Cleaning Spanish Job Market Dataset

Antonio Milán Otero

2 de January 2019

Contents

1. Descripció del dataset

Aquest dataset conté informació sobre ofertes laborals trobades a la web proporcionada per l'estat Espanyol per a tal proposit: https://www.empleate.gob.es/empleo/#/. Les ofertes contingudes tenen data d'entre Juny de 2016 i Novembre del 2018, sent la gran majoria de les ofertes del 2018.

Aquest dataset prové de la pràctica anterior, en la qual, no vaig prestar gens d'atenció a la neteja de les dades, donant com a resultat un dataset molt brut. Això es va fer a proposit per tal de poder aprofitar el dataset en aquesta pràctica.

Per tal de donar a entendre una mica més el contingut d'aquest dataset, pasarem a inspeccionar algunes de les seves característiques.

Començem carregant les dades:

```
offers <- read.csv("../data/offers_dataset.csv")
head(offers)
```

```
##   categoria                          categoriaF
## 1        14    ELECTRICIDAD/ELECTRÓNICA/ENERGÍA
## 2         2 AGRICULTURA/JARDINERÍA/ALIMENTACIÓN
## 3        19     CUIDADOS/ASISTENCIA EN EL HOGAR
## 4        22                       SALUD/DEPORTE
## 5        14    ELECTRICIDAD/ELECTRÓNICA/ENERGÍA
```

```
## 6         1                      ADMINISTRACIÓN
##                            categoriaS ciudad    ciudadF companiaSeleccion
## 1    ELECTRICIDAD/ELECTRÓNICA/ENERGÍA  46165 Massanassa
## 2 AGRICULTURA/JARDINERÍA/ALIMENTACIÓN  28079     Madrid
## 3     CUIDADOS/ASISTENCIA EN EL HOGAR  28106      Parla
## 4                       SALUD/DEPORTE  41091    Sevilla
## 5    ELECTRICIDAD/ELECTRÓNICA/ENERGÍA  10148  Plasencia              True
## 6                     ADMINISTRACIÓN   3058        Cox
##   competenciasReq comunidad           comunidadF consulta1 consulta2
## 1                      10 COMUNIDAD VALENCIANA
## 2                      13               MADRID
## 3                      13               MADRID
## 4                       1             ANDALUCÍA
## 5                      11           EXTREMADURA
## 6                      10 COMUNIDAD VALENCIANA
##   consulta3 consulta4
## 1
## 2
## 3
## 4
## 5
## 6
##
## 1                                                                                Se precisa técnico de el
## 2 Desde Asistencia Familiar buscamos Jardineros/as con experiencia para cuidar el jardín de una casa
## 3                      Desde Asistencia Familiar 24 buscamos personas interesadas en trabajar en 3
## 4
## 5                                Gestión Fuentecaliente (Ciudad Real) y Campo Arañuelo (Cáceres).\nl
## 6
##                  creador cuestionario discapacidad duracion educacion
## 1 CMG METALHIDRAULICA S.L                    False                 6
## 2         VITALSERVIT S.L                    False                 0
## 3         VITALSERVIT S.L                    False                 0
## 4    ALI AL SAAD ALI SAAD                    False                 6
## 5        Axial Ingeniería                    False                 0
## 6         CARGONETWORK SL                    False                 5
##   educacionDes educacionDesF                    educacionF educacionReq
## 1           NA              FP II, Ciclo de Grado Superior         6010
## 2           NA                           Sin especificar            0
## 3           NA                           Sin especificar            0
## 4           NA              FP II, Ciclo de Grado Superior         6024
## 5           NA                           Sin especificar            0
## 6           NA                  FP I, Ciclo de Grado Medio         5004
##                                  educacionReqF
## 1                    Electricidad y Electrónica
## 2                                SIN ESPECIFICAR
## 3                                SIN ESPECIFICAR
## 4 Otras Titulaciones (FP 2, ciclos grado superior)
## 5                                SIN ESPECIFICAR
## 6                                 Administración
##                      educacionS                                email
## 1 FP II, Ciclo de Grado Superior              info@cmghidraulica.com
## 2               Sin especificar alcobendas@asistenciafamiliar24.com
## 3               Sin especificar
```

```
## 4 FP II, Ciclo de Grado Superior
## 5                Sin especificar          tamara@axialingenieria.net
## 6     FP I, Ciclo de Grado Medio
##   empresaSocial ett        fechaCreacion   fechaCreacionBoost
## 1         False     2018-11-02T09:24:44Z 2018-11-02T00:00:00Z
## 2         False     2018-11-02T09:34:00Z 2018-11-02T00:00:00Z
## 3         False     2018-11-02T09:36:21Z 2018-11-02T00:00:00Z
## 4         False     2018-11-02T10:19:16Z 2018-11-02T00:00:00Z
## 5         False     2018-11-02T11:15:56Z 2018-11-02T00:00:00Z
## 6         False     2018-11-02T10:29:39Z 2018-11-02T00:00:00Z
##    fechaCreacionPortal fechaIncorporacion        fechaRevision
## 1 2018-11-02T00:00:00Z                    2019-01-15T01:00:00Z
## 2 2018-11-02T00:00:00Z                    2019-02-02T00:00:00Z
## 3 2018-11-02T00:00:00Z                    2019-02-02T00:00:00Z
## 4 2018-11-02T00:00:00Z                    2019-02-02T00:00:00Z
## 5 2018-11-02T00:00:00Z                    2018-11-06T01:00:00Z
## 6 2018-11-02T00:00:00Z                    2019-02-02T00:00:00Z
##   formacionReq                                        horario         id
## 1                                                8:30-18:30 1734703986
## 2                                                           1734704021
## 3              Martes, Miércoles y Jueves de 9.30 a 14.30 1734704041
## 4                                          lunes a viernes 1734704173
## 5                                4 horas/ Media jornada 1734704643
## 6                                       09:00H A 15:00H 1734704257
##   jornada jornadaF        localizacion minExperiencia nivel noMeInteresa
## 1       1 COMPLETA 39.412597,-0.399604              2
## 2       2  PARCIAL 40.416775,-3.703790              2
## 3       2  PARCIAL 40.232367,-3.768906              2
## 4       2  PARCIAL 37.389092,-5.984459              1
## 5       2  PARCIAL 40.042095,-6.083807              2
## 6       2  PARCIAL 38.143047,-0.890160              2
##   numCandidatos                                          oReq
## 1            NA Trato de cara al público amable, responsable.
## 2            NA
## 3            NA
## 4            NA
## 5            NA                  Experiencia en Electricidad.
## 6            NA
##
## 1
## 2
## 3
## 4
## 5 Coche de empresa. Se trabaja 4 horas, pero los días que no hay trabajo se compensa con mas horas de
## 6
##   origen pais  paisF  paisS provincia provinciaF provinciaLimitrofe
## 1    WEB  724 ESPAÑA ESPAÑA        46   VALENCIA       ['VALENCIA']
## 2    WEB  724 ESPAÑA ESPAÑA        28     MADRID         ['MADRID']
## 3    WEB  724 ESPAÑA ESPAÑA        28     MADRID         ['MADRID']
## 4    WEB  724 ESPAÑA ESPAÑA        41    SEVILLA        ['SEVILLA']
## 5    WEB  724 ESPAÑA ESPAÑA        10    CÁCERES        ['CÁCERES']
## 6    WEB  724 ESPAÑA ESPAÑA         3   ALICANTE       ['ALICANTE']
##   provinciaS publicado1 publicado2 publicado3 publicado4 respuesta1A
## 1   VALENCIA
```

```
## 2     MADRID
## 3     MADRID
## 4    SEVILLA
## 5    CÁCERES
## 6   ALICANTE
##   respuesta1B respuesta2A respuesta2B respuesta2C respuesta3A respuesta3B
## 1
## 2
## 3
## 4
## 5
## 6
##   salarioMax salarioMin    score sector   sectorF sisgarjuv speState
## 1       1600       1000 1.175582      3 INDUSTRIA             Activa
## 2        105        100 1.175582     NA                       Activa
## 3        105        100 1.175582     NA                       Activa
## 4       1000        750 1.175582     NA                       Activa
## 5        600        600 1.175582      3 INDUSTRIA             Activa
## 6       1000        900 1.175582      5 SERVICIOS             Activa
##   speStateId subcategoria       subcategoriaF       subcategoriaS
## 1          1        14002         ELECTRÓNICA         ELECTRÓNICA
## 2          1         2002 FORESTAL/JARDINERÍA FORESTAL/JARDINERÍA
## 3          1        19005  SERVICIO DOMÉSTICO  SERVICIO DOMÉSTICO
## 4          1        22001            MEDICINA            MEDICINA
## 5          1        14001        ELECTRICIDAD        ELECTRICIDAD
## 6          1         1001     ADMINISTRATIVOS     ADMINISTRATIVOS
##   subsector subsectorF tamanoCompania2 telefono tipoContrato tipoContratoN
## 1        NA                        NA                                   2
## 2        NA                        NA                                   3
## 3        NA                        NA                                   3
## 4        NA                        NA                                   5
## 5        NA                        NA                                   3
## 6        NA                        NA                                   2
##                                       titulo trabajosOfertados url
## 1 Técnico electrónico con nociones de hidráulica                1   -
## 2         Jardinero/a para la zona de Valdebebas                1   -
## 3                      Labores Domésticas Parla                 1   -
## 4     PSICÓLOGO CERTIFICADOS MÉDICOS Y LICENCIAS                1   -
## 5                   Electricista con experiencia                1   -
## 6                    ADMINISTRATIVO/A CON INGLES                1   -
##   valor1A valor1B valor2A valor2B valor3B verMail verSalarioMax
## 1      NA      NA      NA      NA      NA                    NA
## 2      NA      NA      NA      NA      NA                    NA
## 3      NA      NA      NA      NA      NA                    NA
## 4      NA      NA      NA      NA      NA                    NA
## 5      NA      NA      NA      NA      NA                     1
## 6      NA      NA      NA      NA      NA                    NA
##   verSalarioMin verTelefono                              web
## 1            NA                      www.cmghidraulica.com
## 2            NA
## 3            NA
## 4            NA
## 5             1             http://www.axialingenieria.net/
## 6            NA
```

```
features_length <- length(offers)
df_length <- length(offers$categoria)
sprintf("Dataset amb %d característiques i %d registres",
        features_length, df_length)
```

```
## [1] "Dataset amb 94 característiques i 40534 registres"
```

Com podem veure, tenim 40534 registres i 94 característiques, moltes de les quals no ens seràn d'utilitat. Podriem mirar ara quines son aquestes 94 variables.

```
names(offers)
```

```
##  [1] "categoria"           "categoriaF"          "categoriaS"
##  [4] "ciudad"              "ciudadF"             "companiaSeleccion"
##  [7] "competenciasReq"     "comunidad"           "comunidadF"
## [10] "consulta1"           "consulta2"           "consulta3"
## [13] "consulta4"           "contenido"           "creador"
## [16] "cuestionario"        "discapacidad"        "duracion"
## [19] "educacion"           "educacionDes"        "educacionDesF"
## [22] "educacionF"          "educacionReq"        "educacionReqF"
## [25] "educacionS"          "email"               "empresaSocial"
## [28] "ett"                 "fechaCreacion"       "fechaCreacionBoost"
## [31] "fechaCreacionPortal" "fechaIncorporacion"  "fechaRevision"
## [34] "formacionReq"        "horario"             "id"
## [37] "jornada"             "jornadaF"            "localizacion"
## [40] "minExperiencia"      "nivel"               "noMeInteresa"
## [43] "numCandidatos"       "oReq"                "oferta"
## [46] "origen"              "pais"                "paisF"
## [49] "paisS"               "provincia"           "provinciaF"
## [52] "provinciaLimitrofe"  "provinciaS"          "publicado1"
## [55] "publicado2"          "publicado3"          "publicado4"
## [58] "respuesta1A"         "respuesta1B"         "respuesta2A"
## [61] "respuesta2B"         "respuesta2C"         "respuesta3A"
## [64] "respuesta3B"         "salarioMax"          "salarioMin"
## [67] "score"               "sector"              "sectorF"
## [70] "sisgarjuv"           "speState"            "speStateId"
## [73] "subcategoria"        "subcategoriaF"       "subcategoriaS"
## [76] "subsector"           "subsectorF"          "tamanoCompania2"
## [79] "telefono"            "tipoContrato"        "tipoContratoN"
## [82] "titulo"              "trabajosOfertados"   "url"
## [85] "valor1A"             "valor1B"             "valor2A"
## [88] "valor2B"             "valor3B"             "verMail"
## [91] "verSalarioMax"       "verSalarioMin"       "verTelefono"
## [94] "web"
```

Veiem també que tenim moltes variables que están duplicades o que no ens proporcionaran informació necessaria per als nostres estudis. Podriem consultar més detalls d'aquest dataset amb la següent commanda, que no executarem per tal d'afavorir la lectura d'aquest document:

```
summary(offers)
```

Veiem que per tal d'estudiar els salaris ofertats, tenim dues característiques a estudiar: salarioMax i salarioMin.

```
sal_min_mean <- mean(offers$salarioMin, na.rm = TRUE)
sal_min_sd <- sd(offers$salarioMin, na.rm = TRUE)
sprintf("Mitjana del salari mínim de les ofertes: %f, amb una desviació estandard de: %f",
```

```
        sal_min_mean, sal_min_sd)
```

```
## [1] "Mitjana del salari mínim de les ofertes: 10092.368558, amb una desviació estandard de: 12983.713
```

```
sal_max_mean <- mean(offers$salarioMax, na.rm = TRUE)
sal_max_sd <- sd(offers$salarioMax, na.rm = TRUE)
sprintf("Mitjana del salari màxim de les ofertes: %f, amb una desviació estandard de: %f",
        sal_max_mean, sal_max_sd)
```

```
## [1] "Mitjana del salari màxim de les ofertes: 16366.986917, amb una desviació estandard de: 286011.30
```

Amb tota aquesta informació, podem enumerar quines son les preguntes que volem respondre:

1. Quines regions d'Espanya generen més ofertes de treball?
2. Podem fer prediccións sobre salaris mínims i màxims?
3. Estudi sobre els salaris en relació a les 5 regions que generen més ofertes. Tenim regions amb salari mínim superior a la resta? I a la categoria d'informàtica i telecomunicacions?

