIPS              county    state met_obj_reduction age_adj_deathrate
## 1 21193               Perry Kentucky                No             125.6
## 2 21197              Powell Kentucky                No             125.3
## 3  2185 North Slope Borough   Alaska                No             124.9
## 4 21189              Owsley Kentucky                No             118.5
## 5 12125               Union  Florida                No             113.5
```

```
## 6 21147            McCreary Kentucky                No             111.1
##   avg_deaths_year trend_recent_deaths trend_5y_deaths population  smokers
## 1              43              stable            -0.6      25758 24.53343
## 2              18              stable             1.7      12359 23.91649
## 3               5                  **              **       9832 23.85109
## 4               8              stable             2.2       4415 26.14039
## 5              19             falling            -2.2      15237 23.26414
## 6              22              rising            22.9      17231 31.87770
##   obesity food_env_index  exercise overcrowding diabetics insuf_sleep
## 1    41.0            7.3  78.84508     3.295711      16.0    40.85610
## 2    32.6            7.5  86.37120     2.815433      17.6    39.07086
## 3    40.1            6.9 100.00000    29.950495       7.6    33.56203
## 4    46.3            6.9  86.68770     2.380952      11.4    39.80378
## 5    36.9            6.5  21.62214     2.313625      13.8    38.70558
## 6    39.8            6.6 100.00000     1.680000      14.3    43.11113
##   traffic_vol  above_65 rural_pop
## 1   61.970027 16.974552  74.07356
## 2   36.796632 15.873654  67.10537
## 3   10.141895  6.817261  59.33192
## 4    1.385804 19.610912 100.00000
## 5   11.574556 14.973226  67.44770
## 6   24.694476 16.044347 100.00000
```

3. Neteja de dades

3.1. Gestió de valors perduts

En primer lloc, explorem les dades mitjançant `str`.

```
str(uscancer)
```

```
## 'data.frame':    3140 obs. of  19 variables:
##  $ FIPS               : int  21193 21197 2185 21189 12125 21147 21131 21159 21165 21109 ...
##  $ county             : chr  "Perry" "Powell" "North Slope Borough" "Owsley" ...
##  $ state              : chr  "Kentucky" "Kentucky" "Alaska" "Kentucky" ...
##  $ met_obj_reduction  : chr  "No" "No" "No" "No" ...
##  $ age_adj_deathrate  : chr  "125.6" "125.3" "124.9" "118.5" ...
##  $ avg_deaths_year    : chr  "43" "18" "5" "8" ...
##  $ trend_recent_deaths: chr  "stable" "stable" "**" "stable" ...
##  $ trend_5y_deaths    : chr  "-0.6" "1.7" "**" "2.2" ...
##  $ population         : int  25758 12359 9832 4415 15237 17231 9877 11195 6489 13329 ...
##  $ smokers            : num  24.5 23.9 23.9 26.1 23.3 ...
##  $ obesity            : num  41 32.6 40.1 46.3 36.9 39.8 41.4 40.5 32.9 43.5 ...
##  $ food_env_index     : num  7.3 7.5 6.9 6.9 6.5 6.6 6.7 7.2 7.8 7.1 ...
##  $ exercise           : num  78.8 86.4 100 86.7 21.6 ...
##  $ overcrowding       : num  3.3 2.82 29.95 2.38 2.31 ...
##  $ diabetics          : num  16 17.6 7.6 11.4 13.8 14.3 15.8 15.5 22.9 17.5 ...
##  $ insuf_sleep        : num  40.9 39.1 33.6 39.8 38.7 ...
##  $ traffic_vol        : num  61.97 36.8 10.14 1.39 11.57 ...
##  $ above_65           : num  16.97 15.87 6.82 19.61 14.97 ...
##  $ rural_pop          : num  74.1 67.1 59.3 100 67.4 ...
```

A continuació convertim les variables *met_obj_reduction* i *trend_recent_deaths* en factors i les variables *age_adj_deathrate*, *avg_deaths_year* i *trend_5y_deaths* en variables numèriques. També substituïm els valors * per NA i avaluem la quantitat de NAs al dataset mitjançant *ColSums* i *VIM::aggr()*.

