Evidencia 1 - Módulo Estadística

Emilia Jácome

2022-08-22

Resumen

A lo largo de este documento, se pretende realizar un análisis estadístico profundo para determinar cuáles son los factores que más inciden en el sueldo de un profesionista de datos en base a un dataset con varios registros de profesionistas de distintas partes del mundo, sus características y el sueldo que ganan. Por lo que, se abordó esta problemática por medio de métodos y herramientas estadísticas avanzadas para identificar a aquellos factores que son significativos para obtener un salario mayor. Específicamente, y dado que los factores de mayor interés son variables de tipo cualitativas, se escogió la técnica de ANOVA, la cuál permite examinar el efecto de factores cualitativos sobre una variable continua en función del análisis de la varianza a través de la comparación de las medias.

Los resultados obtenidos a partir del análisis fueron los siguientes: * El nivel de experiencia de los profesionistas de datos si es influyente sobre el salario promedio que puede llegar a tener un profesionista de datos. * El tamaño de la compañía si influye en el salario promedio que puede aspirar a tener un profesionista de datos. * El tipo de contrato no es un factor influyente sobre el nivel de salario de un profesionista de datos. * Los tipos de contrato que ofrecen más salarios son los de Full-Time y Contract, aunque estadísticamente se ha comprobado que su efecto es mínimo.

Introducción

Identificar las condiciones o factores determinantes que hacen que una persona especialista en analizar datos tenga un mejor sueldo de acuerdo con la base de datos que proporciona Kaggle, en una muestra de personas que se dedican al analisis de datos en diferentes partes del mundo. Las preguntas de investigación son:

¿Influye el nivel de experiencia en el salario? ¿Influye el tamaño de la compañía en el salario que puede ofrecer a un analista de datos? ¿Qué tipo de contrato (parcial, tiempo completo, etc) ofrece mejores salarios? ¿Qué tipo de contrato será el más conveniente?

A. Importar el dataset

```
M=read.csv("ds_salaries.csv") #Leer la base de datos
df <- data.frame(M)
knitr::opts_chunk$set(echo = FALSE)
```

Preview del Data Frame

B. Exploración de las Variables y el significado

Identifica la cantidad de datos y variables presentes.

```
## [1] 607  12
```

```
## [1] "607 filas y 12 columnas"
```

Nombre de las columnas

```
##  [1] "X"                 "work_year"          "experience_level"
##  [4] "employment_type"   "job_title"          "salary"
##  [7] "salary_currency"   "salary_in_usd"      "employee_residence"
## [10] "remote_ratio"      "company_location"   "company_size"
```

Ver 5 primeros registros

```
##   X work_year experience_level employment_type                 job_title
## 1 0      2020               MI              FT             Data Scientist
## 2 1      2020               SE              FT Machine Learning Scientist
## 3 2      2020               SE              FT          Big Data Engineer
## 4 3      2020               MI              FT       Product Data Analyst
## 5 4      2020               SE              FT  Machine Learning Engineer
## 6 5      2020               EN              FT               Data Analyst
##   salary salary_currency salary_in_usd employee_residence remote_ratio
## 1  70000             EUR         79833                 DE            0
## 2 260000             USD        260000                 JP            0
## 3  85000             GBP        109024                 GB           50
## 4  20000             USD         20000                 HN            0
## 5 150000             USD        150000                 US           50
## 6  72000             USD         72000                 US          100
##   company_location company_size
## 1               DE            L
## 2               JP            S
## 3               GB            M
## 4               HN            S
## 5               US            L
## 6               US            L
```

Ver la estructura del Head Count

```
## 'data.frame':    607 obs. of  12 variables:
##  $ X                 : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ work_year         : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ experience_level  : chr  "MI" "SE" "SE" "MI" ...
##  $ employment_type   : chr  "FT" "FT" "FT" "FT" ...
##  $ job_title         : chr  "Data Scientist" "Machine Learning Scientist" "Big Data Engineer" "Product Data Analyst" ...
##  $ salary            : int  70000 260000 85000 20000 150000 72000 190000 11000000 135000 125000 ...
##  $ salary_currency   : chr  "EUR" "USD" "GBP" "USD" ...
##  $ salary_in_usd     : int  79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...
##  $ employee_residence: chr  "DE" "JP" "GB" "HN" ...
##  $ remote_ratio      : int  0 0 50 0 50 100 100 50 100 50 ...
##  $ company_location  : chr  "DE" "JP" "GB" "HN" ...
##  $ company_size      : chr  "L" "S" "M" "S" ...
```

Clasifica las variables de acuerdo a su tipo y escala de medición.

```
## [1] "Variables categóricas (cualitativas) con datos Nominales: "
```

```
## [1] "employment_type"    "job_title"          "salary_currency"
## [4] "employee_residence" "company_location"
```

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

```
## [1] "Variables categóricas (cualitativas) con datos Ordinales: "
```

```
## [1] "work_year"       "experience_level" "company_size"     "remote_ratio"
```

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

```
## [1] "Variables numéricas (cuantitativas) con datos razonales: "
```

```
## [1] "salary"       "salary_in_usd"
```

#C) Exploración de la base de datos

Calcula medidas estadísticas

Variables cuantitativas

Medidas de tendencia central: promedio, media, mediana y moda de los datos.

```
##      salary         salary_in_usd
##  Min.   :    4000   Min.   :  2859
##  1st Qu.:   70000   1st Qu.: 62726
##  Median :  115000   Median :101570
##  Mean   :  324000   Mean   :112298
##  3rd Qu.:  165000   3rd Qu.:150000
##  Max.   :30400000   Max.   :600000
```

```
## Warning: package 'modeest' was built under R version 4.1.3
```

```
## [1] "Moda:"
```

```
## [1]  80000 100000
```

```
## [1] 100000
```

Medidas de dispersión: rango: máximo - mínimo, varianza, desviación estándar.

```
## [1] "Rango para salary : 30396000"
## [1] ""
## [1] "Varianza para salary : 2385040046528.2"
## [1] ""
## [1] "Desviación estándar para salary : 1544357.48663585"
## [1] ""
## [1] "Rango para salary_in_usd : 597141"
## [1] ""
## [1] "Varianza para salary_in_usd : 5034932663.1761"
## [1] ""
## [1] "Desviación estándar para salary_in_usd : 70957.2594113957"
## [1] ""
```

