PRACTICA 2: LIMPIEZA Y VALIDACIÓN DE LOS DATOS

Gines Molina e Iñigo Alvarez

29/12/2020

Contents


```
# Carga de librerías
library(dplyr)
library(VIM)
library(gridExtra)
library(ggplot2)
library(stringi)
library(stringr)
library(tmaptools)
library(polycor)
library(reshape2)
library(wordcloud)
library(wordcloud2)
library(viridis)
library(scales)
library(magrittr)
library(forcats)
```

```
# library(highcharter)
# library(maps)
# library(ggmap)
# library(ggpubr)
# library(tidyverse)
# library(plotly)
```

1. Descripción del dataset

El dataset es un listado de ofertas de trabajo publicadas en Linkedin para ciencia de datos (término "Data Scientist") que fue obtenido en la primera parte de esta práctica.

Se ha decidido la utilización de este dataset ya que se trata de un dataset propio que no ha sido ampliamente explotado a análisis y queríamos tener una experiencia más cercana a lo que sería un proyecto real. Se consiguen resultados y se llevan a cabo las técnicas que queríamos aplicar para obtener las respuestas a nuestras preguntas sobre el panorama de búsqueda de empleo.

Se pueden ver los datos y el código en el siguiente repositorio: https://github.com/InigoAB/DataCleaning

Consta de 6 archivos CSV con las ofertas encontradas a nivel de Escocia, Reino Unido, España, Mundo y remoto que unificaremos, limpiaremos y daremos formato para posteriormente analizar y buscar información relevante.

Trataremos de ir respondiendo a distintas preguntas que nos vayan durgiendo como pueden ser las siguientes: - ¿Qué variables son relevantes? - ¿Que diferencias hay segun la localizacion o paises? - ¿Es una opción los puestos en Remoto? - ¿Hay diferencias entre latitudes positivas y negativas? - ¿Existe correlación entre si hay quick application y si se trata de una empresa grande o no? - ¿Existe correlacion entre el tipo de puesto y el numero de solicitudes? - ¿Qué tipo de puestos de trabajo prioriza LinkedIn en sus resultados? - ¿Qué palabras son las más utilizadas en las descripciones de las ofertas de trabajo?

2. Integración y selección de los datos de interés a analizar

El primer paso es integrar los archivos de las distintas ubicaciones en un mismo dataframe con el que podamos trabajar.

```
Scotland <- read.csv("data/Scotland.csv", header=TRUE, sep=",", stringsAsFactors=FALSE)
Spain <- read.csv("data/Spain.csv", header=TRUE, sep=",", stringsAsFactors=FALSE)
UK <- read.csv("data/UK.csv", header=TRUE, sep=",", stringsAsFactors=FALSE)
WorldWide <- read.csv("data/WorldWide.csv", header=TRUE, sep=",", stringsAsFactors=FALSE)
Remote <- read.csv("data/Remote.csv", header=TRUE, sep=",", stringsAsFactors=FALSE)
Scotland$dataset <- "Scotland"
Spain$dataset <- "Spain"
UK$dataset <- "UK"
WorldWide$dataset <- "WorldWide"
Remote$dataset <- "Remote"
jobs <- rbind(Scotland, Spain, UK, WorldWide, Remote)
dim(jobs)
```

```
## [1] 3456   18
```

Obtenemos un dataframe de 3456 observaciones y 18 variables.

Vemos las primeras observaciones del mismo.

```
head(jobs)
```

```
##   X     Job.ID       Date Company.Name                    Title
## 1 0 2011662890 2020-10-15      Dufrain            Data Engineer
## 2 1 2227320776 2020-10-30 Sopra Steria          User Researcher
## 3 2 2201826818 2020-10-22    Sword ITS            Data Engineer
## 4 3 2284269209 2020-11-05         None Data Scientist - Glasgow
## 5 4 2264114504 2020-10-16     Experian           Data Scientist
## 6 5 2249308521 2020-10-11      Harnham           Data Scientist
##                             Location
## 1 Edinburgh, Scotland, United Kingdom
## 2 Edinburgh, Scotland, United Kingdom
## 3  Aberdeen, Scotland, United Kingdom
## 4                         Glasgow, GB
## 5                         Glasgow, GB
## 6                         Glasgow, GB
##
## 1
leading Data Management, Analytics and BI consultancy with offices across the UK, working with some of
leading Data Management and Analytics Consultancy. We are looking for dynamic individuals, with proven
 (Certified to practitioner level in at least one provider) Experience working with one or more of the
 Spark, Kafka, Snowflake, Hadoop Experience working with both SQL and NoSQL foundational tools and datal
## 2 Bring your User Research expertise to us and in return we'll give you an amazing career with growth
to-end service offerings in the market: Consulting, Systems Integration, Software Development and Busine
to-one interviews, workshops, focus groups, ethnographic research   Essential Skills: (Design and Techni
offs Design Standards - Is confident working within pre-defined standards (e.g. legal, regulatory or des
## 3
class service to our customers. As the pre-eminent provider of Information Technology services to the En
technical people.Hands-on experience in full ETL lifecycle:Source data analysis.Data development process
starter able to work alone and as part of a small team.  For a confidential discussion regards this oppo
line in the first instance.
## 4
hosted machine learning (SaaS) algorithms, the main aim of the role for the Data Scientist will be apply
hoc analysis and presenting results in a clear manner to help inform product direction. Core skills & ex
based services, including model selection, training, configuration and deployment.Experience displaying
## 5
leading data science products and classic bureau data.  Responsibilities  Develop and maintain data proc
## 6
leading technology, using python, scala and spark.  The Company  As a data scientist you will sit within
##             Level             Type
## 1      Intermedio Jornada completa
## 2      Intermedio Jornada completa
## 3 Sin experiencia Jornada completa
## 4 Sin experiencia Jornada completa
## 5 Sin experiencia Jornada completa
## 6 Sin experiencia Jornada completa
##                                          Functions
## 1            Tecnología de la información, Ingeniería
## 2                        Tecnología de la información
## 3 Tecnología de la información, Análisis, Consultoría
## 4            Ingeniería, Tecnología de la información
## 5            Ingeniería, Tecnología de la información
## 6            Ingeniería, Tecnología de la información
##                                                                      Industries
## 1              Servicios y tecnologías de la información, Banca, Servicio de información
## 2                                          Servicios y tecnologías de la información
```

```
## 3 Servicio de información, Servicios y tecnologías de la información, Petróleo y energía
## 4  Servicios y tecnologías de la información, Software, Dotación y selección de personal
## 5             Servicios y tecnologías de la información, Software, Servicios financieros
## 6                                Marketing y publicidad, Software, Servicios financieros
##   Solicitudes     Empleados Quick.Application Emails Visualizaciones
## 1          54        51-200             False                    679
## 2          30 Más de 10.001             False                    169
## 3          71        51-200              True                    325
## 4        None          None             False                      5
## 5          10 Más de 10.001             False                     98
## 6        None        51-200             False                      4
##        Recommended.Flavor  dataset
## 1 ACTIVELY_HIRING_COMPANY Scotland
## 2 ACTIVELY_HIRING_COMPANY Scotland
## 3 ACTIVELY_HIRING_COMPANY Scotland
## 4                    None Scotland
## 5         COMPANY_RECRUIT Scotland
## 6 ACTIVELY_HIRING_COMPANY Scotland
```

Vamos a inspeccionar las variables.

```
str(jobs)
```

```
## 'data.frame':    3456 obs. of  18 variables:
##  $ X                 : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Job.ID            : chr  "2011662890" "2227320776" "2201826818" "2284269209" ...
##  $ Date              : chr  "2020-10-15" "2020-10-30" "2020-10-22" "2020-11-05" ...
##  $ Company.Name      : chr  "Dufrain" "Sopra Steria" "Sword ITS" "None" ...
##  $ Title             : chr  "Data Engineer" "User Researcher" "Data Engineer" "Data Scientist - Glas
##  $ Location          : chr  "Edinburgh, Scotland, United Kingdom" "Edinburgh, Scotland, United Kingd
##  $ Description       : chr  "Data Engineer London, Edinburgh OR Manchester   We are Dufrain. We're a
leading Data Management, Analyti"| __truncated__ "Bring your User Research expertise to us and in retur
##  $ Level             : chr  "Intermedio" "Intermedio" "Sin experiencia" "Sin experiencia" ...
##  $ Type              : chr  "Jornada completa" "Jornada completa" "Jornada completa" "Jornada comple
##  $ Functions         : chr  "Tecnología de la información, Ingeniería" "Tecnología de la información
##  $ Industries        : chr  "Servicios y tecnologías de la información, Banca, Servicio de informaci
##  $ Solicitudes       : chr  "54" "30" "71" "None" ...
##  $ Empleados         : chr  "51-200" "Más de 10.001" "51-200" "None" ...
##  $ Quick.Application : chr  "False" "False" "True" "False" ...
##  $ Emails            : chr  "" "" "" "" ...
##  $ Visualizaciones   : chr  "679" "169" "325" "5" ...
##  $ Recommended.Flavor: chr  "ACTIVELY_HIRING_COMPANY" "ACTIVELY_HIRING_COMPANY" "ACTIVELY_HIRING_COM
##  $ dataset           : chr  "Scotland" "Scotland" "Scotland" "Scotland" ...
```

Quitamos la primera columna ya que no es más que un índice que no aporta ninguna información.

```
jobs$X <- NULL
```

Todas las demás variables han quedado como de tipo carácter así que en el siguiente apartado haremos las transformaciones necesarias, de momento las explicamos.

Explicación de las variables

Nos quedamos con las siguientes columnas - Job_ID: identificador de la oferta de empleo - Date: fecha de publicación - Company_Name: nombre de la empresa - Role: título o puesto de la oferta - Location: ubicación del puesto - Description: descripción del puesto (del que se puede extraer más información) - Level: nivel de experiencia requerida para el puesto - Type: tipo de contrato (jornada completa o parcial) -

Functions: funciones del puesto de la oferta - Industries: sectores que involucra - Solicitudes: número de solicitudes enviadas a la oferta - Empleados: intervalo/número de empleados de la empresa - Quick Application (True/False): método de solicitud rápida por LinkedIn Emails: e-mails de contacto - Visualizaciones: número de visualizaciones que tiene la oferta - Recommended Flavor: tipo de oferta - dataset: dataset de origen de cada variable

3. Limpieza de los datos

Al haber integrado varios archivos de búsquedas de ámbitos territoriales que se solapan lo primero es eliminar las posibles duplicidades que podamos encontrar.

```
sum(duplicated(jobs[1:5]))
```

```
## [1] 429
```

Hay 429 observaciones duplicadas así que las eliminamos. Mantenemos los duplicados para preservar los valores de cada dataset y comparar posteriormente.

```
jobs_with_duplicates <- jobs

jobs <- jobs %>%
  distinct(Job.ID, Company.Name, Location, .keep_all = TRUE)
```

3.1. Valores perdidos

Buscamos valores perdidos a lo largo del dataframe.

```
sum(is.na(jobs))
```

```
## [1] 0
```

```
sum(jobs=="")
```

```
## [1] 2738
```

```
sum(jobs=="None")
```

```
## [1] 2496
```

Sustituimos los valores en blanco y "None" por NA.

```
jobs[jobs==""] <- NA
jobs[jobs=="None"] <- NA
jobs_with_duplicates[jobs_with_duplicates==""] <- NA
jobs_with_duplicates[jobs_with_duplicates=="None"] <- NA
```

Para corregir los tipos de datos vamos a ver qué variables se podrían convertir a tipo factor viendo la cantidad de datos distintos que tiene cada una.

```
sapply(jobs, function(x) length(unique(x)))
```

```
##             Job.ID               Date       Company.Name              Title
##               3026                 71               1916               1321
##           Location        Description              Level               Type
##                828               2913                  7                  7
##          Functions         Industries       Solicitudes          Empleados
##                290                688                444                  9
##  Quick.Application             Emails    Visualizaciones Recommended.Flavor
##                  2                220               1016                  4
```

```
##            dataset
##                  5
```

Todas las variables que tienen menos de 10 valores únicos las convertiremos a tipo factor. Date lo pasamos a formato fecha y Solicitudes y Visualizaciones a tipo numérico.

```
jobs$Date <- as.Date(jobs$Date, format="%Y-%m-%d")
jobs$Solicitudes <- as.numeric(jobs$Solicitudes)
jobs$Visualizaciones <- as.numeric(jobs$Visualizaciones)
jobs$Level <- as.factor(jobs$Level)
jobs$Type <- as.factor(jobs$Type)
jobs$Empleados <- as.factor(jobs$Empleados)
jobs$Quick.Application <- as.factor(jobs$Quick.Application)
jobs$Recommended.Flavor <- as.factor(jobs$Recommended.Flavor)
jobs$dataset <- as.factor(jobs$dataset)

jobs_with_duplicates$Date <- as.Date(jobs_with_duplicates$Date, format="%Y-%m-%d")
jobs_with_duplicates$Solicitudes <- as.numeric(jobs_with_duplicates$Solicitudes)
jobs_with_duplicates$Visualizaciones <- as.numeric(jobs_with_duplicates$Visualizaciones)
jobs_with_duplicates$Level <- as.factor(jobs_with_duplicates$Level)
jobs_with_duplicates$Type <- as.factor(jobs_with_duplicates$Type)
jobs_with_duplicates$Empleados <- as.factor(jobs_with_duplicates$Empleados)
jobs_with_duplicates$Quick.Application <- as.factor(jobs_with_duplicates$Quick.Application)
jobs_with_duplicates$Recommended.Flavor <- as.factor(jobs_with_duplicates$Recommended.Flavor)
jobs_with_duplicates$dataset <- as.factor(jobs_with_duplicates$dataset)

summary(jobs)
```

```
##     Job.ID               Date            Company.Name          Title
##  Length:3026        Min.   :2019-10-10   Length:3026        Length:3026
##  Class :character   1st Qu.:2020-10-12   Class :character   Class :character
##  Mode  :character   Median :2020-10-23   Mode  :character   Mode  :character
##                     Mean   :2020-10-21
##                     3rd Qu.:2020-11-02
##                     Max.   :2020-11-08
##                     NA's   :1
##    Location         Description                              Level
##  Length:3026        Length:3026        Algo de responsabilidad: 667
##  Class :character   Class :character   Director               :  36
##  Mode  :character   Mode  :character   Ejecutivo              :   7
##                                        Intermedio             : 850
##                                        No corresponde         : 229
##                                        Prácticas              :  38
##                                        Sin experiencia        :1199
##                Type       Functions          Industries
##  Contrato por obra: 190   Length:3026        Length:3026
##  Jornada completa :2775   Class :character   Class :character
##  Media jornada    :  21   Mode  :character   Mode  :character
##  Otro             :   3
##  Prácticas        :  22
##  Temporal         :  14
##  Voluntario       :   1
##   Solicitudes              Empleados    Quick.Application    Emails
##  Min.   :   0.00   51-200       : 375   False:2176        Length:3026
##  1st Qu.:   4.75   11-50        : 282   True : 850        Class :character
```

```
##  Median :  25.00   Más de 10.001: 270                    Mode  :character
##  Mean   : 100.16   1001-5000    : 252
##  3rd Qu.:  90.00   201-500      : 204
##  Max.   :3470.00   (Other)      : 330
##  NA's   :198       NA's         :1313
##  Visualizaciones              Recommended.Flavor      dataset
##  Min.   :   1    ACTIVELY_HIRING_COMPANY:1580     Remote   :780
##  1st Qu.:  32    COMPANY_RECRUIT        : 415     Scotland :171
##  Median : 161    JOB_SEEKER_QUALIFIED   : 379     Spain    :500
##  Mean   : 428    NA's                   : 652     UK       :850
##  3rd Qu.: 468                                     WorldWide:725
##  Max.   :9700
##  NA's   :45
```

Cambiamos los niveles para el tipo de jornada ya que la gran mayoría de ofertas son para jornada completa. Dejamos solo los niveles "Jornada completa" y "Otra jornada".

```
Otra <- c("Contrato por obra", "Media jornada", "Otro", "Prácticas", "Temporal", "Voluntario")
jobs <- jobs %>%
  mutate(Type = fct_collapse(Type, "Otra jornada" = Otra))
```

Parece que los niveles de factor de Empleados no están ordenados de una forma coherente así que los ordenamos.

```
levels(jobs$Empleados)
```

```
## [1] "1001-5000"     "11-50"         "2-10"          "201-500"
## [5] "5001-10.000"   "501-1000"      "51-200"        "Más de 10.001"
```

Los ordenamos.

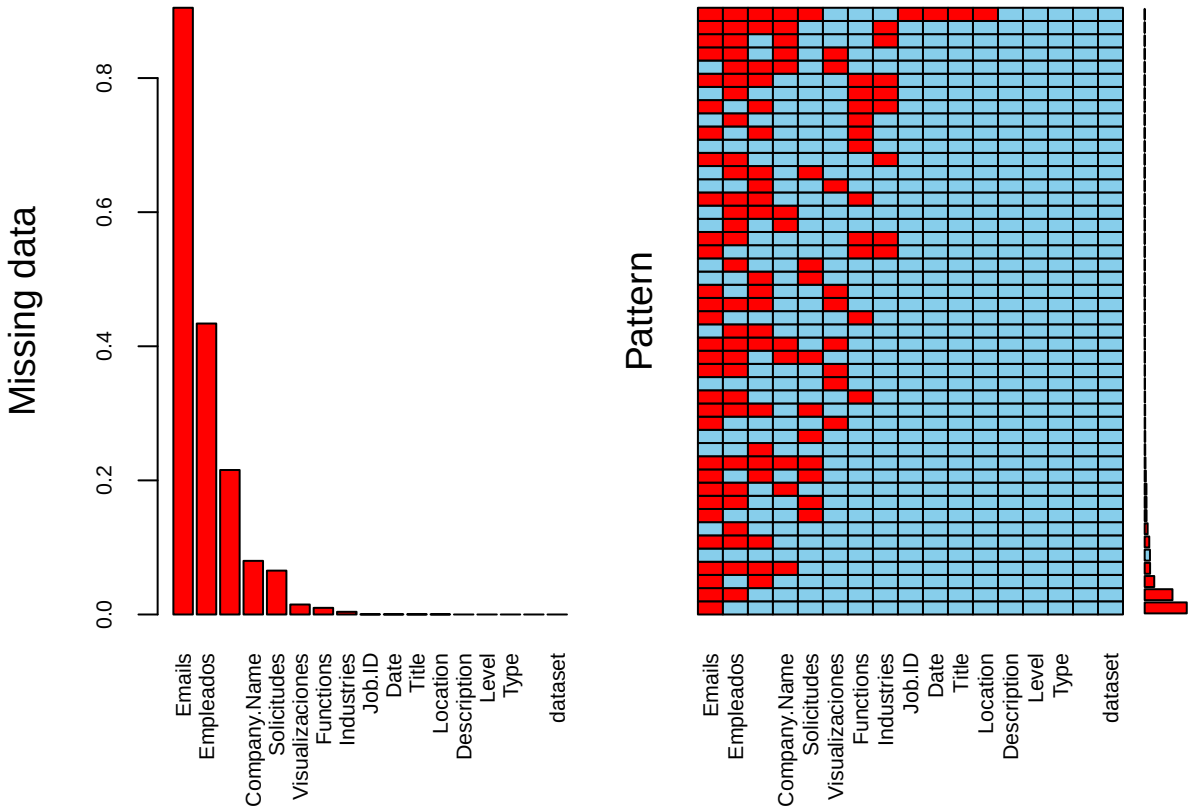
```
jobs$Empleados <- factor(jobs$Empleados, levels = c("2-10", "11-50", "51-200", "201-500", "501-1000", "
```

Mostramos la distribución de los valores perdidos.

```
sum(is.na(jobs))
```

```
## [1] 5234
```

```
aggr(jobs, numbers=TRUE, sortVars=TRUE, labels=names(jobs),
cex.axis=.7, gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##            Variable        Count
##              Emails 0.9048248513
##           Empleados 0.4339061467
##  Recommended.Flavor 0.2154659617
##        Company.Name 0.0799735625
##         Solicitudes 0.0654329147
##     Visualizaciones 0.0148711170
##           Functions 0.0099140780
##          Industries 0.0039656312
##              Job.ID 0.0003304693
##                Date 0.0003304693
##               Title 0.0003304693
##            Location 0.0003304693
##         Description 0.0000000000
##               Level 0.0000000000
##                Type 0.0000000000
##   Quick.Application 0.0000000000
##             dataset 0.0000000000
```

Tratamos los valores perdidos numéricos sustituyéndolos por la mediana.