```
uscancer <- uscancer %>%
  mutate(met_obj_reduction = factor(met_obj_reduction,
                                    levels = c("Yes", "No")),
         trend_recent_deaths = factor(trend_recent_deaths,
                                      levels = c("rising", "stable","falling")),
         age_adj_deathrate = as.numeric(age_adj_deathrate),
         avg_deaths_year = as.numeric(avg_deaths_year),
         trend_5y_deaths = as.numeric(trend_5y_deaths)
         )
```

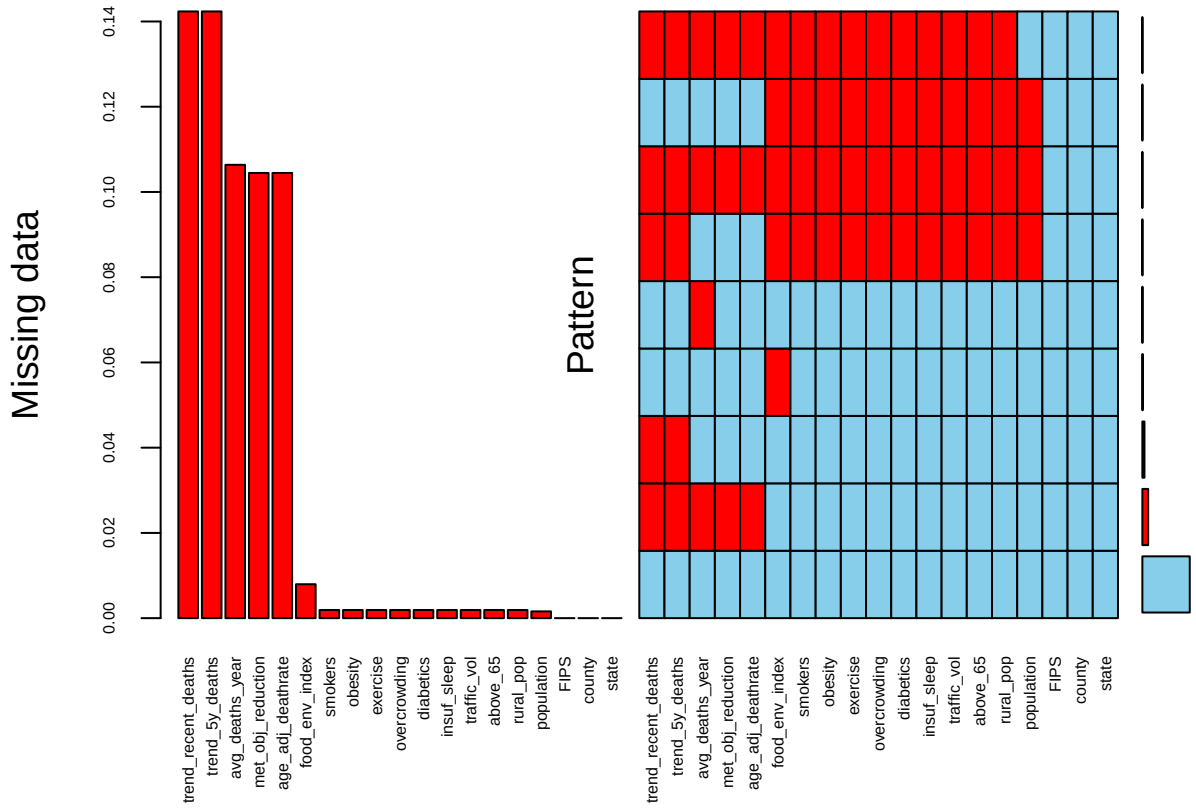
```
## Warning in mask$eval_all_mutate(quo): NAs introducidos por coerción

## Warning in mask$eval_all_mutate(quo): NAs introducidos por coerción

## Warning in mask$eval_all_mutate(quo): NAs introducidos por coerción
```

```
colSums(is.na(uscancer))
```

```
##                FIPS              county               state   met_obj_reduction
##                   0                   0                   0                 328
##   age_adj_deathrate     avg_deaths_year trend_recent_deaths     trend_5y_deaths
##                 328                 334                 447                 447
##          population             smokers             obesity      food_env_index
##                   5                   6                   6                  25
##            exercise        overcrowding           diabetics         insuf_sleep
##                   6                   6                   6                   6
##         traffic_vol           above_65           rural_pop
##                   6                   6                   6
```

```
aggr(uscancer, numbers=TRUE, sortVars=TRUE, labels=names(uscancer),
cex.axis=.5, gap=0, ylab=c("Missing data","Pattern"))
```

```
## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```




```
##
##  Variables sorted by number of missings:
##             Variable       Count
##  trend_recent_deaths 0.142356688
##      trend_5y_deaths 0.142356688
##      avg_deaths_year 0.106369427
##    met_obj_reduction 0.104458599
##    age_adj_deathrate 0.104458599
##       food_env_index 0.007961783
##              smokers 0.001910828
##              obesity 0.001910828
##             exercise 0.001910828
##         overcrowding 0.001910828
##            diabetics 0.001910828
##          insuf_sleep 0.001910828
##          traffic_vol 0.001910828
##             above_65 0.001910828
##            rural_pop 0.001910828
##           population 0.001592357
##                 FIPS 0.000000000
##               county 0.000000000
##                state 0.000000000
```