A la pràctica anterior enumerabem també les següents idees que deixarem obertes per a possibles futurs estudis i que **no** formen part dels objectius d'aquest treball:

- A on trobem un major salari?
- Quin tipus de professional és el més solicitat a Espanya (durant el periode de mostreig)?
- Analitzar els diferents requeriments professionals que tenen les diferents autonomies d'Espanya.
- Identificar el tipus i la qualitat del treball actual al país.
- Analitzar les regions amb més i menys ofertes de treball.
- Analitzar la distribució de les diferents professions en funció de la regió.
- Ajudar a la creació d'un pla per potenciar el mercat laboral basat en el coneixement obtingut a través de les dades.

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

Per aquest apartat ja es va crear un script python que s'encarregaba de compilar les dades obtingudes en diferents dies. La idea darrera d'aquest script era la de recolectar totes les dades disponibles a la web en una primera pasada, i després anar actualitzant el dataset agafant dades diaries i agrupant-les sota el mateix fitxer .csv

Per tant, en aquest apartat considero que no haig de fer més que el ja s'ha fet fins a la data.

El script es pot trobar en la següent URL: [https://github.com/amilan/spanish_job_market/blob/master/src/dataset_merge.py]

També tinc en compte, que la web oficial de la qual es va extreure les dades, ja recompila aquestes dades de diferent fonts, així doncs, no considero que sigui necessari l'integració de dades de diferentes fonts, ja que aquesta ha estat realitzada anteriorment.

Per tot això, en aquest apartat només seleccionarem les dades necessaries per als nostres estudis.

Hem de tenir en compte, que a la pràctica anterior hem vaig limitar a agafar totes les dades possibles i a possar-les en un fitxer .csv. Aquestes dades provenien d'una base de dades NoSQL, ja que vaig detectar que amb les mateixes crides, podiem obtenir dades amb diferents esquemes (schemaless). Així doncs, farem una selecció de les dades que utilitzarem i eliminarem així dades no necessàries o repetides.

Començarem seleccionant les dades d'interés. Recordem que la meva intenció es la de fer un estudi sobre els salaris mínims i màxims de les ofertes de treball a Espanya i en concret a les comunitats autónomes que més ofertes generen. Tot i així, enfocaré aquest primer pas de neteja amb una mirada més amplia i afegiré algunes característiques extra que hem puguin ser d'utilitat en futures revisions o expansions d'aquest treball.

Guardaré aquetes dades netejades en un nou fitxer .csv i després faré una segona neteja per tal de quedar-me només amb les dades d'interés per aquest treball.

Primerament, comprovarem que només tenim dades d'ofertes realitzades a Espanya.

```
levels(offers$paisS)
```

```
## [1] "CONGO"   "ESPAÃ‘A" "ESPAÑA"  "ESPA A"
```

Comprovem dues coses, que tenim ofertes d'Espanya i també al Congo, i que tenim un problema de codificació de caracters, ja que ens troba el país d'Espanya en tres factors diferents. Com que només volem utilitzar les dades de les ofertes a Espanya, podem seleccionar totes les que no siguin al Congo i després eliminar aquesta columna.

Podem corregir les dades erronees de país:

```
offers$paisS <- sub("ESPAÃ‘A", "ESPAÑA", offers$paisS)
offers$paisS <- sub("ESPA A", "ESPAÑA", offers$paisS)
levels(factor(offers$paisS))
```

```
## [1] "CONGO"  "ESPAÑA"
```

També podriem haver canviat la codificació dels caracters, com veurem més endavant.

Seleccionem ara només les ofertes a Espanya.

```
offers <- subset(offers, paisS == "ESPAÑA")
levels(factor(offers$paisS))
```

```
## [1] "ESPAÑA"
```

```
# paisS es ara del tipus chr, hauriem de convertirla de nou a factor
offers$paisS <- factor(offers$paisS)
```

```
class(offers$paisS)
```

```
## [1] "factor"
```

Seguidament, eliminarem les columnes que ofereixen informació duplicada. Ens quedarem amb les característiques:

- categoriaF
- comunidadF
- educacionF
- fechaCreacion
- jornadaF
- provinciaS
- salarioMax
- salarioMin
- subcategoriaS

```
selected_features <- c("categoriaF", "comunidadF", "educacionF",
                       "fechaCreacion", "jornadaF", "provinciaS",
                       "salarioMax", "salarioMin", "subcategoriaS")
offers <- offers[selected_features]
head(offers)
```

```
##                           categoriaF           comunidadF
## 1    ELECTRICIDAD/ELECTRÓNICA/ENERGÍA COMUNIDAD VALENCIANA
## 2 AGRICULTURA/JARDINERÍA/ALIMENTACIÓN               MADRID
## 3     CUIDADOS/ASISTENCIA EN EL HOGAR               MADRID
```

```
## 4                      SALUD/DEPORTE            ANDALUCÍA
## 5    ELECTRICIDAD/ELECTRÓNICA/ENERGÍA          EXTREMADURA
## 6                     ADMINISTRACIÓN COMUNIDAD VALENCIANA
##                       educacionF        fechaCreacion jornadaF provinciaS
## 1 FP II, Ciclo de Grado Superior 2018-11-02T09:24:44Z COMPLETA   VALENCIA
## 2               Sin especificar 2018-11-02T09:34:00Z  PARCIAL     MADRID
## 3               Sin especificar 2018-11-02T09:36:21Z  PARCIAL     MADRID
## 4 FP II, Ciclo de Grado Superior 2018-11-02T10:19:16Z  PARCIAL    SEVILLA
## 5               Sin especificar 2018-11-02T11:15:56Z  PARCIAL    CÁCERES
## 6     FP I, Ciclo de Grado Medio 2018-11-02T10:29:39Z  PARCIAL   ALICANTE
##   salarioMax salarioMin       subcategoriaS
## 1       1600       1000         ELECTRÓNICA
## 2        105        100 FORESTAL/JARDINERÍA
## 3        105        100  SERVICIO DOMÉSTICO
## 4       1000        750            MEDICINA
## 5        600        600        ELECTRICIDAD
## 6       1000        900     ADMINISTRATIVOS
```

3. Neteja de les dades

Com que ens hem adonat abans que hi teniem problemes de codificació amb els strings, lo primer que farem serà corregir aquests problemes. Cal destacar que per tal d'afavorir la lectura del document, no mostrarem tots els factors de les variables, només els sis primers.

```
head(levels(offers$comunidadF))
```

```
## [1] ""           "ANDALUCÃ A" "ANDALUCÍA"  "ARAGÃ"N"    "ARAGÓN"
## [6] "CANTABRIA"
```

```
offers$comunidadF <- sub("ARAGÃ"N", "ARAGÓN", offers$comunidadF)
offers$comunidadF <- sub("CASTILLA Y LE N", "CASTILLA Y LEÓN", offers$comunidadF)
offers$comunidadF <- sub("CASTILLA Y LEÃ"N", "CASTILLA Y LEÓN", offers$comunidadF)
offers$comunidadF <- sub("CATALU A", "CATALUÑA", offers$comunidadF)
offers$comunidadF <- sub("CATALUÃ'A", "CATALUÑA", offers$comunidadF)
offers$comunidadF <- sub("ANDALUCÃ A", "ANDALUCÍA", offers$comunidadF)
offers$comunidadF <- sub("REGI N DE MURCIA", "REGIÓN DE MURCIA", offers$comunidadF)
offers$comunidadF <- sub("REGIÃ"N DE MURCIA", "REGIÓN DE MURCIA", offers$comunidadF)
offers$comunidadF <- sub("PAÃ S VASCO", "PAÍS VASCO", offers$comunidadF)
offers$comunidadF <- sub("Sin especificar", "", offers$comunidadF)
offers$comunidadF <- factor(offers$comunidadF)
head(levels(offers$comunidadF))
```

```
## [1] ""                  "ANDALUCÍA"          "ARAGÓN"
## [4] "CANTABRIA"         "CASTILLA LA MANCHA" "CASTILLA Y LEÓN"
```

```
length(levels(offers$comunidadF))
```

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

Veiem que en aquest cas podem tenir valor buit ("") o **sin especificar**. Ens interessa canviar el valor buit per "Unknown", ja que sabem que la oferta ha d'estar ubicada en alguna comunitat, però no sabem en quina. Aquest valor ens facilitarà futures interpretacions dels resultats.

```
# Convertim els valors buits en NA per reconvertir-los a Unknown.
offers$comunidadF <- NAToUnknown(unknownToNA(offers$comunidadF, unknown = ""), unknown = "Unknown")
```

```
## Warning: new level is introduced: Unknown
```

```
head(levels(offers$comunidadF))
```

```
## [1] "ANDALUCÍA"          "ARAGÓN"             "CANTABRIA"
## [4] "CASTILLA LA MANCHA" "CASTILLA Y LEÓN"    "CATALUÑA"
```

```
length(levels(offers$comunidadF))
```

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

```
head(levels(offers$categoriaF))
```

```
## [1] ""
## [2] "ADMINISTRACIÃ"N"
## [3] "ADMINISTRACIÓN"
## [4] "AGRICULTURA/JARDINERÍA/ALIMENTACIÓN"
## [5] "ALMACENES/REPONEDORES"
## [6] "APRENDICES/PRIMER EMPLEO"
```

En comptes de corregir un a un, transformarem les dades al format latin1.

```
# convertim les dades a encoding latin1
offers$categoriaF <- factor(iconv(offers$categoriaF, to = "latin1"))
length(levels(offers$categoriaF))
```

```
## [1] 24
```

```
# Convertim els valors buits en NA per reconvertir-los a Unknown.
offers$categoriaF <- NAToUnknown(unknownToNA(offers$categoriaF, unknown = ""), unknown = "Unknown")
```

```
## Warning: new level is introduced: Unknown
```

```
head(levels(offers$categoriaF))
```

```
## [1] "ADMINISTRACIÓN"
## [2] "AGRICULTURA/JARDINERÍA/ALIMENTACIÓN"
## [3] "ALMACENES/REPONEDORES"
## [4] "APRENDICES/PRIMER EMPLEO"
## [5] "ARQUITECTURA/DISEÑO"
## [6] "COMERCIAL/VENTAS"
```

```
length(levels(offers$categoriaF))
```

```
## [1] 24
```

```
head(levels(offers$provinciaS))
```

```
## [1] ""         "ÁLAVA"    "ALBACETE" "ALICANTE" "ALMERÍA"  "ASTURIAS"
```

Podem veure que Guipúzcua està repetida degut a la mala codificació.

```
# convertim les dades a encoding latin1
offers$provinciaS <- factor(iconv(offers$provinciaS, to = "latin1"))
head(levels(offers$provinciaS))
```

```
## [1] ""         "ÁLAVA"    "ALBACETE" "ALICANTE" "ALMERÍA"  "ASTURIAS"
```

```
length(levels(offers$provinciaS))
```

```
## [1] 53
```

```r
# Convertim els valors buits en NA per reconvertir-los a Unknown.
offers$provinciaS <- NAToUnknown(unknownToNA(offers$provinciaS, unknown = ""), unknown = "Unknown")
```

```
## Warning: new level is introduced: Unknown
```

```r
head(levels(offers$provinciaS))
```

```
## [1] "ÁLAVA"    "ALBACETE" "ALICANTE" "ALMERÍA"  "ASTURIAS" "ÁVILA"
```

```r
length(levels(offers$provinciaS))
```

```
## [1] 53
```

```r
levels(offers$jornadaF)
```

```
## [1] ""            "COMPLETA"    "INDIFERENTE" "PARCIAL"
```

```r
# Convertim els valors buits en NA per reconvertir-los a Unknown.
offers$jornadaF <- NAToUnknown(unknownToNA(offers$jornadaF, unknown = ""), unknown = "Unknown")
```

```
## Warning: new level is introduced: Unknown
```

```r
levels(offers$jornadaF)
```

```
## [1] "COMPLETA"    "INDIFERENTE" "PARCIAL"     "Unknown"
```

```r
# convertim les dades a encoding latin1
offers$subcategoriaS <- factor(iconv(offers$subcategoriaS, to = "latin1"))
head(levels(offers$subcategoriaS))
```

```
## [1] ""
## [2] "ABOGADOS"
## [3] "ACUICULTURA"
## [4] "ADMINISTRATIVOS"
## [5] "AGENCIA DE VIAJES"
## [6] "AGENTES COMERCIALES/REPRESENTANTES"
```

```r
length(levels(offers$subcategoriaS))
```

```
## [1] 131
```

```r
# Convertim els valors buits en NA per reconvertir-los a Unknown.
offers$subcategoriaS <- NAToUnknown(unknownToNA(offers$subcategoriaS, unknown = ""), unknown = "Unknown
```

```
## Warning: new level is introduced: Unknown
```

```r
head(levels(offers$subcategoriaS))
```

```
## [1] "ABOGADOS"
## [2] "ACUICULTURA"
## [3] "ADMINISTRATIVOS"
## [4] "AGENCIA DE VIAJES"
## [5] "AGENTES COMERCIALES/REPRESENTANTES"
## [6] "AGRICULTURA/GANADERÍA"
```

```r
length(levels(offers$subcategoriaS))
```

```
## [1] 131
```

```r
# convertim les dades a encoding latin1
offers$educacionF <- factor(iconv(offers$educacionF, to = "latin1"))
head(levels(offers$educacionF))
```

```
## [1] ""                                "Bachillerato"
## [3] "Certificados de Profesionalidad" "Diplomado o Ingeniero TÃ©cnico"
## [5] "Diplomado o Ingeniero Técnico"   "Doctor Universitario"
```