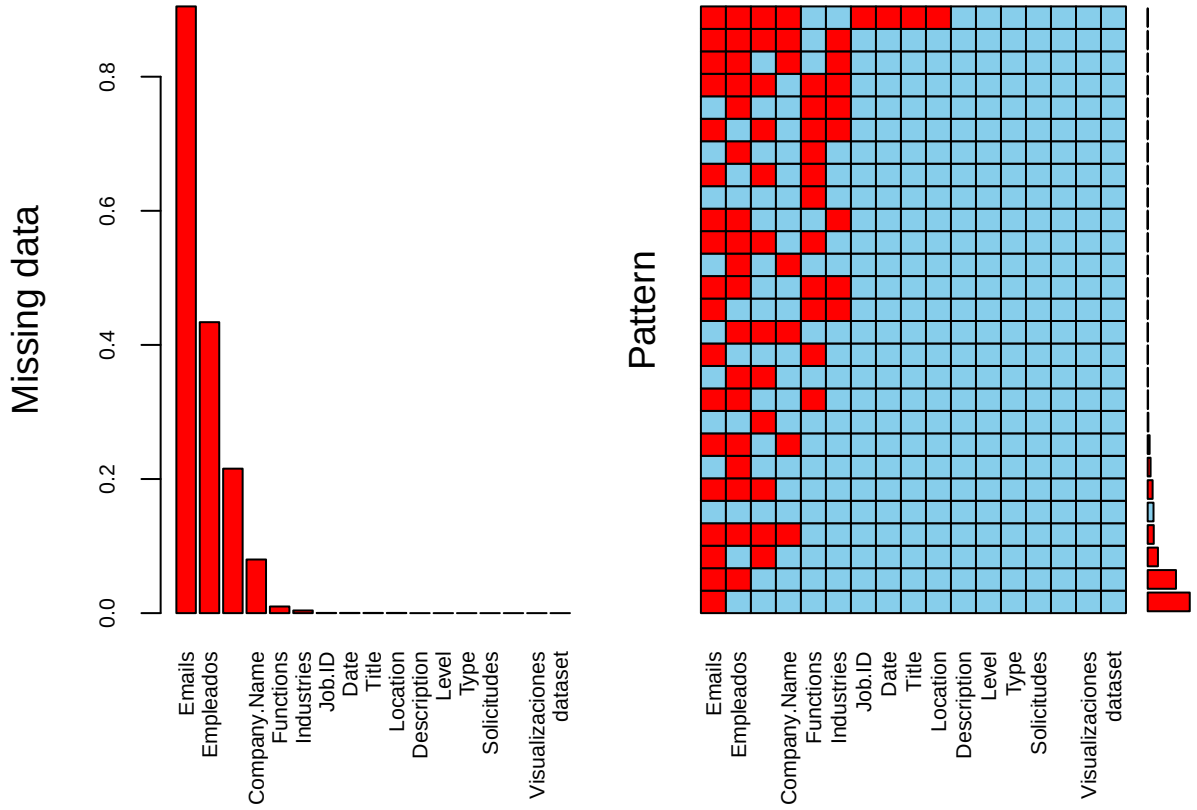
```
jobs <- jobs %>%
  group_by(Level, Quick.Application) %>%
    mutate(Solicitudes = ifelse(is.na(Solicitudes), median(Solicitudes, na.rm = TRUE), Solicitudes))
jobs <- jobs %>%
  group_by(Level, Quick.Application) %>%
    mutate(Visualizaciones = ifelse(is.na(Visualizaciones), median(Visualizaciones, na.rm = TRUE), Visua
```

Comprobamos cómo queda ahora la distribución de valores perdidos.

```
sum(is.na(jobs))
```

```
## [1] 4991
```

```
aggr(jobs, numbers=TRUE, sortVars=TRUE, labels=names(jobs),
cex.axis=.7, gap=3, ylab=c("Missing data","Pattern"))
```



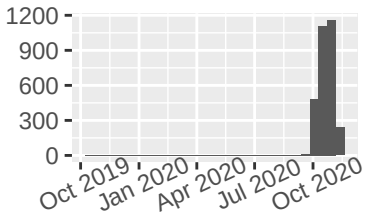

```
##
##  Variables sorted by number of missings:
##            Variable        Count
##              Emails 0.9048248513
##           Empleados 0.4339061467
##  Recommended.Flavor 0.2154659617
##        Company.Name 0.0799735625
##           Functions 0.0099140780
##          Industries 0.0039656312
##              Job.ID 0.0003304693
##                Date 0.0003304693
##               Title 0.0003304693
##            Location 0.0003304693
##         Description 0.0000000000
##               Level 0.0000000000
##                Type 0.0000000000
##         Solicitudes 0.0000000000
##   Quick.Application 0.0000000000
##     Visualizaciones 0.0000000000
##             dataset 0.0000000000
```

Los valores perdidos se han reducido bastante y los que quedan se dejan así porque será importante para posteriores análisis.

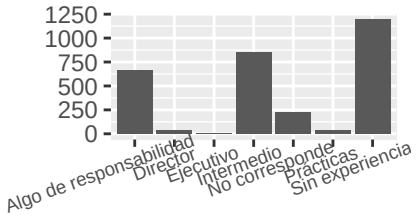
3.2. Identificación y tratamiento de valores extremos

Empezamos haciendo una visualización sencilla de las variables susceptibles de tener valores aislados.

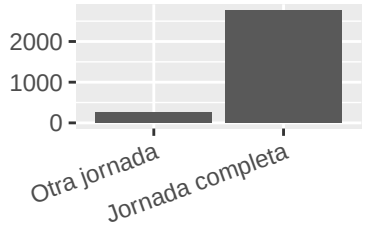
```
grid.arrange(
  qplot(Date, data=jobs)+ theme(axis.text.x = element_text(angle = 25)),
  qplot(Level, data=jobs)+ theme(axis.text.x = element_text(angle = 20, hjust=0.7, size = 7)),
  qplot(Type, data=jobs)+ theme(axis.text.x = element_text(angle = 20, hjust=1)),
  qplot(Solicitudes, data=jobs)+ theme(axis.text.x = element_text(angle = 25)),
  qplot(Empleados, data=jobs)+ theme(axis.text.x = element_text(angle = 30, hjust=0.7, size = 7)),
  qplot(Visualizaciones, data=jobs)+ theme(axis.text.x = element_text(angle = 25)),
  qplot(Recommended.Flavor, data=jobs)+ theme(axis.text.x = element_text(angle = 30, hjust=0.7, size =
)
```



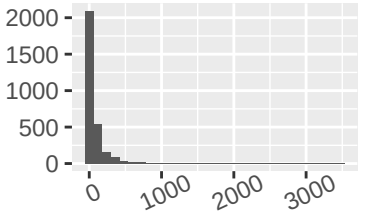
Date



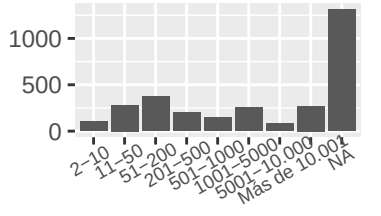
Level



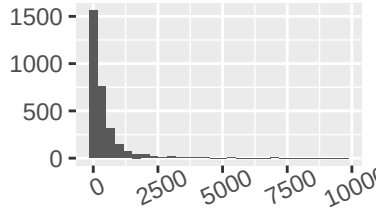
Type



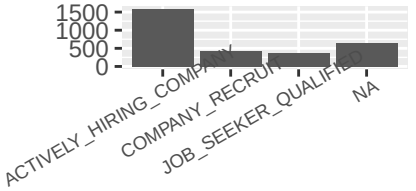
Solicitudes



Empleados



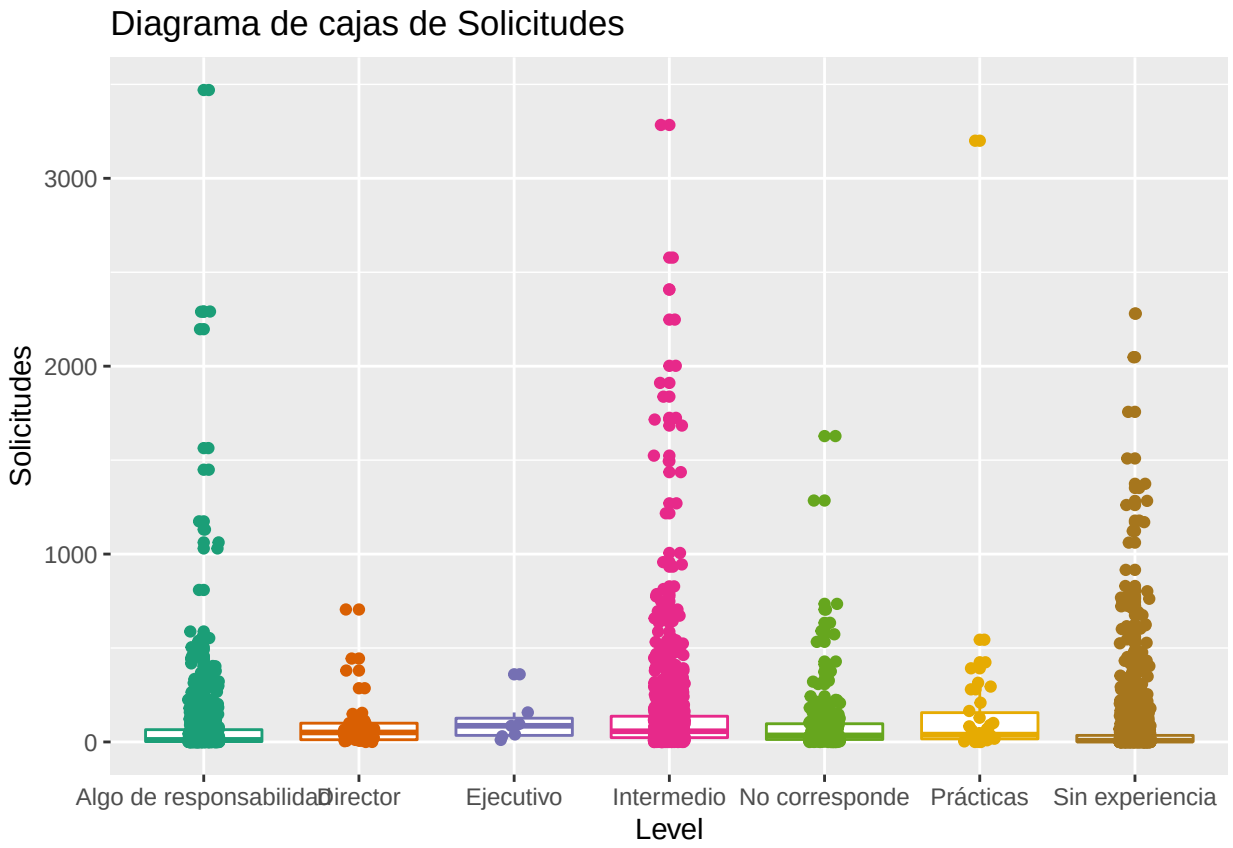
Visualizaciones



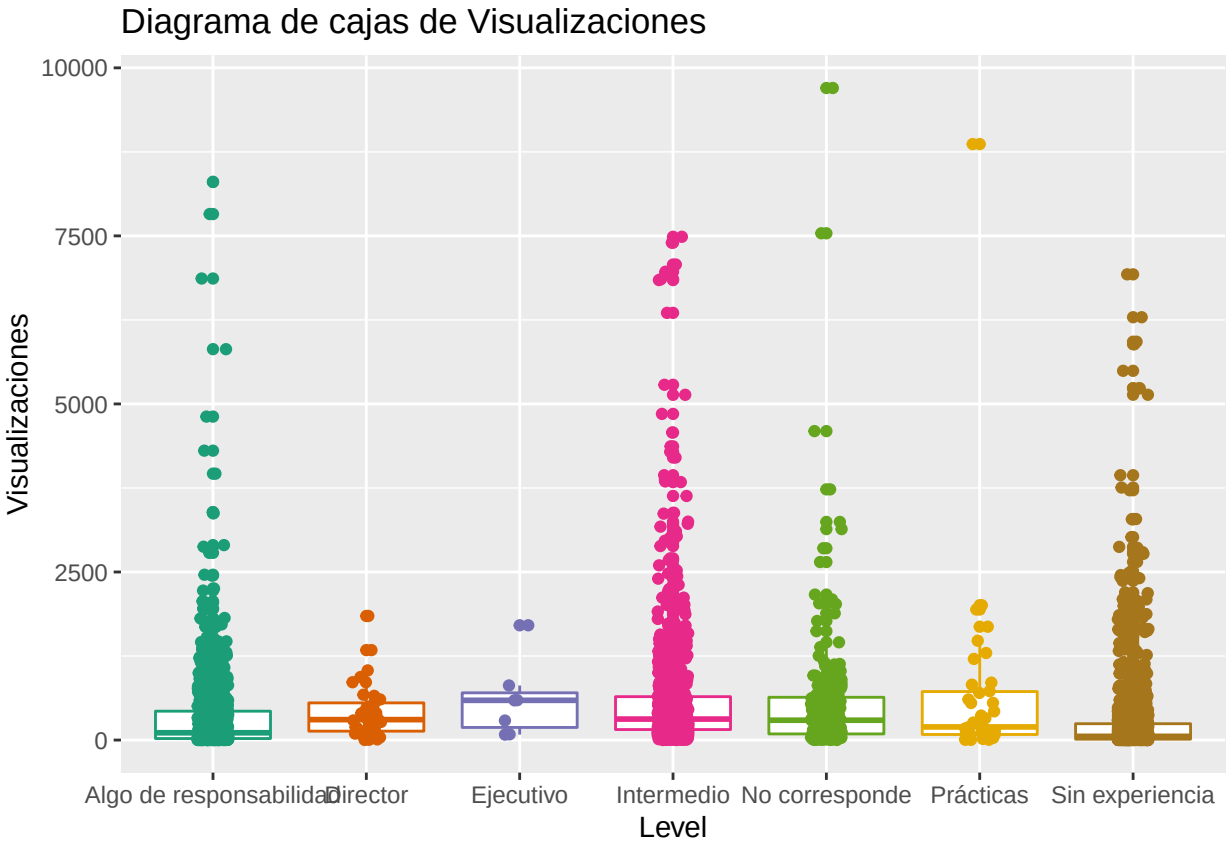
Recommended.Flavor

Hacemos un boxplot para cada una de las variables numéricas agrupándolas en función de la variable "Level" para así visualizar mejor los valores numéricos.

```
ggplot(jobs, aes(x=Level, y=Solicitudes, color=Level)) +
  ggtitle("Diagrama de cajas de Solicitudes") +
  scale_color_brewer(palette="Dark2") +
  geom_boxplot() +
  theme(legend.position = "null") +
  geom_jitter(width = 0.1)
```




```
ggplot(jobs, aes(x=Level, y=Visualizaciones, color=Level)) +
  ggtitle("Diagrama de cajas de Visualizaciones") +
  scale_color_brewer(palette="Dark2") +
  geom_boxplot() +
  theme(legend.position = "null") +
  geom_jitter(width = 0.1)
```



Como ya se veía por las gráficas anteriores son datos con colas muy largas a la derecha y algunos de los valores bastante aislados.

Vamos a mirar los 5 valores más aislados de la variable "Solicitudes".

```
tail(sort(boxplot.stats(jobs$Solicitudes)$out),5)
```

```
## [1] 2408 2578 3200 3284 3470
```

Vamos a observar las tres ofertas de trabajo que superan las 3000 solicitudes.

```
jobs[which(jobs$Solicitudes %in% tail(sort(boxplot.stats(jobs$Solicitudes)$out),3)),]
```

```
## # A tibble: 3 x 17
## # Groups:   Level, Quick.Application [3]
##   Job.ID Date       Company.Name Title Location Description Level Type
##   <chr>  <date>     <chr>        <chr> <chr>    <chr>       <fct> <fct>
## 1 19527~ 2020-10-27 TouchPal     Data~ San Fra~ About the ~ Algo~ Jorn~
## 2 20047~ 2020-10-25 atisfy       Data~ Hyderab~ The ideal ~ Inte~ Jorn~
## 3 21857~ 2020-10-15 LinkedIn     Data~ Sunnyva~ Data Scien~ Prác~ Otra~
## # ... with 9 more variables: Functions <chr>, Industries <chr>,
## #   Solicitudes <dbl>, Empleados <fct>, Quick.Application <fct>, Emails <chr>,
## #   Visualizaciones <dbl>, Recommended.Flavor <fct>, dataset <fct>
```

Aunque sean valores extremos, observándolos en detalles parecen razonables y correctos así que los dejamos.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar)

Para este apartado se realizaran los análsis sobre el dataframe "jobs". Se ha preferido mantener por separado dos dataframes ya que, de eliminar duplicados al mezclar los distintos archivos con sus localizaciones no se realizaría una correcta visualización.

El dataframe "jobs" es el que ha sido limpiado y manipulado previamente para poder realizar los análisis. En primer lugar será necesario encontrar la latitud y la longitud de cada una de las localizaciones. Esto será importante para evaluar posteriormente correlaciones entre variables y aplicar dos tipos de regresión (lineal y logística).

Una vez se obtengan esas nuevas variables para la latitud y la longitud y se comprueben las condiciones de normalidad y homogeneidad se pasará a realizar una correlación de variables para entender dónde podríamos aplicar las regresiones.

El modelo de regresión lineal pretenderá predecir el número de solicitudes respecto a la latitud, longitud y número de visualizaciones. Por otro lado, el modelo de regresión logística viene a predecir si se trata de la oferta cuenta con un enlace de solicitud rápida respecto al nivel de la oferta, el tipo de oferta, el número de solicitudes, el número de visualizaciones y la variable de tipo de ofertante.