Observem que, respecte les dades de mortalitat per càncer, no tenim informació completa sobre el rati de mortalitat per càncer ajustada per edat (*age_adj_deathrate*) en 328 comtats, sobre la mitjana de morts

per any (*avg_deaths_year*) en 334 comtats (entre es quals alguns amb elevada població com San Francisco o Los Ángeles), i tampoc tenim informació sobre les tendències en la mortalitat (*trend_recent_deaths* i *trend_5y_deaths*) en 447 comtats. Per ser una mètrica que facilita la comparació entre comtats, centrarem els anàlisis subsegüents en les dades de rati de mortalitat per càncer ajustada per edat (*age_adj_deathrate*), que seleccionem com a variable d'interès. Observem que els 328 comtats pels què la variable *age_adj_deathrate* no té informació representen un 0.39% de la població total dels EEUU, tal i com indica la taula de sota. Per tant, prioritzant l'exactitud de les dades, hem decidit eliminar els registres que no tenen aquesta informació (`NAs` de *age_adj_deathrate*).

```
uscancer %>%
  mutate(canc_data = ifelse(is.na(age_adj_deathrate), "No", "Yes")) %>%
  group_by(canc_data) %>%
  summarise(population = sum(population, na.rm=T),
            perc_total_pop = (sum(population, na.rm=T)/sum(uscancer$population, na.rm=T))*100)
```

```
##   population perc_total_pop
## 1  328122733            100
```

En paral · lel, tenint en compte les variables demogràfiques, que tenen menys registres incomplets, observem que a banda dels 328 comtats sense dades de *age_adj_deathrate*, 3 comtats no tenen cap dada demogràfica disponible. Per això decidim eliminar aquests registres. A més, 19 registres addicionals no disposen de dades sobre la variable *food_env_index*. Aquesta variable és un indicador de la proximitat a menjar saludable així com de la capacitat econòmica per adquirir-ne de la població. Decidim assumir que els valors podrien ser semblants entre comtats que comparteixin altres característiques del dataset, i per tant s'ha recorregut a una imputació d'aquests 19 valors perduts mitjançant la funció *kNN* del paquet *VIM*.

Finalment, observem que ja no queden valors perduts en les dades.

```
uscancer_clean <- uscancer %>%
  # Eliminem les variables de mortalitat per càncer que no ens interessen
  select(-avg_deaths_year, -trend_recent_deaths, -trend_5y_deaths) %>%
  # Eliminem els registres buits per age_adj_deathrate i
  # els que no tenen cap dada demogràfic (obesity n'és una)
  filter(!is.na(age_adj_deathrate),
         !is.na(obesity)) %>%
  # Imputem els valors NA de food_env_index amb k-Nearest Neighbours
  VIM::kNN(variable = "food_env_index") %>%
  # Eliminem la columna que indica quins registres són imputats amb kNN
  select(-food_env_index_imp)

colSums(is.na(uscancer_clean))
```

```
##              FIPS            county             state met_obj_reduction
##                 0                 0                 0                 0
## age_adj_deathrate        population           smokers           obesity
##                 0                 0                 0                 0
##    food_env_index          exercise      overcrowding         diabetics
##                 0                 0                 0                 0
##       insuf_sleep       traffic_vol          above_65         rural_pop
##                 0                 0                 0                 0
```

3.2. Gestió de valors extrems

Seguidament, avaluem la presència de valors extrems en les dades. Ja que disposem de múltiples variables i ens interessaria poder detectar outliers tenint en compte múltiples dimensions, hem decidit començar avaluant els outliers segons la distància de Mahalanobis per fer-nos una idea de quins podrien ser els casos extrems. Al codi de sota, es computa aquesta distància i es mostren els 30 comtats amb els valors de distància de Mahalanobis més elevats.

```
# Dataframe de les variables numèriques contínues de uscancer_clean
uscancer_cvars <- uscancer_clean %>%
  select(-FIPS, -county, -state,
         -population, -met_obj_reduction)

#
m.dist.order <- order(mahalanobis(uscancer_cvars,
                                  colMeans(uscancer_cvars),
                                  cov(uscancer_cvars)),
                      decreasing=TRUE)

#
m.outliers <- uscancer_clean$county[m.dist.order]

# Mostrem els top 30 outliers
m.outliers[1:30]
```

