PRA 2 - Limpieza y Análisis de Datos

Jonathan Zambrano / Tatiana Piccolomini

Junio 2021

Contents

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset seleccionado contiene datos sobre trabajos de Data scientist escrapeada de Glassdoor, recordemos que Glassdoor es un sitio web estadounidense donde los empleados actuales y anteriores revisan las empresas de forma anónima. Glassdoor también permite a los usuarios enviar y ver salarios de forma anónima, así como buscar y solicitar puestos de trabajo en su plataforma.

Una utilidad de este dataset es poder realizar un analisis de salario por sectores , ver las puntuaciones que los empleados ponen. Un buen analisis de este dataset de forma particular puede ayudar a definir el salario pretendido al comenzar a trabajar como Data Scientist, saber en que sectores se requiere este perfil profesional, que tipo de empresas y cuales de estas esta mejor puntuada. De tal manera, que el análisis que se realizará se enfoca en entender la tendencia respecto de los requerimientos del mercado actual; así como la búsqueda de la relación entre la información empresarial disponible y su remuneración.

- Columnas Data Set Original:

“index” “Job.Title” “Salary.Estimate” “Job.Description” “Rating” “Company.Name” “Location”

“Headquarters” “Size” “Founded” “Type.of.ownership” “Industry” “Sector” “Revenue”

“Competitors”

- Columnas del Data set limpio:

“Job.Title” “Job.Type” “Salary.Estimate” “Salary.Mean” “Rating” “Company.Name” “Location”

“Headquarters” “Size” “Type.of.ownership” “Industry” “Sector” “Revenue” “Competitors”

Detalle de los atributos del dataset (672 observaciones):

- Job.Title: Título del trabajo. Ej: “Sr Data Scientist”.
- Salary.Estimate: Rango del valor del salario anual estimado (k USD). Ej: “$137K-$171K (Glassdoor est.).
- Rating: Valoración que los usuarios de la página dieron al anuncio. Ej: 3.1.
- Company.Name: Nombre de la empresa que posteo el anuncio. Ej: Healthfirst.
- Location: Ubicación de la empresa. Ej: New York, NY.
- Headquarters: Ubicación de la oficina central de la empresa. Ej: Boston, MA.
- Size: Cantidad de empleados que conforman a la empresa. Ej: 1001 to 5000 employees.
- Founded: Año de fundación de la empresa. Ej: 1993.
- Type.of.ownership: Tipo de organización a la que corresponde la empresa. Ej: Nonprofit Organization.
- Industry, Sector: Área de trabajo en la que se desempeñan las actividades de la empresa. Ej: Research & Development – Business Services.
- Revenue: Ganancias totales anuales de la empresa. Ej: $1 to $2 billion (USD).
- Competitors: Principales competidores. Ej: EmblemHealth, UnitedHealth Group, Aetna.

Fuente: https://www.kaggle.com/rashikrahmanpritom/data-science-job-posting-on-glassdoor

---

Lectura del fichero

A continuación, previo al desarrollo del dataset procederemos a realizar la lectura del fichero y la instalación de las librerias a utilizar:

```
# Instalación y llamado a las librerías a utilizar
if(!require(dplyr)){
install.packages('dplyr')
library(dplyr)
}
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
if(!require(stringr)){
install.packages('stringr')
library(stringr)
}
```

```
## Loading required package: stringr
```

```
if(!require(VIM)){
install.packages('VIM')
library(VIM)
}
```

```
## Loading required package: VIM
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
if(!require(ggplot2)){
install.packages('ggplot2')
library(ggplot2)
}
```

```
## Loading required package: ggplot2
```

```
if(!require(car)){
install.packages('car')
library(car)
}
```

```
## Loading required package: car
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
if(!require(kableExtra)){
install.packages('kableExtra')
library(kableExtra)
}
```

```
## Loading required package: kableExtra
```

```
##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
# Lectura de archivo obtenido de la fuente de kaggle
data <- read.csv('Uncleaned_DS_jobs.csv')
```

Así mismo, para tener una visión general del dataset, se presente un resumen y extracto de los datos contenidos:

```
# Presentación del resumen de los datos
summary(data)
```

```
##      index        Job.Title         Salary.Estimate    Job.Description
##  Min.   :  0.0   Length:672         Length:672         Length:672
##  1st Qu.:167.8   Class :character   Class :character   Class :character
##  Median :335.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :335.5
##  3rd Qu.:503.2
##  Max.   :671.0
##      Rating       Company.Name         Location         Headquarters
##  Min.   :-1.000   Length:672         Length:672         Length:672
##  1st Qu.: 3.300   Class :character   Class :character   Class :character
##  Median : 3.800   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 3.519
##  3rd Qu.: 4.300
##  Max.   : 5.000
##      Size              Founded     Type.of.ownership    Industry
##  Length:672         Min.   :  -1   Length:672         Length:672
##  Class :character   1st Qu.:1918   Class :character   Class :character
##  Mode  :character   Median :1995   Mode  :character   Mode  :character
##                     Mean   :1636
##                     3rd Qu.:2009
##                     Max.   :2019
##     Sector            Revenue          Competitors
##  Length:672         Length:672         Length:672
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
# Presentación de un extracto y tipo de variables
str(data)
```

```
## 'data.frame':    672 obs. of  15 variables:
##  $ index            : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Job.Title        : chr  "Sr Data Scientist" "Data Scientist" "Data Scientist" "Data Scientist" ..
##  $ Salary.Estimate  : chr  "$137K-$171K (Glassdoor est.)" "$137K-$171K (Glassdoor est.)" "$137K-$171
##  $ Job.Description  : chr  "Description\n\nThe Senior Data Scientist is responsible for defining, bu
##  $ Rating           : num  3.1 4.2 3.8 3.5 2.9 4.2 3.9 3.5 4.4 3.6 ...
##  $ Company.Name     : chr  "Healthfirst\n3.1" "ManTech\n4.2" "Analysis Group\n3.8" "INFICON\n3.5" ..
##  $ Location         : chr  "New York, NY" "Chantilly, VA" "Boston, MA" "Newton, MA" ...
##  $ Headquarters     : chr  "New York, NY" "Herndon, VA" "Boston, MA" "Bad Ragaz, Switzerland" ...
```

```
##  $ Size             : chr  "1001 to 5000 employees" "5001 to 10000 employees" "1001 to 5000 employees
##  $ Founded          : int  1993 1968 1981 2000 1998 2010 1996 1990 1983 2014 ...
##  $ Type.of.ownership: chr  "Nonprofit Organization" "Company - Public" "Private Practice / Firm" "Co
##  $ Industry         : chr  "Insurance Carriers" "Research & Development" "Consulting" "Electrical & 
##  $ Sector           : chr  "Insurance" "Business Services" "Business Services" "Manufacturing" ...
##  $ Revenue          : chr  "Unknown / Non-Applicable" "$1 to $2 billion (USD)" "$100 to $500 million
##  $ Competitors      : chr  "EmblemHealth, UnitedHealth Group, Aetna" "-1" "-1" "MKS Instruments, Pfe
```

Integración y selección de los datos de interés a analizar.

El presente dataset se encuentra completo, por lo cual, no será necesario realizar trabajos de integración de los datos, de igual manera, se considera que todos los atributos serán necesarios para el análisis, por lo que, posteriormente y en base a los resultados obtenidos se realizarán tareas de selección de datos según sea el caso. En este apartado se incluye la excepción de los atributos Index y Job Description, los cuales se considera que no poseen información relevante para el análisis propuesto; por lo cual, serán eliminados.

```
#Eliminamos Index y descripcion
data <- data[,-c(1,4)]
```

Limpieza de datos

En base a los datos revisamos y mostrados en el punto anterior, se presenta a continuación una serie de trabajos de limpieza de datos en base a cada uno de los atributos del dataset:

- Job.Title: De la revisión realizada a los datos contenidos en este apartado, se verifica que existen diferentes nombres ingresados en función del trabajo solicitado. Así mismo, se evidencia que existen palabras clave dentro de los datos, de los cuales se han seleccionado los siguientes valores:
  - Data Scientist.
  - Data Analyst.
  - Business Intelligence Analyst.
  - Machine Learning.
  - Data Engineer.
  - Other (Cualquier otra descripción). De tal manera, que se creará un nuevo atributo cuantitativo correspondiente al tipo de anuncio.

```
# Filtrado y creación del atributo Job.Type
data <- data %>%
  mutate(Job.Type = case_when(
    str_detect(Job.Title,"Data Scientist") ~ "Data Scientist",
    str_detect(Job.Title,"Data Analyst") ~ "Data Analyst",
    str_detect(Job.Title,"Business Intelligence Analyst") ~ "Business Intelligence Analyst",
    str_detect(Job.Title,"Machine Learning") ~ "Machine Learning Engineer",
    str_detect(Job.Title,"Data Engineer") ~ "Data Engineer",
  ))
# Se asigna el valor Other
data <- mutate_at(data, c("Job.Type"), ~ replace(., is.na(.), "Other"))
```

- Salary.Estimate: Para este atributo se evidencia que existen caracteres adicionales sobre el valor del salario, por lo cual, se procederá a eliminarlos, para finalmente obtener el rango de valores.

```
# Limpieza de la variable Salario
data$Salary.Estimate <- str_extract(data$Salary.Estimate,"[^(]+")
data$Salary.Estimate <- str_remove_all(data$Salary.Estimate,"\\$")
data$Salary.Estimate <- str_remove_all(data$Salary.Estimate,"\\K")
```

- Company.Name: En este apartado se verifica que al final de cada nombre se muestra el valor de su rating, por lo cual, se procede a eliminarlo.

```
# Limpieza del atributo Nombre de la empresa
data$Company.Name <- str_extract(data$Company.Name,"[^\\n]+")
```