Tot i la conversió, encara tenim algun cas que no s'ha codificat correctament. El corregirem manualment.

```
offers$educacionF <- sub("Diplomado o Ingeniero TÃ©cnico",
                         "Diplomado o Ingeniero Técnico",
                         offers$educacionF)
offers$educacionF <- sub("Sin especificar", "",
                         offers$educacionF)
offers$educacionF <- factor(offers$educacionF)
head(levels(offers$educacionF))
```

```
## [1] ""                                "Bachillerato"
## [3] "Certificados de Profesionalidad" "Diplomado o Ingeniero Técnico"
## [5] "Doctor Universitario"            "ESO, EGB, Graduado Escolar"
```

```
length(levels(offers$educacionF))
```

```
## [1] 13
```

```
# Convertim els valors buits en NA per reconvertir-los a Unknown.
offers$educacionF <- NAToUnknown(unknownToNA(offers$educacionF,
                                             unknown = ""),
                                 unknown = "Unknown")
```

```
## Warning: new level is introduced: Unknown
```

```
head(levels(offers$educacionF))
```

```
## [1] "Bachillerato"                    "Certificados de Profesionalidad"
## [3] "Diplomado o Ingeniero Técnico"   "Doctor Universitario"
## [5] "ESO, EGB, Graduado Escolar"      "Estudios primarios"
```

```
length(levels(offers$educacionF))
```

```
## [1] 13
```

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

Com hem vist anteriorment, tenim dades amb element buits en les característiques del tipus factor. Hem convertit també les dades amb valors **sin especificar** en valor buit, ja que aquest, per exemple, representa millor que la oferta està situada a una localització, però no sabem a on. Un cop feta aquesta transformació, hem aplicat una altra per convertir tots els valors buits en "Unknown", ja que aquest valor ens facilitarà més la comprensió dels resultats finals.

Passem ara a mirar si tenim elements nulls (NA).

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

```
##    categoriaF    comunidadF    educacionF fechaCreacion      jornadaF
##             0             0             0             0             0
##    provinciaS    salarioMax    salarioMin subcategoriaS
##             0         38087         31667             0
```

```
length(offers$salarioMax)
```

```
## [1] 40533
```

Passem ara a netejar les característiques numériques. Veiem que aproximadament una quarta part de les dades disposen de valors de salari mínim i máxim. Aquests ens podrien ser suficient per al nostres estudi, sempre i quan tinguem suficient casos d'estudi per a les diferentes regions.

Llavors, ens quedarem amb les dades que tenen un salari màxim, ja que aquest grup és menor que els que tenen salari mínim, i descartarem la resta.

```
offers <- subset(offers, !is.na(offers$salarioMax))
```

Comprovem que ja no tenim cap valor NA.

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

```
##    categoriaF    comunidadF    educacionF fechaCreacion      jornadaF
##             0             0             0             0             0
##    provinciaS    salarioMax    salarioMin subcategoriaS
##             0             0             0             0
```

Com que hi ha menys dades amb salari màxim que mínim, podriem haver seguit alguna de les següents estratègies: - Descartar les dades sense salari màxim. Això reduiría molt el nombre de dades, però encara tindriem prou per al estudi que volem realitzar. Aquesta ha estat la estratégia seguida. - Imputar dades utilitzant KNN. Amb aquesta estratègia podriem obtenir els valors en funció de les dades veines. El problema amb aquesta estratégia es que estariem imputant aproximadament tres quartes parts de les nostres dades, la qual cosa ens portaria a problemes d'anàlisi posterior. Aquesta estratégia la vaig probar en versions diferents d'aquest treball, obtenint resultats erronis. - Imputar els valors en funció de la mitjana poblacional de la mostra.

En cas de voler imputar els valors mitjançant l'algoritme KNN ho fariem de la següent manera.

```
offers$salarioMax <- kNN(offers)$salarioMax
sapply(offers, function(x)(sum(is.na(x))))
```

Seguidament podriem comprovar si les nostres dades tenen el tipus que desitjem.

```
sapply(offers, function(x)(class(x)))
```

```
##    categoriaF    comunidadF    educacionF fechaCreacion      jornadaF
##      "factor"      "factor"      "factor"      "factor"      "factor"
##    provinciaS    salarioMax    salarioMin subcategoriaS
##      "factor"     "numeric"     "numeric"      "factor"
```

Veiem que haurem de tractar el format de la característica fechaCreacion. En aquest moment, tenim la data com a un string amb el format: anys, mes, dia, hora. En el nostre cas, només amb l'any, mes i dia en tindrem prou. A més, haurem de donar-li el tipus de date type.

```
offers$fechaCreacion <- as.Date(gsub("T\\d*:\\d*:\\d*Z",
                                     "",
                                     offers$fechaCreacion))
sapply(offers, function(x)(class(x)))
```

```
##    categoriaF    comunidadF    educacionF fechaCreacion      jornadaF
##      "factor"      "factor"      "factor"        "Date"      "factor"
##    provinciaS    salarioMax    salarioMin subcategoriaS
##      "factor"     "numeric"     "numeric"      "factor"
```

```
summary(offers)
```

```
##                              categoriaF                  comunidadF
##  CONSTRUCCIÓN                     : 298   MADRID              :665
##  COMERCIAL/VENTAS                 : 280   CATALUÑA            :340
```

```
##  ADMINISTRACIÓN                  : 206   ANDALUCÍA           :230
##  ELECTRICIDAD/ELECTRÓNICA/ENERGÍA: 202   COMUNIDAD VALENCIANA:217
##  INFORMÁTICA/TELECOMUNICACIONES  : 182   GALICIA             :177
##  HOSTELERÍA/TURISMO              : 164   CASTILLA Y LEÓN     :156
##  (Other)                         :1114   (Other)             :661
##                            educacionF  fechaCreacion
##  Unknown                         :612   Min.   :2016-06-30
##  ESO, EGB, Graduado Escolar      :343   1st Qu.:2018-08-28
##  FP I, Ciclo de Grado Medio      :334   Median :2018-09-13
##  FP II, Ciclo de Grado Superior :320   Mean   :2018-08-23
##  Licenciado o Ingeniero Superior:161   3rd Qu.:2018-10-11
##  Diplomado o Ingeniero Técnico  :156   Max.   :2018-11-04
##  (Other)                         :520
##         jornadaF         provinciaS     salarioMax        salarioMin
##  COMPLETA   :1822   MADRID    : 653   Min.   :      0   Min.   :    0
##  INDIFERENTE: 171   BARCELONA : 278   1st Qu.:   1000   1st Qu.:  800
##  PARCIAL    : 453   VALENCIA  : 124   Median :   1467   Median : 1100
##  Unknown    :   0   ALICANTE  :  78   Mean   :  16367   Mean   : 6343
##                     CORUÑA (A):  77   3rd Qu.:  15000   3rd Qu.:13000
##                     MURCIA    :  74   Max.   :9999999   Max.   :50000
##                     (Other)   :1162
##                              subcategoriaS
##  AGENTES COMERCIALES/REPRESENTANTES: 152
##  ELECTRICIDAD                      : 122
##  VENDEDORES                        : 119
##  ADMINISTRATIVOS                   : 111
##  CAMAREROS                         :  86
##  ALBAÑILERIA/ACABADOS              :  80
##  (Other)                           :1776
```

Per últim, podem canviar el nombre de les característiques per que tinguin una mica més de sentit i guardem les dades en un nou fitxer csv.

```
names(offers)
```

```
## [1] "categoriaF"    "comunidadF"    "educacionF"    "fechaCreacion"
## [5] "jornadaF"      "provinciaS"    "salarioMax"    "salarioMin"
## [9] "subcategoriaS"
```

```
final_names <- c("Categoria", "Comunidad", "Educacion",
                 "FechaCreacion", "TipoJornada",
                 "Provincia", "SalarioMax", "SalarioMin",
                 "SubCategoria")
names(offers) <- final_names
head(offers)
```

```
##                              Categoria            Comunidad
## 1    ELECTRICIDAD/ELECTRÓNICA/ENERGÍA COMUNIDAD VALENCIANA
## 2 AGRICULTURA/JARDINERÍA/ALIMENTACIÓN               MADRID
## 3     CUIDADOS/ASISTENCIA EN EL HOGAR               MADRID
## 4                       SALUD/DEPORTE             ANDALUCÍA
## 5    ELECTRICIDAD/ELECTRÓNICA/ENERGÍA           EXTREMADURA
## 6                     ADMINISTRACIÓN COMUNIDAD VALENCIANA
##                       Educacion FechaCreacion TipoJornada Provincia
## 1 FP II, Ciclo de Grado Superior    2018-11-02    COMPLETA  VALENCIA
## 2                       Unknown    2018-11-02     PARCIAL    MADRID
```

```
## 3                         Unknown    2018-11-02     PARCIAL    MADRID
## 4 FP II, Ciclo de Grado Superior    2018-11-02     PARCIAL   SEVILLA
## 5                         Unknown    2018-11-02     PARCIAL   CÁCERES
## 6     FP I, Ciclo de Grado Medio    2018-11-02     PARCIAL  ALICANTE
##   SalarioMax SalarioMin        SubCategoria
## 1       1600       1000         ELECTRÓNICA
## 2        105        100 FORESTAL/JARDINERÍA
## 3        105        100  SERVICIO DOMÉSTICO
## 4       1000        750            MEDICINA
## 5        600        600        ELECTRICIDAD
## 6       1000        900     ADMINISTRATIVOS
```

Ara que tenim la variable FechaCreacion com a tipus Date, podriem filtrar les ofertes i quedar-nos només amb les publicades al 2018.

```
offers <- subset(offers, FechaCreacion >= "2018-01-01")
```

Després de fer aquesta neteja, encara podriem comprovar si les nostres dades son consistents. Per fer això podriem mirar si tenim ofertes en les que el salari mínim sigui mallor que el salari màxim, i de ser així, eliminar-les del nostre dataset.

```
length(subset(offers, SalarioMax < SalarioMin)$SalarioMax)
```

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

Veiem que efectivament, tenim ofertes amb dades inconsistents. Procedirem doncs a eliminar-les.

```
offers <- subset(offers, SalarioMax >= SalarioMin)
```

En aquest punt, ens adonem que hi ha un tipus de registres en els quals tenim 0 a salari minim i máxim, lo que vol dir que aquestes ofertes no han introduit un valor real en quant als salaris, o bé son ofertes de pràctiques no remunerades. Ninguna d'aquestes opcions les volem contemplar en el nostre estudi, així que com tenim dades suficients, podem prescindir d'aquestes.

```
offers <- subset(offers, !(SalarioMin == 0 & SalarioMax == 0))
```

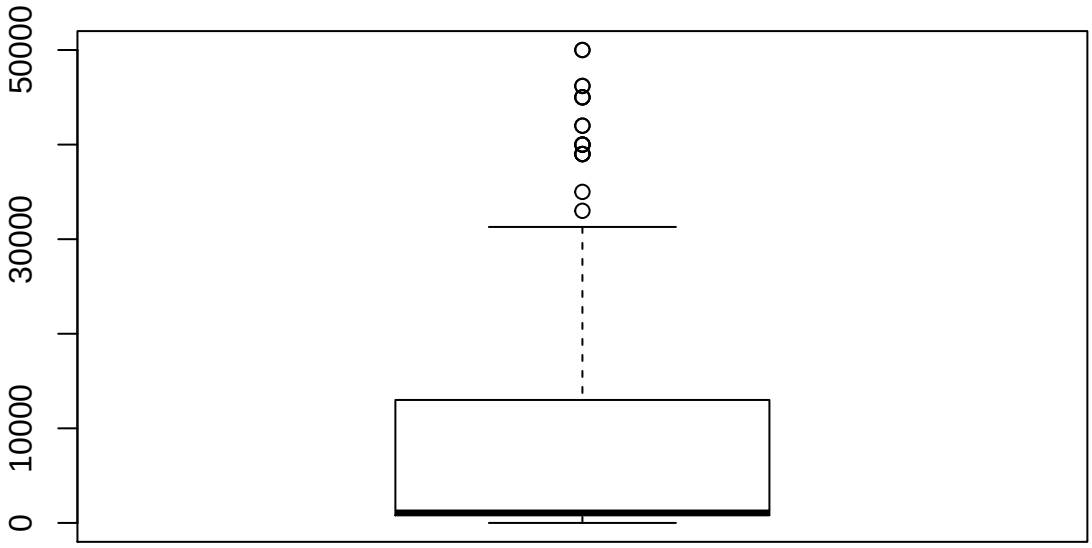
Finalment, podriem exportar el nostre conjunt de dades netejat.

```
write.csv(offers, "../data/spanish_job_offers_clean.csv")
```

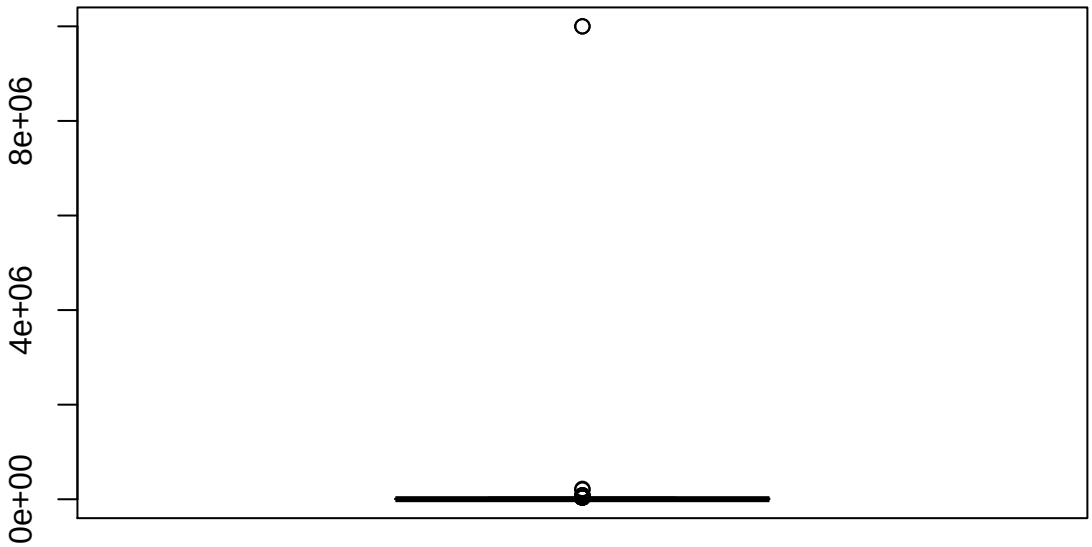
Identificació i tractament de valors extrems.