4.1.1 Preparación de los datos de Longitud y Latitud.

En primer lugar se obtienen las dos nuevas variables haciendo uso de la API de OpenStreetMap. Se tienen en cuenta tan sólo estos dos valores aunque podríamos haber conseguido muchos más datos sobre una localización más precisa (ésto valdría para hacer análisis por localidades donde encontrar piso cercano al mayor número de puestos de trabajo por ejemplo).

```
jobs$Location <- str_remove(jobs$Location, "^*y alrededores.*$")
jobs$Location[which(jobs$Location == "Greater Barcelona Metropolitan Area")] <- "Barcelona"
jobs$Location[which(jobs$Location == "Cracow, Lesser Poland District, Poland")] <- "Cracow, Poland"
jobs$Location[which(jobs$Location == "New York City Metropolitan Area")] <- "New York"
jobs$Location[which(jobs$Location == "Hong Kong, Hong Kong SAR")] <- "Hong Kong"
jobs$Location[which(jobs$Location == "Burnaby (Maywood / Marlborough / Oakalla / Windsor), V5H, CA")] <-
jobs$Location[which(jobs$Location == "District Brno-City, Czech Republic")] <- "Czech Republic"
jobs$Location[which(jobs$Location == "Silkeborg, Middle Jutland, Denmark")] <- "Denmark"
jobs$Location[which(jobs$Location == "Prague, The Capital, Czech Republic")] <- "Prague"
jobs$Location[which(jobs$Location == "Kuala Lumpur, Federal Territory of Kuala Lumpur, Malaysia")] <- "
jobs$Location[which(jobs$Location == "Genève et périphérie")] <- "Genève"
jobs$Location[which(jobs$Location == "Dallas-Fort Worth Metroplex")] <- "Dallas"
jobs$Location[which(jobs$Location == "Raleigh-Durham-Chapel Hill Area")] <- "Raleigh"
jobs$Location[which(jobs$Location == "Herzliyya, Tel Aviv, Israel")] <- "Herzliya"
jobs$Location[which(jobs$Location == "New Territories, Hong Kong SAR")] <- "Hong Kong"
jobs$Location[which(jobs$Location == "Des Moines Metropolitan Area")] <- "Des Moines"
jobs$Location[which(jobs$Location == "Greater Minneapolis-St. Paul Area")] <- "Minneapolis"
jobs$Location[which(jobs$Location == "Greater Munich Metropolitan Area")] <- "Munich"
jobs$Location[which(jobs$Location == "Gurgaon Sub-District, Haryana, India")] <- "Gurgaon"
jobs$Location[which(jobs$Location == "Village of Mayfield, OH, US")] <- "Cleveland"
jobs$Location[which(jobs$Location == "Kraków i okolice")] <- "Kraków"

jobs_with_duplicates$Location <- str_remove(jobs_with_duplicates$Location, "^*y alrededores.*$")
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Greater Barcelona Metropolitan Ar
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Cracow, Lesser Poland District, P
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "New York City Metropolitan Area")
```

```r
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Hong Kong, Hong Kong SAR")] <- "Ho
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Burnaby (Maywood / Marlborough / (
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "District Brno-City, Czech Republic
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Silkeborg, Middle Jutland, Denmark
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Prague, The Capital, Czech Republi
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Kuala Lumpur, Federal Territory of
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Genève et périphérie")] <- "Genève
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Dallas-Fort Worth Metroplex")] <-
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Raleigh-Durham-Chapel Hill Area")]
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Herzliyya, Tel Aviv, Israel")] <-
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "New Territories, Hong Kong SAR")]
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Des Moines Metropolitan Area")] <-
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Greater Minneapolis-St. Paul Area
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Greater Munich Metropolitan Area")
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Gurgaon Sub-District, Haryana, Ind
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Village of Mayfield, OH, US")] <-
jobs_with_duplicates$Location[which(jobs_with_duplicates$Location == "Kraków i okolice")] <- "Kraków"

OSM_jobs <- geocode_OSM(jobs$Location, details = TRUE, as.data.frame = TRUE)
jobs_merged<-merge(x=jobs,y=OSM_jobs[2:3], by = 0, all= TRUE)

write.csv(jobs_merged,"jobs_merged.csv")

OSM_jobs_duplicates <- geocode_OSM(jobs_with_duplicates$Location, details = TRUE, as.data.frame = TRUE)
jobs_merged_with_duplicates<-merge(x=jobs_with_duplicates,y=OSM_jobs_duplicates[2:3], by = 0, all= TRUE)

write.csv(jobs_merged_with_duplicates,"jobs_merged_with_duplicates.csv")

head(OSM_jobs)
```

```
##                                query      lat       lon  lat_min  lat_max
## 1 Edinburgh, Scotland, United Kingdom 55.95335 -3.188375 55.81879 56.00408
## 2 Edinburgh, Scotland, United Kingdom 55.95335 -3.188375 55.81879 56.00408
## 3  Aberdeen, Scotland, United Kingdom 57.14824 -2.092809 57.07619 57.23531
## 4                        Glasgow, GB 55.86098 -4.248879 55.70098 56.02098
## 5                        Glasgow, GB 55.86098 -4.248879 55.70098 56.02098
## 6                        Glasgow, GB 55.86098 -4.248879 55.70098 56.02098
##     lon_min   lon_max  place_id osm_type   osm_id place_rank
## 1 -3.449533 -3.074951 257101080 relation  1920901         12
## 2 -3.449533 -3.074951 257101080 relation  1920901         12
## 3 -2.360940 -2.016151 257105109 relation  1900654         12
## 4 -4.408879 -4.088879    107701     node 11127374         16
## 5 -4.408879 -4.088879    107701     node 11127374         16
## 6 -4.408879 -4.088879    107701     node 11127374         16
##                                          display_name    class
## 1             City of Edinburgh, Scotland, United Kingdom boundary
## 2             City of Edinburgh, Scotland, United Kingdom boundary
## 3                 Aberdeen City, Scotland, United Kingdom boundary
## 4 Glasgow, Glasgow City, Scotland, G2 9SD, United Kingdom    place
## 5 Glasgow, Glasgow City, Scotland, G2 9SD, United Kingdom    place
## 6 Glasgow, Glasgow City, Scotland, G2 9SD, United Kingdom    place
##            type       importance
## 1 administrative  1.0867042570635
## 2 administrative  1.0867042570635
```

```
## 3 administrative  1.0099117188774
## 4           city 0.79905249303697
## 5           city 0.79905249303697
## 6           city 0.79905249303697
##                                                                                 icon
## 1 https://nominatim.openstreetmap.org/ui/mapicons//poi_boundary_administrative.p.20.png
## 2 https://nominatim.openstreetmap.org/ui/mapicons//poi_boundary_administrative.p.20.png
## 3 https://nominatim.openstreetmap.org/ui/mapicons//poi_boundary_administrative.p.20.png
## 4              https://nominatim.openstreetmap.org/ui/mapicons//poi_place_city.p.20.png
## 5              https://nominatim.openstreetmap.org/ui/mapicons//poi_place_city.p.20.png
## 6              https://nominatim.openstreetmap.org/ui/mapicons//poi_place_city.p.20.png
```

```
head(jobs_merged)
```

```
##   Row.names     Job.ID       Date               Company.Name
## 1         1 2011662890 2020-10-15                    Dufrain
## 2        10 2249373089 2020-11-05              Accenture UK
## 3       100 2249373063 2020-11-05              Accenture UK
## 4      1000 2187582068 2020-10-16 Opus Recruitment Solutions
## 5      1001 2198827173 2020-10-21              Checkout.com
## 6      1002 2176271964 2020-10-13            Venture Search
##                         Title                             Location
## 1               Data Engineer Edinburgh, Scotland, United Kingdom
## 2 User Researcher - Newcastle                          Glasgow, GB
## 3         Cloud Data Engineer                          Glasgow, GB
## 4              Data Scientist     Oxford, England, United Kingdom
## 5         Lead Data Scientist     London, England, United Kingdom
## 6     Quantitative Researcher     London, England, United Kingdom
##
## 1
leading Data Management, Analytics and BI consultancy with offices across the UK, working with some of
leading Data Management and Analytics Consultancy. We are looking for dynamic individuals, with proven
 (Certified to practitioner level in at least one provider) Experience working with one or more of the
 Spark, Kafka, Snowflake, Hadoop Experience working with both SQL and NoSQL foundational tools and datal
## 2
Tyne  Salary: £28,000 - £48,000  Career Level: (Accenture will be recruiting at the following levels: S
generation technology to each business challenge.  We believe in inclusion and diversity and supporting
centred designs for digital products and services.  This is a great opportunity to work across multiple
knit UX and Design team, you will champion the value of User-Centred Design across all products and ser
to-face, session analysis, report writing and playback to your team  Collaborate with product owners an
centred design process  Some experience working with GDS standards would be beneficial but not essentia
class services we are known for.   About Accenture  Accenture is a leading global professional services
## 3 Cloud Data Engineer  Location: London  Salary: £45,000- £55,000  Career Level:  (Accenture will be
generation technology to each business challenge.  We believe in inclusion and diversity and supporting
to-end solutions for our client - from data strategy/governance to Core Engineering, enabling them to t
edge technologies and will have the opportunity to develop a wide range of new skills on the job.  In ou
facing / consulting environment to build trusted relationships with client stakeholders and act as a tru
location team-oriented environment.  Proven ability in delivering high-quality deliverables to tight ti
functional requirements and operations) management.  Regulatory and Compliance work in Data Management
functional requirements and operations) management.  E2E Solution Design skills - Prototyping, Usability
class services we are known for.   About Accenture  Accenture is a leading global professional services
## 4
learn / and other packagesRelational and/or non-relational databasesExperience with Big DataAWS and clou
## 5
real-time transaction risk predictions, which Checkout.com's merchants use to make smart payment routing
```

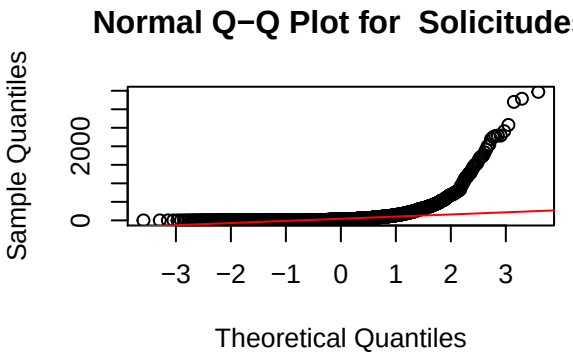
```
world problemsExperience working on machine learning for fraud detectionStrong expertise in: machine lea
computing (e.g. Spark, Dask, Hadoop)Solid software engineering skills and able to write high-
quality Python codeExperience working with scientific Python stack (e.g. pandas, scikit-
learn, XGBoost, SciPy)Experience with SQL databases and key-value stores (NoSQL)Experience with Docker
compose) for development and deploymentExperience with AWS or at least another common cloud platform (G
quality Python for feature engineering and model training
## 6
starter who takes a proactive approach to work
##                     Level             Type
## 1              Intermedio Jornada completa
## 2 Algo de responsabilidad Jornada completa
## 3         Sin experiencia Jornada completa
## 4              Intermedio Jornada completa
## 5              Intermedio Jornada completa
## 6 Algo de responsabilidad Jornada completa
##                                          Functions
## 1           Tecnología de la información, Ingeniería
## 2                       Tecnología de la información
## 3                       Tecnología de la información
## 4           Ingeniería, Tecnología de la información
## 5 Ingeniería, Tecnología de la información, Análisis
## 6                           Investigación, Análisis
##                                                                  Industries
## 1  Servicios y tecnologías de la información, Banca, Servicio de información
## 2 Servicios y tecnologías de la información, Software, Servicios financieros
## 3 Servicios y tecnologías de la información, Software, Servicios financieros
## 4                                        Dotación y selección de personal
## 5                                                   Servicios financieros
## 6                             Servicios financieros, Gestión de inversiones
##   Solicitudes   Empleados Quick.Application Emails Visualizaciones
## 1          54      51-200             False   <NA>             679
## 2           4 5001-10.000             False   <NA>              56
## 3           3 5001-10.000             False   <NA>              29
## 4         246     201-500              True   <NA>             534
## 5          58    501-1000              True   <NA>             232
## 6         540        2-10              True   <NA>            1683
##        Recommended.Flavor  dataset      lat        lon
## 1 ACTIVELY_HIRING_COMPANY Scotland 55.95335 -3.1883749
## 2 ACTIVELY_HIRING_COMPANY Scotland 55.86098 -4.2488787
## 3 ACTIVELY_HIRING_COMPANY Scotland 55.86098 -4.2488787
## 4 ACTIVELY_HIRING_COMPANY       UK 51.75201 -1.2578499
## 5 ACTIVELY_HIRING_COMPANY       UK 51.50732 -0.1276474
## 6 ACTIVELY_HIRING_COMPANY       UK 51.50732 -0.1276474
```

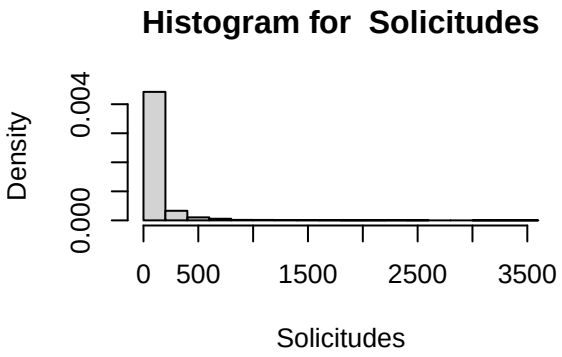
4.2. Comprobación de la normalidad y homogeneidad de la varianza.

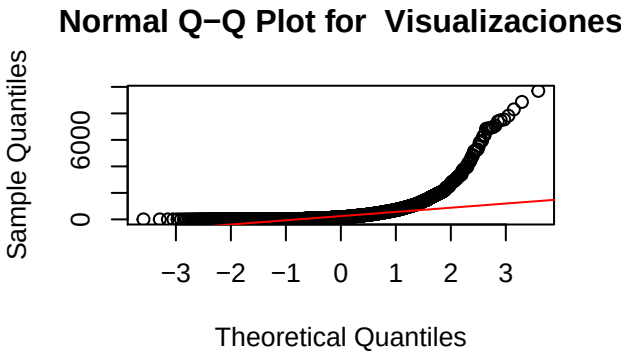
En el dataset se obtienen 19 variables. En este apartado se busca verificar si las variables numéricas (lat, lon, Solicitudes y Visualizaciones) siguen una distribución normal.

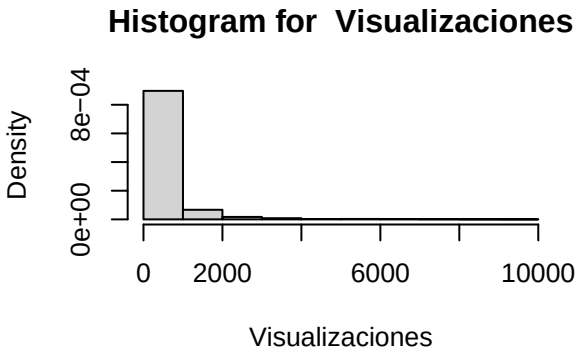
Para verificarlo, primero se realizará un estudio visual de la normalidad mediante la gráfica quantile-quantile (Q-Q) que dibuja la correlación entre una muestra dada y la distribución normal. En esta primera gráfica se podrá observar el grado de aproximación o similitud con la línea de referencia de 45 grados. Además, también se realizará el estudio visual sobre el histograma de la variable

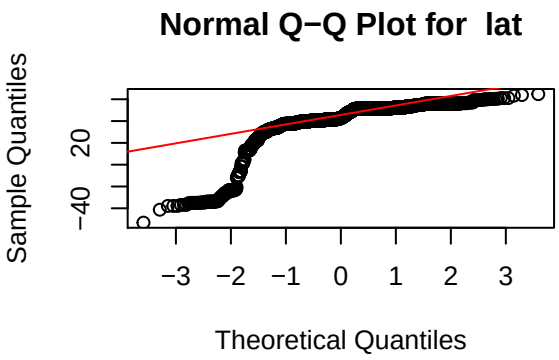
```
par(mfrow=c(2,2))
for(i in 1:ncol(jobs_merged)) {
  if (is.numeric(jobs_merged[,i])){
    qqnorm(jobs_merged[,i],main = paste("Normal Q-Q Plot for ",colnames(jobs_merged)[i]))
    qqline(jobs_merged[,i],col="red")
    hist(jobs_merged[,i],
      main=paste("Histogram for ", colnames(jobs_merged)[i]),
      xlab=colnames(jobs_merged)[i], freq = FALSE)
  }
}
```

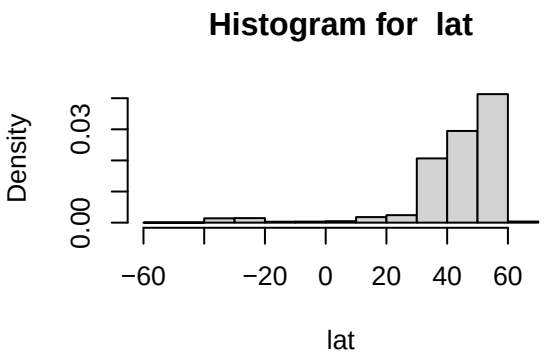


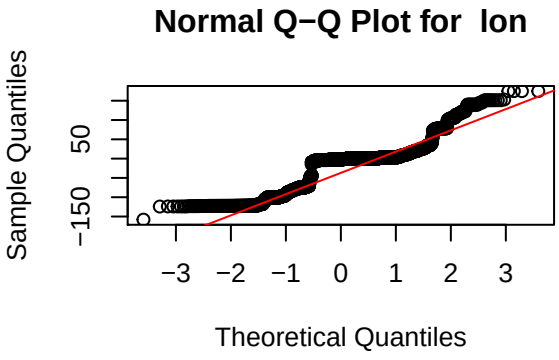


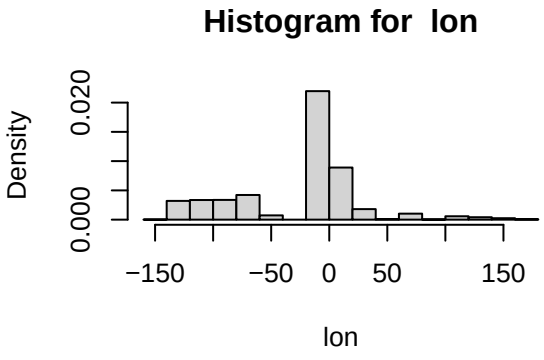












No parece que las gráficas representen que claramente las variables siguen una distribución normal. Para verificarlo, a continuación se realiza la comprobación mediante el test de Shapiro-Wilk y así no dar lugar a errores.

```
for(i in 1:ncol(jobs_merged)) {
  if (is.numeric(jobs_merged[,i])){
    result <- shapiro.test(jobs_merged[,i])
    print(paste("Resultados para: ", colnames(jobs_merged)[i]))
    print(result)
  }
}
```

```
## [1] "Resultados para:  Solicitudes"
##
##  Shapiro-Wilk normality test
##
## data:  jobs_merged[, i]
## W = 0.38417, p-value < 2.2e-16
##
## [1] "Resultados para:  Visualizaciones"
##
##  Shapiro-Wilk normality test
##
## data:  jobs_merged[, i]
## W = 0.49682, p-value < 2.2e-16
##
## [1] "Resultados para:  lat"
##
##  Shapiro-Wilk normality test
##
```

```
## data:  jobs_merged[, i]
## W = 0.66142, p-value < 2.2e-16
##
## [1] "Resultados para:  lon"
##
##  Shapiro-Wilk normality test
##
## data:  jobs_merged[, i]
## W = 0.83991, p-value < 2.2e-16
```

Se puede observar que a través de los test aplicados a las variables cuantitativas, el p-value es menor que 0.05, por lo que se deberá rechazar la hipótesis nula y aceptar que las variables no siguen una distribución normal. Por lo tanto, se deberán utlizar métodos para el análisis que no supongan que las variables cuantitativas siguen una distribución normal.

No obstante, cuando el número de observaciones es mayor o igual a 30, como es el caso de estas variables cuantitativas y debido al teorema central del límite, se podrán utilizar pruebas paramétricas asumiendo que con un aumento de observaciones, la distribución se volvería normal y en forma de campana. Así las variables se podrían aproximar como una distribución normal de media 0 y desviación estándar 1.

A continuación, se lleva a cabo la comprobación de homogeneidad de las varianzas. Se utilizará el test de Flinge-Killneen que resulta apropiado para variables no paramétricas que no siguen una distribución normal.

```
fligner.test(Visualizaciones ~ lat, jobs_merged)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Visualizaciones by lat
## Fligner-Killeen:med chi-squared = 1214.6, df = 629, p-value < 2.2e-16
```

```
fligner.test(Visualizaciones ~ lon, jobs_merged)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Visualizaciones by lon
## Fligner-Killeen:med chi-squared = 1215.1, df = 630, p-value < 2.2e-16
```

```
fligner.test(Solicitudes ~ lat, jobs_merged)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Solicitudes by lat
## Fligner-Killeen:med chi-squared = 1397.2, df = 629, p-value < 2.2e-16
```

```
fligner.test(Solicitudes ~ lon, jobs_merged)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Solicitudes by lon
## Fligner-Killeen:med chi-squared = 1397.4, df = 630, p-value < 2.2e-16
```

```
fligner.test(Visualizaciones ~ Solicitudes, jobs_merged)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
```

```
##
## data:  Visualizaciones by Solicitudes
## Fligner-Killeen:med chi-squared = 1662.6, df = 444, p-value < 2.2e-16
```

Se obtiene como resultado que las variables numéricas no son homogéneas según su varianza ya que en todos estos test se consigue un p-value menor de 0.05. Se deberá tener en cuenta para la aplicación de métodos analíticos que asuman homogeneidad de las varianzas.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos.