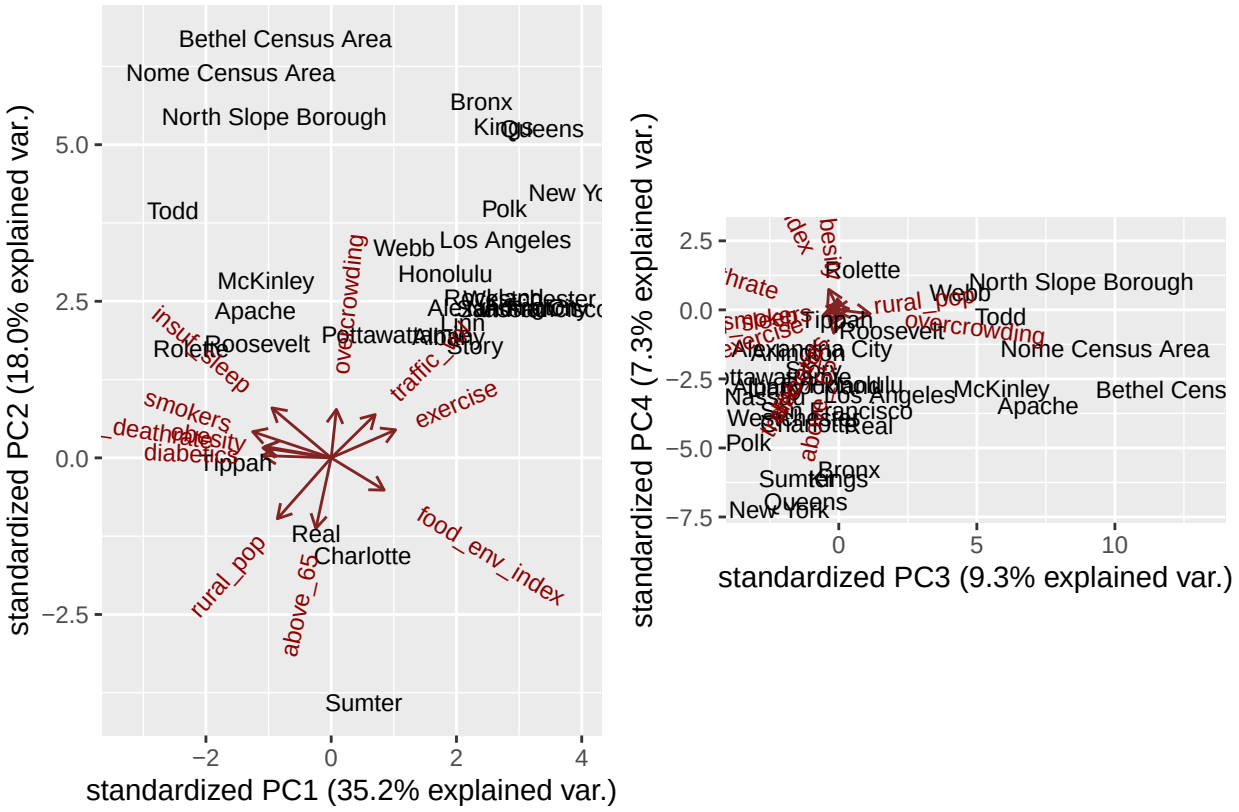
```
##  [1] "Bethel Census Area"  "North Slope Borough" "Queens"
##  [4] "New York"            "Nome Census Area"    "Polk"
##  [7] "Bronx"               "Kings"               "Todd"
## [10] "Sumter"              "Linn"                "Westchester"
## [13] "Nassau"              "Pottawattamie"       "Apache"
## [16] "Rolette"             "McKinley"            "Los Angeles"
## [19] "San Francisco"       "Roosevelt"           "Honolulu"
## [22] "Story"               "Webb"                "Arlington"
## [25] "Albany"              "Tippah"              "Alexandria City"
## [28] "Rockland"            "Charlotte"           "Real"
```

Seguidament, mostrem com es distribueixen aquests 30 comtats més extrems sobre els 4 primers components principals.

```
uscancer.pca <- prcomp(as.matrix(uscancer_cvars),
                       center = TRUE,
                       scale. = TRUE)

labels_top30_outliers <- uscancer_clean$county
labels_top30_outliers[-m.dist.order[1:30]] <- NA

# Representem els top 30 outliers sobre els 6 primers components principals
cowplot::plot_grid( ncol=2, nrow=1,
  ggbiplot(uscancer.pca,
           alpha=0.2,
           labels = labels_top30_outliers,
           choices=1:2
           )
```

```
  ,
  ggbiplot(uscancer.pca,
           alpha=0.2,
           labels = labels_top30_outliers,
           choices=3:4
           )
)
```