Donem ara un cop d'ull a les dades per tal d'identificar valors extrems.

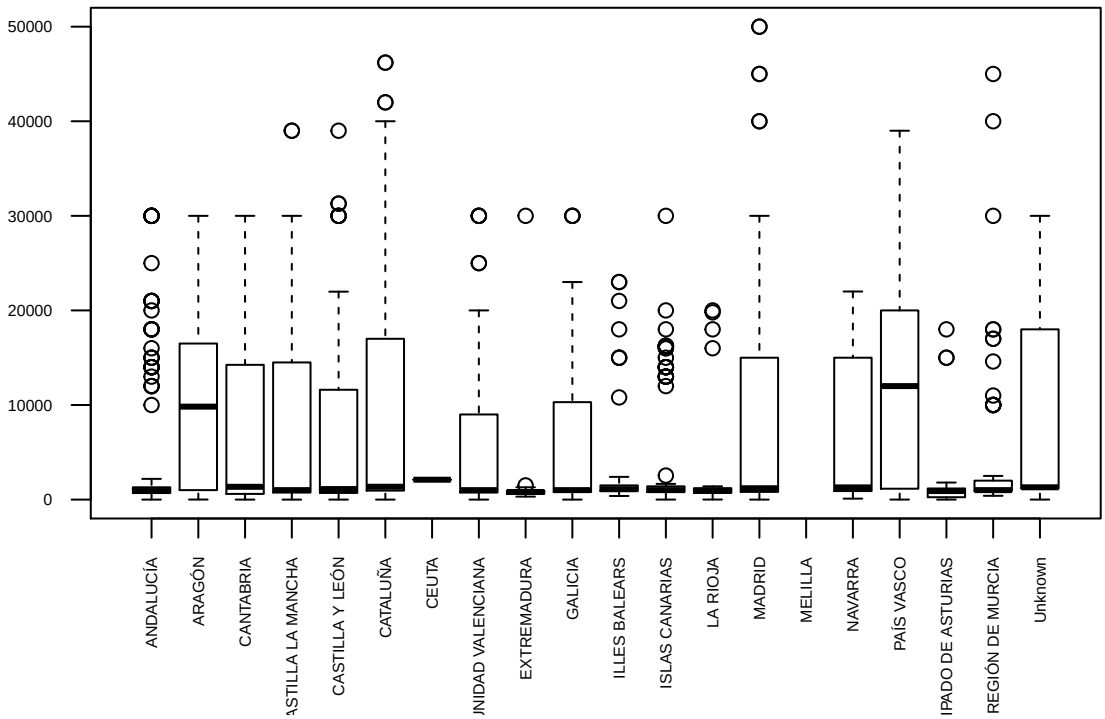
```
boxplot(offers$SalarioMin)
```



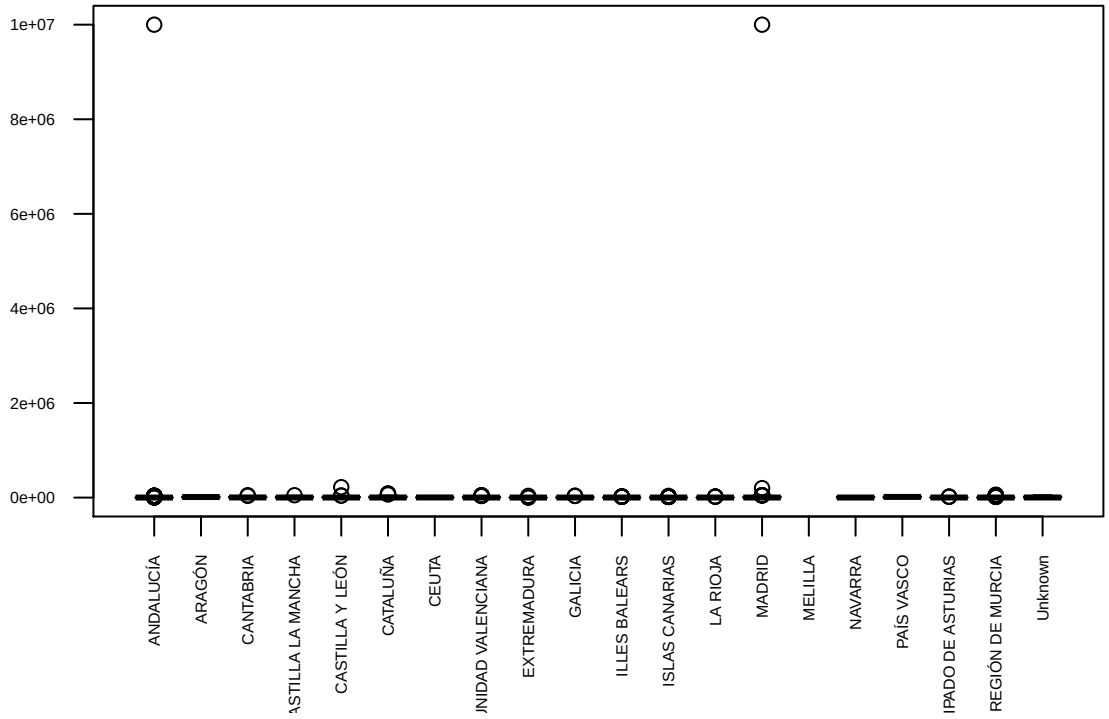

```
boxplot(offers$SalarioMax)
```




```
boxplot(offers$SalarioMin ~ offers$Comunidad, las=2, cex.axis= 0.50)
```

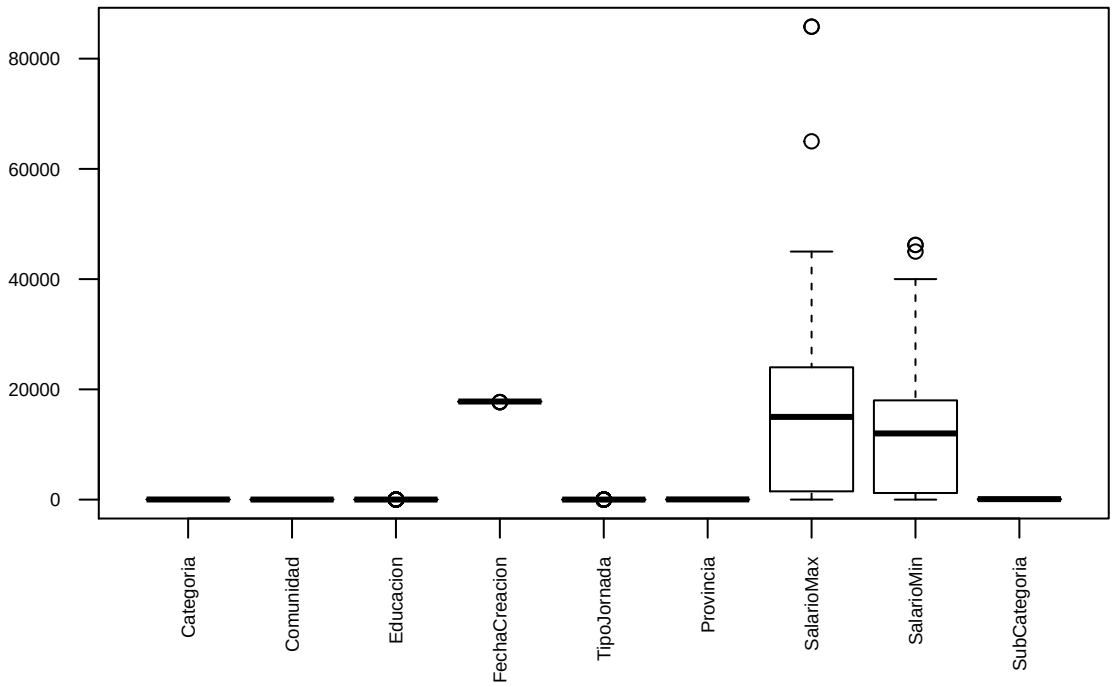



```
boxplot(offers$SalarioMax ~ offers$Comunidad, las=2, cex.axis= 0.50)
```



Com a comprovació extra, podem mirar si existeixen valors extrems al filtrar per la categoria: Informática/telecomunicacions.

```
boxplot(subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES"), las=2, cex.axis= 0.60)
```



Seguidament, podem veure també quins son els valors extrems que trobem a les nostres dades.

```
boxplot.stats(offers$SalarioMin)$out
```

```
##  [1] 46200 33000 40000 45000 45000 50000 50000 40000 40000 39000 39000
## [12] 40000 45000 42000 42000 39000 39000 35000 46200
```

```
boxplot.stats(offers$SalarioMax)$out
```

```
##  [1]   85800   50000   50000   60000   40000   45000   50000   50000
##  [9]   40000   40000   42000   50000   42000   50000   39000   50000
## [17]   50000   65000   50000   50000   40000 9999999   40000  200000
## [25]   45000 9999999   50000   43000   50000   42000   40000   40000
## [33]   40000   40000   40000   39000   50000   43000   50000   42000
## [41]   40000   40000   40000   40000   40000   40000   39000   40000
## [49]   39000   40000   39000  220000   38000   40000   42000   85800
```

Veiem que tenim valors extrems tant en els salaris màxims com en els mínims. En el cas dels salaris mínims, son valor raonables, i crec que els hauriem de deixar tal qual son. En canvi, trobem dos valors extrems molt curiosos, que semblen ser alguna mena de valor prefixat per a no donar un límit superior. En aquest cas, ja que son només dos valors i tenim suficient dades per al nostre estudi, considero que lo millor sería treure les dades corresponents. Així doncs, ho farem de la següent manera.

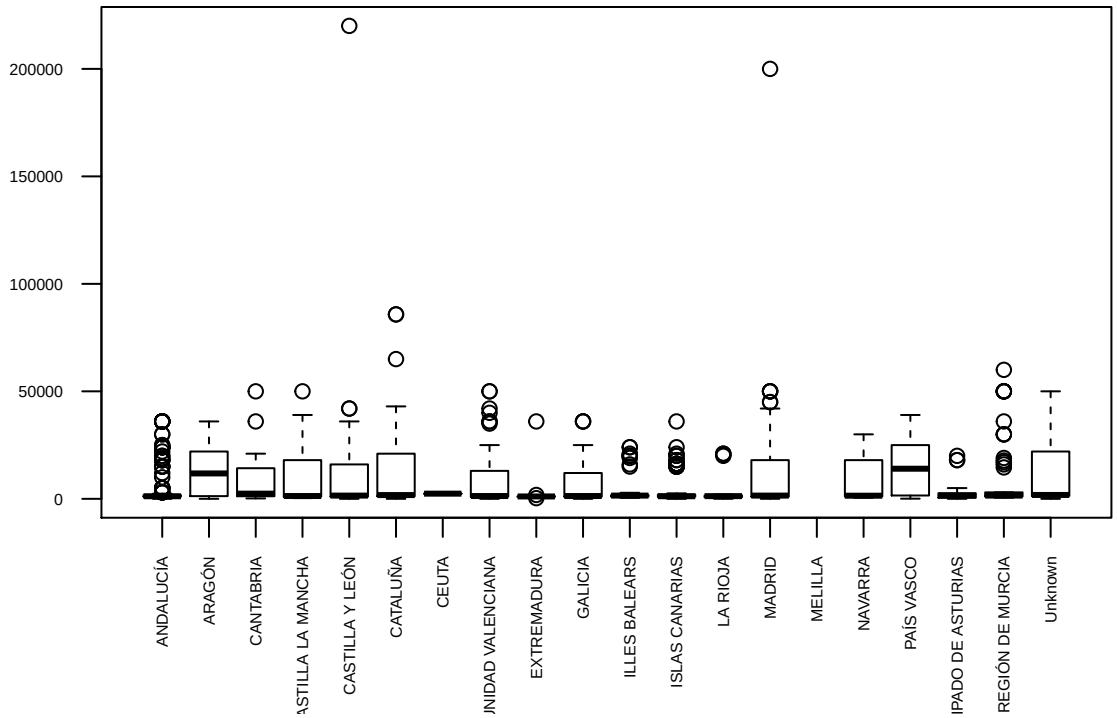
```
subset(offers, SalarioMax == 9999999)
```

```
##             Categoria Comunidad Educacion FechaCreacion TipoJornada
## 17853 COMERCIAL/VENTAS    MADRID   Unknown    2018-09-11 INDIFERENTE
## 18691 COMERCIAL/VENTAS ANDALUCÍA   Unknown    2018-09-11 INDIFERENTE
##       Provincia SalarioMax SalarioMin                      SubCategoria
## 17853    MADRID    9999999          0 AGENTES COMERCIALES/REPRESENTANTES
## 18691   SEVILLA    9999999          0 AGENTES COMERCIALES/REPRESENTANTES
```

```
offers <- subset(offers, !(SalarioMax==9999999))
```

```
offers$SalarioMin <- as.numeric(offers$SalarioMin)
```

```
offers$SalarioMax <- as.numeric(offers$SalarioMax)
boxplot(offers$SalarioMax ~ offers$Comunidad, las=2, cex.axis= 0.50)
```