Variables cualitativas

Tabla de distribución de frecuencia

```
## [1] "Tabla de Frecuencias para work_year"
##
## 2020 2021 2022
##   72  217  318
## [1] ""
## [1] "Tabla de Frecuencias para experience_level"
##
##  EN  EX  MI  SE
##  88  26 213 280
## [1] ""
## [1] "Tabla de Frecuencias para employment_type"
##
##  CT  FL  FT  PT
##   5   4 588  10
## [1] ""
## [1] "Tabla de Frecuencias para job_title"
##
##            3D Computer Vision Researcher
##                                        1
##                              AI Scientist
##                                        7
##                      Analytics Engineer
##                                        4
##                   Applied Data Scientist
##                                        5
##       Applied Machine Learning Scientist
##                                        4
##                          BI Data Analyst
##                                        6
##                       Big Data Architect
##                                        1
##                        Big Data Engineer
##                                        8
##                     Business Data Analyst
##                                        5
##                       Cloud Data Engineer
##                                        2
##                  Computer Vision Engineer
##                                        6
##        Computer Vision Software Engineer
##                                        3
##                             Data Analyst
##                                       97
##                   Data Analytics Engineer
##                                        4
##                       Data Analytics Lead
##                                        1
##                    Data Analytics Manager
##                                        7
##                           Data Architect
##                                       11
##                            Data Engineer
##                                      132
##                  Data Engineering Manager
##                                        5
##                   Data Science Consultant
##                                        7
##                     Data Science Engineer
##                                        3
##                     Data Science Manager
##                                       12
##                           Data Scientist
##                                      143
##                          Data Specialist
##                                        1
##              Director of Data Engineering
##                                        2
##                  Director of Data Science
##                                        7
##                            ETL Developer
##                                        2
##                     Finance Data Analyst
##                                        1
##                    Financial Data Analyst
##                                        2
##                             Head of Data
##                                        5
##                     Head of Data Science
##                                        4
##                  Head of Machine Learning
##                                        1
##                       Lead Data Analyst
```

```
##                                       3
##                         Lead Data Engineer
##                                       6
##                        Lead Data Scientist
##                                       3
##              Lead Machine Learning Engineer
##                                       1
##                  Machine Learning Developer
##                                       3
##                   Machine Learning Engineer
##                                      41
## Machine Learning Infrastructure Engineer
##                                       3
##                    Machine Learning Manager
##                                       1
##                  Machine Learning Scientist
##                                       8
##                     Marketing Data Analyst
##                                       1
##                                ML Engineer
##                                       6
##                               NLP Engineer
##                                       1
##                     Principal Data Analyst
##                                       2
##                    Principal Data Engineer
##                                       3
##                   Principal Data Scientist
##                                       7
##                      Product Data Analyst
##                                       2
##                        Research Scientist
##                                      16
##                      Staff Data Scientist
##                                       1
## [1] ""
## [1] "Tabla de Frecuencias para salary_currency"
##
## AUD BRL CAD CHF CLP CNY DKK EUR GBP HUF INR JPY MXN PLN SGD TRY USD
##   2   2  18   1   1   2   2  95  44   2  27   3   2   3   2   3 398
## [1] ""
## [1] "Tabla de Frecuencias para employee_residence"
##
##  AE  AR  AT  AU  BE  BG  BO  BR  CA  CH  CL  CN  CO  CZ  DE  DK  DZ  EE  ES  FR
##   3   1   3   3   2   1   1   6  29   1   1   1   1   1  25   2   1   1  15  18
##  GB  GR  HK  HN  HR  HU  IE  IN  IQ  IR  IT  JE  JP  KE  LU  MD  MT  MX  MY  NG
##  44  13   1   1   1   2   1  30   1   1   4   1   7   1   1   1   1   2   1   2
##  NL  NZ  PH  PK  PL  PR  PT  RO  RS  RU  SG  SI  TN  TR  UA  US  VN
##   5   1   1   6   4   1   6   2   1   4   2   2   1   3   1 332   3
## [1] ""
## [1] "Tabla de Frecuencias para remote_ratio"
##
##   0  50 100
## 127  99 381
## [1] ""
## [1] "Tabla de Frecuencias para company_location"
##
##  AE  AS  AT  AU  BE  BR  CA  CH  CL  CN  CO  CZ  DE  DK  DZ  EE  ES  FR  GB  GR
##   3   1   4   3   2   3  30   2   1   2   1   2  28   3   1   1  14  15  47  11
##  HN  HR  HU  IE  IL  IN  IQ  IR  IT  JP  KE  LU  MD  MT  MX  MY  NG  NL  NZ  PK
##   1   1   1   1   1  24   1   1   2   6   1   3   1   1   3   1   2   4   1   3
##  PL  PT  RO  RU  SG  SI  TR  UA  US  VN
##   4   4   1   2   1   2   3   1 355   1
## [1] ""
## [1] "Tabla de Frecuencias para company_size"
##
##   L   M   S
## 198 326  83
## [1] ""
```

Moda

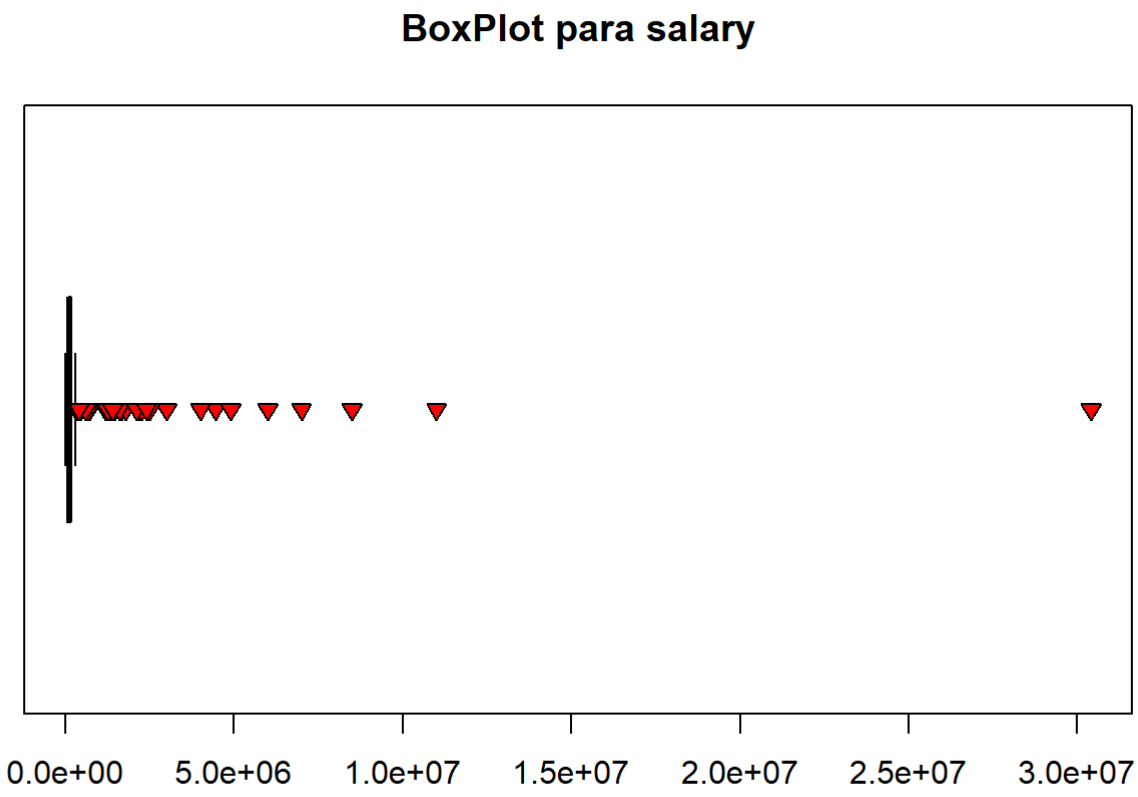
```
## [1] "Moda de work_year : 2022"
## [1] "Moda de experience_level : SE"
## [1] "Moda de employment_type : FT"
## [1] "Moda de job_title : Data Scientist"
## [1] "Moda de salary_currency : USD"
## [1] "Moda de employee_residence : US"
## [1] "Moda de remote_ratio : 100"
## [1] "Moda de company_location : US"
## [1] "Moda de company_size : M"
```