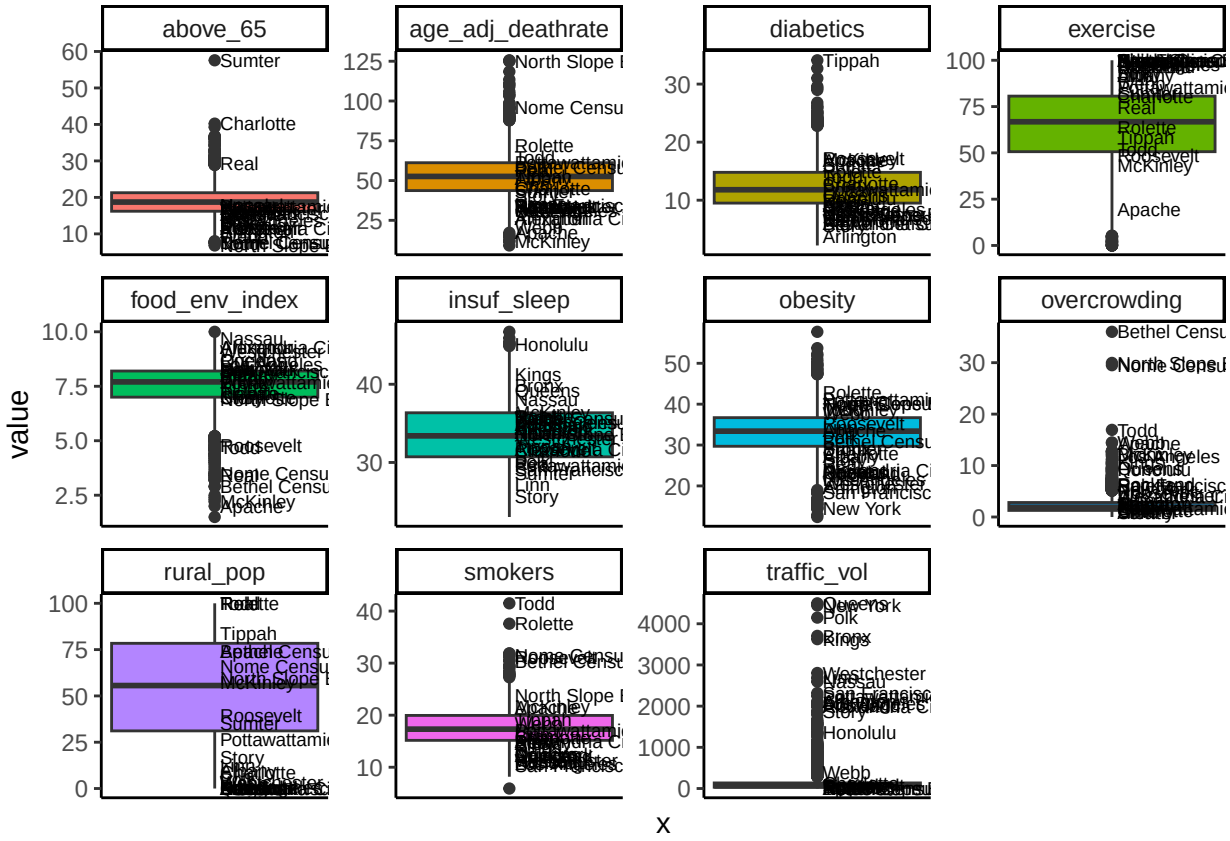
```
## Warning: Removed 2779 rows containing missing values (`geom_text()`).
## Removed 2779 rows containing missing values (`geom_text()`).
```



A sota mostrem la distribució de cada variable contínua en format boxplot, i també s'hi indiquen les posicions dels 30 comtats amb valors de Mahalanobis més elevats.

```
uscancer_cvars %>%
  mutate(top30_outliers = labels_top30_outliers) %>%
  gather(var, value, -top30_outliers) %>%
  ggplot(aes(x=1,
             y=value,
             fill=var)) +
  geom_boxplot() +
  geom_text(aes(x=1+0.02,
                y=value,
                label=top30_outliers),
            size=2.5,
```

```
            hjust = 0
            ) +
  facet_wrap(~var, scales="free") +
  theme_classic() +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank(),
        legend.position = "none")
```



De les dades anteriors, destaquem 4 variables amb valors molt extrems:

1. *above_65*: Observem que hi ha comtats que presenten una proporció de persones majors de 65 anys molt elevada, essent el màxim el comtat de Sumter, Florida, amb 57.6% de la població major de 65 anys. Considerem que aquests valors extrems són correctes ja que tenen una explicació sociodemogràfica: corresponen a comtats que són llocs de residència populars per a gent jubilada. Per tant, els mantenim en el dataset d'estudi.

2. *overcrowding*: Es tracta d'una variable que identifica el percentatge de població que viu en espais amb una quantitat excessivament elevada de persones. En aquest cas observem valors elevats esperables per a comtats en grans ciutats (com Nova York, San Francisco o Los Angeles). Però observem que diversos comtats rurals presenten valors extrems (entre ells Bethel Census Area, North Slope Borough i Nome Census Area a Alaska). Tot i ser sorprenent d'entrada, sembla que l'overcrowding en regions rurals és un problema real, especialment associat a pobresa i a una població predominantment de natius americans. Per tant, decidim mantenir aquests valors extrems.