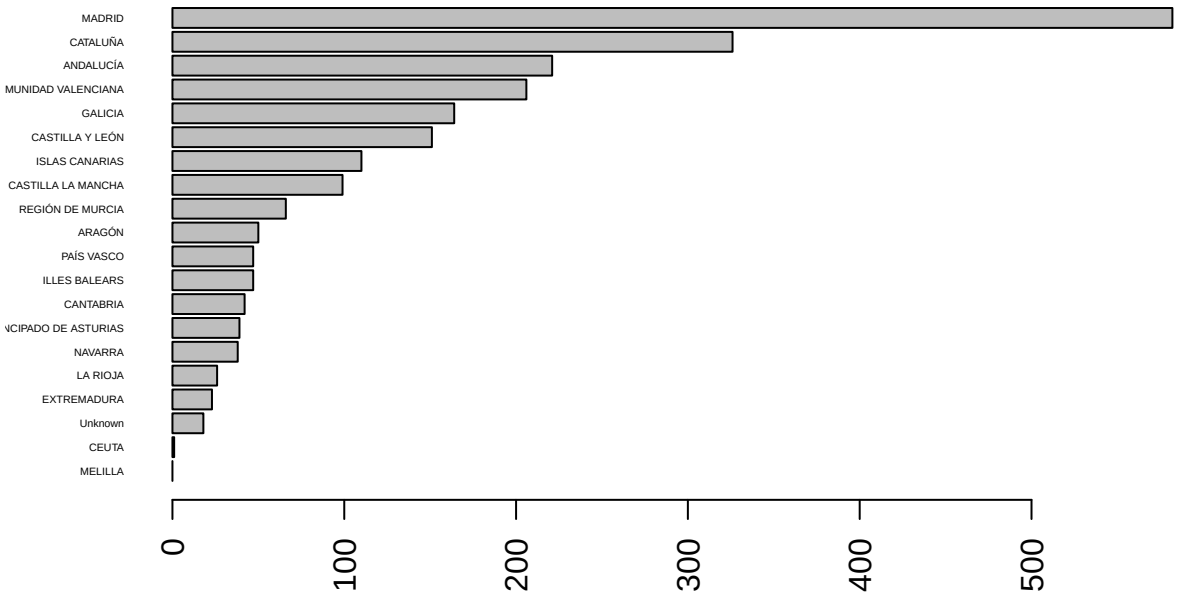
Veiem que els valors extrems que tenim ara son més raonables, i considero que els podriem deixar tal qual.

4. Anàlisi de les dades

Selecció dels grups de dades que es volen analitzar/comparar

Per al nostre estudi ens interesen les dades totals dels salaris máxims i mínims, però també estudiarem les característiques d'aquests salaris ofertats en les 4 communitats autónomes amb més ofertes.

```
barplot(sort(summary(offers$Comunidad)), horiz = TRUE, las=2, cex.names = 0.30)
```



Amb aquest gràfic ja podem donar resposta a la primera de les preguntes plantejades: Quines regions d'Espanya generen més ofertes de treball?

Podem veure que les 5 primeres regions amb més ofertes de treball son: **Madrid, Catalunya, Andalucía, Comunitat Valenciana i Galicia**.

Creem doncs els grups a estudiar.

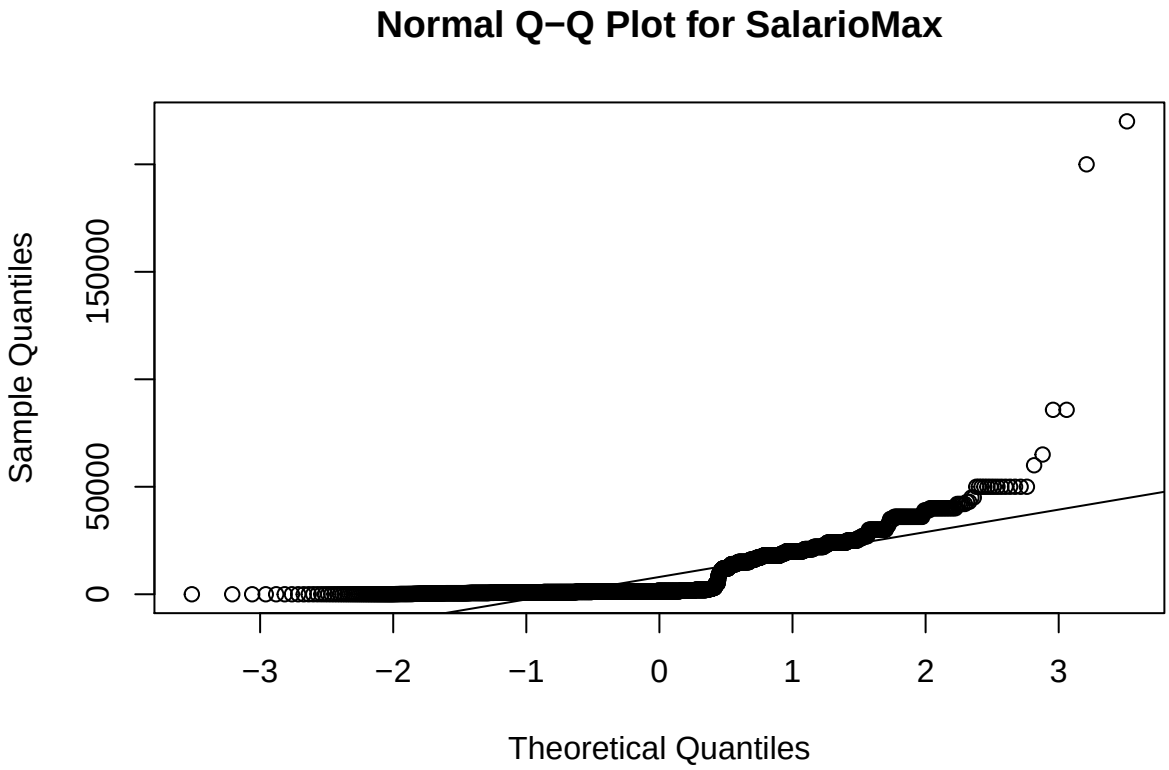
```
offers_mad <- subset(offers, Comunidad=="MADRID")
offers_cat <- subset(offers, Comunidad=="CATALUÑA")
offers_and <- subset(offers, Comunidad=="ANDALUCÍA")
offers_val <- subset(offers, Comunidad=="COMUNIDAD VALENCIANA")
offers_gal <- subset(offers, Comunidad=="GALICIA")
```

Comprovació de la normalitat i homogeneïtat de la variància.

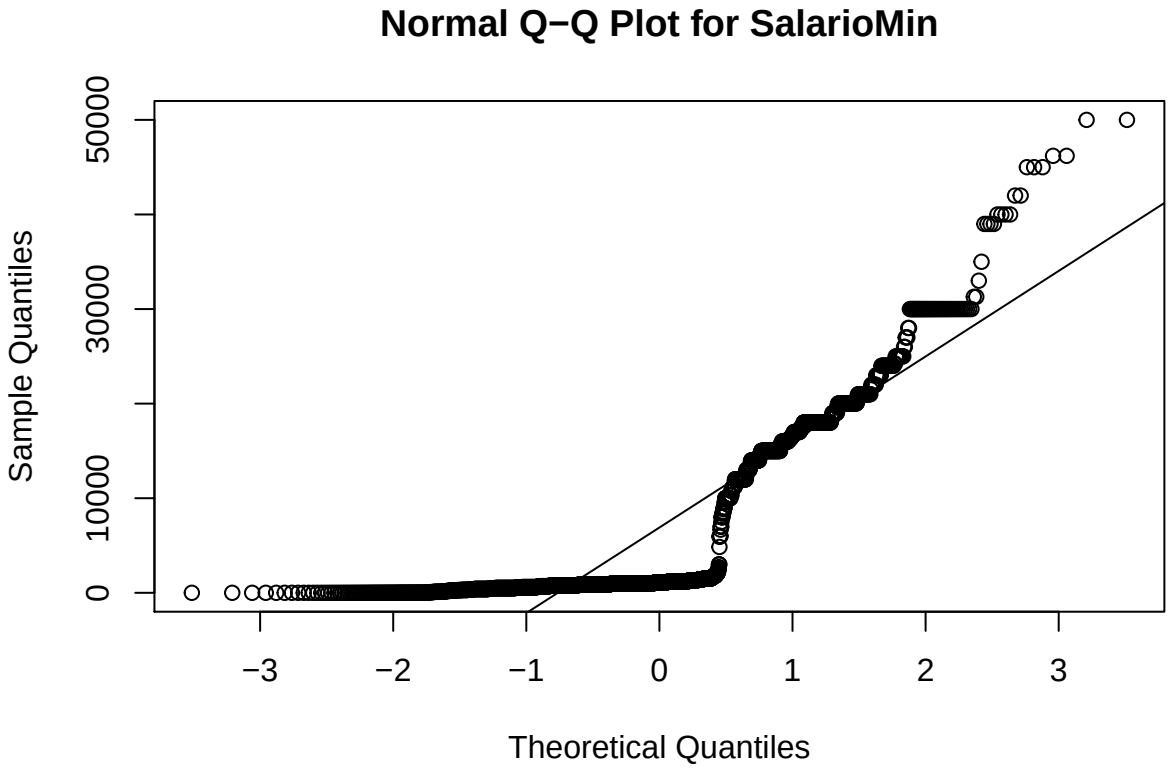
Comprovació de la Normalitat

Comprovarem ara si les nostres variables d'interés pertanyen a una distribució normal. Començarem amb una inspecció visual als salaris mínims i màxims.

```
qqnorm(offers$SalarioMax, main = "Normal Q-Q Plot for SalarioMax")
qqline(offers$SalarioMax)
```




```
qqnorm(offers$SalarioMin, main = "Normal Q-Q Plot for SalarioMin")
qqline(offers$SalarioMin)
```

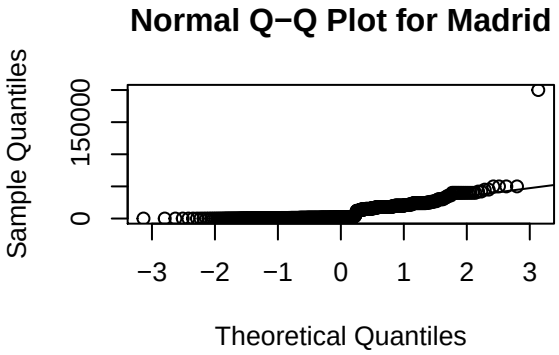


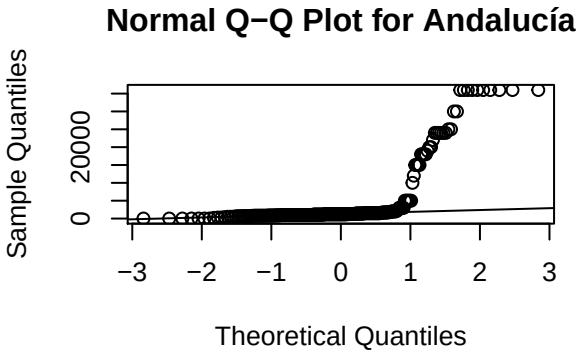
Veiem que tant els salaris mínims com màxims no s'ajusten gaire bé a la normalitat. Mirarem ara com es comporten els grups a estudiar.