- Size: Se evidencia que al final de la descripción del atributo se encuentra la palabra "employees", la cual, será eliminada.

```
# Limpieza del atributo Size
data$Size <- str_remove(data$Size, "\\ employees")
```

- Revenue: Ganancias totales anuales de la empresa. Ej: $1 to $2 billion (USD). Sobre estos valores se han eliminado los caracteres correspondientes a la moneda, considerando que los valores se encuentran en USD Dolars.

```
# Limpieza del atributo Revenue
data$Revenue <- str_remove_all(data$Revenue,"\\$")
data$Revenue <- str_remove(data$Revenue,fixed(" (USD)"))
```

```
# Ordenamiento de los atributos
data <- data[,c(1,14,2,3,4,5,6,7,8,9,10,11,12,13)]
# Presentación de los datos limpios
str(data)
```

```
## 'data.frame':    672 obs. of  14 variables:
##  $ Job.Title        : chr  "Sr Data Scientist" "Data Scientist" "Data Scientist" "Data Scientist" ..
##  $ Job.Type         : chr  "Data Scientist" "Data Scientist" "Data Scientist" "Data Scientist" ...
##  $ Salary.Estimate  : chr  "137-171 " "137-171 " "137-171 " "137-171 " ...
##  $ Rating           : num  3.1 4.2 3.8 3.5 2.9 4.2 3.9 3.5 4.4 3.6 ...
##  $ Company.Name     : chr  "Healthfirst" "ManTech" "Analysis Group" "INFICON" ...
##  $ Location         : chr  "New York, NY" "Chantilly, VA" "Boston, MA" "Newton, MA" ...
##  $ Headquarters     : chr  "New York, NY" "Herndon, VA" "Boston, MA" "Bad Ragaz, Switzerland" ...
##  $ Size             : chr  "1001 to 5000" "5001 to 10000" "1001 to 5000" "501 to 1000" ...
##  $ Founded          : int  1993 1968 1981 2000 1998 2010 1996 1990 1983 2014 ...
##  $ Type.of.ownership: chr  "Nonprofit Organization" "Company - Public" "Private Practice / Firm" "Co
##  $ Industry         : chr  "Insurance Carriers" "Research & Development" "Consulting" "Electrical & 
##  $ Sector           : chr  "Insurance" "Business Services" "Business Services" "Manufacturing" ...
##  $ Revenue          : chr  "Unknown / Non-Applicable" "1 to 2 billion" "100 to 500 million" "100 to 
##  $ Competitors      : chr  "EmblemHealth, UnitedHealth Group, Aetna" "-1" "-1" "MKS Instruments, Pfe
```

¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Luego del análisis del dataset, se verifica que, si bien todos los campos presentan valores completos existen datos con un valor numérico de "-1", el cual, se ha considerado como un valor o elemento vacío. Para lo cual, se reasignará el valor de estos datos como "NA".

```r
# Asiganación de valor NA a valores -1
data[data == -1] <- NA
```

De igual manera, se encuentra que existen valores denominados como desconocidos ("Unknown") que no se consideran elementos vacíos, sino como la denominación de valores no conocidos dentro de cada atributo.

Para la gestión de los valores o elementos vacíos se presenta una estadística de la cantidad de valores vacíos presentes en cada grupo de datos:

```r
#Busqueda de NAs
print('Cantidad de NA s')
```

```
## [1] "Cantidad de NA s"
```

```r
apply(is.na(data), 2, sum)
```

```
##         Job.Title          Job.Type   Salary.Estimate            Rating
##                 0                 0                 0                50
##      Company.Name          Location      Headquarters              Size
##                 0                 0                31                27
##           Founded Type.of.ownership          Industry            Sector
##               118                27                71                71
##           Revenue       Competitors
##                27               501
```

En primer lugar, se gestionarán los datos vacíos del atributo Rating, en el cual, considerando que se trata de una variable cualitativa, se utilizará el algoritmo de k-vecinos mas cercanos (k-NN) para aproximar el valor de los datos faltantes, en donde, tenemos lo siguiente:

```r
# Reemplazo de NAs numéricos
data$Rating <- kNN(data)$Rating
```

Como siguiente paso en la gestión de los elementos vacíos, luego de una inspección a los registros del dataset, se evidencia que existen registros que no cuentan con información completa de la empresa solicitante, donde, los únicos datos disponibles corresponden al del nombre y ubicación de la empresa; por lo que se considera que estos registros (27) no cuentan con información fiable. De tal manera, que procederemos a identificar y posteriormente eliminar estos registros.

Así se muestra un extracto de los datos que se pretende eliminar:

```r
# Identificación de filas sin información de compañia
missing <- data[rowSums(is.na(data[ , 8:11])) == 4, ]
# Presentación de un extracto de las filas a eliminar
head(missing)
```

```
##                           Job.Title                  Job.Type Salary.Estimate
## 155 ELISA RESEARCH SCIENTIST (CV-15)                     Other         90-109
## 159        Machine Learning Engineer Machine Learning Engineer        101-165
## 352                   Data Scientist            Data Scientist        122-146
## 358                   Data Scientist            Data Scientist        122-146
## 359                   Data Scientist            Data Scientist        122-146
## 360                   Data Scientist            Data Scientist        122-146
##     Rating             Company.Name          Location Headquarters Size Founded
## 155    5.0 Covid-19 Search Partners     Hauppauge, NY         <NA> <NA>      NA
## 159    4.3      Radical Convergence        Reston, VA         <NA> <NA>      NA
## 352    4.4         Point72 Ventures     Palo Alto, CA         <NA> <NA>      NA
## 358    4.4           Hatch Data Inc San Francisco, CA         <NA> <NA>      NA
## 359    4.4           Hatch Data Inc San Francisco, CA         <NA> <NA>      NA
```

```
## 360    4.4           Hatch Data Inc San Francisco, CA         <NA> <NA>      NA
##     Type.of.ownership Industry Sector Revenue Competitors
## 155              <NA>     <NA>   <NA>    <NA>        <NA>
## 159              <NA>     <NA>   <NA>    <NA>        <NA>
## 352              <NA>     <NA>   <NA>    <NA>        <NA>
## 358              <NA>     <NA>   <NA>    <NA>        <NA>
## 359              <NA>     <NA>   <NA>    <NA>        <NA>
## 360              <NA>     <NA>   <NA>    <NA>        <NA>
```

Posteriormente se eliminan los datos indicados:

```
# Eliminación de filas sin información de compañia
index <- which(rowSums(is.na(data[ , 8:11])) == 4)
data <- data[-index,]
```

Finalmente, se verifica que el resto de los datos faltantes se encuentran en los campos de Headcuarters, Founded, Industry, Sector y Competitors, los cuales, no pueden ser aproximados o establecidos debido a que corresponden a datos específicos de cada empresa, por lo cual, el tratamiento que daremos a estos datos será el de asignarles la etiqueta de "Unknown". Así mismo, es importante recalcar que el atributo Competitors presenta un 73% de valores pérdidos, por lo que posteriormente, este atributo no será considerado para el análisis.

```
# Cambio de NAs a Unknown
data <- mutate_at(data, c(7,9,11,12,14),~ replace(., is.na(.), "Unknown"))
```

Así ya contamos con un dataset son valores vacíos como se muestra a continuación:

```
#Busqueda de NAs
print('Cantidad de NA s')
```

```
## [1] "Cantidad de NA s"
```

```
apply(is.na(data), 2, sum)
```

```
##         Job.Title          Job.Type   Salary.Estimate            Rating
##                 0                 0                 0                 0
##      Company.Name          Location      Headquarters              Size
##                 0                 0                 0                 0
##           Founded Type.of.ownership          Industry            Sector
##                 0                 0                 0                 0
##           Revenue       Competitors
##                 0                 0
```

Considerando que los valores de salario se presentan en un atributo categórico, se procede a generar un nuevo atributo que permita obtener un valor estimado/promedio del salario percibido por cada trabajo, para lo cual, calculará el valor promedio de salario en base al rango categórico presentado:

```
# Calculo de Salario Promedio, ya que salario toma valores por rango.
data <- data %>% rowwise() %>%
  mutate(Salary.Mean = mean(c(as.numeric(strsplit(Salary.Estimate,"\\-")[[1]][1]),
                              as.numeric(strsplit(Salary.Estimate,"\\-")[[1]][2]))))
# Deagrupación de datos usada en el paso previo
data <- ungroup(data)
```

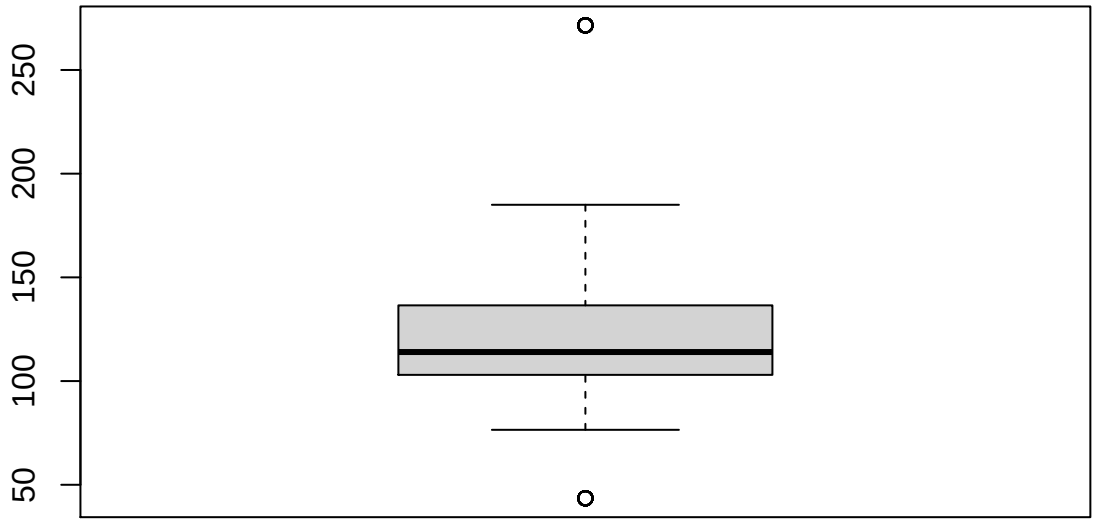
Identificación y tratamiento de valores extremos.