3. *smokers*: En aquest cas observem que els comtats outliers amb taxes molt elevades de fumadors (per sobre del 35% de la població) s'associen a poblacions majoritàriament de natius americans, mentre que

els valors molt baixos es troben majoritàriament a l'estat de Utah, possiblement associats a població de religió mormona. El comtat de Utah, a Utah, amb un 5% de fumadors entre la població i un 82% són mormons. Atenent a aquests fets, decidim mantenir aquestes dades.

4. *traffic_vol*: els valors més extrems corresponen a grans ciutats, especialment a Nova York. Per tant, també decidim mantenir aquests valors.

Observant els valors extrems de les altres variables sociodemogràfiques, trobem que totes són explicables segons les particularitats de cada comtat, com ara la taxa de diabètics del 34% de la població al comtat de Tippah, Mississippi. Per tant, tots els valors extrems trobats són explicables a causa de la diversitat existent entre regions del país, i els mantenim per als anàlisis subsegüents.

4. Anàlisi de les dades

Una vegada tenim les dades netejades, ordenades i seleccionades procedirem a l'anàlisi d'aquestes. Concretament hem decidit estudiar-les a través de tres perspectives: Primerament estudiarem la mortalitat del càncer segons si es tracta de comtats urbans o rurals; després estudiarem com influeixen les diferents variables a la mortalitat per càncer; i finalment tractarem de crear un model que puga predir la mortalitat del càncer segons les diferents variàbles donades.

4.1 Estudi de la mortalitat segons ruralitat

Com hem comentat, en aquest apartat estudiarem la mortalitat del càncer segons si es tracta de comtats rurals o urbans. Aquest estudi és interessant perquè es comenta habitualment que els entorns urbans afavoreixen la mortalitat per càncer. Per tant és interessant saber si la mitjana poblacional dels comtats urbans és major a la mitjana dels entorns rurals.

Per fer aquest anàlisi, primerament discretitzem les dades en una nova columna, *isurban*, on guardarem els comtats que tinguen una *rural_pop* major a 50 com a 1, i la resta com a 0. Interpretem que els comtats amb 1 equivalen a entorns urbans i els que tenen valor 0 com entorns rurals.

```
# Discretització de la columna rural_pop en una nova variable
uscancer_clean$is_urban <- uscancer_clean$rural_pop
uscancer_clean$is_urban[uscancer_clean$rural_pop >= 50] <- 1
uscancer_clean$is_urban[uscancer_clean$rural_pop < 50] <- 0
```

```
## Hi ha 1217 comtats rurals i 1592 comtats urbans.
## Els comtats urbans suposen el 43.33 % dels registres.
## Mentre que els comtats urbans representen el 56.67 %.
```

Una vegada tenim la variable discretitzada, cal estudiar la normalitat i l'homogeneitat de les dades per determinar quin test aplicar-hi. Primerament estudiarem la normalitat de les dades amb el test de Shapiro-Wilk:

```
# Comprovem la normalitat de les dades:
shapiro.test(uscancer_clean$age_adj_deathrate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  uscancer_clean$age_adj_deathrate
## W = 0.98345, p-value < 2.2e-16
```

Tot i que obtenim un valor estadístic elevat, el p-valor és molt menut i per tant s'ha de descartar l'hipòtesi nul · la de normalitat. Amb aquest resultat hem de concloure que les dades no segueixen una distribució normal. Així i tot, procurarem de millorar la normalitat abans de continuar. Primerament eliminarem els valors extrems que superen la ràtio del 100%.

```
# Eliminem els outliers majors a 100 del deathrate
uscancer_clean2 <- uscancer_clean
uscancer_clean2<-uscancer_clean2[!(uscancer_clean2$age_adj_deathrate > 100),]
shapiro.test(uscancer_clean2$age_adj_deathrate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  uscancer_clean2$age_adj_deathrate
## W = 0.99468, p-value = 1.66e-08
```