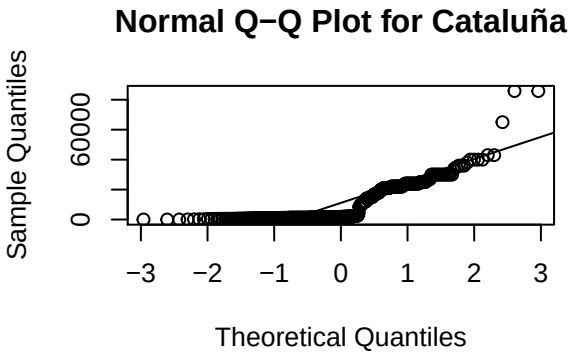
```
comunity_names <- c("Madrid", "Cataluña", "Andalucía", "Valencia", "Galicia")
comunity_df_list <- list(offers_mad, offers_cat, offers_and, offers_val, offers_gal)
i <- 1
layout(matrix(c(1,2,3,4,5),2,2))
```

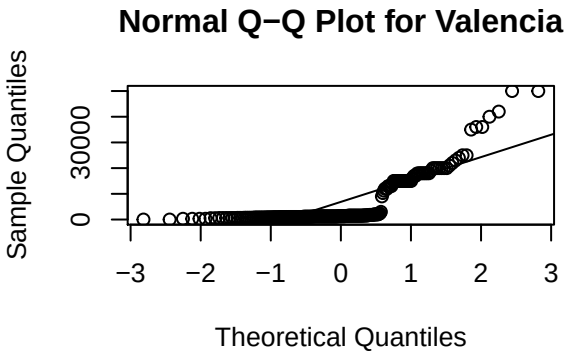
```
## Warning in matrix(c(1, 2, 3, 4, 5), 2, 2): data length [5] is not a sub-
## multiple or multiple of the number of rows [2]
```

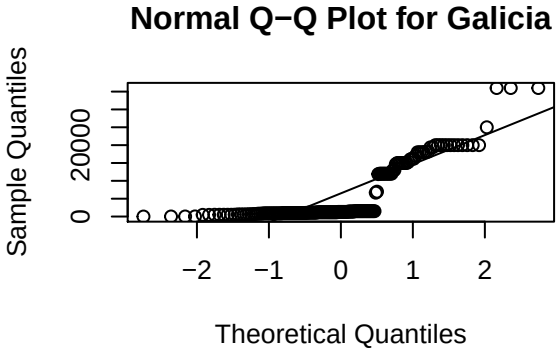
```
for(data in comunity_df_list){
  qqnorm(data$SalarioMax, main=paste("Normal Q-Q Plot for", comunity_names[i]))
  qqline(data$SalarioMax)
  i<-i+1
}
```







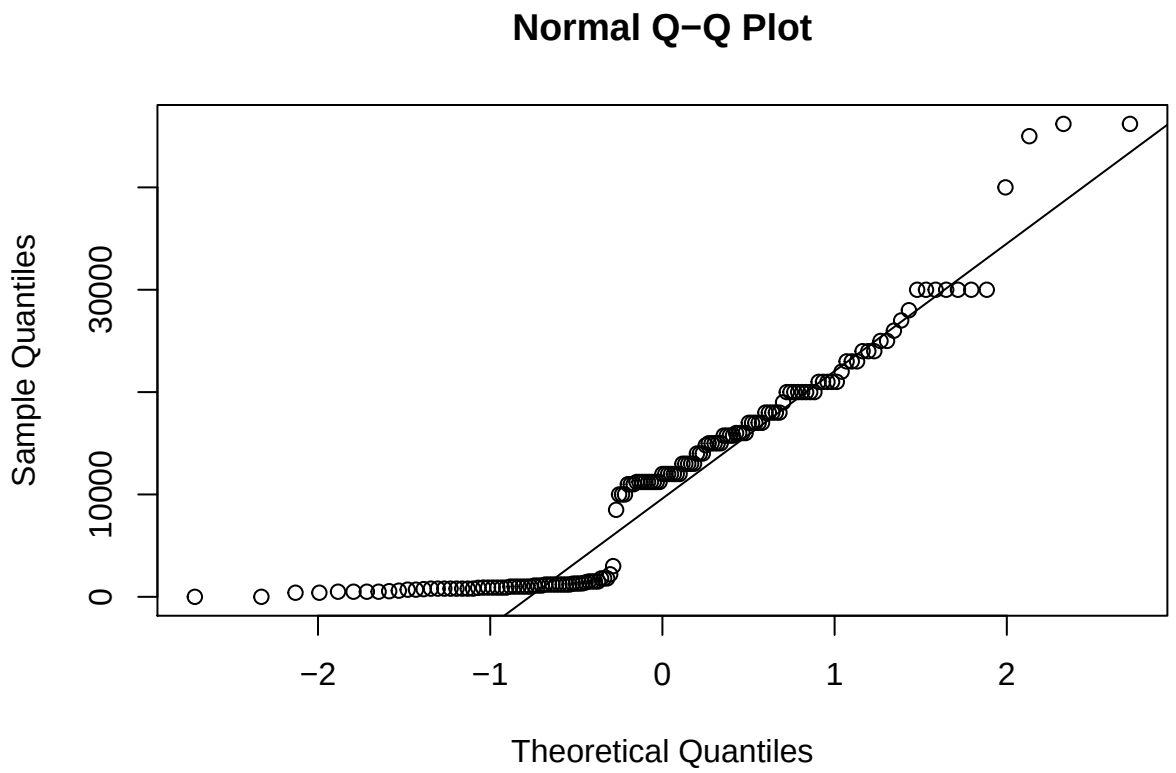




Veiem a les gràfiques que tampoc s'ajusten gairebé a la normalitat.

Igual que a l'apartat anterior, podem comprovar ara si la distribució es normal un cop filtrades les ofertes per a la categoria d'informàtica/telecomunicacions.

```r
qqnorm(subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES")$SalarioMin)
qqline(subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES")$SalarioMin)
```




```r
shapiro.test(subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES")$SalarioMin)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES")$SalarioMin
## W = 0.88047, p-value = 1.096e-09
```

Veiem que tampoc s'ajusten gairebé a la normalitat.

Podriem també realitzar els càlculs mitjançant un test de Shapiro-Wilk.

```r
shapiro.test(offers$SalarioMax)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  offers$SalarioMax
## W = 0.60708, p-value < 2.2e-16
```

```r
shapiro.test(offers$SalarioMin)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  offers$SalarioMin
## W = 0.70678, p-value < 2.2e-16
```

Veiem que els resultats dels tests ens ofereixen les mateixes conclusions, i es que no tenim una distribució

normal per aquestes dues variables.

Passem ara a comprovar també la normalitat mitjançant el mateix test sobre els salaris mínims i màxims de les ofertes publicades a les 5 regions amb més ofertes.

```
salarioMax_pvalues <- numeric(5)
i <- 1

for(data in comunity_df_list){
  p_val <- shapiro.test(data$SalarioMax)
  salarioMax_pvalues[i] <- p_val$p.value
  i <- i+1
}
```

Fem el mateix test per al salari mínim.

```
salarioMin_pvalues <- numeric(5)
i <- 1

for(data in comunity_df_list){
  p_val <- shapiro.test(data$SalarioMin)
  salarioMin_pvalues[i] <- p_val$p.value
  i <- i+1
}
```

I mostrem els resultats en la següent taula.

```
salaries_pvalues <- cbind(pval_min=salarioMin_pvalues, pval_max=salarioMax_pvalues)
rownames(salaries_pvalues) <- comunity_names
salaries_pvalues
```

```
##               pval_min     pval_max
## Madrid    4.434943e-28 1.127035e-33
## Cataluña  9.359195e-22 6.452120e-23
## Andalucía 1.305102e-24 1.845279e-24
## Valencia  2.076444e-20 1.172722e-20
## Galicia   4.720774e-17 3.615451e-17
```

En tots els casos veiem que el valor de p_value no supera el 0.05, amb la qual cosa no passen el test de normalitat, es a dir, no segueixen una distribució normal.

Un cop més, com a cas extra, podriem veure com es comporta el conjunt de dades per a la categoria d'informàtica/telecomunicacions.

```
shapiro.test(subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES")$SalarioMin )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES")$SalarioMin
## W = 0.88047, p-value = 1.096e-09
```

```
shapiro.test(subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES")$SalarioMax )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES")$SalarioMax
## W = 0.82213, p-value = 2.901e-12
```

Veiem que en aquest cas tampoc tenim un p-value superior a 0.05, amb la qual cosa podem dir que tampoc segueix una distribució normal.

Homogeneïtat de la Variància

Seguidament podriem mirar la homogeneïtat de la variància. Per a tal efecte, podriem aplicar un test de Fligner-Killeen.

```
fligner.test(SalarioMin ~ SalarioMax, data=offers )
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SalarioMin by SalarioMax
## Fligner-Killeen:med chi-squared = 1028.6, df = 282, p-value <
## 2.2e-16
```

```
fligner.test(SalarioMax ~ SalarioMin, data=offers )
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SalarioMax by SalarioMin
## Fligner-Killeen:med chi-squared = 853.04, df = 278, p-value <
## 2.2e-16
```

Com que el valor de p_value es menor que 0.05, no podem acceptar la hipòtesi nul · la de que les variàncies son homogenies. Això es compleix en tots dos casos.

Fem ara els mateixos tests per a les diferents regions.

```
salarioMin_vector <- numeric(5)
i <- 1

for(data in comunity_df_list){
  p_val <- fligner.test(SalarioMin ~ SalarioMax, data=data)
  salarioMin_vector[i] <- p_val$p.value
  i <- i+1
}
```

```
salarioMax_vector <- numeric(5)
i <- 1
for(data in comunity_df_list){
  p_val <- fligner.test(SalarioMax ~ SalarioMin, data=data)
  salarioMax_vector[i] <- p_val$p.value
  i <- i+1
}
```

Podem crear una taula per visualitzar millor els resultats

```
sal_results <- cbind(pval_min=salarioMin_vector, pval_max=salarioMax_vector, deparse.level = 2)
rownames(sal_results) <- comunity_names
sal_results
```

```
##               pval_min     pval_max
## Madrid    1.251128e-19 1.108972e-18
## Cataluña  8.389376e-08 7.466924e-03
## Andalucía 6.131947e-05 2.441518e-03
```

```
## Valencia  1.636001e-01 1.075477e-05
## Galicia   7.421507e-03 4.421103e-06
```

Amb aquests resultats podem veure que no tenim homogeneïtat a les variables Salari{Min,Max} i tampoc a les mateixes variables per regions a estudiar amb una única excepció, la variable de salari mínim a la comunitat valenciana.

Com a la resta de la pràctica, podem mirar com es comporta el conjunt de dades un cop filtrat per la categoria d'informàtica i telecomunicacions.

```
fligner.test(SalarioMin ~ SalarioMax, data=subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES"
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SalarioMin by SalarioMax
## Fligner-Killeen:med chi-squared = 61.092, df = 43, p-value =
## 0.03599
```

Veiem que es comporta d'igual manera i que tampoc hi ha homogeneïtat de la variància.

Aplicació de proves estadístiques per comparar els grups de dades.

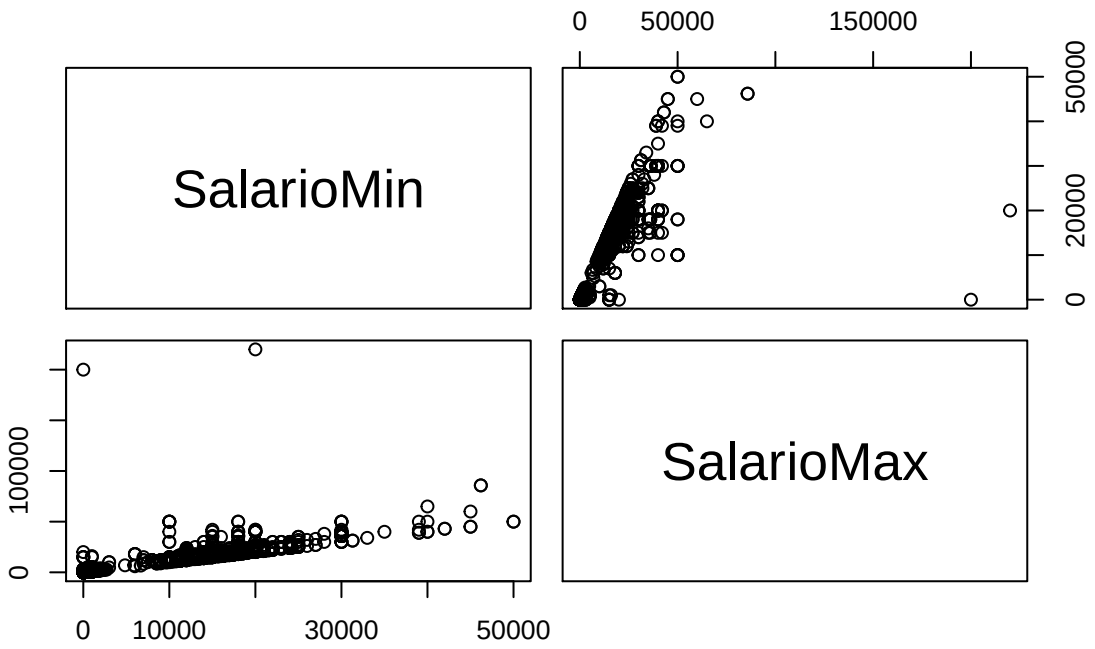
Correlacions

Lógicament, la correlació del salari mínim i màxim ha de ser positiva, en la que a major salari mínim tenim també un major salari màxim. De totes maneres, per il · lustrar aquesta relació, podem calcular el valor de correlació d'aquestes dues variables.

```
cor_matrix <- cor(offers$SalarioMin, offers$SalarioMax)
round(cor_matrix, 2)
```

```
## [1] 0.84
```

```
pairs(~SalarioMin+SalarioMax,data=offers,
   main="Simple Salary Scatterplot Matrix")
```

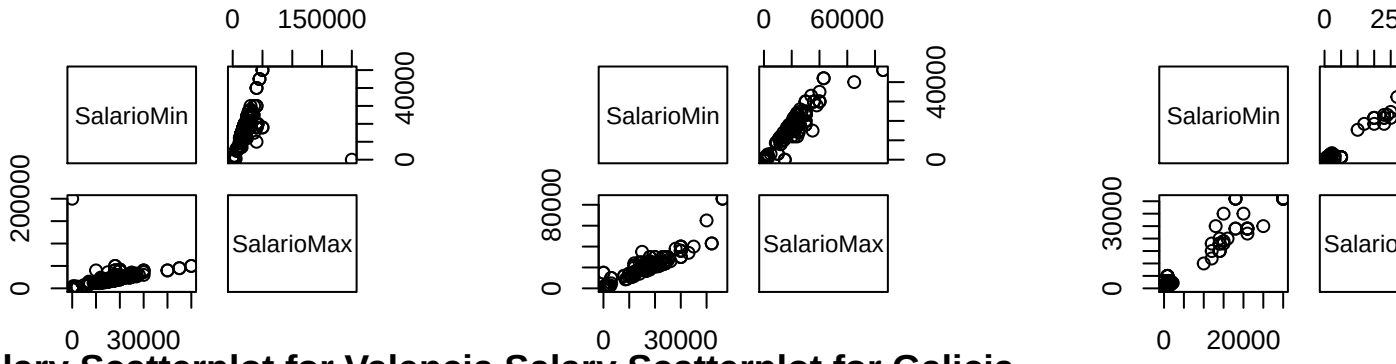


Veiem que el coeficient de correlació entre les variables de salari mínim i màxim s'aproxima molt a 1, donant-nos així el resultat esperat, es a dir, que hi existeix correlació.