Para la identificación de los valores extremos o outliners utilizaremos la representación del gráfico de cajas (boxplots) con el objetivo de detectar aquellos valores que se encuentran mas alla de 3 desviaciones estándar de la media. Para nuestro caso, se evaluarán los outliners para los atributos de Salary.Mean y Rating.

- Valores Extremos en Salary.Mean

Para el análisis de valores extremos se presentará en primer lugar la gráfica de boxplot, y posteriormente evaluaremos los valores que se encuentra alejados tanto por arriba como por abajo.

```
# Evaluación del Diagrama de Cajas de Salary.Mean
boxplot(data$Salary.Mean) -> out_salary
```




```
# Obtención de los valores extremos
unique(out_salary$out)
```

```
## [1]  43.5 271.5
```

Del diagrama de cajas encontrado se obtienen que existen dos valores (mínimo y máximo) en salarios ofertados para Data Science. Para determinar si se pueden considerar como valores extremos se evaluará si existe alguna empresa que se encuentra ofertando salarios por fuera de la media, por lo cual, se presenta el listado de empresas que presentan un salario mínimo y máximo.

```
# Listado de Empresas que ofertan el salario mínimo
print("Empresas que ofertan salario mínimo:")
```

```
## [1] "Empresas que ofertan salario mínimo:"
```

```
data$Company.Name[data$Salary.Mean == 43.5]
```

```
##  [1] "Tempus Labs"
```

```
##  [2] "Grid Dynamics"
##  [3] "7Park Data"
##  [4] "Protolabs"
##  [5] "LinQuest"
##  [6] "Western Digital"
##  [7] "Trexquant Investment"
##  [8] "Ameritas Life Insurance Corp"
##  [9] "Fleetcor"
## [10] "Radiant Digital"
## [11] "Child Care Aware of America"
## [12] "IntelliPro Group Inc."
## [13] "Quest Integrity"
## [14] "Praxis Engineering"
## [15] "USI"
## [16] "Apex Systems"
## [17] "Pragmatics, Inc."
## [18] "Johns Hopkins University Applied Physics Laboratory"
## [19] "Cambridge Mobile Telematics"
## [20] "Phoenix Operations Group"
```

```
# Listado de Empresas que ofertan el salario máximo
print("Empresas que ofertan salario máximo:")
```

```
## [1] "Empresas que ofertan salario máximo:"
```

```
data$Company.Name[data$Salary.Mean == 271.5]
```

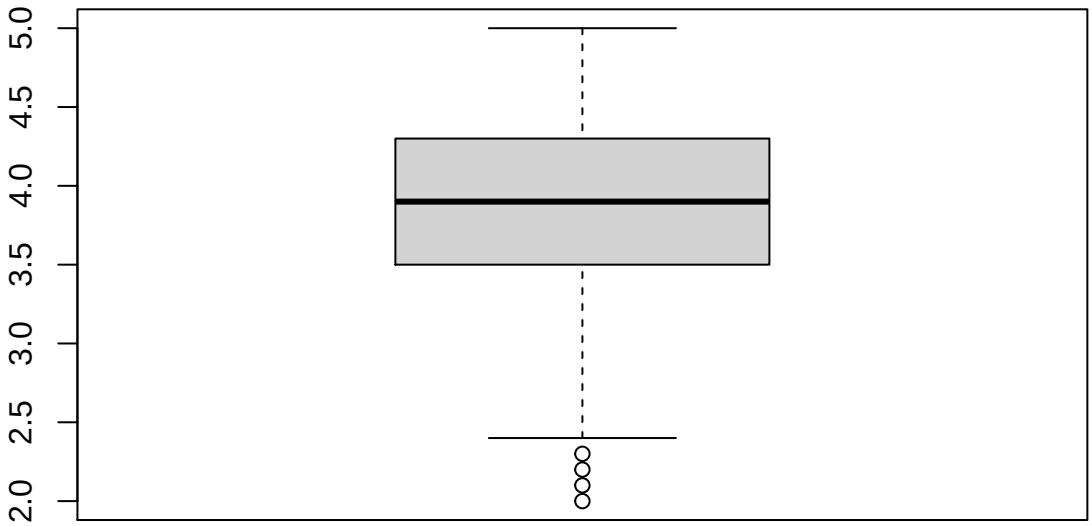
```
##  [1] "Roche"
##  [2] "AstraZeneca"
##  [3] "Creative Circle"
##  [4] "Blue Horizon Tek Solutions"
##  [5] "Maxar Technologies"
##  [6] "Sharpedge Solutions Inc"
##  [7] "Maxar Technologies"
##  [8] "Alaka`ina Foundation Family of Companies"
##  [9] "Southwest Research Institute"
## [10] "Hexagon US Federal"
## [11] "Klaviyo"
## [12] "Comtech Global Inc"
## [13] "Aveshka, Inc."
## [14] "10x Genomics"
## [15] "Southwest Research Institute"
## [16] "CompuForce"
## [17] "Smith Hanley Associates"
## [18] "Allen Institute"
## [19] "1-800-Flowers"
## [20] "Aptive"
```

De la evaluación realizada al listado de empresas se puede concluir que los valores encontrados no pueden considerarse extremos; así como se determina que no se encuentra una tendencia respecto a Job.Type o Revenue que indique el tipo de empresa o empleo que represente un salario máximo o mínino; por lo que, se puede concluir que en el mercado laboral los salarios correspondientes a Data Science se encuentran distribuidos uniformemente.

- Valores Extremos en Rating

Así mismo para el análisis de valores extremos se presentará la gráfica de boxplot, y los valores detectados.

```
# Evaluación del Diagrama de Cajas de Rating
boxplot(data$Rating) -> out_rating
```




```
unique(out_rating$out)
```

```
## [1] 2.2 2.3 2.1 2.0
```

En este caso, se obtuvo que se presentan valores extremos por debajo de la media, por lo cual, se analizará si existen empresas que presenten un rating menor a 2.5, valor límite de las 3 desviaciones estándar de la muestra.

```
# Listado de Empresas que presentan menor Rating
print("Empresas que presentan menor Rating:")
```

```
## [1] "Empresas que presentan menor Rating:"
```

```
data$Company.Name[data$Rating < 2.5]
```

```
## [1] "Great-Circle Technologies" "Crown Bioscience"
## [3] "United BioSource"          "Hive (CA)"
## [5] "Conagen"
```

En este caso, al existir algunas empresas por debajo de este umbral no se considera como un valor extremo, si no mas bien una representación de las empresas que, según el criterio de los usuarios, ofertan puestos no atractivos en el campo de Data Science.

Una vez que contamos con el dataset limpio, solo resta realizar la redefinición de las variables correspondientes al tipo factor.

```
# Cambio a Factor las variables categoricas.
data$Salary.Estimate = factor(data$Salary.Estimate)
data$Size = factor(data$Size)
data$Type.of.ownership = factor(data$Type.of.ownership)
data$Sector = factor(data$Sector)
data$Revenue = factor(data$Revenue)
data$Job.Type = factor(data$Job.Type)
```

Así, para culminar la etapa de limpieza de datos, se presenta el resumen y extracto de los datos limpios.

```
# Ordenamiento de los atributos
data <- data[,c(1,2,3,15, 4,5,6,7,8,9,10,11,12,13,14)]
# Presentación del resumen de los datos
summary(data)
```

```
##   Job.Title                                    Job.Type   Salary.Estimate
##  Length:645         Business Intelligence Analyst:  6   75-131  : 32
##  Class :character   Data Analyst                 : 47   79-131  : 32
##  Mode  :character   Data Engineer                : 47   99-132  : 32
##                     Data Scientist               :432   137-171 : 30
##                     Machine Learning Engineer    : 35   90-109  : 29
##                     Other                        : 78   56-97   : 22
##                                                         (Other) :468
##   Salary.Mean        Rating     Company.Name         Location
##  Min.   : 43.5   Min.   :2.00   Length:645         Length:645
##  1st Qu.:103.0   1st Qu.:3.50   Class :character   Class :character
##  Median :114.0   Median :3.90   Mode  :character   Mode  :character
##  Mean   :123.4   Mean   :3.89
##  3rd Qu.:136.5   3rd Qu.:4.30
##  Max.   :271.5   Max.   :5.00
##
##  Headquarters                Size       Founded
##  Length:645         51 to 200   :135   Length:645
##  Class :character   1001 to 5000:104   Class :character
##  Mode  :character   1 to 50     : 86   Mode  :character
##                     201 to 500  : 85
##                     10000+      : 80
##                     501 to 1000 : 77
##                     (Other)     : 78
##                       Type.of.ownership   Industry
##  Company - Private             :397     Length:645
##  Company - Public              :153     Class :character
##  Nonprofit Organization        : 36     Mode  :character
##  Subsidiary or Business Segment: 28
##  Government                    : 10
##  Other Organization            :  5
##  (Other)                       : 16
##                         Sector                          Revenue
##  Information Technology   :188   Unknown / Non-Applicable:213
##  Business Services        :120   100 to 500 million      : 94
##  Biotech & Pharmaceuticals: 66   10+ billion             : 63
##  Aerospace & Defense      : 46   2 to 5 billion          : 45
##  Unknown                  : 44   10 to 25 million        : 41
##  Finance                  : 33   1 to 2 billion          : 36
##  (Other)                  :148   (Other)                 :153
```

```
##  Competitors
##  Length:645
##  Class :character
##  Mode  :character
##
##
##
##
```

```
# Presentación de un extracto y tipo de variables
str(data)
```

```
## tibble [645 x 15] (S3: tbl_df/tbl/data.frame)
##  $ Job.Title        : chr [1:645] "Sr Data Scientist" "Data Scientist" "Data Scientist" "Data Scienti
##  $ Job.Type         : Factor w/ 6 levels "Business Intelligence Analyst",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Salary.Estimate  : Factor w/ 30 levels "101-165 ","105-167 ",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ Salary.Mean      : num [1:645] 154 154 154 154 154 154 154 154 154 154 ...
##  $ Rating           : num [1:645] 3.1 4.2 3.8 3.5 2.9 4.2 3.9 3.5 4.4 3.6 ...
##  $ Company.Name     : chr [1:645] "Healthfirst" "ManTech" "Analysis Group" "INFICON" ...
##  $ Location         : chr [1:645] "New York, NY" "Chantilly, VA" "Boston, MA" "Newton, MA" ...
##  $ Headquarters     : chr [1:645] "New York, NY" "Herndon, VA" "Boston, MA" "Bad Ragaz, Switzerland"
##  $ Size             : Factor w/ 8 levels "1 to 50","10000+",..: 3 5 3 6 7 7 2 3 5 7 ...
##  $ Founded          : chr [1:645] "1993" "1968" "1981" "2000" ...
##  $ Type.of.ownership: Factor w/ 12 levels "College / University",..: 7 3 9 3 2 2 3 3 3 2 ...
##  $ Industry         : chr [1:645] "Insurance Carriers" "Research & Development" "Consulting" "Electr
##  $ Sector           : Factor w/ 23 levels "Accounting & Legal",..: 13 5 5 14 5 12 4 19 12 12 ...
##  $ Revenue          : Factor w/ 13 levels "1 to 2 billion",..: 13 1 5 5 13 13 4 1 6 13 ...
##  $ Competitors      : chr [1:645] "EmblemHealth, UnitedHealth Group, Aetna" "Unknown" "Unknown" "MKS
```