En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

4.3.1 Correlaciones.

En primer lugar, se analizan las posibles relaciones entre las variables.

```
borrar <- c("Date","Row.names")
jobs_corr <- jobs_merged[ , !(names(jobs_merged) %in% borrar)]

corr_matrix<-hetcor(jobs_corr, ML=FALSE, std.err=FALSE)
```

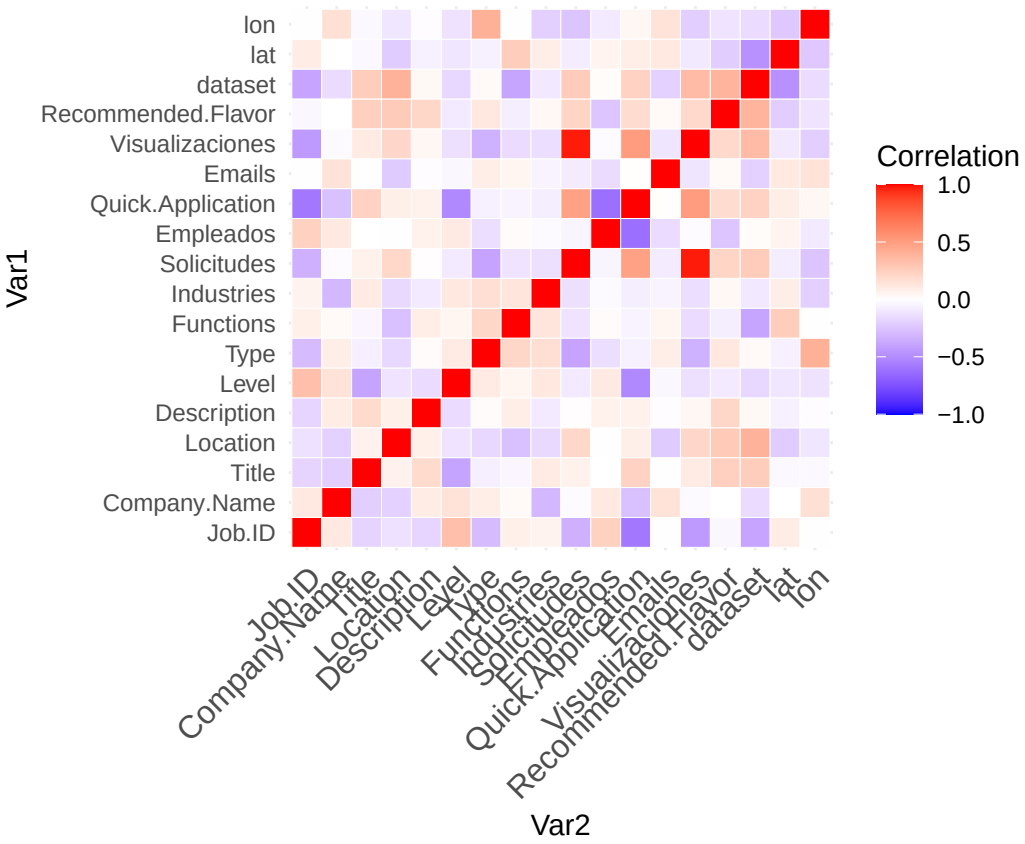
```
## data contain one or more character variables
## the values of which are ordered alphabetically
## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 row with zero marginal
## removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 row with zero marginal
## removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 row with zero marginal
## removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 row with zero marginal
## removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 row with zero marginal
## removed
## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 column with zero
## marginal removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 column with zero
## marginal removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 column with zero
## marginal removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 column with zero
## marginal removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 column with zero
## marginal removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 column with zero
## marginal removed
```

```
## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 column with zero
## marginal removed
## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed
## Warning in polychor(x, y, ML = ML, std.err = std.err): 1 column with zero
## marginal removed
## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed

## Warning in polychor(x, y, ML = ML, std.err = std.err): 2 rows with zero
## marginals removed
corr_matrix$correlations

##                        Job.ID  Company.Name        Title     Location
## Job.ID             1.000000000  1.177427e-01 -0.182066417 -0.125832648
## Company.Name       0.117742697  1.000000e+00 -0.204866517 -0.197812801
## Title             -0.182066417 -2.048665e-01  1.000000000  0.065882785
## Location          -0.125832648 -1.978128e-01  0.065882785  1.000000000
## Description       -0.176712232  9.824716e-02  0.191150144  0.081320435
## Level              0.330888615  1.455492e-01 -0.392652430 -0.118576467
## Type              -0.291671295  8.961448e-02 -0.067520473 -0.170079398
```

```
## Functions           0.079254301  2.511600e-02 -0.041412330 -0.266757093
## Industries          0.061440316 -3.025750e-01  0.104913660 -0.159877784
## Solicitudes        -0.339436762 -1.868993e-02  0.075857276  0.204141674
## Empleados           0.240099227  1.163025e-01 -0.001813957 -0.008080275
## Quick.Application  -0.584195135 -2.714189e-01  0.234683156  0.083564993
## Emails             -0.008915290  1.469190e-01 -0.006433509 -0.219981235
## Visualizaciones    -0.436574483 -2.086866e-02  0.103546469  0.207027044
## Recommended.Flavor -0.035161927  1.049451e-03  0.247619934  0.273280114
## dataset            -0.392180013 -1.466451e-01  0.263056200  0.398125696
## lat                 0.098287048  3.020288e-05 -0.031044196 -0.219391527
## lon                -0.005589807  1.635337e-01 -0.027966289 -0.106748162
##                     Description       Level        Type   Functions  Industries
## Job.ID             -0.176712232  0.33088862 -0.29167129  0.07925430  0.06144032
## Company.Name        0.098247162  0.14554921  0.08961448  0.02511600 -0.30257499
## Title               0.191150144 -0.39265243 -0.06752047 -0.04141233  0.10491366
## Location            0.081320435 -0.11857647 -0.17007940 -0.26675709 -0.15987778
## Description         1.000000000 -0.15231896  0.02019992  0.08946535 -0.09041754
## Level              -0.152318957  1.00000000  0.10372067  0.04964986  0.12313831
## Type                0.020199921  0.10372067  1.00000000  0.20914303  0.17083141
## Functions           0.089465353  0.04964986  0.20914303  1.00000000  0.13704713
## Industries         -0.090417536  0.12313831  0.17083141  0.13704713  1.00000000
## Solicitudes         0.009042817 -0.09002080 -0.39429181 -0.11897601 -0.12990029
## Empleados           0.067248712  0.10860342 -0.14070824  0.02068281 -0.02395602
## Quick.Application   0.071053575 -0.51499313 -0.05906716 -0.04997408 -0.07263795
## Emails             -0.011271159 -0.03213824  0.09194105  0.04902759 -0.05222230
## Visualizaciones     0.042720269 -0.13152498 -0.33126156 -0.15492748 -0.13890436
## Recommended.Flavor  0.210174962 -0.08757995  0.12128933 -0.07282805  0.03909756
## dataset             0.033441880 -0.17106396  0.03240679 -0.38737377 -0.09843934
## lat                -0.063463199 -0.10685084 -0.06258805  0.26200865  0.08918268
## lon                -0.011250163 -0.12386647  0.40573169  0.00385023 -0.19818768
##                     Solicitudes    Empleados Quick.Application       Emails
## Job.ID             -0.339436762  0.240099227      -0.584195135 -0.008915290
## Company.Name       -0.018689934  0.116302531      -0.271418878  0.146918961
## Title               0.075857276 -0.001813957       0.234683156 -0.006433509
## Location            0.204141674 -0.008080275       0.083564993 -0.219981235
## Description         0.009042817  0.067248712       0.071053575 -0.011271159
## Level              -0.090020805  0.108603424      -0.514993134 -0.032138236
## Type               -0.394291809 -0.140708235      -0.059067158  0.091941052
## Functions          -0.118976008  0.020682813      -0.049974078  0.049027590
## Industries         -0.129900293 -0.023956019      -0.072637950 -0.052222302
## Solicitudes         1.000000000 -0.042694313       0.487250714 -0.085776220
## Empleados          -0.042694313  1.000000000      -0.623636819 -0.147341771
## Quick.Application   0.487250714 -0.623636819       1.000000000  0.006894546
## Emails             -0.085776220 -0.147341771       0.006894546  1.000000000
## Visualizaciones     0.975463341 -0.014831305       0.510811507 -0.109827235
## Recommended.Flavor  0.220703186 -0.241848871       0.181070217  0.025738048
## dataset             0.265578459  0.017197953       0.232912425 -0.195878457
## lat                -0.080436013  0.058687660       0.092054800  0.115096737
## lon                -0.250102324 -0.087326099       0.043617483  0.148278110
##                    Visualizaciones Recommended.Flavor     dataset           lat
## Job.ID                 -0.43657448       -0.035161927 -0.39218001  9.828705e-02
## Company.Name           -0.02086866        0.001049451 -0.14664509  3.020288e-05
## Title                   0.10354647        0.247619934  0.26305620 -3.104420e-02
## Location                0.20702704        0.273280114  0.39812570 -2.193915e-01
```

```
## Description             0.04272027        0.210174962  0.03344188 -6.346320e-02
## Level                  -0.13152498       -0.087579954 -0.17106396 -1.068508e-01
## Type                   -0.33126156        0.121289325  0.03240679 -6.258805e-02
## Functions              -0.15492748       -0.072828048 -0.38737377  2.620086e-01
## Industries             -0.13890436        0.039097563 -0.09843934  8.918268e-02
## Solicitudes             0.97546334        0.220703186  0.26557846 -8.043601e-02
## Empleados              -0.01483131       -0.241848871  0.01719795  5.868766e-02
## Quick.Application       0.51081151        0.181070217  0.23291243  9.205480e-02
## Emails                 -0.10982724        0.025738048 -0.19587846  1.150967e-01
## Visualizaciones         1.00000000        0.199847847  0.35238735 -9.238855e-02
## Recommended.Flavor      0.19984785        1.000000000  0.38618914 -2.114904e-01
## dataset                 0.35238735        0.386189144  1.00000000 -4.725544e-01
## lat                    -0.09238855       -0.211490430 -0.47255443  1.000000e+00
## lon                    -0.20405167       -0.118924398 -0.15268885 -2.354341e-01
##                                lon
## Job.ID              -0.005589807
## Company.Name         0.163533687
## Title               -0.027966289
## Location            -0.106748162
## Description         -0.011250163
## Level               -0.123866466
## Type                 0.405731687
## Functions            0.003850230
## Industries          -0.198187682
## Solicitudes         -0.250102324
## Empleados           -0.087326099
## Quick.Application    0.043617483
## Emails               0.148278110
## Visualizaciones     -0.204051672
## Recommended.Flavor  -0.118924398
## dataset             -0.152688848
## lat                 -0.235434120
## lon                  1.000000000
```

```
ggplot(
  melt(corr_matrix$correlations),
  aes(Var2, Var1, fill = value)
)+
geom_tile(color = "white")+
scale_fill_gradient2(
  low = "blue",
  high = "red",
  mid = "white",
  midpoint = 0,
  limit = c(-1,1),
  space = "Lab",
  name="Correlation") +
theme_minimal()+ # minimal theme
theme(
  axis.text.x = element_text(
  angle = 45, vjust = 1,
  size = 12, hjust = 1))+
coord_fixed()
```



Se observa que:

- Tal y como era de suponer, existe una fuerte relación entre la variable Solicitudes y Visualizaciones.
- También existe fuerte relación entre el númnero de solicitudes y visualizaciones con el tipo de oferta. Esto tiene sentido, posteriormente se verá como quedaría reflejada la distribución según cada una de las categorías.
- Hay una relación interesante entre si la oferta tiene enlace de aplicación rápida y el número de solicitudes, visualizaciones y número de empleados.

4.3.2 Modelo de regresión lineal

Entendiendo las distintas variables que pueden formar un buen modelo se decide la aplicación de un modelo de regresión lineal en el que se pueda predecir el número de solicitudes respecto al número de personas que lo hayan visualizado y de su localización.

```
remote_lat_positive<-jobs_merged[which(jobs_merged$lat>0),]
ntrain <- nrow(remote_lat_positive)*0.8
ntest <- nrow(remote_lat_positive)*0.2
set.seed(12312)
index_train<-sample(1:nrow(remote_lat_positive),size = ntrain)
train<-remote_lat_positive[index_train,]
test<-remote_lat_positive[-index_train,]
modelo<-lm(formula = Solicitudes ~ lat + lon + Visualizaciones, data=train)
summary(modelo)
```

```
##
## Call:
## lm(formula = Solicitudes ~ lat + lon + Visualizaciones, data = train)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1107.09   -15.55     7.97    21.79  1410.16
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     12.188240  11.591630   1.051 0.293152
## lat             -0.714894   0.245797  -2.908 0.003666 **
## lon             -0.150482   0.041722  -3.607 0.000317 ***
## Visualizaciones  0.262781   0.002632  99.822  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104.2 on 2331 degrees of freedom
## Multiple R-squared:  0.8162, Adjusted R-squared:  0.816
## F-statistic:  3451 on 3 and 2331 DF,  p-value: < 2.2e-16
```

```
prob_sl<-predict(modelo, test, type="response")
mc_sl<-data.frame(
  real=test$Solicitudes,
  predicted= prob_sl,
  dif=ifelse(test$Solicitudes>prob_sl, -prob_sl*100/test$Solicitudes,prob_sl*100/test$Solicitudes)
  )
colnames(mc_sl)<-c("Real","Predecido","Dif%")
knitr::kable(mc_sl)
```

	Real	Predecido	Dif%
7	59	93.1109204	157.8151192
16	2	-17.5197481	875.9874046
18	29	14.1114604	-48.6602082
22	28	17.6928637	-63.1887990
29	12	10.3350045	-86.1250378
36	9	-16.8214514	186.9050155
42	7	-23.1739299	331.0561410
44	29	44.4964936	153.4361848
46	290	241.3192268	-83.2135265
48	41	28.2333023	-68.8617129
52	35	74.4534917	212.7242620
54	0	-22.1394305	Inf
60	0	-24.3520459	Inf
66	2	-21.4614584	1073.0729184
70	6	-8.8479855	147.4664247
76	84	65.2561677	-77.6859140
82	51	33.1984156	-65.0949326
85	33	23.6528856	-71.6754109
86	10	16.6432324	166.4323244
87	59	32.1458014	-54.4844092
96	7	10.3364960	147.6642285
106	94	46.6279502	-49.6042024
108	113	49.4908181	-43.7971841
112	26	12.8711392	-49.5043815
113	0	-26.0559398	Inf
117	21	24.7879422	118.0378201
118	23	77.6068599	337.4211301

	Real	Predecido	Dif%
123	3	-17.2569674	575.2322469
126	13	20.5834513	158.3342404
141	0	-24.1938037	Inf
144	143	69.9862201	-48.9414126
167	9	-5.6946173	63.2735251
176	19	15.5906182	-82.0558855
178	24	2.7435758	-11.4315659
182	39	16.3804518	-42.0011583
187	13	18.8415127	144.9347132
193	5	-12.5269151	250.5383015
200	18	22.4229160	124.5717558
202	0	-26.5443367	Inf
210	10	-10.1618889	101.6188891
218	20	25.3261980	126.6309900
222	16	-3.3295911	20.8099443
223	11	1.4004612	-12.7314658
224	12	3.1498807	-26.2490056
230	14	8.7583204	-62.5594316
231	2	-18.3080901	915.4045074
232	31	53.9565983	174.0535428
233	8	-0.1762229	2.2027859
234	5	-10.6754026	213.5080514
236	7	-20.2376003	289.1085758
238	21	38.2872834	182.3203972
243	3	-20.4103356	680.3445209
247	0	-26.8442819	Inf
249	222	509.6183063	229.5577956
251	3	-19.6099459	653.6648622
262	33	201.1137819	609.4357029
267	37	3.7654874	-10.1769930
281	5	-12.5269151	250.5383015
283	17	63.4167029	373.0394290
295	18	1.3924092	-7.7356068
296	47	52.6426948	112.0057337
301	13	-7.9070315	60.8233191
302	6	-24.4420913	407.3681878
313	0	-26.8442819	Inf
319	2	2.1888033	109.4401647
325	128	120.1773309	-93.8885398
326	26	-22.5125811	86.5868504
329	18	42.7673987	237.5966595
330	28	7.4444170	-26.5872036
332	5	-8.3304761	166.6095225
333	4	-18.4182589	460.4564722
334	98	177.9890816	181.6215119
337	19	8.4554120	-44.5021683
340	130	75.9846137	-58.4497029
341	185	196.1224404	106.0121299
344	7	-20.4583458	292.2620827
351	50	68.9350973	137.8701946
354	13	34.7736083	267.4892942
356	3	-15.0443520	501.4784004

	Real	Predecido	Dif%
361	17	27.7226854	163.0746199
366	72	34.5123190	-47.9337764
368	45	53.6938176	119.3195946
377	160	129.6389270	-81.0243294
385	5	-11.7598730	235.1974597
402	0	-14.9819866	Inf
404	0	-23.3009232	Inf
413	0	-26.0559398	Inf
422	0	-23.1379984	Inf
423	0	-25.4916480	Inf
428	0	-12.2641344	Inf
434	13	26.3938375	203.0295191
441	0	-23.1905533	Inf
453	5	3.2399260	-64.7985207
454	104	124.2165674	119.4390072
456	24	25.3135036	105.4729316
459	428	955.2943483	223.1996141
460	84	46.0330500	-54.8012500
468	0	-22.4238591	Inf
471	116	112.4465223	-96.9366572
472	5	-4.3229620	86.4592391
478	8	37.0744646	463.4308070
482	7	-21.2594157	303.7059381
485	91	92.5853590	101.7421527
489	45	73.1395883	162.5324184
491	0	-14.9819866	Inf
493	7	-13.3152571	190.2179590
496	22	45.5476163	207.0346197
497	32	-9.4136746	29.4177331
499	221	110.1931563	-49.8611567
501	5	-9.8991082	197.9821645
512	3	-13.3233091	444.1103048
513	4	-18.0653654	451.6341355
514	63	37.4029066	-59.3696930
515	55	49.7521073	-90.4583769
518	29	135.6813913	467.8668667
520	17	19.5323285	114.8960501
523	32	-24.6549543	77.0467323
527	7	-21.7331518	310.4735978
532	0	-22.3496563	Inf
534	83	95.5142535	115.0774139
538	0	-24.3520459	Inf
540	23	57.1099665	248.3042021
545	13	10.5977852	-81.5214248
548	54	38.4525378	-71.2084034
555	7	-22.6924797	324.1782820
559	44	20.3206706	-46.1833422
563	2	-21.6234210	1081.1710519
566	0	-21.4614584	Inf
571	29	107.3010773	370.0037150
572	7	-0.7017842	10.0254892
574	4	-12.6370838	315.9270954

	Real	Predecido	Dif%
581	0	-21.1986777	Inf
582	3	-14.8919412	496.3980413
585	376	458.9277873	122.0552626
588	79	95.0240206	120.2835704
590	32	51.9378877	162.3058992
602	0	-27.0698981	Inf
606	59	71.5510105	121.2728991
613	2	-23.6537492	1182.6874606
615	18	61.7048337	342.8046315
618	2048	1378.9614038	-67.3320998
619	150	124.7084296	-83.1389530
628	675	613.0990675	-90.8294915
629	3	-7.5392916	251.3097196
630	746	508.9761652	-68.2273680
633	244	246.5146310	101.0305865
634	672	610.7893686	-90.8912751
647	4	-9.7007749	242.5193735
653	12	14.7418768	122.8489735
663	207	542.6682733	262.1585861
664	72	124.2005508	172.5007650
667	12	26.0824010	217.3533420
668	2	-25.5303785	1276.5189230
670	828	986.7163093	119.1686364
671	147	183.7252119	124.9831374
673	207	396.2994316	191.4490008
675	43	64.5780818	150.1815855
677	308	251.1350804	-81.5373638
680	0	-14.9819866	Inf
681	46	60.3122233	131.1135290
687	486	653.2734639	134.4184082
689	340	345.0910281	101.4973612
709	19	17.1022799	-90.0119997
710	172	181.1679846	105.3302236
720	0	-22.6526691	Inf
721	418	491.4214166	117.5649322
723	12	9.9416464	-82.8470530
727	6	8.3322481	138.8708009
732	234	310.7928166	132.8174430
745	70	147.3767959	210.5382798
746	0	-14.9819866	Inf
752	1352	1344.0959298	-99.4153794
759	5	-14.2013707	284.0274142
766	144	168.8727759	117.2727610
767	6	-23.3156577	388.5942942
773	215	293.1685314	136.3574565
780	0	-14.9819866	Inf
802	252	217.7276878	-86.3998761
803	250	201.3697964	-80.5479186
808	131	187.5486753	143.1669277
810	308	246.4569944	-80.0185047
819	5	-3.7918599	75.8371985
828	54	56.6412368	104.8911792