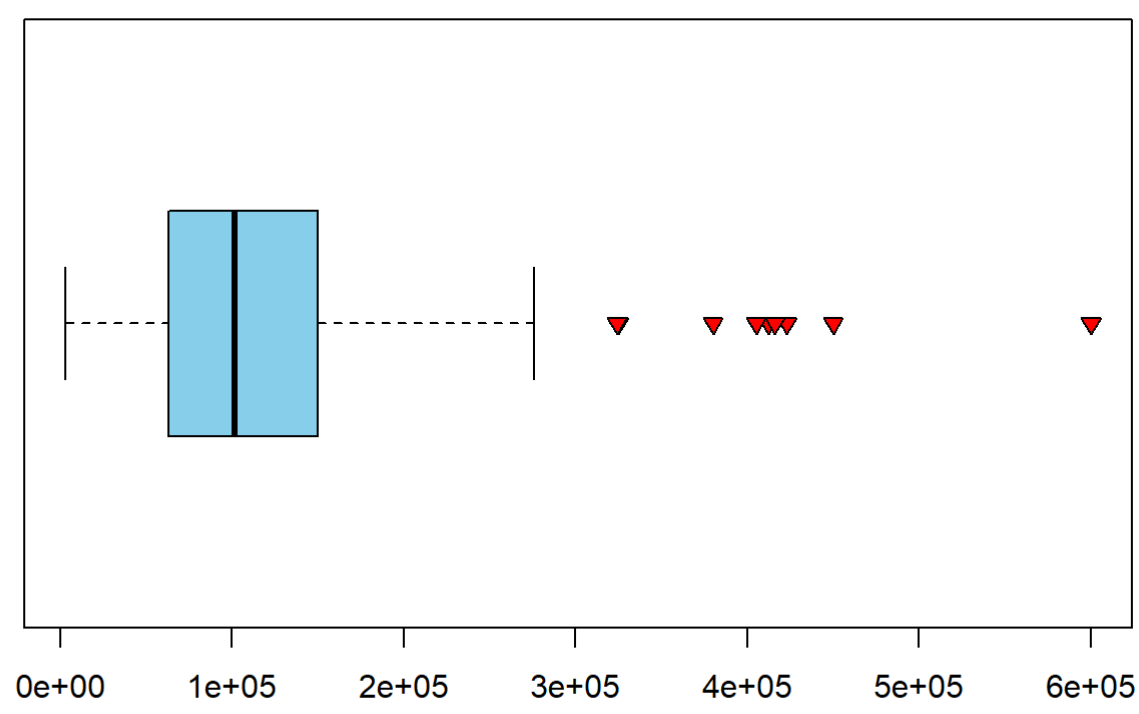
Explora los datos usando herramientas de visualización

Variables cuantitativas:

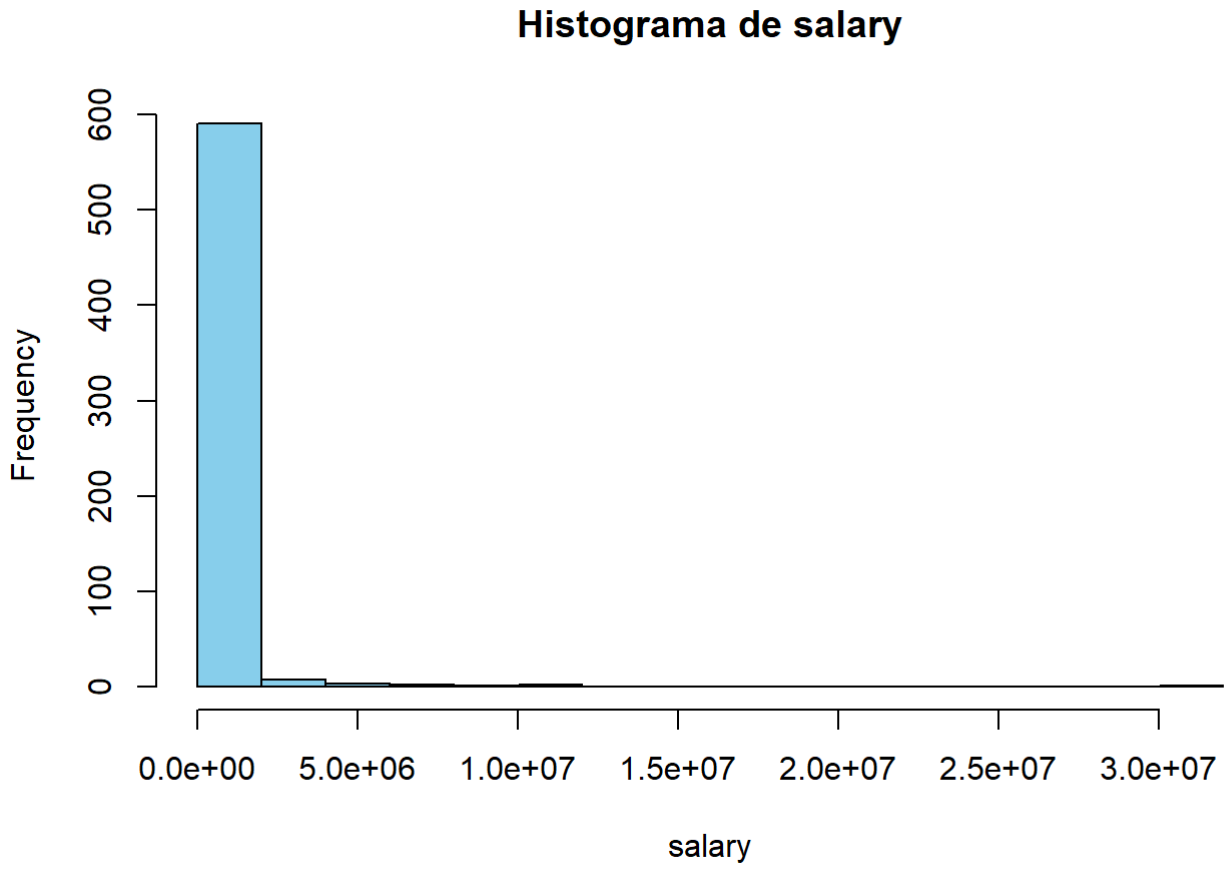
Medidas de posición: cuartiles, outlier (valores atípicos), boxplots