Se guarda el dataset limpio.

```
# Escritura del Fichero
write.csv(data,"Cleaned_DS_jobs.csv")
```

A continuacion realizaremos las consignas 4, 5 y 6 del la PRA2:

4. Análisis de los datos.
5. Representación de los resultados a partir de tablas y gráficas.
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Análisis de los datos

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

Para continuar con el análisis consiguiente se ha determinado el uso de las siguientes variables, considerando que las variables no seleccionadas no tienen inflencia en las respuestas a las preguntas planteadas.

Variables a utilizar:

- Job.Type (factor)
- Salary.Mean (numeric)
- Rating (numeric)
- Size (factor)
- Type.of.Ownership (factor)
- Sector (factor)
- Revenue (factor)

Preguntas:

- Hay diferencia por sector en relacion al salario?, se realizara un analisis bivariado de las variables Salary.Mean y Sector.
- Que tipo de empresa tiene mejor Rating?, se realizara un analisis bivariado de las variables Rating y Type.of.ownership.

Comprobación de normalidad y homogeneidad de la varianza

En este apartado se aplicarán las siguientes pruebas:

- Test de normalidad de Shapiro para las variables Salary.Mean y Rating:

```
# Se aplica el test de normalidad de Shapiro para las variables Salary.Mean y Rating
shapiro.test(data$Salary.Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Salary.Mean
## W = 0.85918, p-value < 2.2e-16
```

```
shapiro.test(data$Rating)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Rating
## W = 0.98107, p-value = 2.092e-07
```

En el test de Shapiro-Wilks se plantea como hipótesis nula que una muestra x1, . . . , xn proviene de una población normalmente distribuida.

Interpretación del test de Shapiro: Siendo la hipótesis nula que la población está distribuida normalmente, si el p-valor es menor a alfa (nivel de significancia) entonces la hipótesis nula es rechazada (se concluye que los datos no vienen de una distribución normal). Si el p-valor es mayor a alfa, se concluye que no se puede rechazar dicha hipótesis.

De tal manera, que debido a que los test de Normalidad de Shapiro-Wilks dieron p-valores muy pequeños para las variables Rating y Salary.Mean, no podemos decir que tienen una distribucion normal.

Los siguientes correspoden a tests no paramétricos que comparan las varianzas basándose en la mediana. Es también una alternativa cuando no se cumple la condición de normalidad en las muestras.

- Test de Levene para la variable Rating agrupado por Size, para ver si la varianza es similar o no en cada grupo:

```r
# Se aplica el test de Levene de homogeneidad de la varianza de Rating agrupado por Size
leveneTest(Rating ~ Size, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   7   8.754 2.7e-10 ***
##       637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El test de Levene presenta un resultado significativo

- Test de Fligner de diferencia de varianza para Salary.Mean por Rating:

```r
# Se aplica el test de homogeneidad de la varianza para las variables Salary.Mean y Rating
fligner.test(Salary.Mean ~ Rating, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Salary.Mean by Rating
## Fligner-Killeen:med chi-squared = 32.388, df = 30, p-value = 0.3497
```

El resultado del test de Fligner para homegeneidad de la varianza de Salary.Mean vs Rating no fue significativa.

Aplicación de pruebas estadísticas para comparar los grupos de datos

---

- Correlacion entre las variables Rating y Salary.Mean.

```r
# Correlación entre variables Salary.Mean y Rating
cor.test(data$Salary.Mean, data$Rating, method = "spearman", exact = FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$Salary.Mean and data$Rating
## S = 43568203, p-value = 0.5129
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## 0.02581196
```

En base a la valor de rho obtenido se podría pensar que el Rating no estaría linealmente correlacionado con el Salario.

- Test de Chi cuadrado para ver independencia de Salary.Mean con las variables Rating , Size , Sector , Job.Type , Revenue , Type.of.ownership

```r
# Test Chi Cuadrado (Job.Type vs Salary Mean)
Job.Type <- table(data$Job.Type, data$Salary.Mean)
chisq.test(Job.Type)
```

```
## Warning in chisq.test(Job.Type): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  Job.Type
## X-squared = 193.35, df = 125, p-value = 8.539e-05
```

```
# Test Chi Cuadrado (Rating vs Salary Mean)
Rating <- table(data$Rating, data$Salary.Mean)
chisq.test(Rating)
```

```
## Warning in chisq.test(Rating): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Rating
## X-squared = 733.82, df = 750, p-value = 0.6567
```

```
# Test Chi Cuadrado (Size vs Salary Mean)
Size <- table(data$Size, data$Salary.Mean)
chisq.test(Size)
```

```
## Warning in chisq.test(Size): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Size
## X-squared = 205.19, df = 175, p-value = 0.05884
```

```
# Test Chi Cuadrado (Sector vs Salary Mean)
Sector<- table(data$Sector, data$Salary.Mean)
chisq.test(Sector)
```

```
## Warning in chisq.test(Sector): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Sector
## X-squared = 564.03, df = 550, p-value = 0.3302
```

```
# Test Chi Cuadrado (Revenue vs Salary Mean)
Revenue <- table(data$Revenue, data$Salary.Mean)
chisq.test(Revenue)
```

```
## Warning in chisq.test(Revenue): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Revenue
## X-squared = 302.77, df = 300, p-value = 0.4444
```

```
# Test Chi Cuadrado (Ownership vs Salary Mean)
own <- table(data$Type.of.ownership, data$Salary.Mean)
chisq.test(own)
```

```
## Warning in chisq.test(own): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  own
## X-squared = 283.28, df = 275, p-value = 0.3527
```

Mirando los resultados la unica variable en donde se rechaza la hipotesis nula es Job.Type frente a Salary.Mean, por lo que se puede concluir que existe una asociación significativa entre el tipo de trabajo y su salario.

- Regresion lineal con variable objetivo (target) Salary.Mean y variables independientes Rating, Size, Sector, Job.Type, Revenue y Type.of.ownership.

```
# Regresión lineal
m1 = lm(Salary.Mean ~ Rating + Size + Sector + Job.Type + Revenue + Type.of.ownership, data = data)
summary(m1)
```

```
##
## Call:
## lm(formula = Salary.Mean ~ Rating + Size + Sector + Job.Type +
##     Revenue + Type.of.ownership, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.366 -21.364  -6.333  16.229 148.446
##
## Coefficients: (1 not defined because of singularities)
##                                                Estimate Std. Error t value
## (Intercept)                                    111.0906    40.2294   2.761
## Rating                                          -0.6211     2.9822  -0.208
## Size10000+                                      -9.9457    11.0203  -0.902
## Size1001 to 5000                                -1.0499     8.4230  -0.125
## Size201 to 500                                   5.5307     7.6577   0.722
## Size5001 to 10000                               -4.3079    10.3538  -0.416
## Size501 to 1000                                  4.8786     8.0624   0.605
## Size51 to 200                                    8.1469     6.8418   1.191
## SizeUnknown                                     17.4845    11.6871   1.496
## SectorAerospace & Defense                       -5.2576    26.5456  -0.198
## SectorAgriculture & Forestry                   -35.7561    46.9845  -0.761
## SectorBiotech & Pharmaceuticals                -15.6013    26.3663  -0.592
## SectorBusiness Services                         -5.7898    26.0637  -0.222
## SectorConstruction, Repair & Maintenance       -58.5836    38.1812  -1.534
## SectorConsumer Services                         70.7747    39.4834   1.793
## SectorEducation                                -11.0647    40.2386  -0.275
## SectorFinance                                  -20.4986    26.8546  -0.763
## SectorGovernment                                -2.8459    28.2789  -0.101
## SectorHealth Care                              -21.4576    27.2094  -0.789
## SectorInformation Technology                   -15.2177    25.8040  -0.590
## SectorInsurance                                -23.3474    26.9022  -0.868
## SectorManufacturing                             -9.8170    27.3013  -0.360
## SectorMedia                                     24.8175    30.4486   0.815
## SectorNon-Profit                               -99.1952    48.1270  -2.061
## SectorOil, Gas, Energy & Utilities             -34.6190    29.1110  -1.189
## SectorReal Estate                              -25.0876    34.3836  -0.730
## SectorRetail                                    11.1932    29.7921   0.376
## SectorTelecommunications                       -24.8541    29.9527  -0.830
## SectorTransportation & Logistics               -13.9306    30.6400  -0.455
## SectorTravel & Tourism                         -10.8959    34.4666  -0.316
## SectorUnknown                                  -12.1946    25.8570  -0.472
```