	Real	Predecido	Dif%
830	109	272.0864967	249.6206392
832	120	92.3338406	-76.9448671
836	25	47.5848394	190.3393577
838	0	5.9672219	Inf
841	2	-9.3124488	465.6224423
847	106	211.3877634	199.4224183
857	146	201.9598705	138.3286784
860	283	287.9887609	101.7628130
867	5	-14.9467700	298.9353992
875	1061	978.2377286	-92.1995974
877	120	144.7321164	120.6100970
879	79	125.2288867	158.5175781
880	10	15.1580252	151.5802517
881	52	49.1053555	-94.4333759
894	603	477.3773002	-79.1670481
902	14	19.5900750	139.9291075
906	9	0.0813483	-0.9038698
923	16	31.5343629	197.0897680
929	5	11.6446727	232.8934544
943	306	278.8726972	-91.1348684
944	86	138.3410785	160.8617191
955	716	446.1601422	-62.3128690
956	63	313.8705857	498.2072789
960	25	37.0176824	148.0707294
969	221	220.3962610	-99.7268149
972	31	27.6237010	-89.1087130
973	9	25.4553322	282.8370245
974	3	-0.4612307	15.3743571
975	36	43.9723748	122.1454854
979	81	115.7079280	142.8492939
980	116	125.9957719	108.6170447
987	1684	1016.3939562	-60.3559356
1005	0	-23.6882620	Inf
1010	27	37.0527573	137.2324343
1012	150	285.6006641	190.4004427
1015	16	16.6672757	104.1704732
1016	1911	1657.9214017	-86.7567452
1022	442	466.9619581	105.6475018
1024	5	3.0358681	-60.7173618
1030	36	79.8860089	221.9055804
1039	32	59.8894876	187.1546487
1042	131	122.1936992	-93.2776330
1050	52	55.1399242	106.0383158
1065	155	136.2898615	-87.9289429
1069	70	99.7840141	142.5485916
1072	1725	1121.0230835	-64.9868454
1075	120	114.5874269	-95.4895224
1078	0	-2.0590072	Inf
1083	29	12.7110090	-43.8310654
1088	17	24.2220717	142.4827746
1089	0	-23.4958543	Inf
1090	74	183.8280033	248.4162206

	Real	Predecido	Dif%
1091	7	-16.5917990	237.0256993
1092	1716	1262.8455235	-73.5923965
1094	50	48.2916050	-96.5832099
1095	18	24.8874312	138.2635068
1099	16	6.7521005	-42.2006283
1103	7	-5.8251948	83.2170679
1104	52	63.7234554	122.5451066
1114	19	16.6994875	-87.8920395
1117	153	110.2745736	-72.0748847
1119	8	21.2437744	265.5471794
1131	298	159.9388099	-53.6707416
1133	433	432.5564851	-99.8975716
1135	42	30.0032468	-71.4363018
1136	214	329.6722597	154.0524578
1142	380	776.6680735	204.3863351
1152	0	-15.1451475	Inf
1160	97	81.8928116	-84.4255790
1161	370	324.9645028	-87.8282440
1166	16	65.5885795	409.9286218
1171	26	42.8795003	164.9211551
1177	346	265.0591905	-76.6067025
1183	4	10.4344935	260.8623384
1186	5	-1.2420801	24.8416020
1190	5	18.1751055	363.5021107
1195	239	294.4823859	123.2143874
1201	2290	1508.3472334	-65.8666914
1204	310	427.1762089	137.7987771
1213	540	743.1872486	137.6272683
1231	1261	751.7966702	-59.6190857
1232	46	84.2147867	183.0756233
1238	7	-15.9457623	227.7966039
1245	342	290.5422937	-84.9538870
1246	122	204.3317413	167.4850338
1248	147	190.8885552	129.8561600
1259	14	15.5503091	111.0736362
1267	170	236.9514441	139.3832024
1269	339	219.9120858	-64.8708218
1277	79	105.1936223	133.1564840
1285	4	-8.7132238	217.8305938
1287	24	35.4277011	147.6154214
1291	10	24.3201274	243.2012738
1292	24	31.1585057	129.8271069
1302	6	-10.6297037	177.1617290
1325	73	123.9296634	169.7666622
1328	7	-23.6095439	337.2791990
1341	65	82.8653157	127.4851011
1344	4	-13.3801824	334.5045602
1347	112	157.8127154	140.9042102
1348	7	6.1467154	-87.8102206
1352	353	194.3822795	-55.0658016
1355	10	1.5261773	-15.2617734
1381	41	13.6231345	-33.2271573

	Real	Predecido	Dif%
1385	9	-1.7139643	19.0440480
1391	303	351.9984098	116.1710923
1399	134	165.0856455	123.1982429
1404	444	278.0709002	-62.6285811
1405	77	41.1095559	-53.3890336
1406	112	90.7568725	-81.0329218
1415	5	17.2995055	345.9901108
1427	180	254.0900525	141.1611403
1428	27	41.0084007	151.8829655
1430	762	534.1087005	-70.0930053
1439	42	73.4618062	174.9090623
1441	75	81.6316490	108.8421986
1447	128	185.4046874	144.8474120
1448	404	281.0888123	-69.5764387
1451	111	102.3600349	-92.2162476
1453	355	240.4279801	-67.7261916
1457	44	101.1516938	229.8902131
1481	133	117.4636469	-88.3185315
1483	1178	646.4411661	-54.8761601
1488	18	3.7654874	-20.9193745
1489	185	153.4199004	-82.9296759
1495	211	202.7232264	-96.0773585
1496	242	415.3795293	171.6444336
1498	24	59.7656922	249.0237174
1508	22	48.7015413	221.3706423
1513	315	231.2904740	-73.4255473
1517	122	74.3285117	-60.9250096
1524	50	127.3301560	254.6603119
1528	34	39.1550027	115.1617728
1531	100	86.4683515	-86.4683515
1532	19	74.5559690	392.3998369
1534	50	59.2656526	118.5313052
1538	76	139.1872423	183.1411083
1563	28	6.1092136	-21.8186200
1566	71	111.3647845	156.8518091
1579	7	-23.5637038	336.6243407
1589	106	74.8484756	-70.6117694
1594	68	64.3534037	-94.6373583
1595	217	190.5166124	-87.7956739
1599	52	32.9356349	-63.3377595
1602	14	34.8323310	248.8023642
1603	48	64.2881238	133.9335912
1604	44	55.0974231	125.2214161
1605	67	46.1628798	-68.8998206
1608	201	162.3119426	-80.7522103
1612	10	25.6659863	256.6598634
1617	265	207.2387559	-78.2033041
1618	2	0.0006807	-0.0340330
1620	15	40.9170530	272.7803534
1624	41	55.0664437	134.3083994
1629	156	424.8746340	272.3555346
1634	3	20.3929074	679.7635796

	Real	Predecido	Dif%
1642	14	31.1941677	222.8154833
1644	8	21.8200295	272.7503691
1647	15	6.7353024	-44.9020163
1657	103	109.3602361	106.1749865
1658	38	35.9144331	-94.5116659
1663	17	32.2355035	189.6206086
1666	18	64.3045340	357.2474111
1667	108	133.0662908	123.2095285
1673	68	71.2313118	104.7519292
1681	11	25.1404250	228.5493179
1686	37	56.6240185	153.0378879
1687	16	26.9798898	168.6243110
1699	6	14.1134231	235.2237188
1711	59	57.4985586	-97.4551841
1712	28	30.6588194	109.4957834
1713	32	33.2546707	103.9208459
1722	4	3.1540489	-78.8512221
1723	23	43.9084574	190.9063364
1724	4	-11.3481099	283.7027471
1739	3	11.0801294	369.3376473
1744	115	131.2971985	114.1714769
1750	55	43.7978536	-79.6324611
1752	126	119.0257511	-94.4648818
1753	13	70.1341353	539.4933483
1758	55	206.8372725	376.0677682
1771	32	30.7312827	-96.0352583
1772	3	13.7206113	457.3537115
1774	74	85.5799825	115.6486251
1780	45	40.3507254	-89.6682786
1781	35	45.6305251	130.3729289
1786	6	37.7287869	628.8131148
1794	72	91.3611576	126.8904967
1798	171	130.1144983	-76.0903499
1809	38	97.4051134	256.3292457
1812	14	26.4641153	189.0293952
1822	35	52.3136682	149.4676234
1823	5	6.7457770	134.9155402
1835	125	-10.6297037	8.5037630
1837	39	27.1425676	-69.5963273
1842	50	59.2640025	118.5280050
1843	110	79.4150424	-72.1954931
1847	42	110.0185863	261.9490149
1851	24	27.0819542	112.8414757
1853	52	46.4289267	-89.2863975
1857	208	241.0026257	115.8666470
1872	0	-0.0542260	Inf
1879	43	150.6687027	350.3923318
1881	149	126.6439434	-84.9959352
1884	9	4.4679523	-49.6439145
1889	72	51.6472726	-71.7323230
1890	13	16.8144763	129.3421254
1901	6	21.0196796	350.3279936

	Real	Predecido	Dif%
1909	6	15.0914391	251.5239855
1923	53	98.7027177	186.2315427
1926	116	52.2493367	-45.0425316
1930	180	329.7032390	183.1684661
1940	53	86.0477221	162.3541927
1945	89	360.7211468	405.3046594
1956	314	228.2093946	-72.6781512
1983	225	221.4375968	-98.4167097
1991	21	0.7752515	-3.6916739
1998	53	56.1856895	106.0107348
2017	352	240.6420158	-68.3642090
2024	54	78.9785099	146.2564998
2028	41	106.6730035	260.1780573
2029	2	2.1392038	106.9601880
2037	0	-11.8665920	Inf
2040	18	30.1332580	167.4069888
2044	104	108.1954877	104.0341228
2054	0	-19.3592129	Inf
2057	7	5.9254794	-84.6497060
2064	17	22.7658651	133.9168535
2067	90	154.5680806	171.7423118
2076	0	-18.0197747	Inf
2077	132	183.3441844	138.8971094
2082	10	32.5910949	325.9109490
2084	3	6.2333971	207.7799035
2100	14	53.7584087	383.9886335
2101	0	-17.4610254	Inf
2103	9	22.2662036	247.4022627
2105	4	3.6600596	-91.5014900
2109	57	67.2591903	117.9985795
2110	105	142.8991625	136.0944405
2116	26	0.1823728	-0.7014339
2126	8	9.7727810	122.1597629
2136	14	-14.0157246	100.1123185
2139	65	250.0904784	384.7545821
2149	30	0.6413303	-2.1377678
2159	7	-21.4614584	306.5922624
2160	6	-1.5955541	26.5925684
2162	0	8.3915589	Inf
2170	0	-2.8899069	Inf
2182	7	15.5565987	222.2371250
2189	323	309.2063456	-95.7295188
2195	29	13.8042761	-47.6009519
2196	3	18.9948735	633.1624489
2199	10	19.5450348	195.4503479
2207	33	25.4032057	-76.9794111
2215	2	-21.8062577	1090.3128868
2235	50	32.9909751	-65.9819502
2239	0	-1.1854513	Inf
2250	0	-23.3009232	Inf
2258	4	-11.4170696	285.4267399
2259	45	27.7362763	-61.6361695

	Real	Predecido	Dif%
2262	4	-23.2916740	582.2918502
2265	10	33.8472130	338.4721301
2270	59	64.5315189	109.3754558
2279	6	57.9570313	965.9505214
2282	36	109.6230167	304.5083799
2285	20	46.3946811	231.9734057
2287	8	-4.0592104	50.7401301
2288	6	0.6758124	-11.2635401
2291	22	33.5174514	152.3520516
2294	5	-2.2910436	45.8208730
2297	22	20.5726318	-93.5119628
2298	32	56.1175665	175.3673953
2301	55	66.6278177	121.1414867
2307	5	11.4448500	228.8970003
2308	0	-12.5873556	Inf
2310	0	-16.9354640	Inf
2311	6	-0.6440377	10.7339618
2315	9	-7.4753593	83.0595480
2316	0	-6.9507741	Inf
2326	4	-12.2054116	305.1352912
2327	3	-6.1614559	205.3818630
2334	4	7.2403591	181.0089763
2350	3	-21.5515037	718.3834577
2356	30	43.5031174	145.0103913
2362	7	-19.3191584	275.9879772
2367	7	-10.3659469	148.0849551
2368	4	-1.7687098	44.2177453
2376	125	156.4988120	125.1990496
2396	92	127.0796087	138.1300095
2411	7	-13.6067584	194.3822632
2415	7	-14.3789552	205.4136452
2416	29	32.4785621	111.9950417
2417	17	14.4494501	-84.9967655
2425	7	-14.6417358	209.1676550
2426	16	15.9233789	-99.5211183
2429	0	-14.8341947	Inf
2437	0	-17.1982447	Inf
2442	0	-15.8853174	Inf
2445	13	10.1309466	-77.9303584
2451	0	-16.6726833	Inf
2455	16	-2.4835025	15.5218906
2457	19	14.3354375	-75.4496713
2459	4	-13.5193151	337.9828769
2466	4	-10.8915082	272.2877056
2467	0	-15.6225368	Inf
2469	0	-15.3597561	Inf
2471	0	-16.9354640	Inf
2474	35	15.6483648	-44.7096136
2475	0	-17.4610254	Inf
2477	10	4.3487953	-43.4879531
2483	2	-6.9497979	347.4898973
2499	0	-15.6215606	Inf

	Real	Predecido	Dif%
2511	4	12.4959728	312.3993188
2513	0	-17.5406279	Inf
2514	0	-16.4099026	Inf
2521	0	-16.6726833	Inf
2533	0	-15.8843412	Inf
2548	0	-17.9766540	Inf
2553	6	-16.9354640	282.2577330
2572	0	-27.2274597	Inf
2573	0	-18.0237100	Inf
2586	2	-13.2565344	662.8267195
2588	5	-10.8915082	217.8301645
2589	2	-10.1031662	505.1583084
2595	58	100.1159535	172.6137129
2596	0	-16.9354640	Inf
2604	0	-17.1982447	Inf
2626	2	-4.2909305	214.5465254
2634	0	-5.3731138	Inf
2639	2	-17.1213965	856.0698268
2643	0	-2.1212036	Inf
2651	0	-17.4610254	Inf
2655	0	-16.4099026	Inf
2657	0	-16.5603924	Inf
2660	0	-13.7510352	Inf
2671	12	-9.1107662	75.9230514
2673	33	49.7962233	150.8976464
2678	64	115.1844979	179.9757780
2679	109	174.8357134	160.3997371
2680	40	17.9475924	-44.8689810
2688	42	10.8204382	-25.7629480
2689	67	59.7377733	-89.1608557
2698	254	460.2155375	181.1872195
2706	0	-24.4792557	Inf
2709	12	1.9260226	-16.0501884
2716	129	117.5495241	-91.1236621
2718	43	85.2053184	198.1519032
2720	332	279.9479875	-84.3216830
2736	60	120.5927235	200.9878726
2738	97	72.3512462	-74.5889136
2749	11	-4.3807138	39.8246712
2753	13	4.8166101	-37.0508473
2754	8	-15.6802833	196.0035412
2757	13	-3.4397598	26.4596910
2764	32	73.9230769	231.0096152
2777	38	145.4042767	382.6428334
2781	11	-2.5412490	23.1022640
2783	0	-15.3774481	Inf
2786	6	2.4804947	-41.3415783
2791	7	-13.0605285	186.5789780
2792	0	-17.7825288	Inf
2799	0	-21.9870197	Inf
2802	11	-4.1179331	37.4357559
2812	7	-21.1986777	302.8382526

	Real	Predecido	Dif%
2821	56	34.5108276	-61.6264778
2822	155	144.3531540	-93.1310671
2833	4	-17.7383731	443.4593286
2837	205	134.9222604	-65.8157368
2839	0	-15.3774481	Inf
2843	13	-5.1690559	39.7619684
2848	40	67.0956325	167.7390813
2852	33	62.6283609	189.7829117
2853	22	4.0282681	-18.3103095
2855	104	142.5136892	137.0323934
2858	96	97.8424641	101.9192335
2859	52	33.7239770	-64.8538019
2860	86	39.0073104	-45.3573376
2868	43	86.0077898	200.0181159
2884	18	40.8175640	226.7642445
2889	15	20.5849427	137.2329515
2899	150	116.2371121	-77.4914081
2908	5	-14.6291606	292.5832111
2924	32	58.0454260	181.3919564
2931	87	119.5246856	137.3846961
2935	5	16.0279532	320.5590640
2937	41	41.3598341	100.8776442
2943	9	-8.5852048	95.3911644
2952	532	419.4845313	-78.8504758
2955	3	-15.5476274	518.2542450
2959	108	111.5055683	103.2458966
2962	0	-22.7753618	Inf
2974	160	127.0111202	-79.3819501
2979	10	19.2695478	192.6954783
2989	5	-16.7314060	334.6281207
2991	32	13.4022892	-41.8821539
2992	7	-22.2498004	317.8542918
2994	6	-0.0034875	0.0581258
3005	4	-9.7263729	243.1593225
3010	139	176.6751782	127.1044448
3015	235	433.1491269	184.3187774
3022	2	-11.5040793	575.2039634

```r
# Predicción

newdata <- data.frame(
  lat = 50,
  lon = 4,
  Visualizaciones = 300
)

# Predecir el número de solicitudes respecto a la latitud, longitud y el número de visualizaciones.
predict(modelo, newdata)
```

```
##        1
## 54.67582
```

Se obtiene un modelo con un coeficiente de R^2 ajustado en torno al 80%. Se trata de un valor aceptable.

La tabla tambien representa las diferencias entre los valores reales y predecidos. Algunos de ellos incluso son tomados como número negativos, esto refleja que el modelo podría ser mejorado teniendo en cuenta otras nuevas variables.

4.3.3 Modelo de Regresión logística

Hacemos un modelo de regresión logística con el que poder predecir si una oferta dispondrá de un sistema rápido para mandar la candidatura "Quick.Application".

```
lgm <- glm(formula = Quick.Application ~ Level + Type + Solicitudes + Visualizaciones + Recommended.Fla
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(lgm)
```

```
##
## Call:
## glm(formula = Quick.Application ~ Level + Type + Solicitudes +
##     Visualizaciones + Recommended.Flavor, family = binomial(link = logit),
##     data = jobs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5909  -0.6409  -0.3587   0.6140   2.8540
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.2915949  0.2179132  -1.338  0.18086
## LevelDirector                      1.8044767  0.4632006   3.896 9.79e-05
## LevelEjecutivo                     1.9319727  0.9663333   1.999  0.04558
## LevelIntermedio                    1.2943331  0.1398127   9.258  < 2e-16
## LevelNo corresponde               -0.5222445  0.2361947  -2.211  0.02703
## LevelPrácticas                    -1.0772465  0.5822032  -1.850  0.06427
## LevelSin experiencia              -1.5931064  0.1752669  -9.090  < 2e-16
## TypeJornada completa              -0.8178515  0.1985983  -4.118 3.82e-05
## Solicitudes                        0.0163629  0.0013057  12.532  < 2e-16
## Visualizaciones                   -0.0027469  0.0002851  -9.634  < 2e-16
## Recommended.FlavorCOMPANY_RECRUIT -0.8053258  0.1689682  -4.766 1.88e-06
## Recommended.FlavorJOB_SEEKER_QUALIFIED  0.4076369  0.1579011   2.582  0.00983
##
## (Intercept)
## LevelDirector                     ***
## LevelEjecutivo                    *
## LevelIntermedio                   ***
## LevelNo corresponde               *
## LevelPrácticas                    .
## LevelSin experiencia              ***
## TypeJornada completa              ***
## Solicitudes                       ***
## Visualizaciones                   ***
## Recommended.FlavorCOMPANY_RECRUIT ***
## Recommended.FlavorJOB_SEEKER_QUALIFIED **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3016.1  on 2373  degrees of freedom
## Residual deviance: 1998.3  on 2362  degrees of freedom
##   (652 observations deleted due to missingness)
## AIC: 2022.3
##
## Number of Fisher Scoring iterations: 6
```

Este es el resumen del modelo y para la matriz de confusion:

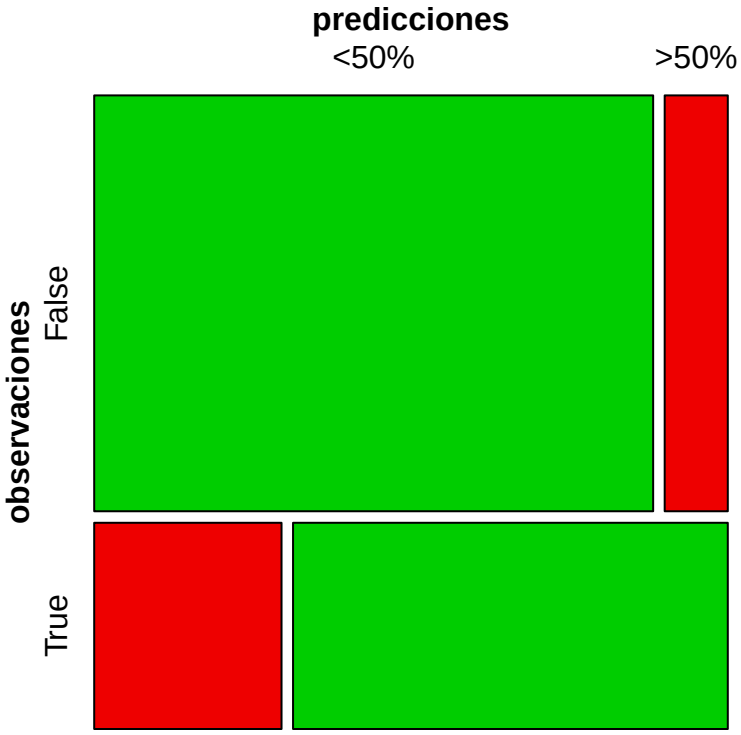
```
predicciones <- ifelse(test = lgm$fitted.values >= 0.5, yes = ">50%", no = "<50%")
matriz_confusion <- table(lgm$model$Quick.Application, predicciones,
                          dnn = c("observaciones", "predicciones"))
matriz_confusion
```

```
##              predicciones
## observaciones <50% >50%
##         False 1426  161
##         True   237  550
```

```
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.0.2
```

```
mosaic(matriz_confusion, shade = T, colorize = T,
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```



Podemos ver que hay 161 falsos positivos y 237 falsos negativos.