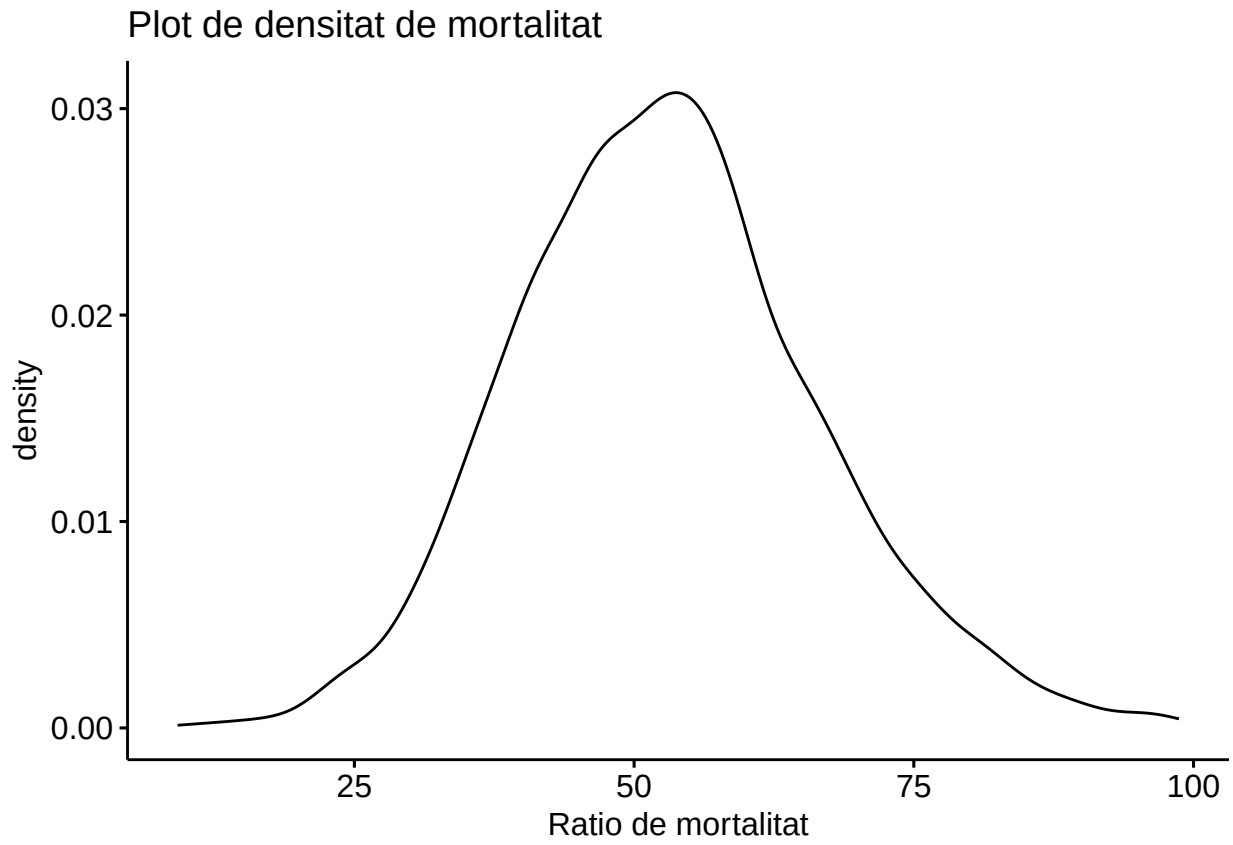
Tot i que es millora la normalitat de les dades, segueix essent insuficient, ja que també ens dóna un p-valor molt menut que ens obliga a descartar la normalitat. Provarem finalment a fer una transformació per millorar encara més la seua normalitat:

```
# Després d'haver fet la distribució més endavant, sabem que té una desviació
# lleugera positiva. Per tant la millor transformació és l'arrel quadrada.
dades_trans <- sqrt(uscancer_clean2$age_adj_deathrate)
shapiro.test(dades_trans)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades_trans
## W = 0.99716, p-value = 4.565e-05
```

Una vegada més, tot i que obtenim una millora en la normalitat de les dades, seguim amb un p-valor inferior a 0.05. Com la transformació no ens ha ajudat per arreglar la normalitat de les dades la descartarem per evitar complicar l'interpretació dels test. Al veure que no es pot millorar la normalitat amb test formals, hem decidit visualitzar la distribució de la mortalitat per veure si s'aproxima a una distribució normal. A la gràfica de sota s'hi pot veure amb claritat que la nostra distribució és prou semblant a una distribució normal. És per això que decidim concloure que les dades tenen una distribució normal, però ens calen més mostres per poder afirmar-ho formalment. És a dir, amb més dades la distribució tendirà a una distribució normal. El que volem dir és que ens basem exclusivament en el Teorema del Limit Central i l'observació de la densitat per afirmar que les dades segueixen una distribució normal. Asumim per tant la seua normalitat.

```
ggdensity(uscancer_clean2$age_adj_deathrate,
          main = "Plot de densitat de mortalitat",
          xlab = "Ratio de mortalitat")
```