```
## Job.TypeData Analyst                                14.2505    18.1962   0.783
## Job.TypeData Engineer                                9.5703    18.3393   0.522
## Job.TypeData Scientist                              17.5802    17.5067   1.004
## Job.TypeMachine Learning Engineer                    8.9929    19.0197   0.473
## Job.TypeOther                                       20.1122    18.0830   1.112
## Revenue1 to 5 million                               -2.3464    12.1422  -0.193
## Revenue10 to 25 million                             -6.0746    11.0533  -0.550
## Revenue10+ billion                                   5.8458    10.7158   0.546
## Revenue100 to 500 million                            0.1887     9.0512   0.021
## Revenue2 to 5 billion                                3.4746     9.4356   0.368
## Revenue25 to 50 million                             -6.1601    11.3303  -0.544
## Revenue5 to 10 billion                              19.6274    16.1140   1.218
## Revenue5 to 10 million                              -8.9937    14.4549  -0.622
## Revenue50 to 100 million                           -23.9650    11.3435  -2.113
## Revenue500 million to 1 billion                      9.0409    12.3580   0.732
## RevenueLess than 1 million                          -9.0519    15.8569  -0.571
## RevenueUnknown / Non-Applicable                      2.0928     8.8985   0.235
## Type.of.ownershipCompany - Private                   7.6357    20.7995   0.367
## Type.of.ownershipCompany - Public                   16.6306    21.1207   0.787
## Type.of.ownershipContract                          -11.1714    34.9631  -0.320
## Type.of.ownershipGovernment                         22.2254    25.8974   0.858
## Type.of.ownershipHospital                           86.1862    45.9659   1.875
## Type.of.ownershipNonprofit Organization             13.7767    22.4167   0.615
## Type.of.ownershipOther Organization                 37.6828    28.3110   1.331
## Type.of.ownershipPrivate Practice / Firm             6.5004    29.4020   0.221
## Type.of.ownershipSelf-employed                      21.8306    53.3395   0.409
## Type.of.ownershipSubsidiary or Business Segment -12.6593    22.2002  -0.570
## Type.of.ownershipUnknown                                NA         NA      NA
##                                                  Pr(>|t|)
## (Intercept)                                       0.00593 **
## Rating                                            0.83509
## Size10000+                                        0.36717
## Size1001 to 5000                                  0.90085
## Size201 to 500                                    0.47044
## Size5001 to 10000                                 0.67751
## Size501 to 1000                                   0.54535
## Size51 to 200                                     0.23423
## SizeUnknown                                       0.13518
## SectorAerospace & Defense                         0.84307
## SectorAgriculture & Forestry                      0.44695
## SectorBiotech & Pharmaceuticals                   0.55427
## SectorBusiness Services                           0.82428
## SectorConstruction, Repair & Maintenance          0.12548
## SectorConsumer Services                           0.07356 .
## SectorEducation                                   0.78343
## SectorFinance                                     0.44558
## SectorGovernment                                  0.91987
## SectorHealth Care                                 0.43066
## SectorInformation Technology                      0.55559
## SectorInsurance                                   0.38583
## SectorManufacturing                               0.71929
## SectorMedia                                       0.41537
## SectorNon-Profit                                  0.03973 *
## SectorOil, Gas, Energy & Utilities                0.23484
```

```
## SectorReal Estate                                0.46590
## SectorRetail                                     0.70727
## SectorTelecommunications                         0.40700
## SectorTransportation & Logistics                 0.64953
## SectorTravel & Tourism                           0.75202
## SectorUnknown                                    0.63738
## Job.TypeData Analyst                             0.43385
## Job.TypeData Engineer                            0.60198
## Job.TypeData Scientist                           0.31570
## Job.TypeMachine Learning Engineer                0.63652
## Job.TypeOther                                    0.26650
## Revenue1 to 5 million                            0.84683
## Revenue10 to 25 million                          0.58282
## Revenue10+ billion                               0.58560
## Revenue100 to 500 million                        0.98337
## Revenue2 to 5 billion                            0.71282
## Revenue25 to 50 million                          0.58686
## Revenue5 to 10 billion                           0.22370
## Revenue5 to 10 million                           0.53406
## Revenue50 to 100 million                         0.03505 *
## Revenue500 million to 1 billion                  0.46472
## RevenueLess than 1 million                       0.56832
## RevenueUnknown / Non-Applicable                  0.81415
## Type.of.ownershipCompany - Private               0.71367
## Type.of.ownershipCompany - Public                0.43136
## Type.of.ownershipContract                        0.74945
## Type.of.ownershipGovernment                      0.39113
## Type.of.ownershipHospital                        0.06129 .
## Type.of.ownershipNonprofit Organization          0.53907
## Type.of.ownershipOther Organization              0.18370
## Type.of.ownershipPrivate Practice / Firm         0.82510
## Type.of.ownershipSelf-employed                   0.68249
## Type.of.ownershipSubsidiary or Business Segment  0.56874
## Type.of.ownershipUnknown                             NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.94 on 587 degrees of freedom
## Multiple R-squared:  0.1295, Adjusted R-squared:  0.04493
## F-statistic: 1.532 on 57 and 587 DF,  p-value: 0.009271
```

Solo se entreno una regresion lineal con los datos. Se hizo con un fin de relacion lineal entre variables y no como entrenamiento de un modelo de Machine learning que sirva para predecir el Salary.Mean, pues este otro objetivo requeriria otra metodologia.

El R cuadrado obtenido es bajo. Lo cual era esperable porque pareciera no haber relacion nivel aparente entre la variable objetivo y las regresoras.

- Análisis Comparativo de Valores Promedio de Salary y Rating por atributo

En este apartado se pretende evaluar a través de la comparativa de los atributos Salary.Mean y Rating y el resto de atributos, cuales son los campos que tienen los mejores salarios y ratings, con el objetivo de obtener una idea mas global de las áreas que lideran el mercado laboral en Data Science.

```
# Creación de la función para cálculo de valores promedio
promedios <- function(x) mean(x, na.rm = TRUE)
```

var	Col_Max	Salary_Mean_Max
Size	Unknown	143.3824
Ownership	Hospital	183.0000
Sector	Consumer Services	203.7500
Revenue	5 to 10 billion	137.1875

var	Col_Max	Rating_Mean_Max
Size	1 to 50	4.329070
Ownership	Unknown	4.325000
Sector	Media	4.280000
Revenue	Less than 1 million	4.321429

```
# Generación de dataframe con valores máximos promedios de Salary.Mean
out_Salary <- data.frame(var=c("Size","Ownership","Sector","Revenue"),
                  Col_Max = c(names(which.max(tapply(data$Salary.Mean, data$Size, promedios))),
                              names(which.max(tapply(data$Salary.Mean, data$Type.of.ownership, promedios
                              names(which.max(tapply(data$Salary.Mean, data$Sector, promedios))),
                              names(which.max(tapply(data$Salary.Mean, data$Revenue, promedios)))),
                  Salary_Mean_Max = c(max(tapply(data$Salary.Mean, data$Size, promedios)),
                                      max(tapply(data$Salary.Mean, data$Type.of.ownership, promedios)),
                                      max(tapply(data$Salary.Mean, data$Sector, promedios)),
                                      max(tapply(data$Salary.Mean, data$Revenue, promedios))))
# Creación de la tabla
out_Salary %>% kable() %>% kable_styling()
```

Así, en base al promedio de Salary se puede obtener que el Sector de Servicio al Cliente es el área que mejor paga dentro del Data Science y así mismo, aunque en menor cantidad se puede ver que la empresas Hospitalarias también ofrecen un nivel de salario alto.

Otro punto a resaltar en la tabla previa, es que el área del Data Science presenta un nivel de ingreso alto (aproximadamente 12000 (USD/mes)).

De igual manera, procedemos con el mismo análisis ahora respecto del rating de los anuncios:

```
# Generación de dataframe con valores máximos promedios de Rating
out_Rating <- data.frame(var=c("Size","Ownership","Sector","Revenue"),
                  Col_Max = c(names(which.max(tapply(data$Rating, data$Size, promedios))),
                              names(which.max(tapply(data$Rating, data$Type.of.ownership, promedios))),
                              names(which.max(tapply(data$Rating, data$Sector, promedios))),
                              names(which.max(tapply(data$Rating, data$Revenue, promedios)))),
                  Rating_Mean_Max = c(max(tapply(data$Rating, data$Size, promedios)),
                                      max(tapply(data$Rating, data$Type.of.ownership, promedios)),
                                      max(tapply(data$Rating, data$Sector, promedios)),
                                      max(tapply(data$Rating, data$Revenue, promedios))))
# Creación de la tabla
out_Rating %>% kable() %>% kable_styling()
```

En esta comparativa se puede observar que los valores de rating en cada uno de los atributos presentan valores altos (mayor a 4), en donde, es llamativo observar que es las organizaciones que tienen bajo número de empleados son la que mejor rating tiene, pudiendo llevar a concluir que las empresas más pequeñás son las que se preocupan mayormente en el ambiente laboral. Así mismo

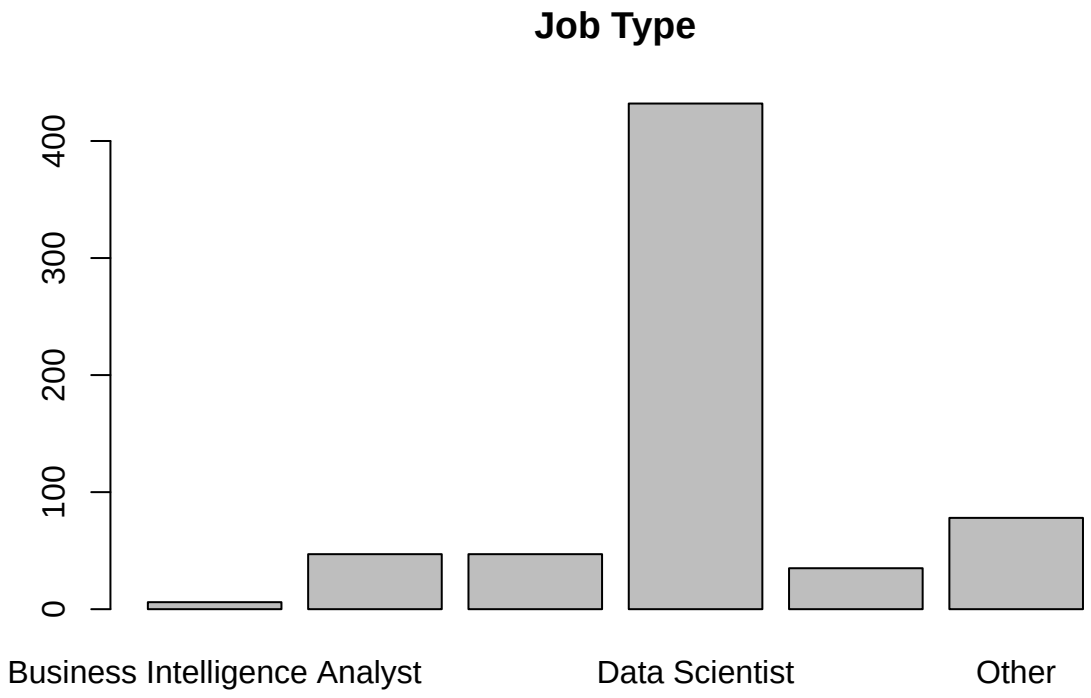
Representación de los resultados a partir de tablas y gráficas.

- Graficos univariados

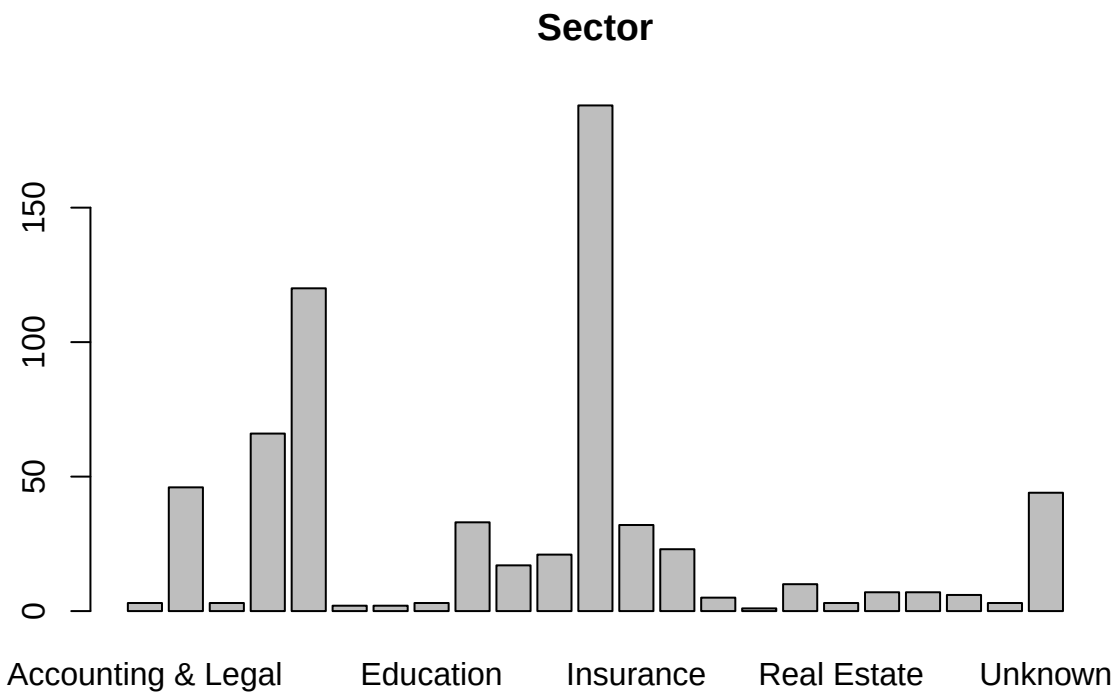
Comenzamos observando graficos univariados de las variables a analizar, en donde, se observará que cantidad de veces toman cada valor cada una de las variables categoricas.

- GRafico de Barras de variables categoricas:
  - "Job.Type", "Location", "Size", "Type.of.ownership", "Sector".

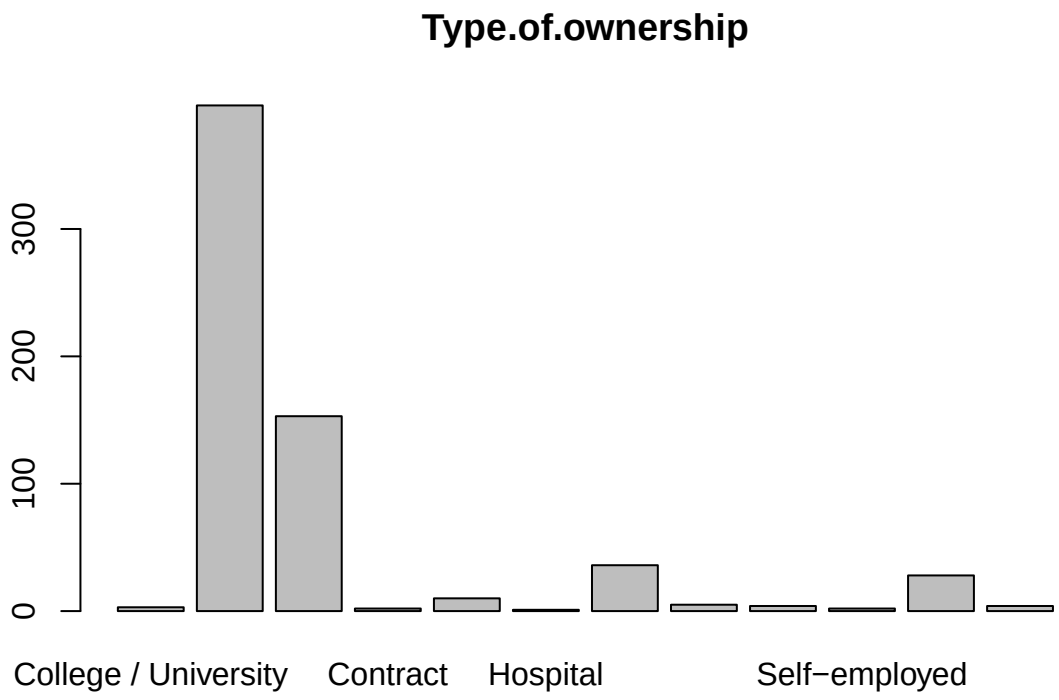
```
# Gráfico univariado de Job.Type
plot(data$Job.Type, main='Job Type')
```



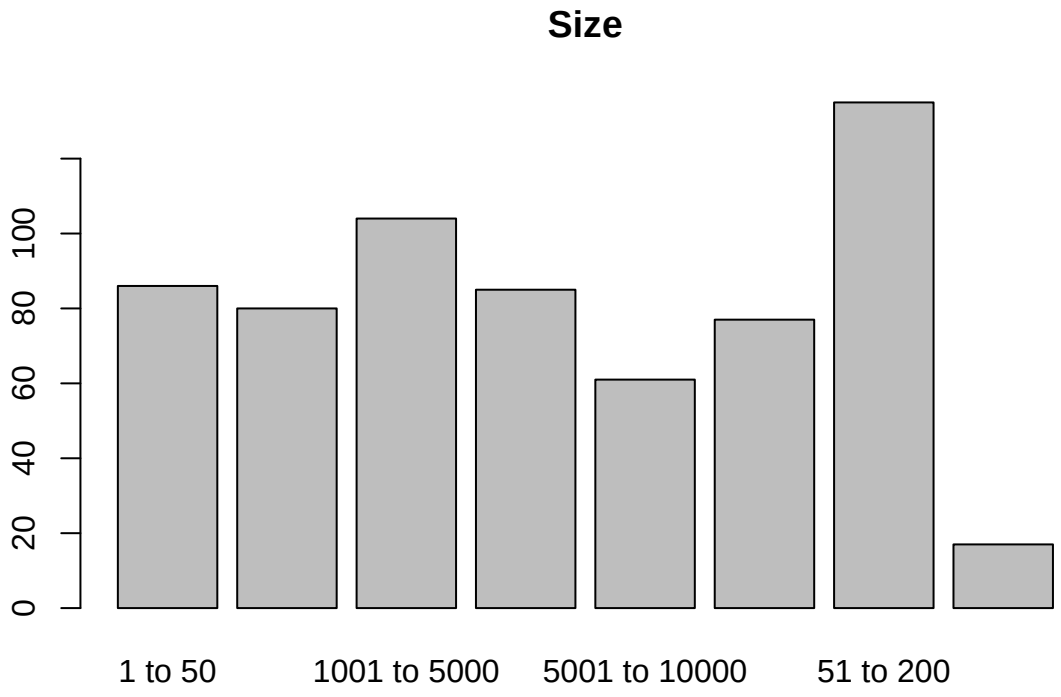