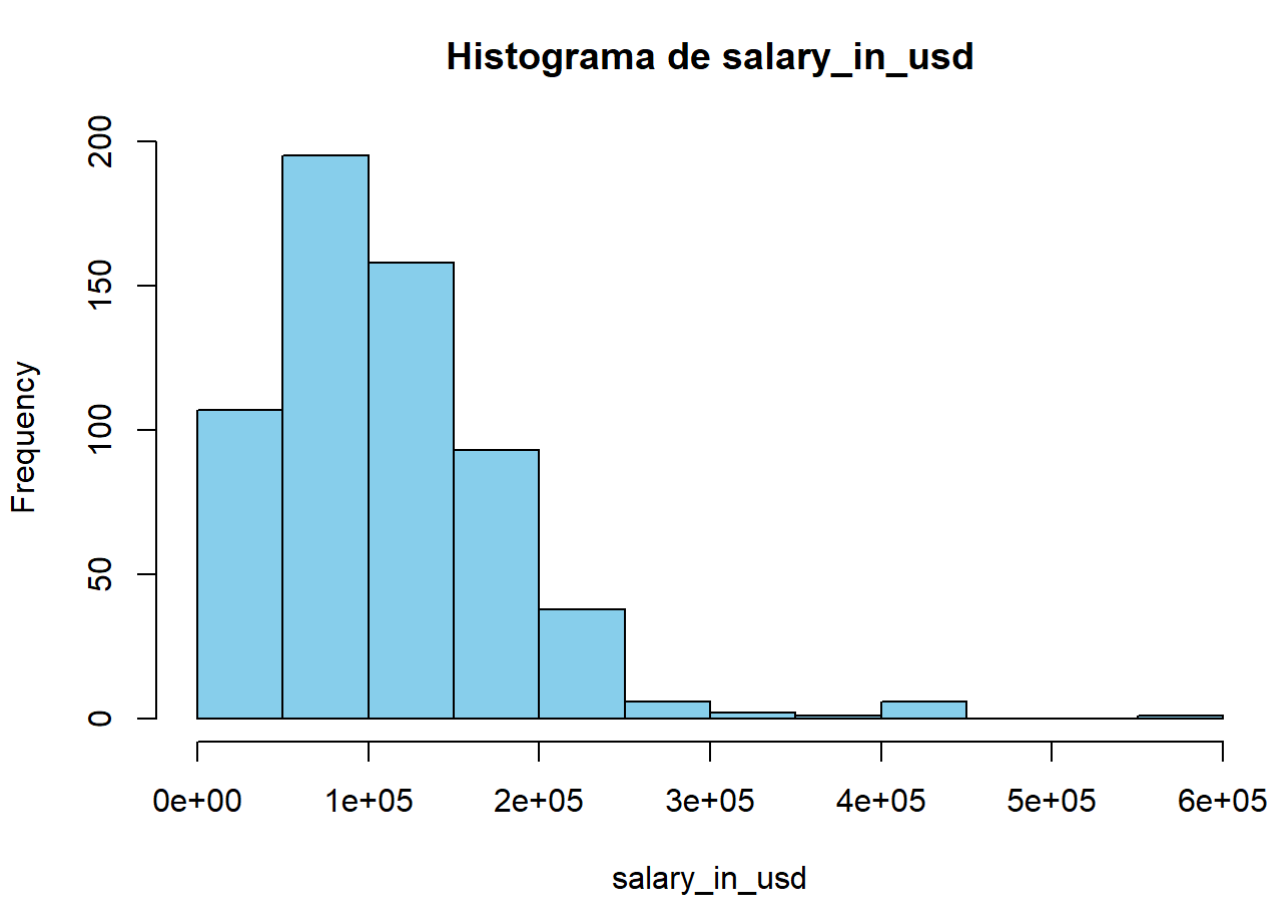
```
## [1] "Cuartiles de salary"
##       0%      25%      50%      75%     100% 
##     4000    70000   115000   165000 30400000 
## [1] ""
## [1] "Cuartiles de salary_in_usd"
##     0%    25%    50%    75%   100% 
##   2859  62726 101570 150000 600000 
## [1] ""
```





Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asimétrica

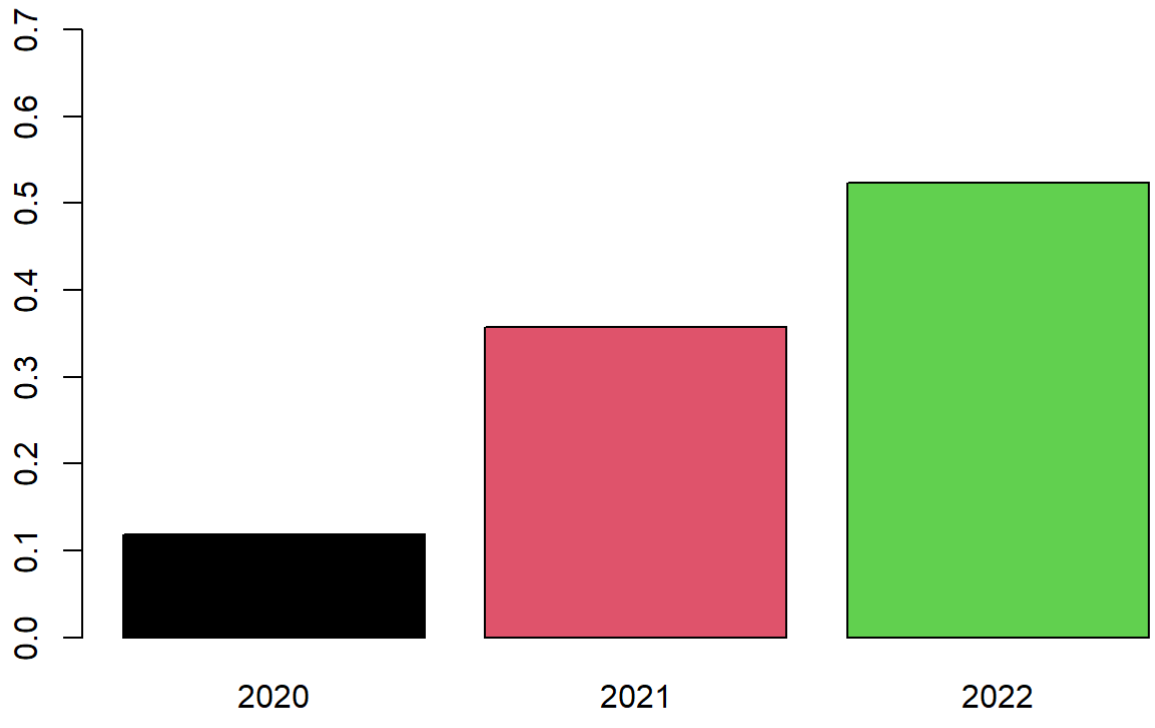


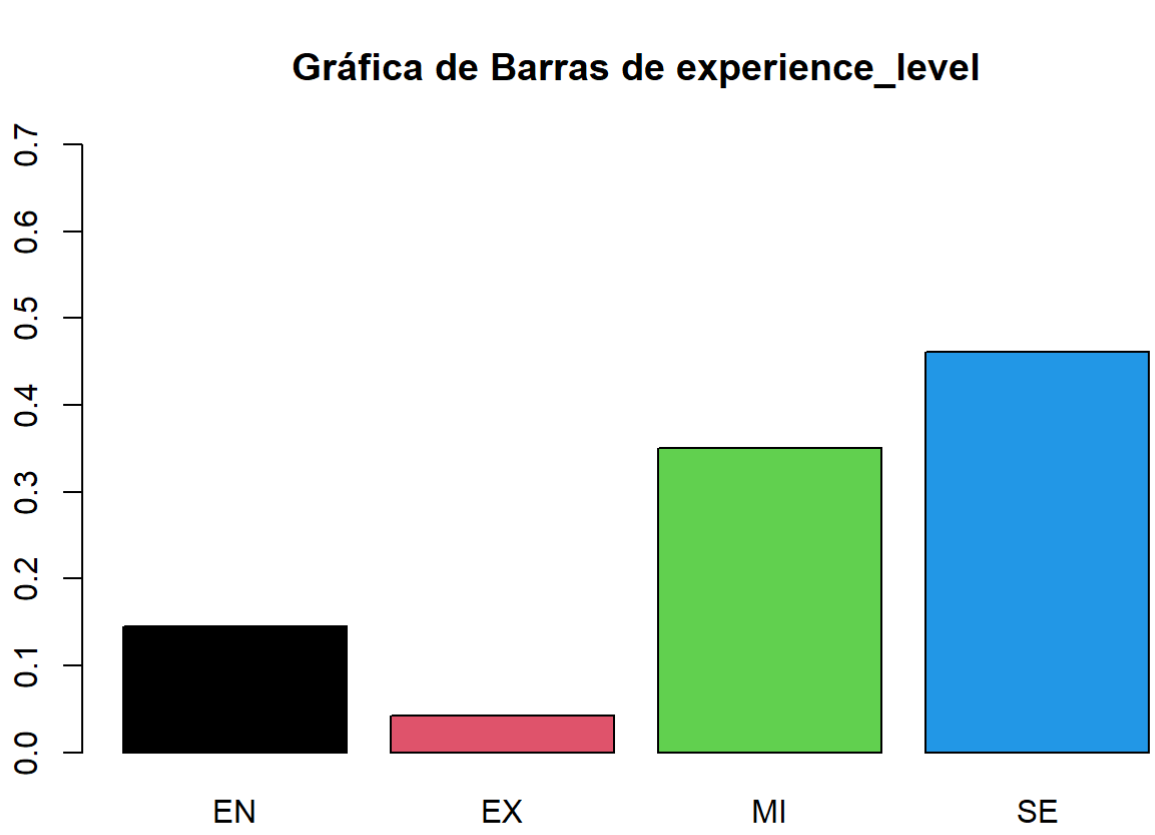


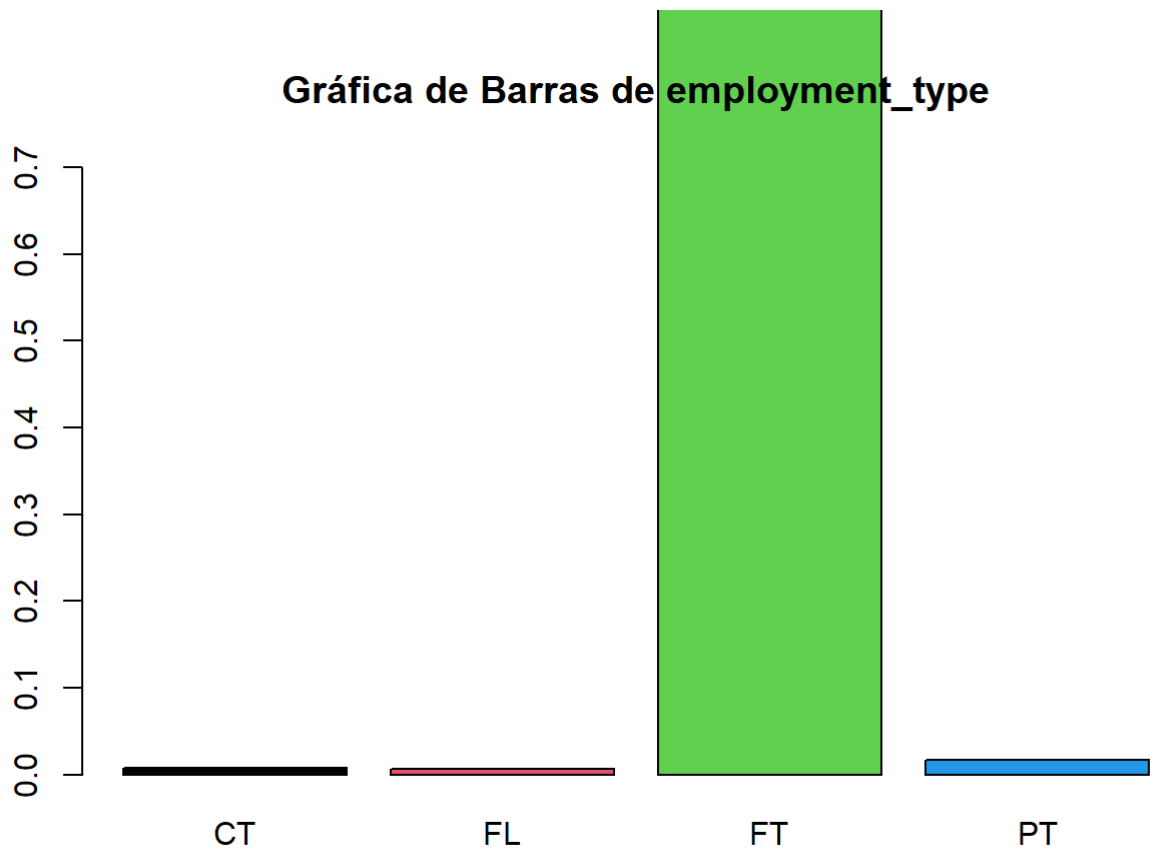

```
## [1] "Ambas distribuciones son asimétricas"
```