També podem comprovar aquesta relació a la gràfica d'amunt, a on es mostra gràficament aquesta correlació entre salari mínim i màxim.

Podriem comprovar ara les mateixes correlacions per a les regions que estem analitzant.

```
cor_vector <- numeric(5)
i <- 1

layout(matrix(c(1,2,3,4,5),1,1))
for(data in comunity_df_list){
  cor_matrix <- cor(data$SalarioMin, data$SalarioMax)
  cor_vector[i] <- round(cor_matrix, 2)
  pairs(~SalarioMin+SalarioMax,data=data,
        main=paste("Salary Scatterplot for",
                   comunity_names[i]))
  i <- i+1
}
```

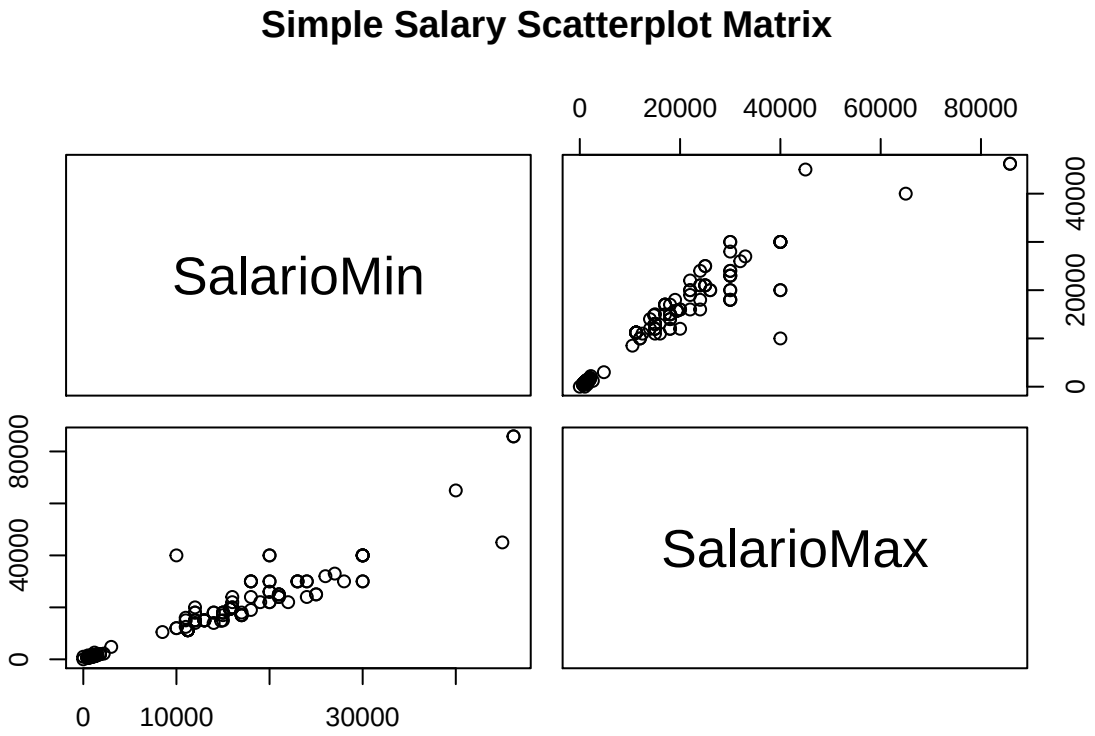



```
cor_table <- cbind(correlation=cor_vector)
rownames(cor_table) <- comunity_names
cor_table
```

```
##           correlation
## Madrid           0.76
## Cataluña         0.96
## Andalucía        0.97
## Valencia         0.94
## Galicia          0.99
```

Podem veure tant amb els valors obtinguts com a les gràfiques, que aquesta correlació existeix també quan analitzem els conjunts de dades per comunitats. Cal comentar també que a la comunitat de Madrid es a on trobem un valor de correlació inferior a la resta.

```
pairs(~SalarioMin+SalarioMax,data=subset(offers, Categoria == "INFORMÁTICA/TELECOMUNICACIONES"),
   main="Simple Salary Scatterplot Matrix")
```



D'igual manera, veiem que existeix una correlació en aquestes variables quan filtrem per categoria Informática/telecomunicaciones.

Regressió Lineal

Podriem passar ara a crear diferents models de regressió lineal per tal de predir els salaris màxims.

Començem amb un model en el que intentarem obtenir el salari màxim en funció del mínim i la comunitat autónoma de la oferta.

```
# Creem el nostre model lineal
model_1 <- lm(SalarioMax ~ SalarioMin + Comunidad, data=offers)
summary(model_1)$r.squared
```

```
## [1] 0.7101318
```

Podem veure que la qualitat del model obtingut no es massa bona amb un R squared de 0.582. Tot i així, amb aquest model podriem fer prediccions com les següents.

```
# Salari Màxim si sabem que el mínim es 20000 i l'oferta es a Catalunya.
data_pred_cat <- data.frame(Comunidad="CATALUÑA", SalarioMin=20000)
data_pred_mad <- data.frame(Comunidad="MADRID", SalarioMin=20000)
data_pred_and <- data.frame(Comunidad="ANDALUCÍA", SalarioMin=20000)
data_pred_val <- data.frame(Comunidad="COMUNIDAD VALENCIANA", SalarioMin=20000)
data_pred_gal <- data.frame(Comunidad="GALICIA", SalarioMin=20000)
```

```
predictions_model1_vector <- numeric(5)
i <- 1

for(data_pred in list(data_pred_mad, data_pred_cat, data_pred_and, data_pred_val, data_pred_gal)){
  prediction <- predict(model_1, data_pred)
  predictions_model1_vector[i] <- prediction
  i <- i+1
```

```
}

pred_m1_table <- rbind(madrid=predictions_model1_vector[1],
                       cataluña=predictions_model1_vector[2],
                       andalucia=predictions_model1_vector[3],
                       valencia=predictions_model1_vector[4],
                       galicia=predictions_model1_vector[5])
colnames(pred_m1_table) <- c("Salary_prediction_model_1")
pred_m1_table
```

```
##           Salary_prediction_model_1
## madrid                    24892.88
## cataluña                  24619.58
## andalucia                 24803.69
## valencia                  24918.34
## galicia                   24268.85
```

Podem mirar si obtenim un model millor utilitzant els dataset amb les regions amb més ofertes de treball.

```
# Creem els nostres models de regressió lineal
model_cat <- lm(SalarioMax ~ SalarioMin, data=offers_cat)
model_mad <- lm(SalarioMax ~ SalarioMin, data=offers_mad)
model_and <- lm(SalarioMax ~ SalarioMin, data=offers_and)
model_val <- lm(SalarioMax ~ SalarioMin, data=offers_val)
model_gal <- lm(SalarioMax ~ SalarioMin, data=offers_gal)

r_square_coms <- rbind(madrid=summary(model_mad)$r.squared,
                       cataluña=summary(model_cat)$r.squared,
                       andalucia=summary(model_and)$r.squared,
                       valencia=summary(model_val)$r.squared,
                       galicia=summary(model_gal)$r.squared)
colnames(r_square_coms) <- c("R squared")
r_square_coms
```

```
##           R squared
## madrid    0.5704384
## cataluña  0.9240381
## andalucia 0.9445220
## valencia  0.8796393
## galicia   0.9871338
```

En aquest cas, la qualitat del nostre model es bastant alta, amb lo qual, les nostres prediccions seràn més fiables.

Comprovem també que el model de regressió lineal per a la comunitat de Madrid té una qualitat molt més baixa, això es degut a que, com hem vist abans, el coeficient de correlació entre el salari mínim i màxim es més baix que a la resta. Això ens donarà unes pitjors prediccions per aquest conjunt de dades.

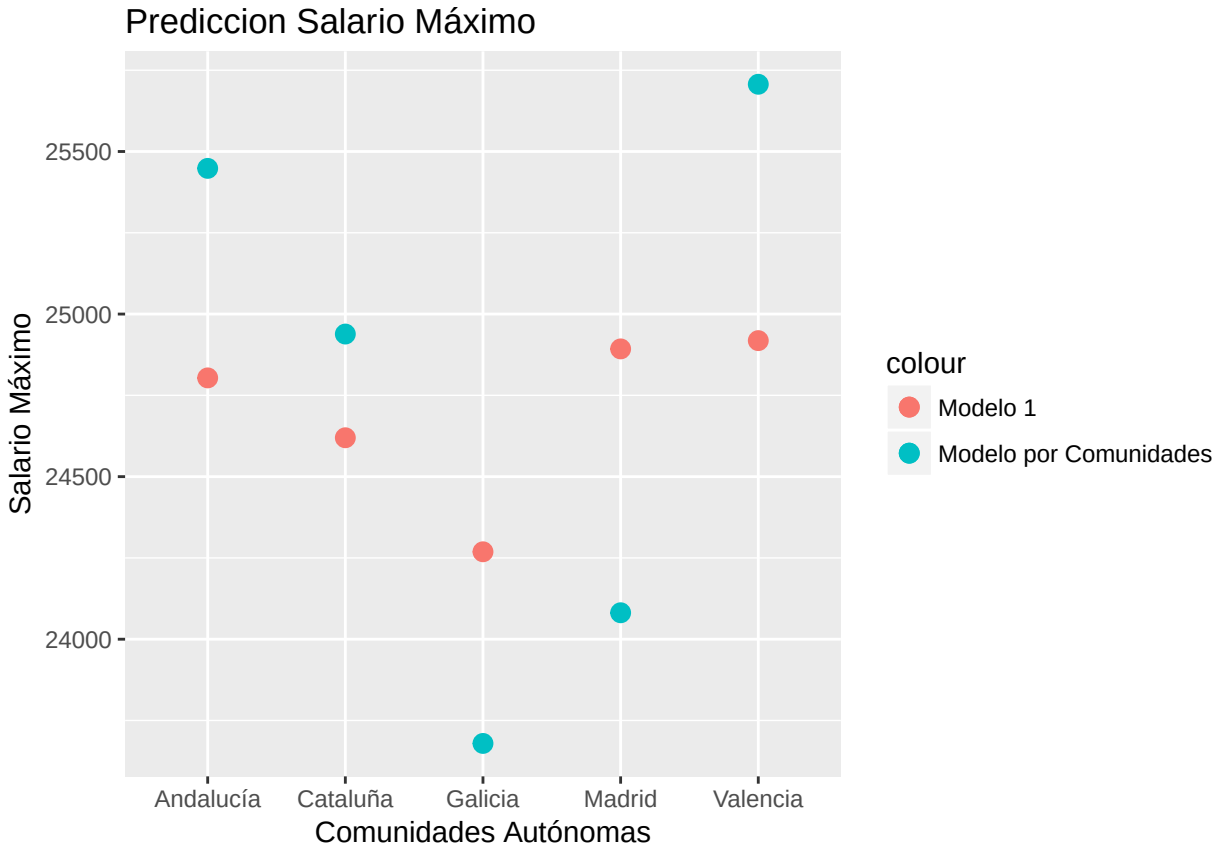
Mirem ara quina sería la predicció del salari màxim per aquest nou model.

```
data_pred <- data.frame(SalarioMin=20000)

predictions_vector <- numeric(5)
i <- 1

for(model_lm in list(model_mad,model_cat,model_and,model_val,model_gal)){
  prediction <- predict(model_lm, data_pred)
```

```
  predictions_vector[i] <- prediction
  i <- i+1
}

pred_table <- rbind(madrid=predictions_vector[1],
                    cataluña=predictions_vector[2],
                    andalucia=predictions_vector[3],
                    valencia=predictions_vector[4],
                    galicia=predictions_vector[5])
colnames(pred_table) <- c("Salary_prediction")
pred_table
```

```
##           Salary_prediction
## madrid              24081.34
## cataluña            24938.56
## andalucia           25448.12
## valencia            25706.81
## galicia             23679.43
```

```
ggplot() +
  geom_point(data = as.data.frame(pred_table),
             aes(comunity_names,
                 Salary_prediction,
                 colour='Modelo por Comunidades'),
             size = 3) +
  geom_point(data = as.data.frame(pred_m1_table),
             aes(comunity_names,
                 Salary_prediction_model_1,
                 colour = 'Modelo 1'),
             size = 3) +
  labs(x="Comunidades Autónomas", y="Salario Máximo",
       title="Prediccion Salario Máximo") +
  theme(legend.position = "right")
```



Veiem que amb aquests models podem fer prediccions sobre quant podria arribar a ser el salari màxim en funció del mínim, la comunitat i la categoria de la oferta.

Podriem mirar ara si som capaços de predir el salari màxim sense fer ús del salari mínim.

```
# Creem el nostre model lineal
model_no_salmin <- lm(SalarioMax ~ Categoria + Comunidad + Educacion + TipoJornada + FechaCreacion, data
summary(model_no_salmin)$r.squared
```

```
## [1] 0.2159268
```

Veiem que la qualitat del model obtingut és molt baixa, amb la qual cosa, si fem ús d'aquest model, les nostres prediccions serien molt poc fiables.