```
# Gráfico univariado de Sector
plot(data$Sector, main='Sector')
```



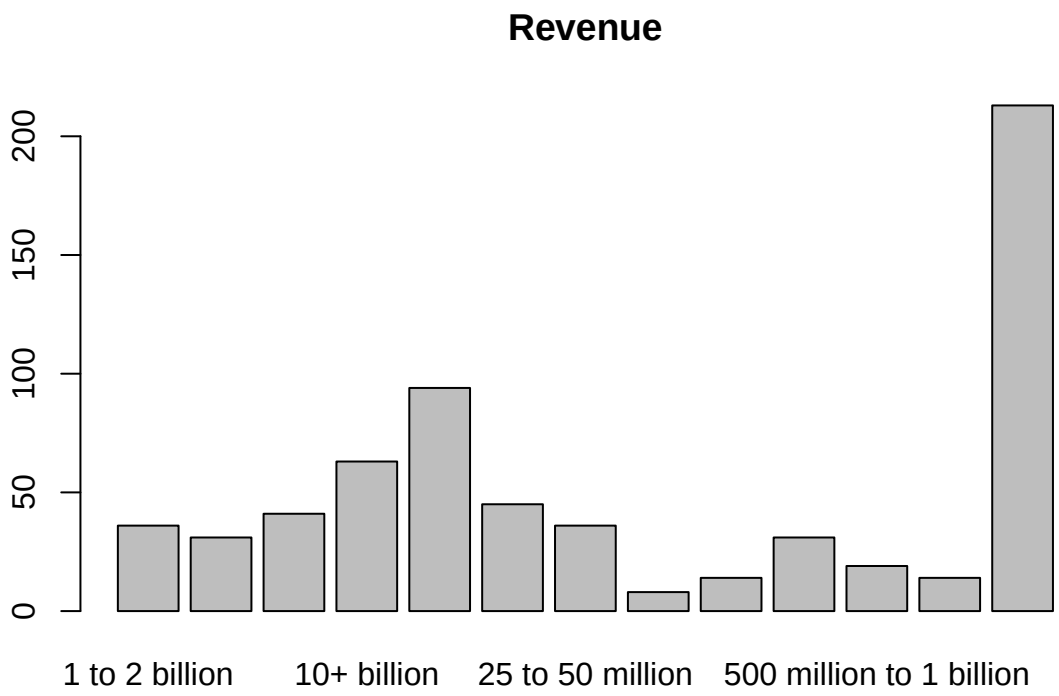

```
# Gráfico univariado de OwnerShip
plot(data$Type.of.ownership, main='Type.of.ownership')
```




```
# Gráfico univariado de Size
plot(data$Size, main='Size')
```



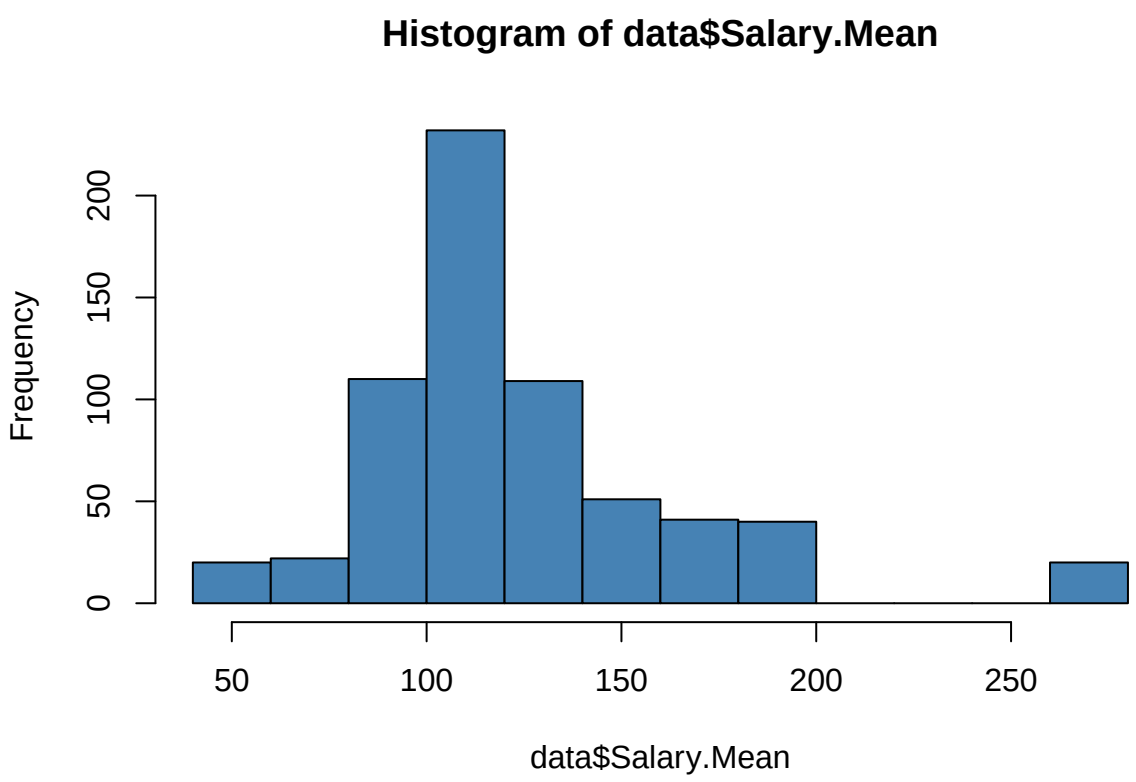

```
# Gráfico univariado de Revenue
plot(data$Revenue, main='Revenue')
```



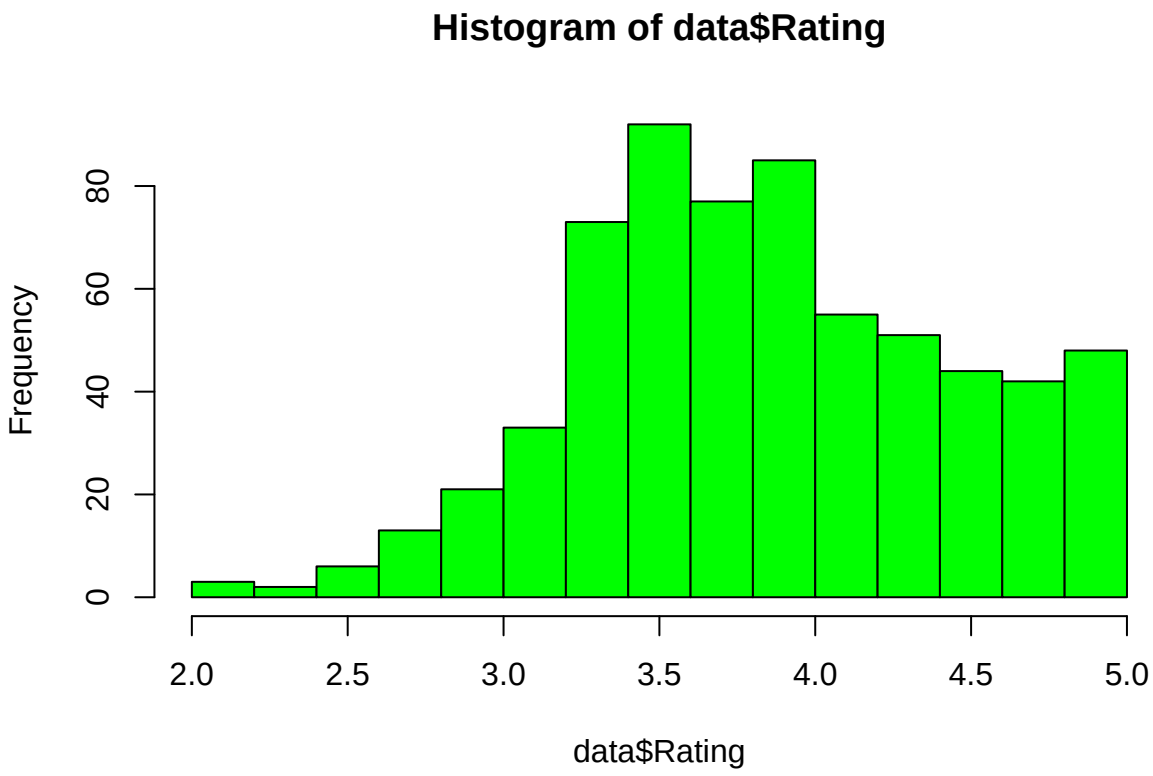
- Grafico densidad de las variables Salary y Rating

Se presenta a continuación un histograma para identificar las frecuencias de los datos respecto de Salario y Rating de las ofertas.

```
# Histograma de Salario
hist(data$Salary.Mean, col = "steelblue", breaks=10)
```




```r
# Histograma de Rating
hist(data$Rating, col = "green")
```



- Graficos de densidad y test de Levene para las variables Salary.Mean y Sector

El test de normalidad de la variable Salary.Mean dio negativo mas arriba.