```
sensibilidad <- label_percent(accuracy = 0.01)(matriz_confusion[4]/(matriz_confusion[4]+matriz_confusion
especificidad <- label_percent(accuracy = 0.01)(matriz_confusion[1]/(matriz_confusion[1]+matriz_confusi
cat("sensibilidad: ",  sensibilidad)
```

```
## sensibilidad:  69.89%
```

```
cat("\nespecificidad: ", especificidad)
```

```
##
## especificidad:  89.86%
```

La sensibilidad es del 69.89% y la especificidad del 89.86%.

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.0.2
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following object is masked from 'package:colorspace':
##
##     coords
```

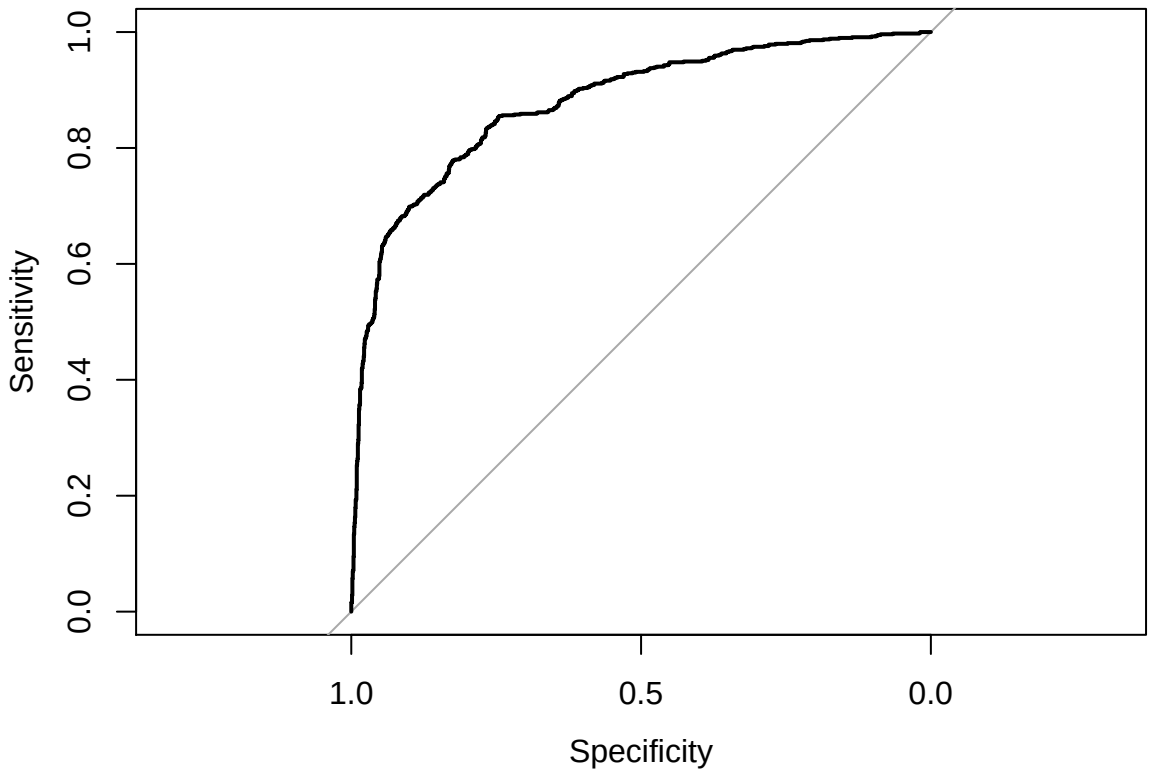
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
prob=predict(lgm, newdata=jobs, type="response")
r=roc(response=jobs$Quick.Application, predictor=prob, data=data)
```

```
## Setting levels: control = False, case = True
```

```
## Setting direction: controls < cases
```

```
plot (r)
```



```
auc(r)
```

```
## Area under the curve: 0.8777
```

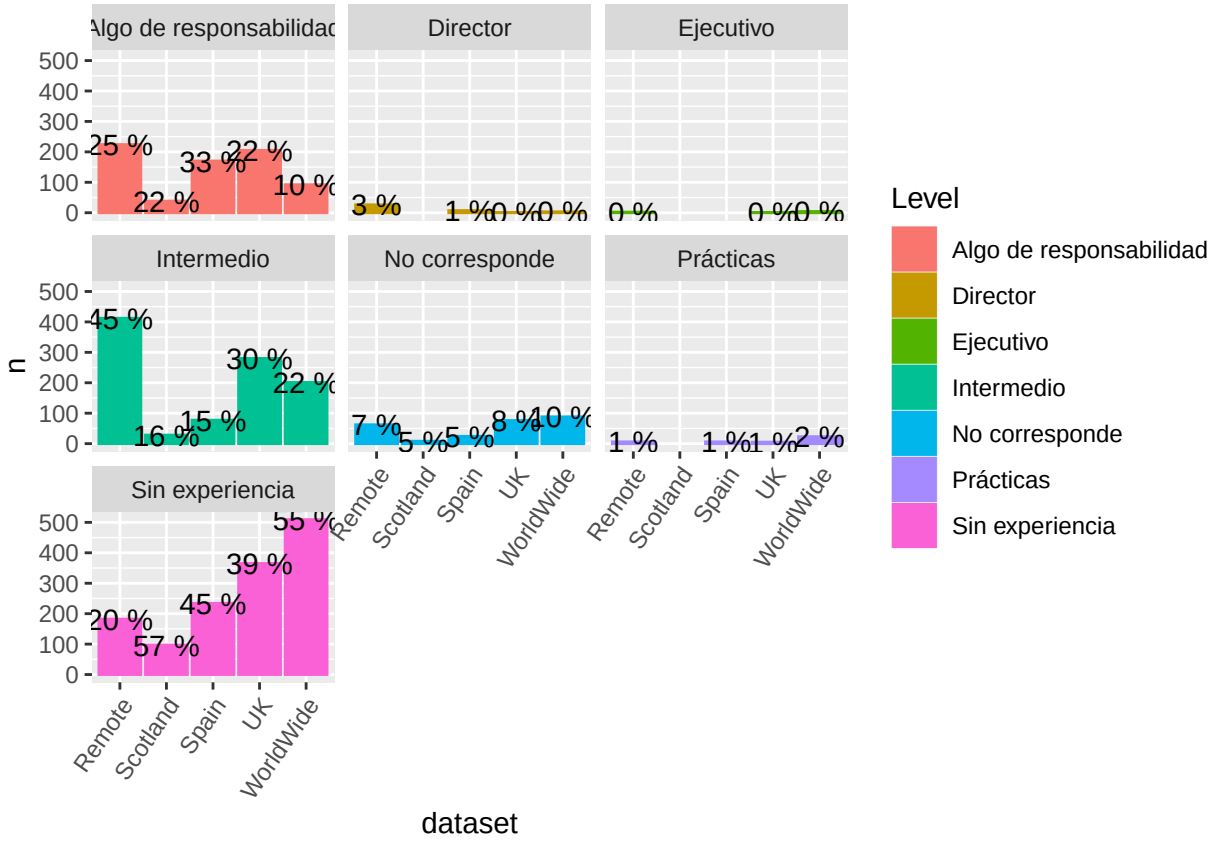
El modelo obtenido es bastante bueno ya que cuenta con un AUC de 0.8777312 y con una sensibilidad y especificidad aceptables por lo que se puede decir que se puede determinar si habrá Quick.Application en función de la oferta.

5. Representación de los resultados a partir de las tablas y gráficas.

A continuación la idea es obtener una representación de los resultados según la localización y las distintas variables para poder encontrar patrones interesantes. Como se ha podido comprobar en apartados anteriores, parece que hay una diferencia notable entre latitudes positivas y negativas. En cuanto a la variable longitud también cabe destacar que cobra especial importancia sobre todo suponiendo que para ciertas longitudes significará grupo de países como Europa, América y Asia y también hay océanos y mares entre ellos.

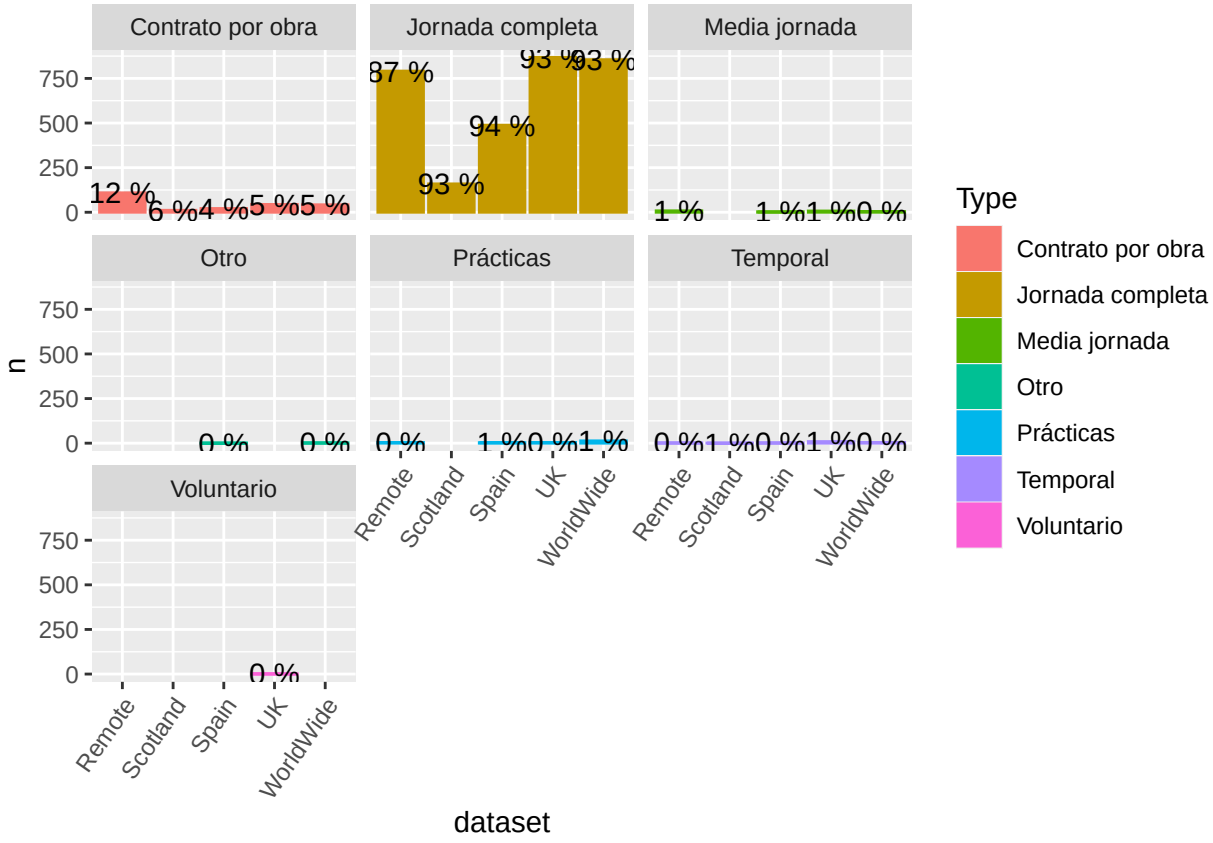
5.1 Principales diferencias.

```
jobs_with_duplicates %>%
  group_by(dataset) %>%
  count(Level) %>%
  mutate(freq = round(n / sum(n) * 100, 0)) %>%
  ggplot(mapping = aes(y = n, x = dataset, color=Level, fill=Level)) + geom_bar( stat="identity") + geo
```

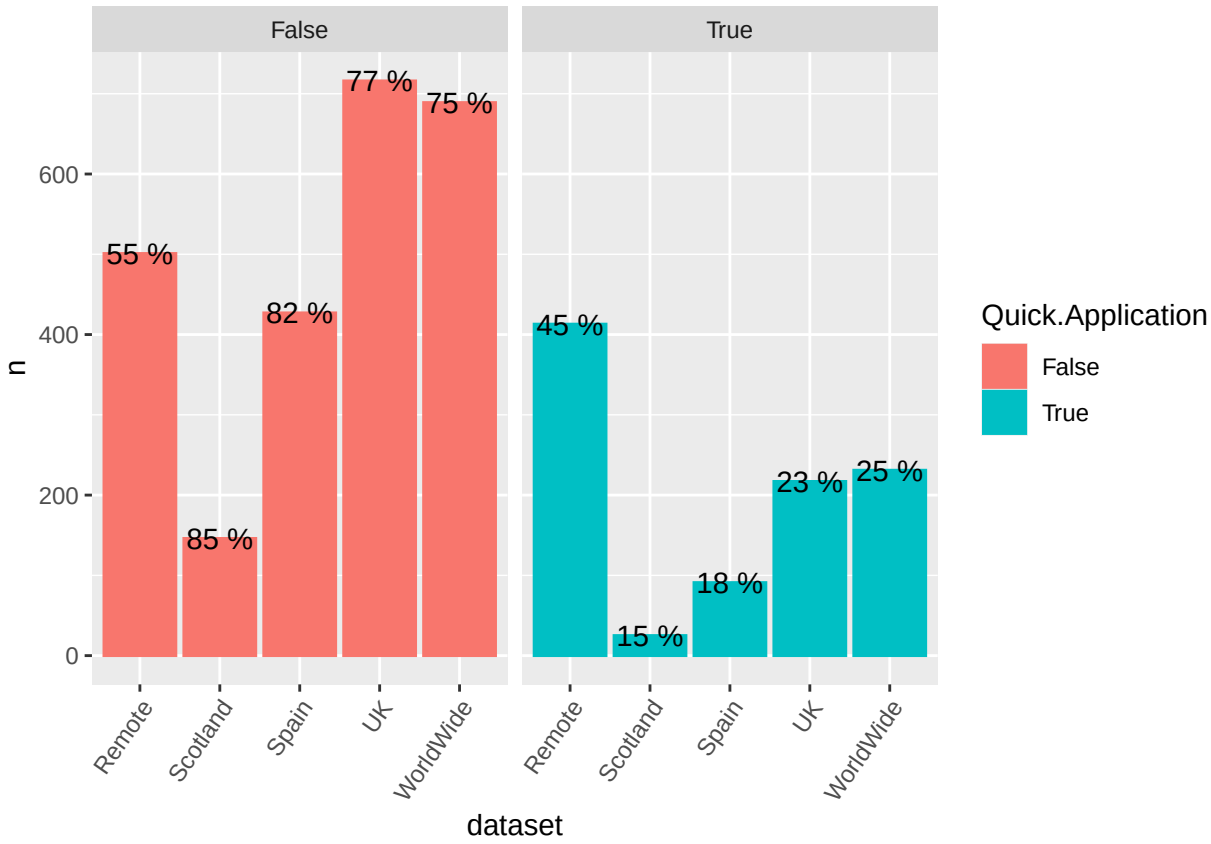



```
jobs_with_duplicates %>%
  group_by(dataset) %>%
```

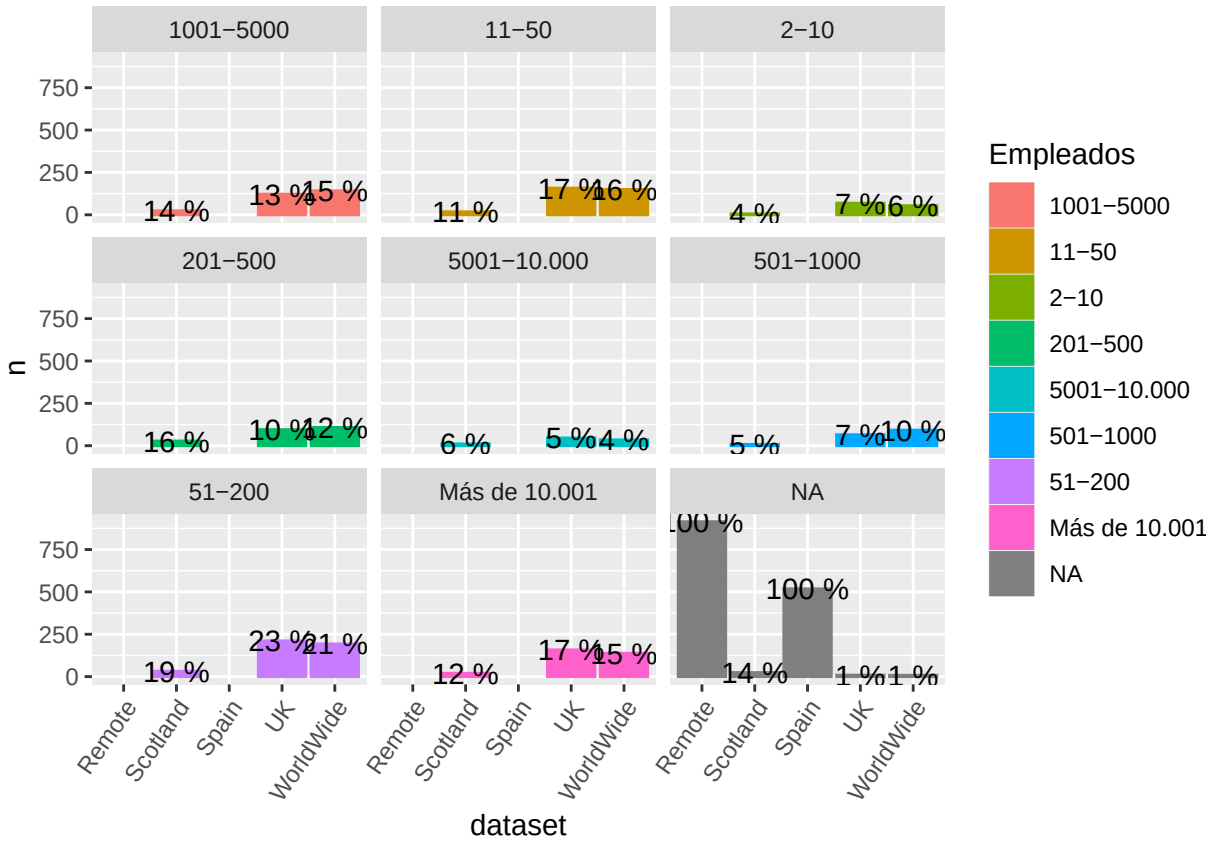
```
  count(Type) %>%
  mutate(freq = round(n / sum(n) * 100, 0)) %>%
  ggplot(mapping = aes(y = n, x = dataset, color=Type, fill=Type)) + geom_bar( stat="identity") + geom_
```