Per últim, podriem fer les mateixes operacions filtrant només per ofertes d'informàtica i telecomunicacions.

```
# Creem els nous models de regressió lineal
model_cat_cat <- lm(SalarioMax ~ SalarioMin + Categoria, data=offers_cat)
model_mad_cat <- lm(SalarioMax ~ SalarioMin + Categoria, data=offers_mad)
model_and_cat <- lm(SalarioMax ~ SalarioMin + Categoria, data=offers_and)
model_val_cat <- lm(SalarioMax ~ SalarioMin + Categoria, data=offers_val)
model_gal_cat <- lm(SalarioMax ~ SalarioMin + Categoria, data=offers_gal)

# Creem una taula amb els valors de r squared
r_square_coms_cat <- rbind(madrid=summary(model_mad_cat)$r.squared,
                       cataluña=summary(model_cat_cat)$r.squared,
                       andalucia=summary(model_and_cat)$r.squared,
                       valencia=summary(model_val_cat)$r.squared,
                       galicia=summary(model_gal_cat)$r.squared)
```

```r
colnames(r_square_coms_cat) <- c("R squared")
#r_square_coms_cat

# Creem el nostre dataframe per fer prediccions
data_pred_cat <- data.frame(SalarioMin=20000, Categoria="INFORMÁTICA/TELECOMUNICACIONES")

predictions_vector_cat <- numeric(5)
i <- 1

# Fem les prediccions
for(model_lm_cat in list(model_mad_cat,
                         model_cat_cat,
                         model_and_cat,
                         model_val_cat,
                         model_gal_cat)){
  prediction_cat <- predict(model_lm_cat, data_pred_cat)
  predictions_vector_cat[i] <- prediction_cat
  i <- i+1
}

# Recollim resultats en una taula
pred_table_cat <- rbind(madrid=predictions_vector_cat[1],
                    cataluña=predictions_vector_cat[2],
                    andalucia=predictions_vector_cat[3],
                    valencia=predictions_vector_cat[4],
                    galicia=predictions_vector_cat[5])
colnames(pred_table_cat) <- c("Salary_prediction_cat")
pred_table_cat <- cbind(R_squared=r_square_coms_cat, salary_predictions_cat=pred_table_cat)
pred_table_cat
```

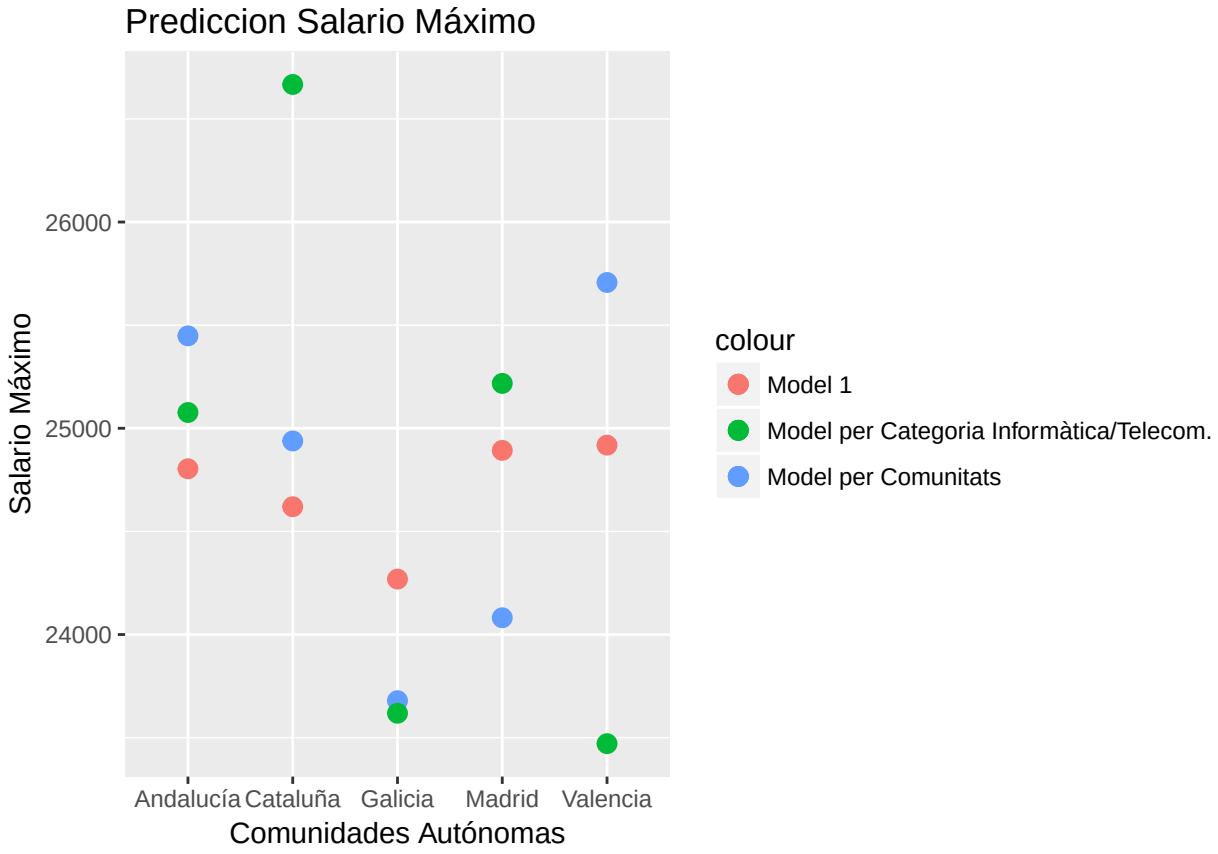
```
##           R squared Salary_prediction_cat
## madrid    0.5906161              25217.63
## cataluña  0.9300864              26666.90
## andalucia 0.9572046              25076.56
## valencia  0.9023168              23471.06
## galicia   0.9898928              23618.46
```

Veiem que la qualitat dels models son una mica millors en aquest cas.

Per últim, podem comparar gràficament tots els resultats obtinguts en les nostres prediccions.

```r
ggplot() +
  geom_point(data = as.data.frame(pred_table),
             aes(comunity_names,
                 Salary_prediction,
                 colour='Model per Comunitats'),
             size = 3) +
  geom_point(data = as.data.frame(pred_m1_table),
             aes(comunity_names,
                 Salary_prediction_model_1,
                 colour = 'Model 1'),
             size = 3) +
  geom_point(data = as.data.frame(pred_table_cat),
             aes(comunity_names,
                 Salary_prediction_cat,
```

```
                colour = 'Model per Categoria Informàtica/Telecom.'),
            size = 3) +
labs(x="Comunidades Autónomas", y="Salario Máximo",
     title="Prediccion Salario Máximo") +
theme(legend.position = "right")
```



Resulta interessant veure com les nostres prediccions fetes amb el primer model (amb una qualitat inferior) ens dona com a resultat unes prediccions amb un rang bastant similar, mentres que en els models més especialitats s'observen majors diferències entre regions.

També es interessant veure com les nostres prediccions per a la categoria d'informàtica i telecomunicacions ens donen diferències positives molt grans per a Catalunya i Madrid, i negatives per a la resta de regions.

Contrasts d'Hipòtesi

Seguidament podriem fer un contrast entre regions per tal de donar resposta a la pregunta: a Catalunya es generen ofertes de treball amb un salari mínim superior a la resta d'Espanya?

Per a tal motiu utilitzarem necessitarem preparar el conjunt de dades de mostres de comunitats diferents a Catalunya.

```
offers_no_cat <- subset(offers, !(Comunidad=="CATALUÑA"))
```

Per aquest estudi utilitzarem la següent hipòtesi nul · la i alternativa:

$$H0 : \mu_1 - \mu_2 = 0$$

$$H1 : \mu_1 - \mu_2 > 0$$

Farem servir un contrast sobre la diferència de mitjanes. Com que les nostres dues mostres tenen més de 30 observacions, gràcies al teorema del límit central podrem considerar-les com distribucions normals.

```
t_test_no_cat <- t.test(offers_cat$SalarioMin, offers_no_cat$SalarioMin, alternative = "greater")
t_test_no_cat
```

```
##
##  Welch Two Sample t-test
##
## data:  offers_cat$SalarioMin and offers_no_cat$SalarioMin
## t = 3.5063, df = 403.5, p-value = 0.0002527
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1105.215      Inf
## sample estimates:
## mean of x mean of y
##  8102.466  6016.414
```

Com que el p-value es menor que 0.05, podem rebutjar la hipòtesi nul · la a favor de la hipòtesi alternativa, lo que vol dir que podem afirmar que el salari mínim les ofertes de treball a Catalunya sigui més alt que a la resta d'Espanya.

Mirem ara si això es compleix també en la resta de comunitats que volem estudiar.

```
offers_no_mad <- subset(offers, !(Comunidad=="MADRID"))
t_test_no_mad <- t.test(offers_mad$SalarioMin, offers_no_mad$SalarioMin, alternative = "greater")
offers_no_and <- subset(offers, !(Comunidad=="ANDALUCÍA"))
t_test_no_and <- t.test(offers_and$SalarioMin, offers_no_and$SalarioMin, alternative = "greater")
offers_no_val <- subset(offers, !(Comunidad=="COMUNIDAD VALENCIANA"))
t_test_no_val <- t.test(offers_val$SalarioMin, offers_no_val$SalarioMin, alternative = "greater")
offers_no_gal <- subset(offers, !(Comunidad=="GALICIA"))
t_test_no_gal <- t.test(offers_val$SalarioMin, offers_no_gal$SalarioMin, alternative = "greater")
```

```
results_coms <- rbind(t_test_no_mad$p.value, t_test_no_cat$p.value, t_test_no_and$p.value, t_test_no_val
rownames(results_coms) <- comunity_names
colnames(results_coms) <- c("p_values")
results_coms
```

```
##               p_values
## Madrid    6.820997e-05
## Cataluña  2.526702e-04
## Andalucía 1.000000e+00
## Valencia  9.993536e-01
## Galicia   9.990527e-01
```

Podem veure que tant a Catalunya com a Madrid, tenim p-values menors a 0.05 que es el nostre valor de significació, amb la qual cosa rebutjem les hipòtesis nul · les en favor de les alternatives, i això ens porta a concloure que la mitjana en els salaris mínims de les ofertes de treball a Catalunya i Madrid estàn per sobre de les ofertes a la resta del país. En canvi, a Andalucía, València i Galicia tenim els casos contraris.

Per últim, podem fer el mateix exercici per veure si el salari mínim a la categoria d'informàtica i telecomunicacions es superior a la resta de categories.

```
offers_inf <- subset(offers, (Categoria=="INFORMÁTICA/TELECOMUNICACIONES"))
offers_no_inf <- subset(offers, !(Categoria=="INFORMÁTICA/TELECOMUNICACIONES"))
```

```r
t_test_inf <- t.test(offers_inf$SalarioMin, offers_no_inf$SalarioMin, alternative = "greater")
t_test_inf
```

```
##
##  Welch Two Sample t-test
##
## data:  offers_inf$SalarioMin and offers_no_inf$SalarioMin
## t = 6.6344, df = 163.77, p-value = 2.267e-10
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4409.181      Inf
## sample estimates:
## mean of x mean of y
## 11798.437  5924.712
```

Com que també obtenim un valor p-value menor a 0.05, rebutjem l'hipòtesi nul · la en favor de l'alternativa, amb la qual cosa podem dir que els salaris a la categoria d'informàtica i telecomunicacions estàn per sobre de la resta de categories.

5. Representació dels resultats a partir de taules i gràfiques

Per a una millor comprensió dels resultats, s'han anat afegint diferentes taules i gràfiques al llarg d'aquesta pràctica. Com que aquest treball dona resposta a diferentes preguntes, considero que el millor lloc per a aquestes gràfiques és al costat dels seus exercicis previs. Per tant, no tornaré a generar-les en aquest apartat, ja que supondria una repetició que no aportaría nova informació.

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

Les tres preguntes plantejades a l'inici de la pràctica eran les següents:

1. Quines regions d'Espanya generen més ofertes de treball?
2. Podem fer prediccións sobre salaris mínims i màxims?
3. Estudi sobre els salaris en relació a les 5 regions que generen més ofertes. Tenim regions amb salari mínim superior a la resta? I a la categoria d'informàtica i telecomunicacions?

Per tal de respondre a la primera pregunta, hem pogut veure tant gràficament com al sumari de les nostres dades, que les regions amb més ofertes publicades son Madrid, Catalunya, Andalusia, Comunitat Valenciana i Galícia.

Com hem pogut veure al llarg de la pràctica, amb les dades existents som capaços d'elaborar un model predictiu bassat en regressió lineal per tal de predir els salaris màxims (o mínims si volguessim), en funció dels salaris mínims, donant resposta així a la segona pregunta. També hem provat a generar un model de regressió lineal que ens permetés predir el salari màxim en funció d'altres variables que no siguin el salari mínim, però la qualitat d'aquest models son molt dolentes i les seves prediccions no serien molt correctes.

Per últim, per tal de resoldre la tercera pregunta, hem pogut comprobar que de les cinc regions amb més ofertes de treball públicades, a les regions de Catalunya i Madrid trobem salaris mínims més elevats que a la resta del país, mentres que a Andalusia, València i Galícia no passa igual. També dintre d'aquesta pregunta hem observat que els salaris de la categoria d'informàtica i telecomunicacions estàn per sobre de la resta de categories.

7. Referències

Bibliografia:

- Dalgaard, Peter (2002). Introductory Statistics with R. Verlang, New York. Springer.
- Jarman, Kristin (2017). The Art of Data Analysis. New Jersey. Wiley.
- Osborne, Jason (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. North Carolina State. Elsevier.

Links:

- Quick-R. https://www.statmethods.net/index.html
- R-Bloggers. https://www.r-bloggers.com/
- Statistical Data Analysis. https://stat.ethz.ch/R-manual/
- Cookbook for R. http://www.cookbook-r.com/
- Working with Unknown Values. https://cran.r-project.org/web/packages/gdata/vignettes/unknown.pdf
- ggplot2. https://ggplot2.tidyverse.org/