```
leveneTest(Salary.Mean ~ Sector, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group  22  0.9833 0.4837
##       622
```

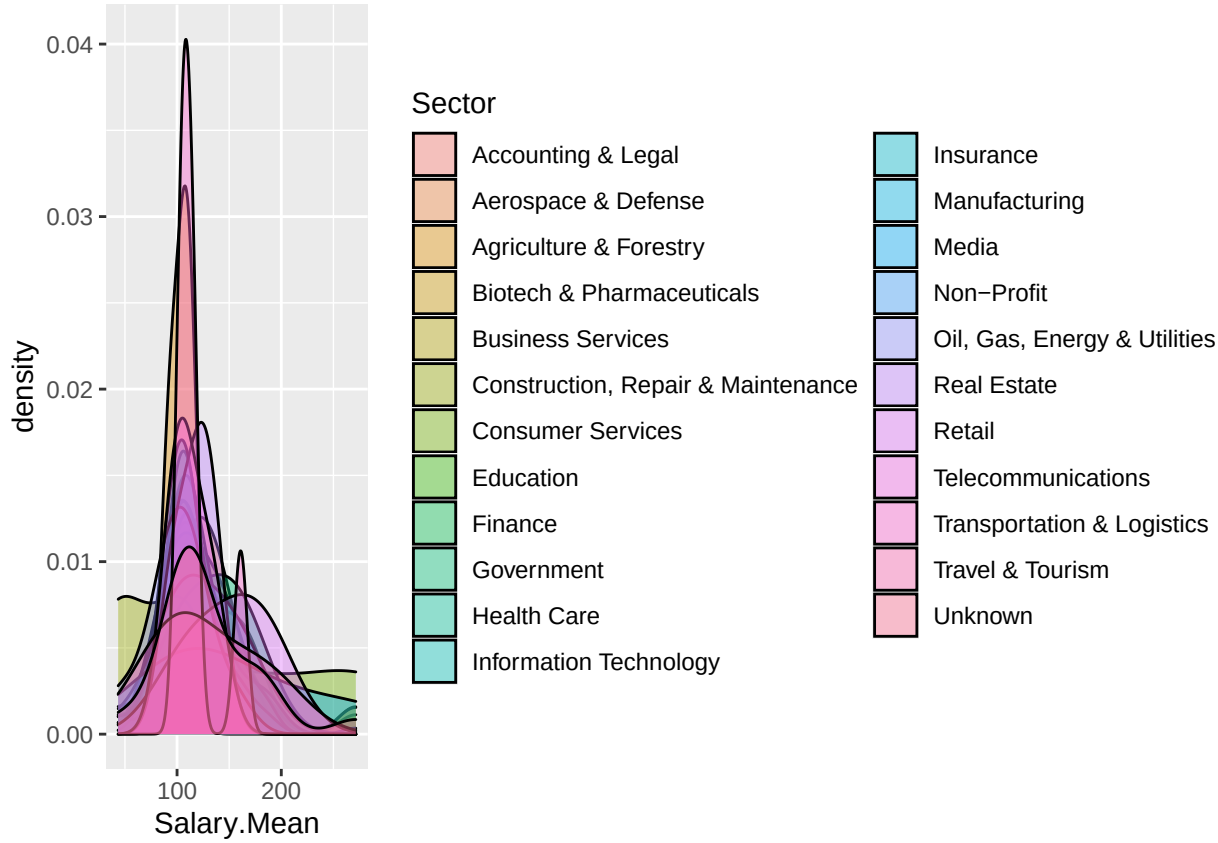
```
print('Distribucion de Salary.Mean eparado por sector')
```

```
## [1] "Distribucion de Salary.Mean eparado por sector"
```

```
p2 <- ggplot(data=data, aes(x=Salary.Mean, group=Sector, fill=Sector)) +
    geom_density(adjust=1.5, alpha=.4)
p2
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```



Vemos que la distribucion de Salary.Mean cambia su distribucon segun cada sector. Como el p-valor del test de levene es aprox. 0,4 las varianzas del Salario parecieran no ser diferentes segun el sector en el que nos encontremos (se acepto la hipotesis nula).

Obs: El test de Levene se puede aplicar con la función leveneTest() del paquete car. Se caracteriza, además de por poder comparar 2 o más poblaciones, por permitir elegir entre diferentes estadísticos de centralidad :mediana (por defecto), media, media truncada. Esto es importante a la hora de contrastar la homocedasticidad dependiendo de si los grupos se distribuyen de forma normal o no.

- Graficos de densidad y test de Levene para las variables Rating y Type.of.ownership

```
leveneTest(Rating ~ Type.of.ownership, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group  11  7.7557 1.132e-12 ***
##       633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

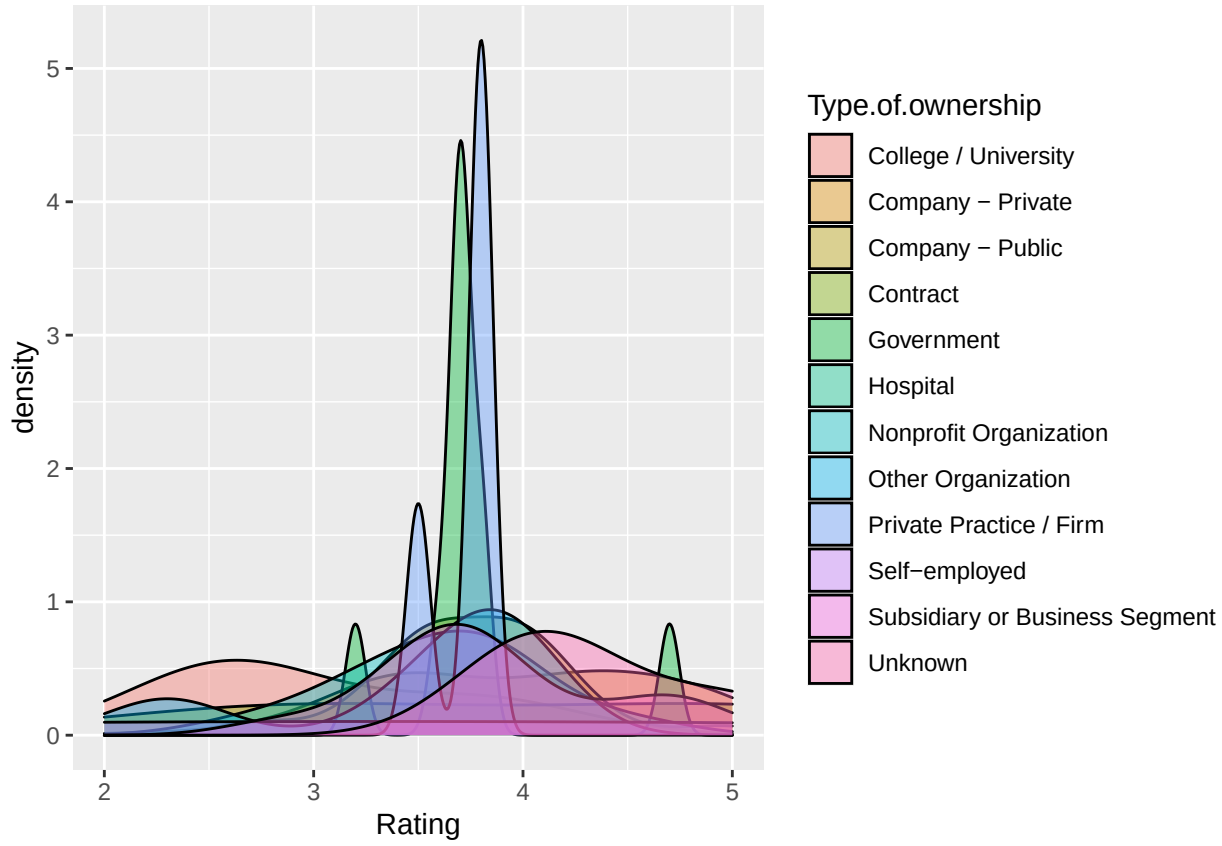
```
print('Distribucion de Rating separado por Type.of.ownership')
```

```
## [1] "Distribucion de Rating separado por Type.of.ownership"
```

```
p3 <- ggplot(data=data, aes(x=Rating, group=Type.of.ownership, fill=Type.of.ownership)) +
    geom_density(adjust=1.5, alpha=.4)
p3
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```



Pareciera que la distribucion de Rating se diferencia en Goverment y Private practice. Y el test de Levene esta indicando que hay diferencia significativas de las varianzas del Rating en los distintos Type of ownership de las compañias empleadoras ya que el p-valor del test de Levene dio muy pequeño.

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

En el anális realizado previamente se ha considerado principalmente los atributos de Salary y Rating, tomando estos atributos como un medidor de calidad de las ofertas laborales, en donde se han obtenido las siguientes conclusiones:

- Mediante el analísis de los valores extremos tomados por las variables Salary y Rating, se obtiene que ambos atributos presentan valores distribuidos en todas los atributos analizados, de tal manera, que no se puede establecer un Sector o tipo de empresa que presente un mejor/peor salario o una mejor/peor calificación.
- En base a las pruebas estadísticas realizadas se obtuvo que las variables Salary y Rating no siguen una distribución normal y adicionalmente no presentan una correlación lineal.

- Al evaluar la relación entre el salario y el resto de atributos se puede concluir en base a la prueba estadística chi cuadrado que el salario tiene una asociación estadistica únicamente con el tipo de trabajo que en el caso de nuestra muestra tiene una frecuencia mas alta respecto a trabajos generales de Ciencia de Datos.
- Respecto del salario se pudo observar que el sector que mejor retribuye es el de Servicio al Cliente, aunque de manera general se puede concluir que los salarios en el área de Ciencia de Datos presentan un valor alto.
- Así mismo respecto del tipo de empresa y el Rating se pudo observar que en base a los datos recolectados el mejor Rating se presentaba entre los datos de los cuales no se dispone información del tipo de empresa, siendo las empresas Hospitalarias el siguente grupo mejor calificado.
- De manera gráfica se pudo concluir que la densidad del salario varía según cada sector, aunque de igual forma se observa que las varianzas no presentan una mayor diferencia entre los sectores.
- Con respecto a la densidad entre el tipo de empresa y su rating se observó una diferencia marcada entre los tipos de empresa relacionados al Gobierno y la empresa privada.

Código

El presente proyecto de limpieza y análisis de datos se encuentra disponible en el siguiente portal de github:

https://github.com/thanry89/PRA_2_Limpieza_de_Datos_Piccolomini_Zambrano

Contribuciones	Firma
Investigación Previa	Tatiana Piccolomini, Jonathan Zambrano
Redacción de las respuestas	Tatiana Piccolomini, Jonathan Zambrano
Desarrollo código	Tatiana Piccolomini, Jonathan Zambrano