```
jobs_with_duplicates %>%
  group_by(dataset) %>%
  count(Quick.Application) %>%
  mutate(freq = round(n / sum(n) * 100, 0)) %>%
  ggplot(mapping = aes(y = n, x = dataset, color=Quick.Application, fill=Quick.Application)) + geom_bar
```



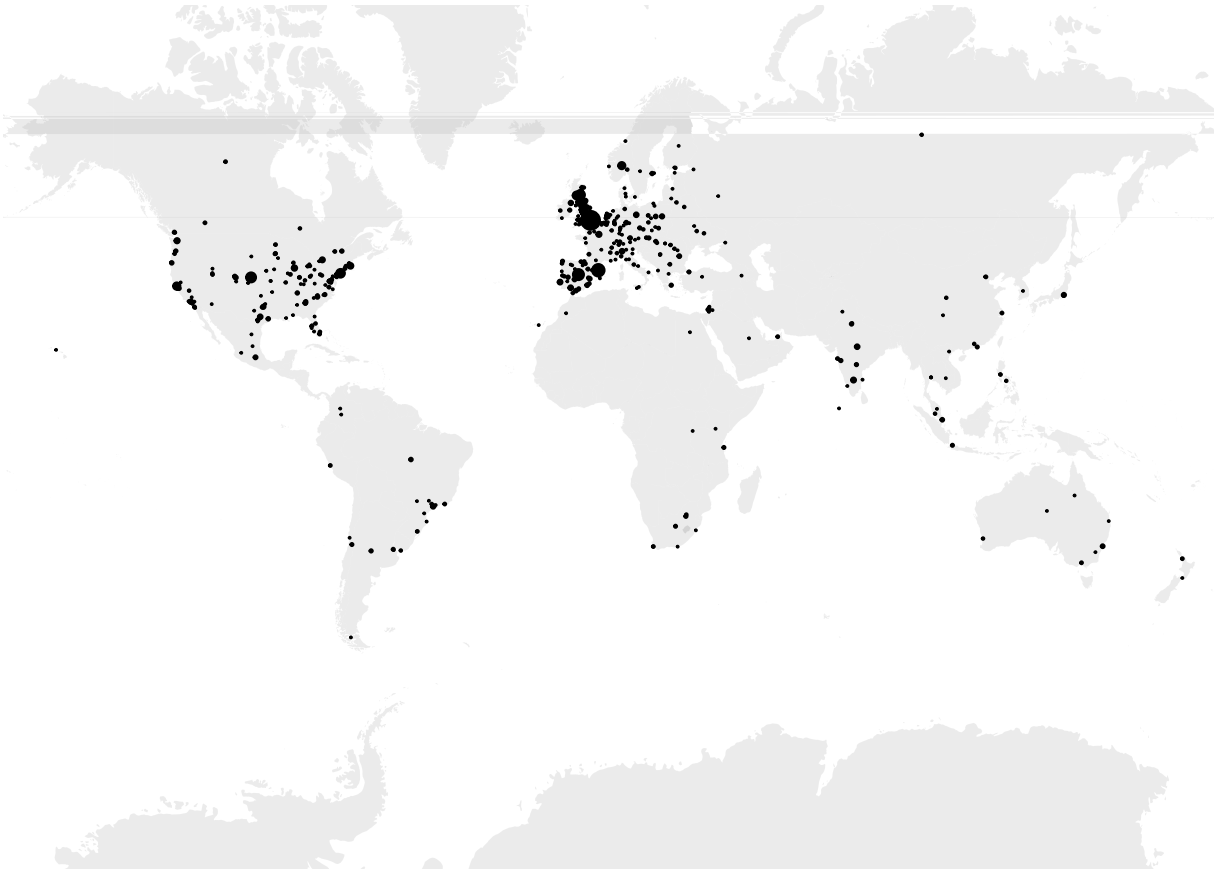

```
jobs_with_duplicates %>%
  group_by(dataset) %>%
  count(Empleados) %>%
  mutate(freq = round(n / sum(n) * 100, 0)) %>%
  ggplot(mapping = aes(y = n, x = dataset, color=Empleados, fill=Empleados)) + geom_bar( stat="identity
```



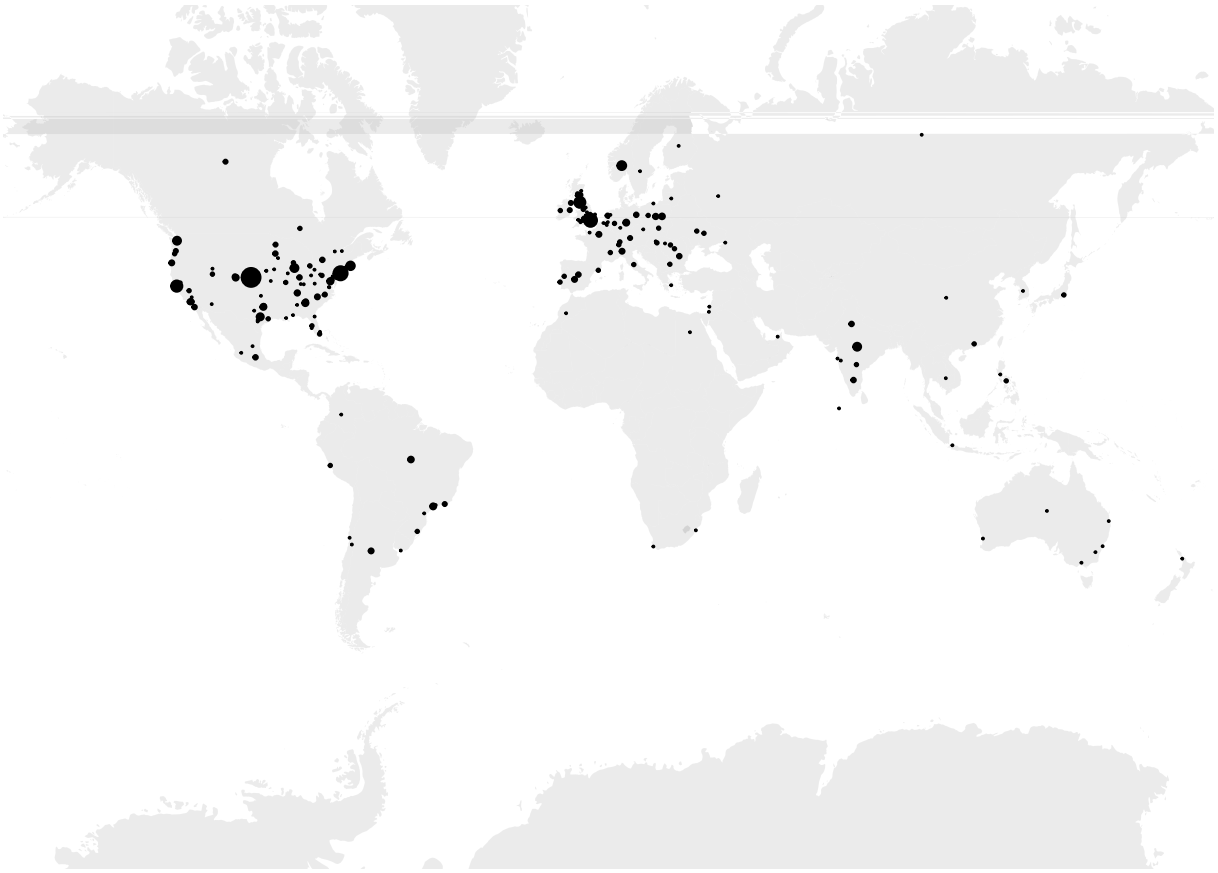
5.2 Mapas por localización y a nivel global

En este caso lo interesante será ver las distribuciones de las ofertas según los distintos mapas.

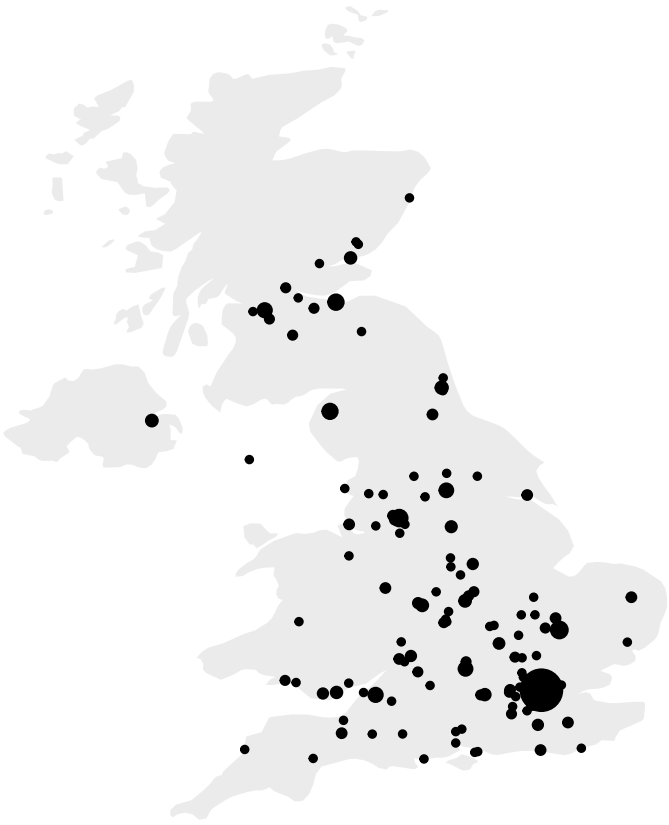
```
NI_world <- map_data("world")

p_world <- jobs_merged_with_duplicates %>%
  ggplot() +
    geom_polygon(data = NI_world, aes(x=long, y = lat, group = group), fill="grey", alpha=0.3) +
    geom_count(aes(x=lon, y=lat)) +
    scale_size_continuous(range=c(0.01,3)) +
    scale_color_viridis(option="inferno") +
    theme_void() +
    coord_map() +
    theme(legend.position = "none")
p_world
```



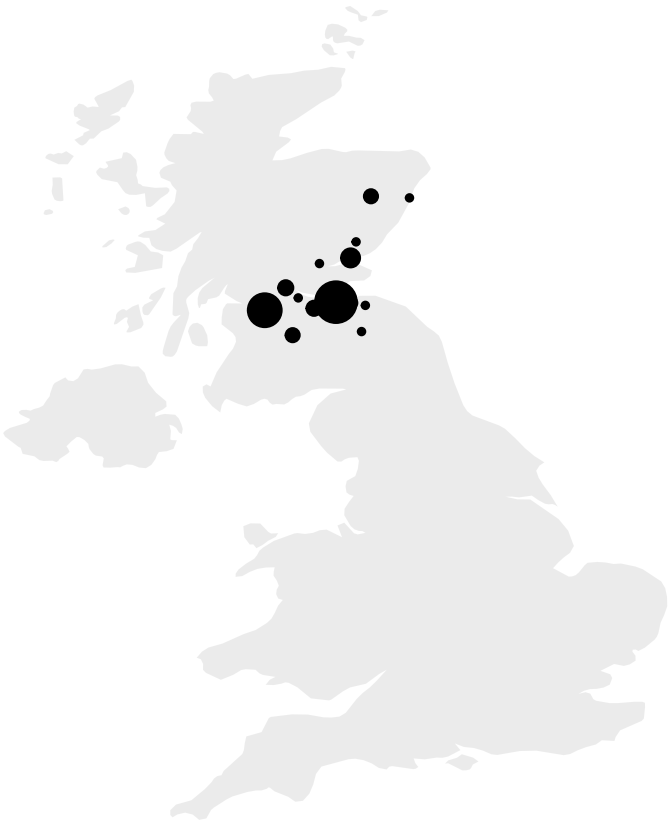
```
p_remote <- jobs_merged_with_duplicates[which(jobs_merged_with_duplicates$dataset=="Remote"),] %>%
  ggplot() +
    geom_polygon(data = NI_world, aes(x=long, y = lat, group = group), fill="grey", alpha=0.3) +
    geom_count(aes(x=lon, y=lat)) +
    scale_size_continuous(range=c(0.01,3)) +
    scale_color_viridis(option="inferno") +
    theme_void() +
    coord_map() +
    theme(legend.position = "none")
p_remote
```



```
NI_uk <- map_data("world") %>%
  filter(region == "UK")
p_uk <- jobs_merged_with_duplicates[which(jobs_merged_with_duplicates$dataset=="UK"),] %>%
  ggplot() +
    geom_polygon(data = NI_uk, aes(x=long, y = lat, group = group), fill="grey", alpha=0.3) +
    geom_count(aes(x=lon, y=lat)) +
    scale_size_continuous(range=c(1,7)) +
    scale_color_viridis(option="inferno") +
    theme_void() +
    coord_fixed(ratio = 1.3,
              xlim = c(-10,3),
              ylim = c(50, 59)) +
    theme(legend.position = "none")
p_uk
```



```
NI_scotland <- map_data("world") %>%
  filter(region == "UK")
p_scotland <- jobs_merged_with_duplicates[which(jobs_merged_with_duplicates$dataset=="Scotland"),] %>%
  ggplot() +
    geom_polygon(data = NI_scotland, aes(x=long, y = lat, group = group), fill="grey", alpha=0.3) +
    geom_count(aes(x=lon, y=lat)) +
    scale_size_continuous(range=c(1,7)) +
    scale_color_viridis(option="inferno") +
    theme_void() +
    coord_fixed(ratio = 1.3,
                xlim = c(-10,3),
                ylim = c(50, 59)) +
    theme(legend.position = "none")
p_scotland
```



```
NI_spain <- map_data("world") %>%
  filter(region == "Spain")
p_spain <- jobs_merged_with_duplicates[which(jobs_merged_with_duplicates$dataset=="Spain"),] %>%
  ggplot() +
    geom_polygon(data = NI_spain, aes(x=long, y = lat, group = group), fill="grey", alpha=0.3) +
    geom_count(aes(x=lon, y=lat)) +
    scale_size_continuous(range=c(1,15)) +
    scale_color_viridis(option="inferno") +
    theme_void() +
    ylim(35,45) +
    coord_map() +
    theme(legend.position = "none")
p_spain
```

```
## Warning: Removed 24 rows containing non-finite values (stat_sum).
```

Se puede observar las grandes potencias y ciudades que obtienen el mayor número de ofertas en el portal de trabajo. Esto resulta interesante para poder decidir en qué ciudades enfocarse y sobre todo, comprobar en qué lugares se puede optar por el teletrabajo respecto a otras.

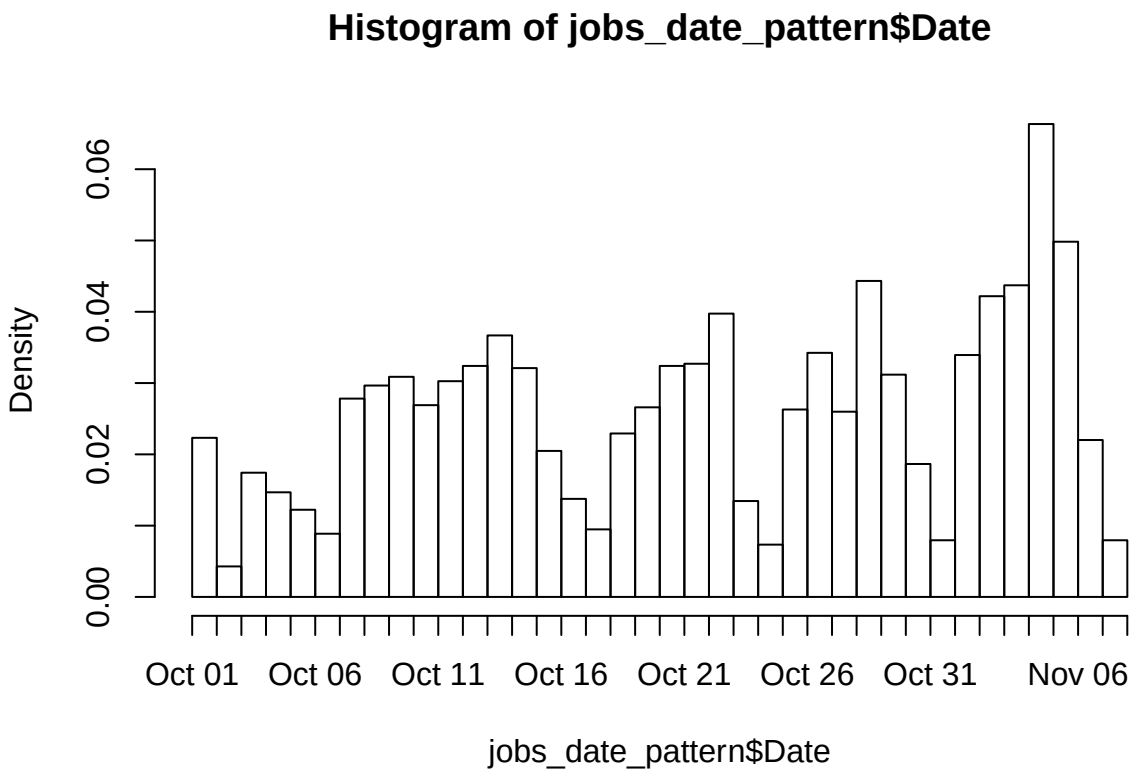
5.3 Patrones de publicación de trabajos por fecha

A continuación se realiza una representación de las fechas obtenidas en cada localización observando un claro patrón semanal. Esto es interesante para conocer cuándo son publicadas el mayor número de ofertas.

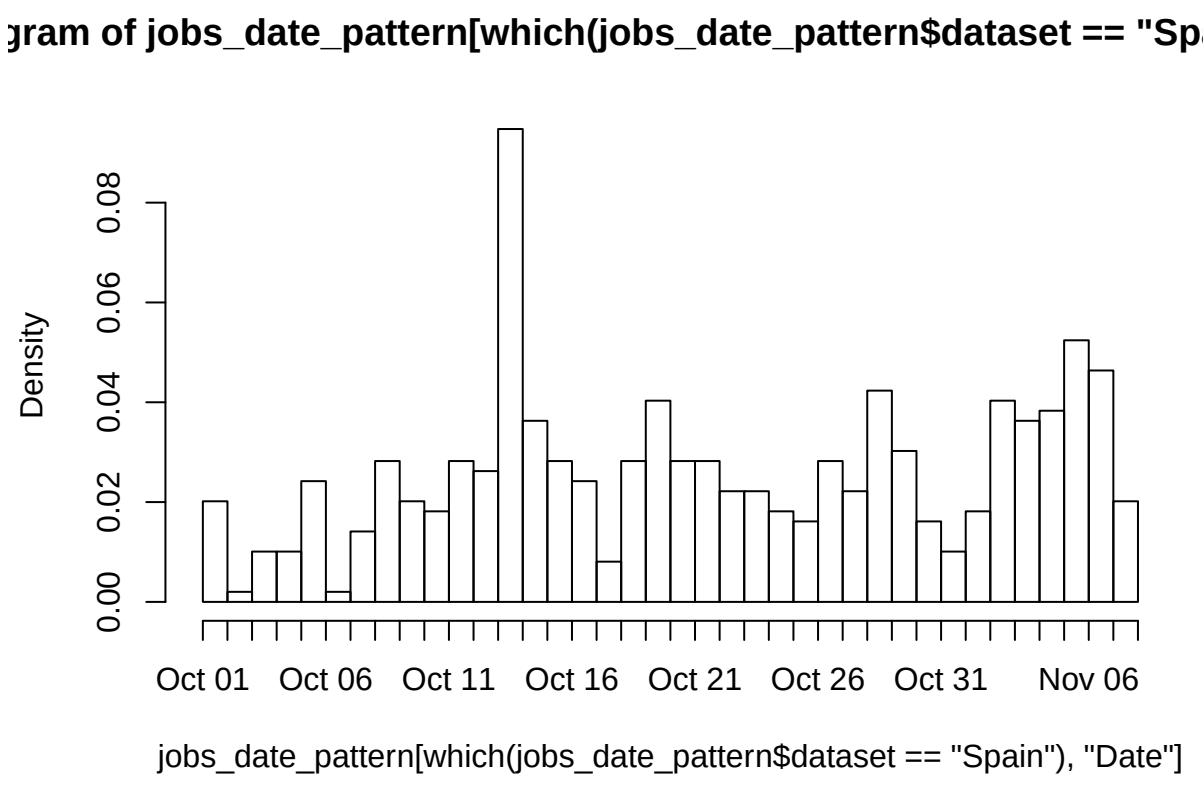
```
jobs_date_pattern <- jobs_merged_with_duplicates

jobs_date_pattern <- jobs_date_pattern %>%
  filter(Date >= "2020-10-01")