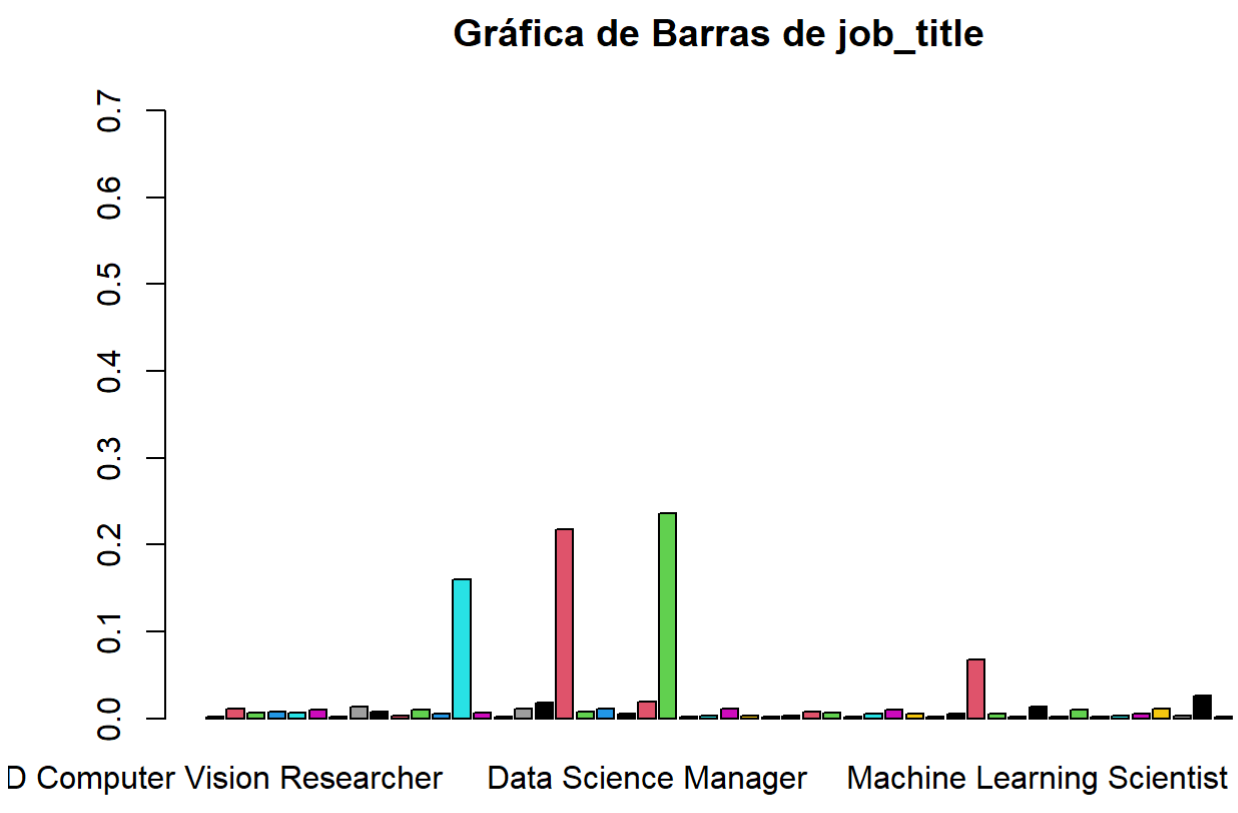
Variables categóricas

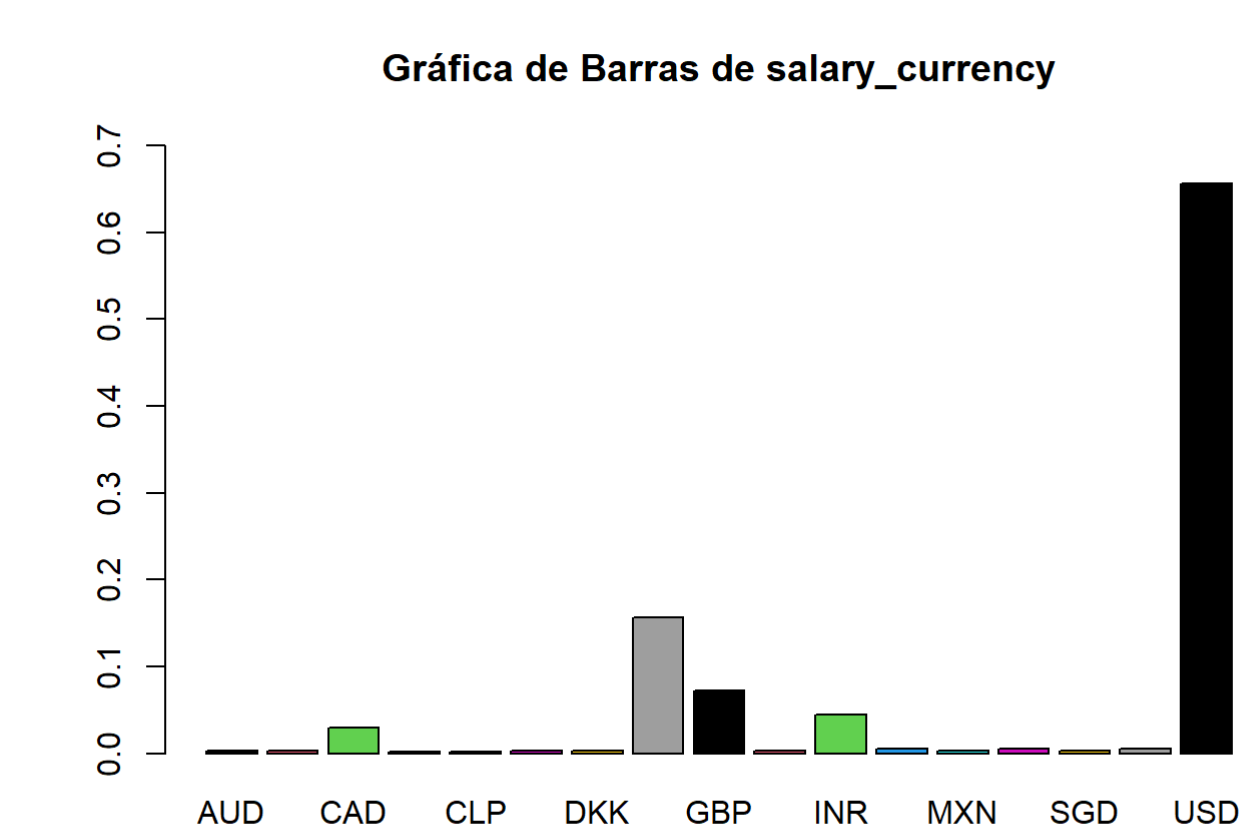
Distribución de los datos (diagramas de barras, diagramas de pastel)

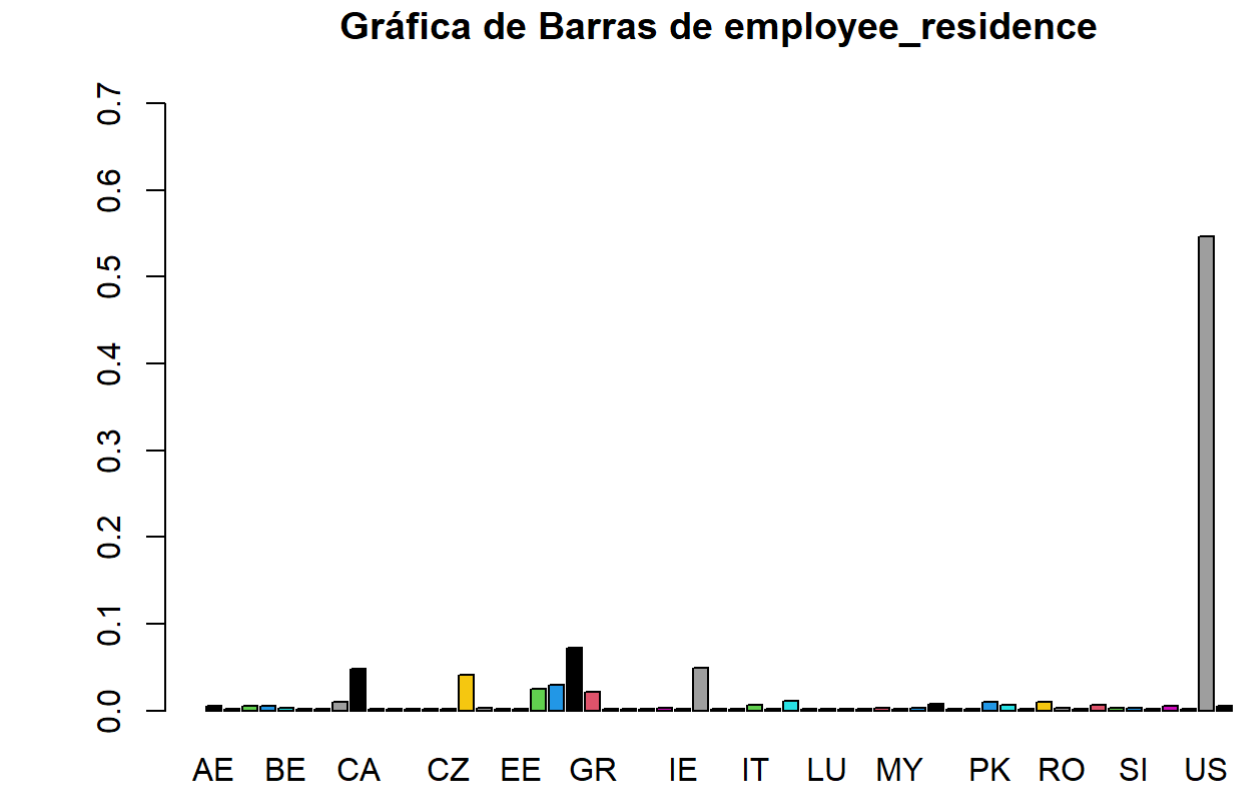

Gráfica de Barras de work_year
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2020
2021
2022