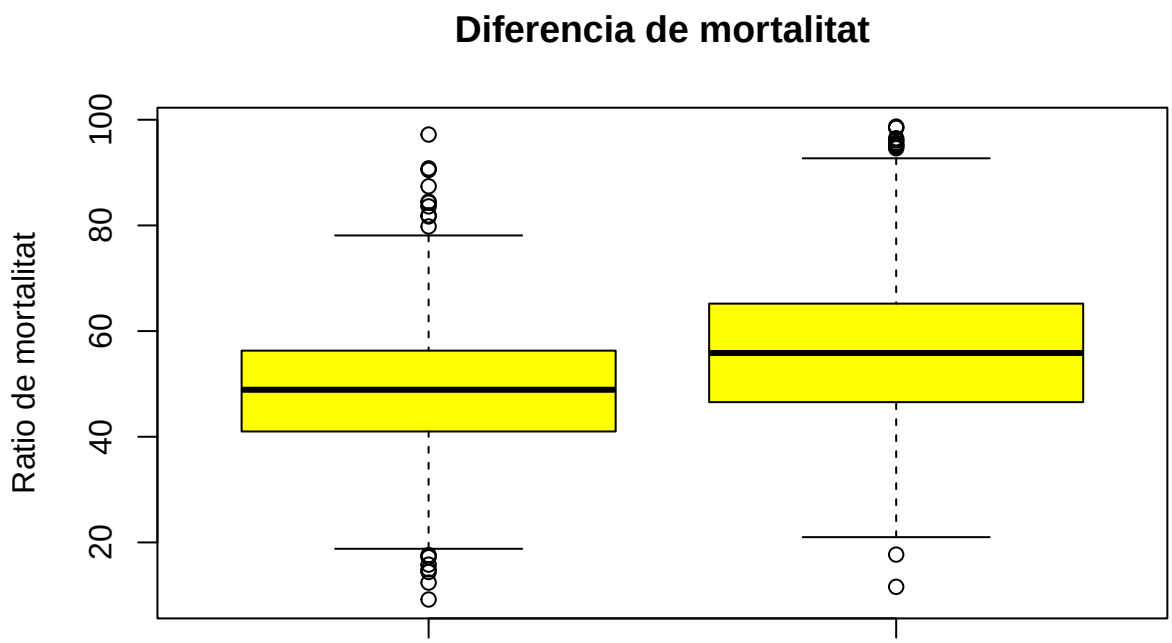


Ara cal revisar que els dos paràmetres que volem comparar tinguen una variància igual per la ràtio de mortalitat. Com sabem que es sol dir que en els entorns urbans augmenta la mortalitat, podem intuïr que la variància no serà igual. Així i tot, comprovem l'homoscedasticitat:

```
# Separació de les dades
rural <- uscancer_clean2[!(uscancer_clean2$is_urban == 1),]
urban <- uscancer_clean2[!(uscancer_clean2$is_urban == 0),]


var.test(rural$age_adj_deathrate,
         urban$age_adj_deathrate,
         conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  rural$age_adj_deathrate and urban$age_adj_deathrate
## F = 0.7103, num df = 1216, denom df = 1579, p-value = 3.5e-10
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6392276 0.7898611
## sample estimates:
## ratio of variances
##          0.7103022
```

```
boxplot(rural$age_adj_deathrate,
        urban$age_adj_deathrate,
        col="yellow",ylab="Ratio de mortalitat",main="Diferencia de mortalitat")
```



A través d'una inspecció visual, i del var.test, podem concloure que ambdues poblacions tenen variàncies diferents. Primerament al var.test veiem que el p-valor no és superior a 0.05, i després podem observar les distribucions d'ambdós paràmetres i s'hi veu com les dades dels entorns urbans tendeixen a ser pitjors (major mortalitat).

Ara, finalment, aplicarem un test per comparar dues poblacions independents, que asumim que seran normals, amb variàncies desconegudes i diferents. La nostra hipòtesi nul · la és que les mitjanes són iguals per a ambdues poblacions; i la nostra hipòtesi alternativa és que la mitjana dels entorns urbans és major a la rural. És a dir:

H0 = y(urban) = y(rural)

H1 = y(urban) > y(rural)

```
t.test(urban$age_adj_deathrate,
       rural$age_adj_deathrate,
       alternative="greater",var.equal=FALSE,conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  urban$age_adj_deathrate and rural$age_adj_deathrate
## t = 15.193, df = 2772.5, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.57144     Inf
## sample estimates:
## mean of x mean of y
##  56.13861  48.76902
```

Amb una confinança del 95% podem concloure que la mitjana del ràtio de mortalitat en entorns urbans és major que la mitjana als entorns rurals. És a dir: y(urban) > y(rural). Extraiem aquesta conclusió perquè el p-valor és menor que la significància escollida, 0.05, i per tant s'ha de descartar l'hipòtesi nul · la i acceptar l'alternativa.

Tot això ens ve a dir que els factors ambientals influèixen a la mortalitat del càncer. Concretament, les persones que viuen en entorns urbans tenen més risc de morir per càncer que les persones que viuen a entorns rurals.

4.2 Estudi de la mortalitat segons variables de control

4.3 Model predictiu de la mortalitat per càncer