nrow(jobs_date_pattern)
```

```
## [1] 3271
```

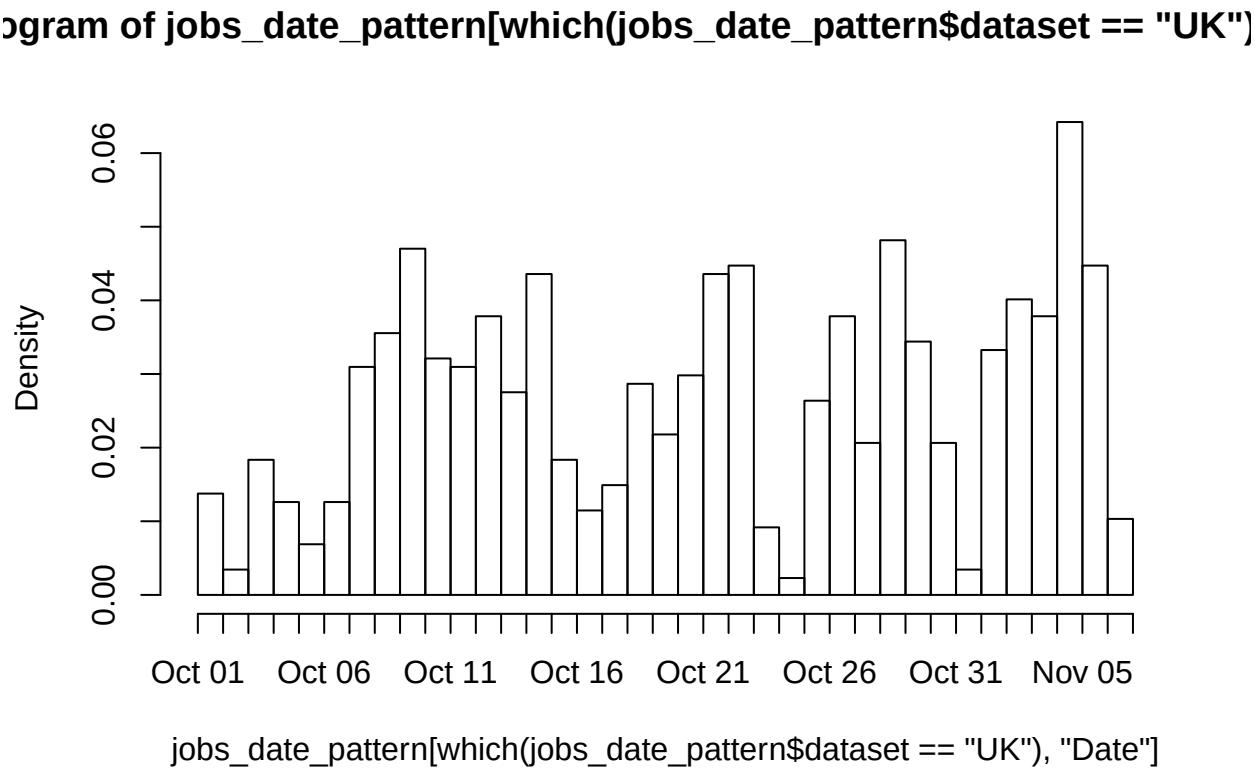
```
hist(jobs_date_pattern$Date,"days")
```



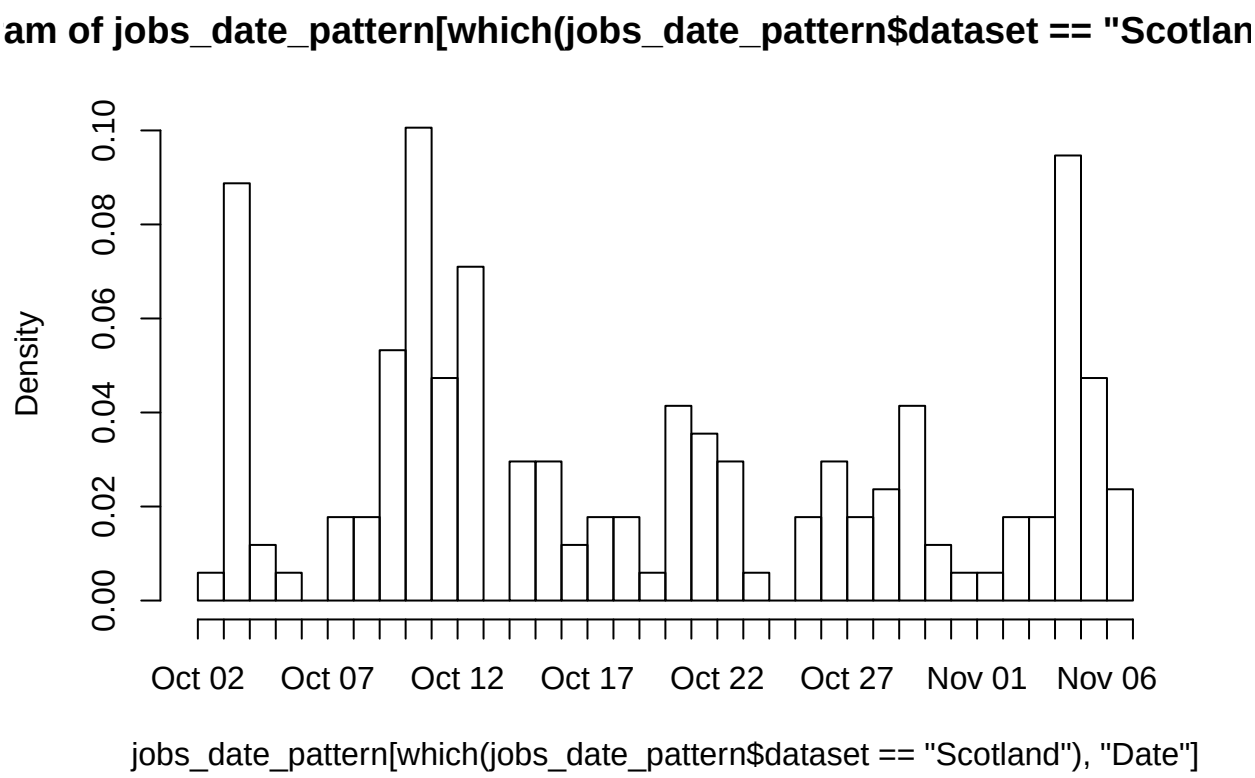

```
hist(jobs_date_pattern[which(jobs_date_pattern$dataset=="Spain"),"Date"],"days")
```



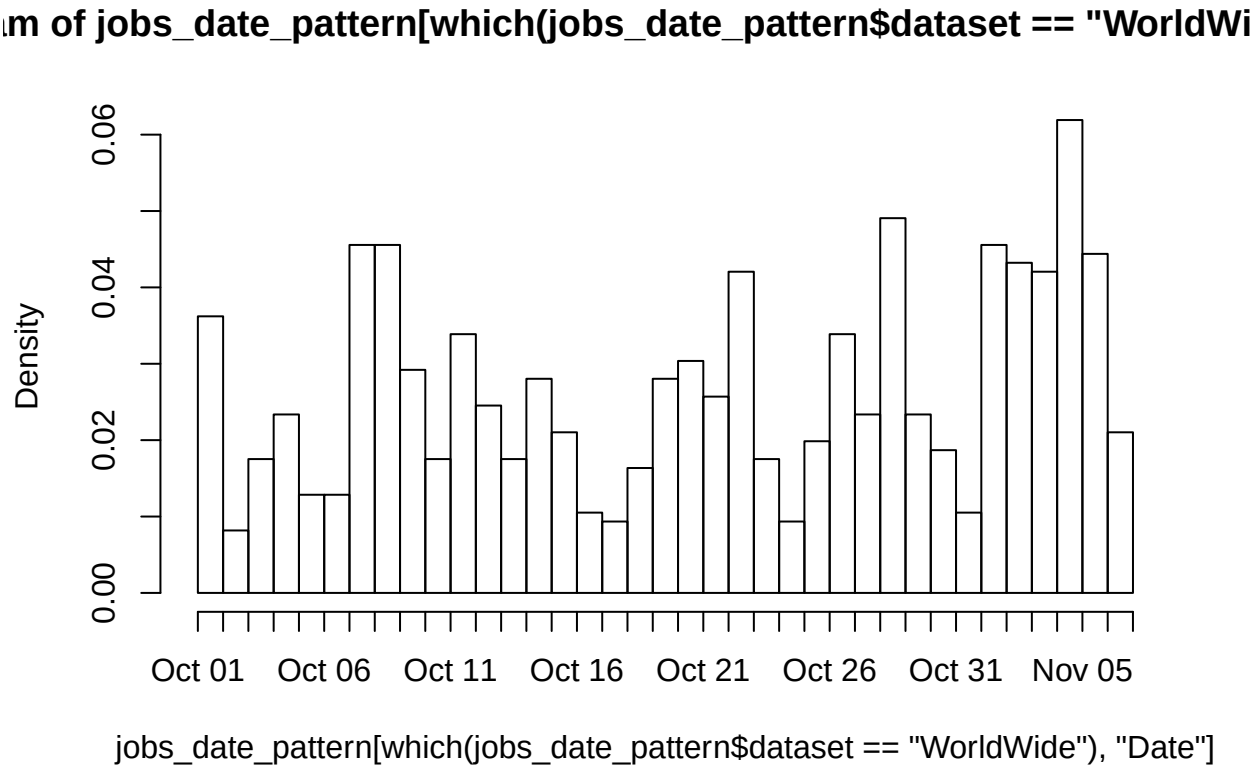

```
hist(jobs_date_pattern[which(jobs_date_pattern$dataset=="UK"),"Date"],"days")
```



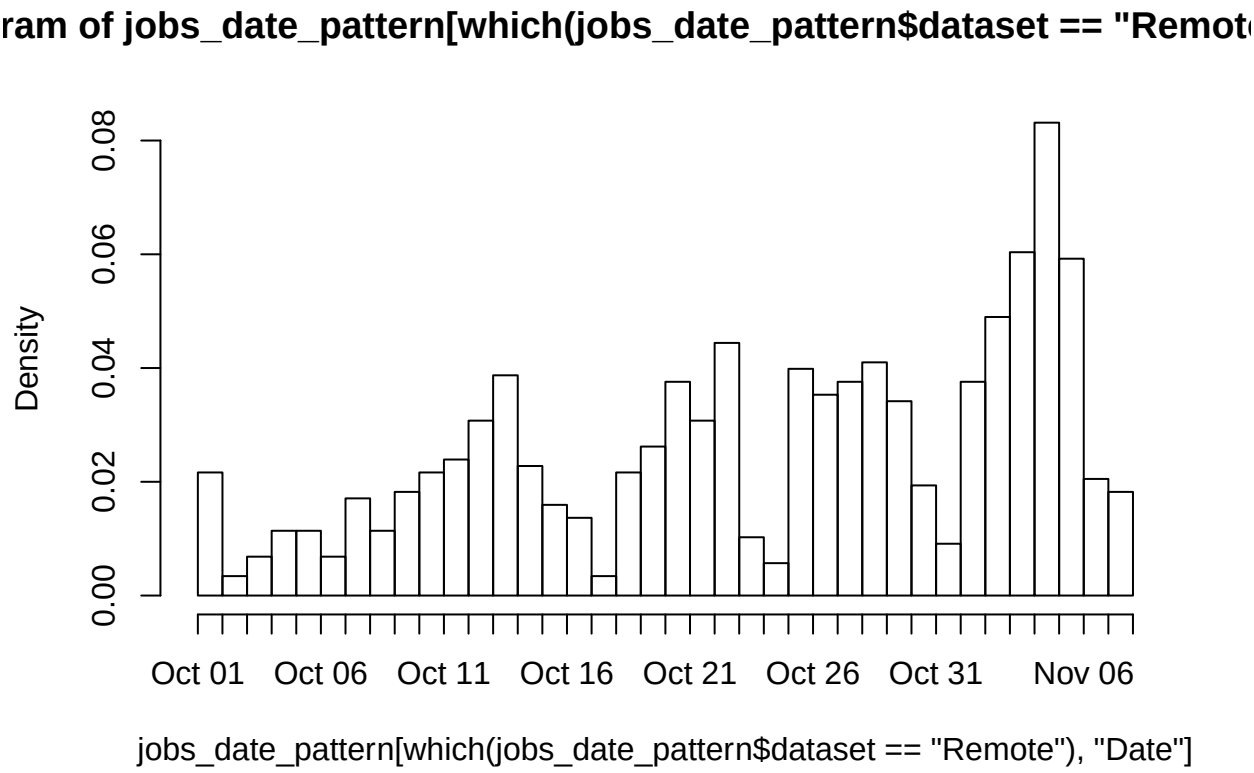

```
hist(jobs_date_pattern[which(jobs_date_pattern$dataset=="Scotland"),"Date"],"days")
```




```
hist(jobs_date_pattern[which(jobs_date_pattern$dataset=="WorldWide"),"Date"],"days")
```




```
hist(jobs_date_pattern[which(jobs_date_pattern$dataset=="Remote"),"Date"],"days")
```



Se observa un dato interesante en algunas de estas gráficas en torno al 12 de Octubre que coincide con el día de la Hispanidad. Puede ser que estos días sean aprovechados por las empresas y agencias para realizar un mayor número de ofertas. Obtenemos esa información de valor para el futuro. Será interesante chequear el

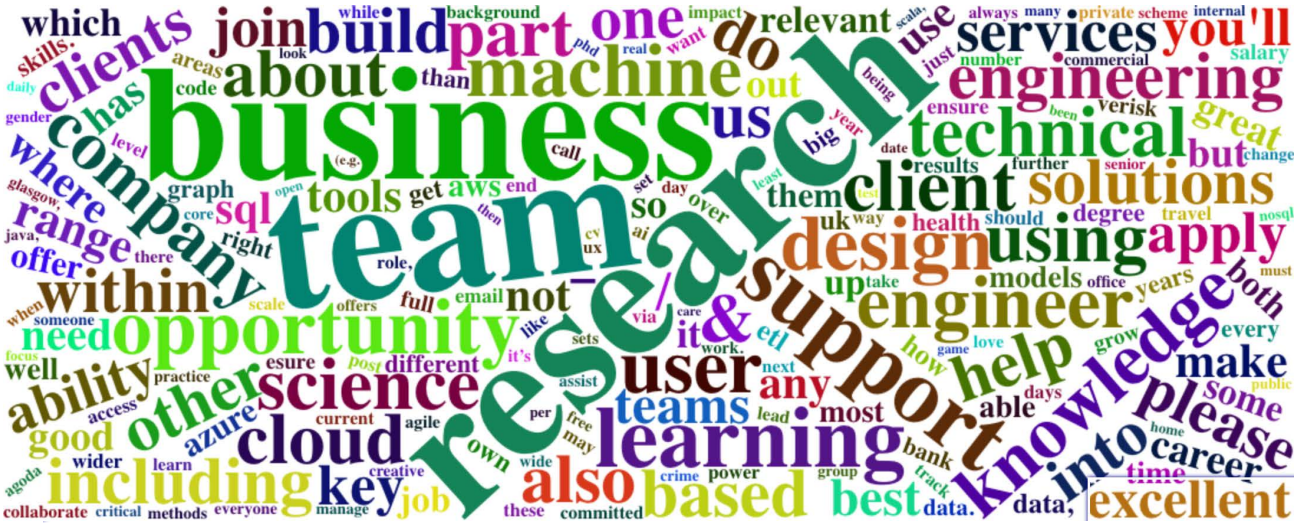
portal de empleo en torno a las fechas de fiestas nacionales.

5.4 Nube de palabras más usadas en UK

Por último se realiza otra de las representaciones que consideramos de gran importancia para evaluar que herramientas o lenguajes son más solicitados en las descripciones de los puestos de trabajo. En este caso se decide utilizar una única localización por el idioma.

```
UK$Description <- gsub(",", "", UK$Description, fixed = TRUE)
UK$Description <- gsub(".", "", UK$Description, fixed = TRUE)
UK$Description <- gsub("/", "", UK$Description, fixed = TRUE)

df1 <- data.frame(table(unlist(strsplit(tolower(UK$Description), " "))))

borrar <- c("and"," ","","to","the","of","a","in","with","for","you","our","is","as","we","on","or","ar
df1 <- df1[!(df1$Var1 %in% borrar),]

set.seed(1234)

wordcloud(words = df1$Var1, freq = df1$Freq, min.freq = 1, max.words=200, random.order=FALSE, rot.per=0
```


```
word_dataframe <- df1[order(-df1$Freq),]

wordcloud2(data=word_dataframe[2:400,], size=1.6, color='random-dark')
```




```
word_dataframe[1:25,]
```

```
##              Var1 Freq
## 6829         data 8058
## 9374   experience 2954
## 22147        team 2498
## 19014    research 1719
## 4691     business 1710
## 13180    learning 1609
## 19875     science 1287
## 13871     machine 1190
## 7602  development 1045
## 16425      people  894
## 7475       design  881
## 20871   solutions  853
## 11208        help  850
## 17574     product  800
## 3048    analytics  772
## 5827      company  758
## 22400  technology  758
## 12940   knowledge  738
## 8698  engineering  737
## 19921   scientist  715
## 21843     support  714
## 22218       teams  711
```

```
## 18102      python  708
## 15787 opportunity  683
## 14680      models  653
```

Entre los lenguajes de programación, herramientas y plataformas/servicios cabe destacar estas coincidencias:

- R: 223 coincidencias
- SQL: 451 coincidencias
- AI: 308 coincidencias
- Python: 708 coincidencias
- BI: 136 coincidencias
- ETL: 183 coincidencias
- ML: 236 coincidencias
- Design: 881 coincidencias
- Azure: 225 coincidencias
- Cloud: 473 coincidencias
- AWS: 258 coincidencias

Lugares como: - London: 371 coincidencias - home: 152 coincidencias - remote: 176 coincidencias

En UK podríamos entender entonces en mayor profundidad si realmente 176 ofertas en remoto nos parecen suficientes respecto a otras localizaciones.

6. Resolución del problema. Conclusiones.

Estas serían las respuestas a las preguntas inicialmente hechas:

- ¿Qué variables son relevantes?

Si una oferta permite el trabajo en remoto. La localización de la oferta. El tipo de oferta. Las herramientas o lenguajes de programación más usados. El número de Solicitudes, Visualizaciones y si tiene enlace de aplicación rápida o no. A final, lo importante es entender si una oferta según el tipo de oferta o empresa será posiblemente muy solicitada o no. También es muy importante el nivel pedido.

- ¿Que diferencias hay segun la localización o paises?

Se puede afirmar que a latitudes positivas y por grupos de longitud haciendo referencia a grupo de países como grandes potencias son lugares donde más oportunidades laborales hay. Es algo obvio que en ciudades grandes hay un mayor número de ofertas, pero podemos ver que por ejemplo, Londres podría ser una ciudad europea interesante para un científico de datos.

- ¿Es una opción los puestos en Remoto?

No hay evidencia de que el Remoto sea una opción más allá de la situación de alarma social actual. Aún hay muchos puestos de trabajo que no permiten esta modalidad aunque se puede comprobar que según el tipo de oferta y localización dependerá mucho si hay opciones para el teletrabajo.

- ¿Hay diferencias entre latitudes positivas y negativas?

Si, existen diferencias a diferentes latitudes. También cabe destacar que en el ecuadro con latitud 0, a penas hay opciones.

- ¿Existe correlación entre si hay quick application y si se trata de una empresa grande o no?

Si, influye el númnero de empleados con la posibilidad de tener un enlace de aplicación rápida.

- ¿Existe correlación entre el tipo de puesto y el número de solicitudes?

Si, existe una relación directa entre el tipo de oferta y el número de solicitudes que reciben.

- ¿Qué tipo de puestos de trabajo prioriza LinkedIn en sus resultados?

Esta es una de las preguntas más interesantes que hemos podido constatar. Pues según el número total de ofertas en todo el mundo, será limitado, y en este caso, LinkedIn reflejará ofertas de trabajo más cercanas. Esto es porque existe una diferencia notable entre los resultados obtenidos en UK y US respecto al número total de ofertas que se pueden encontrar en la plataforma y con la limitación que nos encontramos al hacer web scraping. Esto no quiere decir que no se hayan podido obtener respuestas globales, pero si que claramente la localización del perfil influirá en que no se mostrarán ofertas aleatorias de todo el mundo.

- ¿Qué palabras son las más utilizadas en las descripciones de las ofertas de trabajo?

Para una localización como ha sido el caso de UK podemos observar como "team" y "research" han sido dos de las más repedidas. Sin tener en cuenta otras palabras como "data", que claramente sería la más utilizada con diferencia.

Contribuciones	Firma
Investigación previa	GMA, IAB
Redacción de las respuestas	GMA, IAB
Desarrolllo código	GMA, IAB