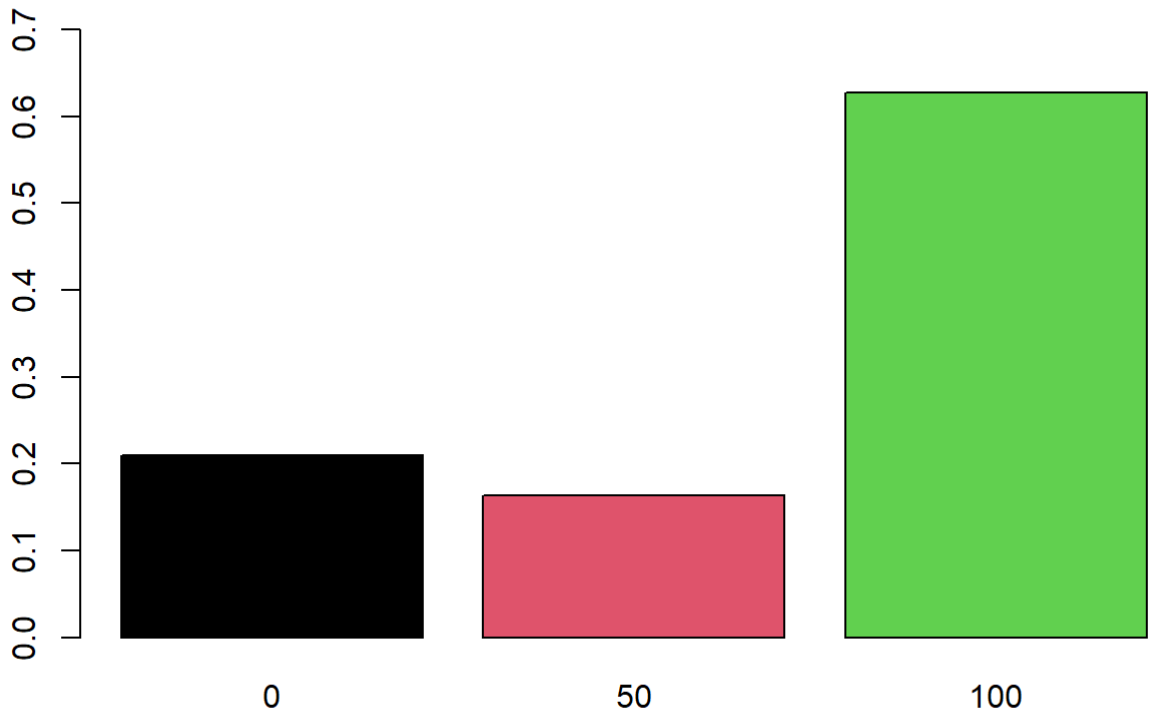

Gráfica de Barras de experience_level
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
EN
EX
MI
SE

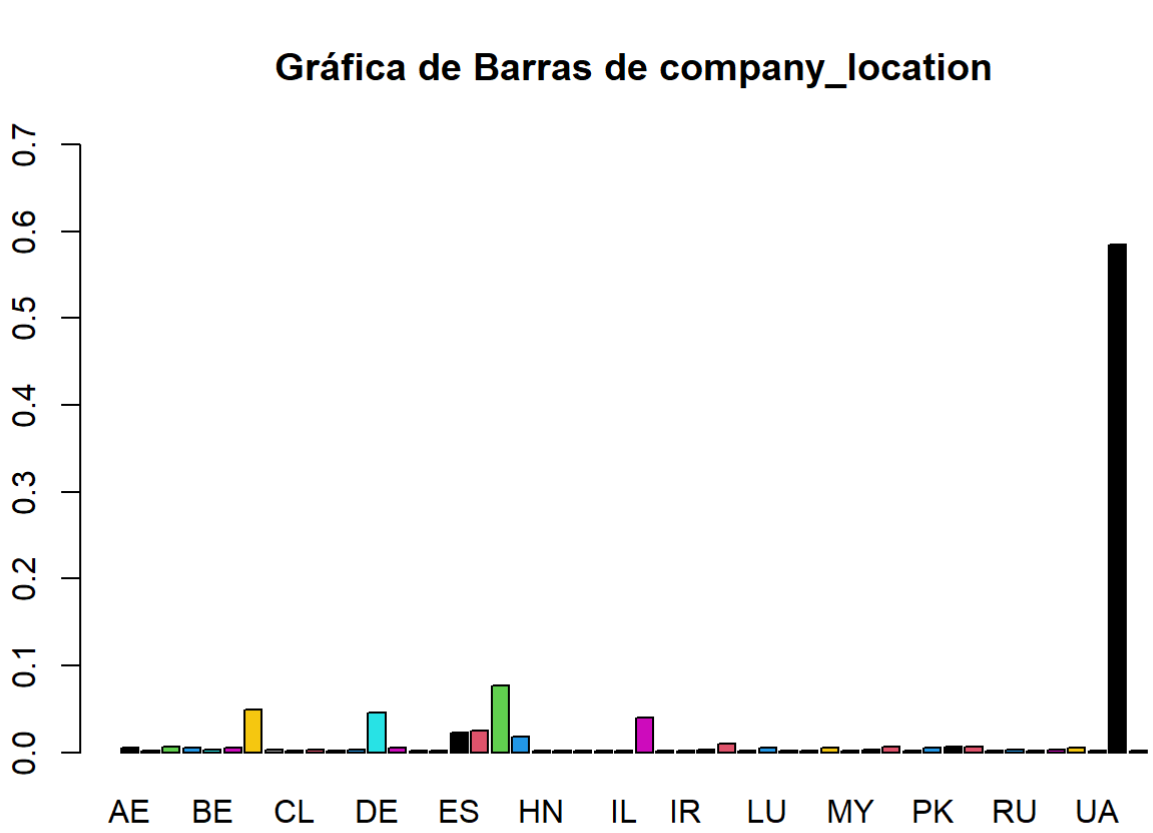

Gráfica de Barras de employment_type
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
CT
FL
FT
PT

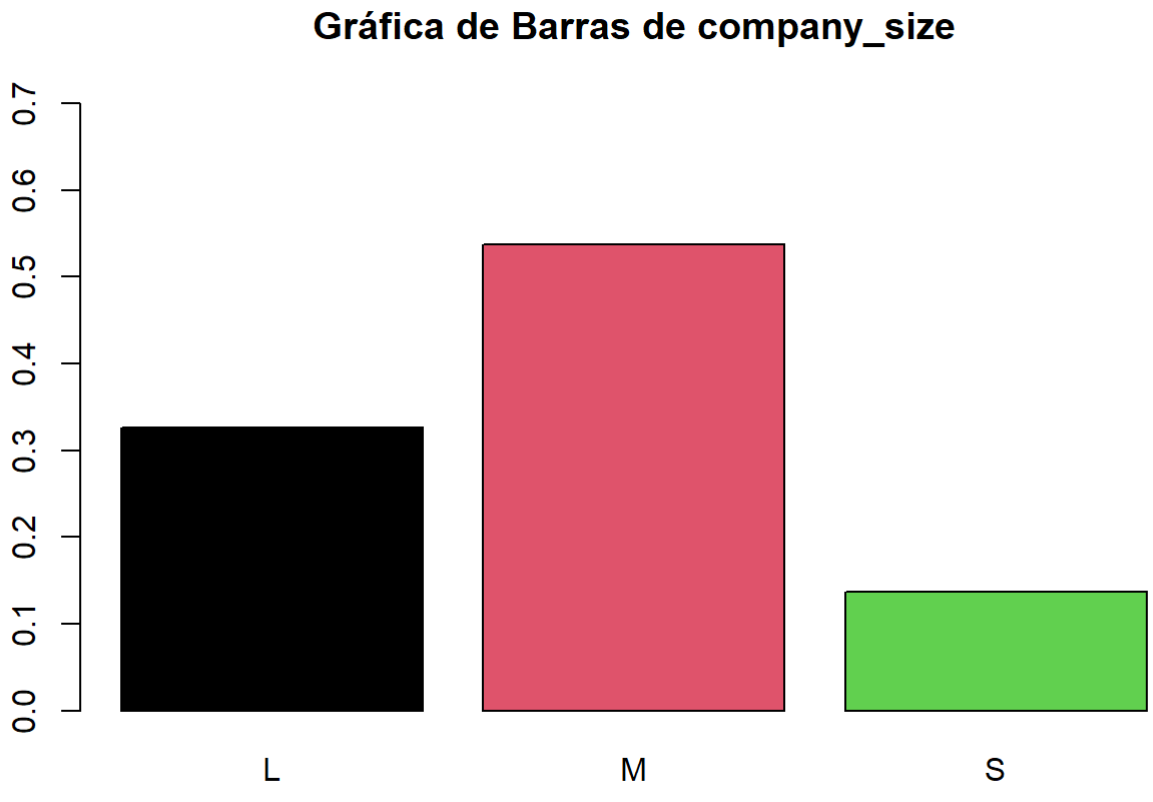

Gráfica de Barras de job_title
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
D Computer Vision Researcher
Data Science Manager
Machine Learning Scientist


Gráfica de Barras de salary_currency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
AUD
CAD
CLP
DKK
GBP
INR
MXN
SGD
USD


Gráfica de Barras de employee_residence
0.0
0.1
0.2
0.3
0.4
0.5
0.6
AE
BE
CA
CZ
EE
GR
IE
IT
LU
MY
PK
RO
SI
US







Identifica problemas de calidad

de datos (registros duplicados, valores faltantes, outliers, etc).

Detectar datos nulos variables numéricas

```
## [1] "La columna salary tiene valores nulos: FALSE"
## [1] "La columna salary_in_usd tiene valores nulos: FALSE"
```

Detectar datos nulos variables categóricas

```
## [1] "La columna work_year tiene valores nulos: FALSE"
## [1] "La columna experience_level tiene valores nulos: FALSE"
## [1] "La columna employment_type tiene valores nulos: FALSE"
## [1] "La columna job_title tiene valores nulos: FALSE"
## [1] "La columna salary_currency tiene valores nulos: FALSE"
## [1] "La columna employee_residence tiene valores nulos: FALSE"
## [1] "La columna remote_ratio tiene valores nulos: FALSE"
## [1] "La columna company_location tiene valores nulos: FALSE"
## [1] "La columna company_size tiene valores nulos: FALSE"
```

Encontrar outliers

```
## [1] "Outliers para columna salary"
##  [1] 11000000  3000000  4450000   423000   450000   325000   720000   450000
##  [9]   450000   412000   400000  1450000  2200000   450000 11000000  2250000
## [17]   700000  3000000  7000000  8500000   423000 30400000   420000  1672000
## [25]  1799997  4000000   435000  2500000   416000  1200000  1600000  1335000
## [33]   600000  2100000  1250000  4900000  7000000  6000000  1400000  2400000
## [41]  1400000   324000   380000   405000
## [1] ""
##       X work_year experience_level employment_type
## 8     7      2020               MI              FT
## 12   11      2020               MI              FT
## 17   16      2020               EN              FT
## 19   18      2020               EN              FT
## 22   21      2020               MI              FT
## 26   25      2020               EX              FT
## 28   27      2020               SE              FT
## 34   33      2020               MI              FT
## 51   50      2020               EN              FT
## 64   63      2020               SE              FT
## 78   77      2021               MI              PT
## 93   92      2021               MI              FT
## 95   94      2021               EN              FT
## 98   97      2021               MI              FT
## 103 102      2021               MI              FT
## 110 109      2021               EN              FT
## 128 127      2021               MI              FT
## 130 129      2021               SE              FT
## 137 136      2021               MI              FT
## 138 137      2021               MI              FT
## 158 157      2021               MI              FT
## 178 177      2021               MI              FT
## 180 179      2021               MI              FT
## 181 180      2021               MI              FT
## 198 197      2021               SE              FT
## 199 198      2021               SE              FT
## 214 213      2021               EN              FT
## 223 222      2021               MI              FT
## 226 225      2021               EX              CT
## 231 230      2021               EN              FT
## 240 239      2021               EN              FT
## 245 244      2021               EN              FT
## 253 252      2021               EX              FT
## 254 253      2021               EN              FT
## 263 262      2021               MI              FT
## 264 263      2021               SE              FT
## 286 285      2021               SE              FT
## 385 384      2022               EX              FT
## 459 458      2022               MI              FT
## 460 459      2022               MI              FT
## 464 463      2022               EN              FT
## 483 482      2022               EX              FT
## 520 519      2022               SE              FT
## 524 523      2022               SE              FT
##                           job_title   salary salary_currency salary_in_usd
## 8                     Data Scientist 11000000             HUF         35735
## 12                    Data Scientist  3000000             INR         40481
## 17                    Data Engineer  4450000             JPY         41689
## 19            Data Science Consultant   423000             INR          5707
## 22               Product Data Analyst   450000             INR          6072
## 26           Director of Data Science   325000             USD        325000
## 28                    Data Engineer   720000             MXN         33511
## 34                 Research Scientist   450000             USD        450000
## 51                       Data Analyst   450000             INR          6072
## 64                    Data Scientist   412000             USD        412000
## 78       3D Computer Vision Researcher   400000             INR          5409
## 93                  Lead Data Analyst  1450000             INR         19609
## 95                    Data Scientist  2200000             INR         29751
## 98               Financial Data Analyst   450000             USD        450000
## 103                   BI Data Analyst 11000000             HUF         36259
## 110                   Data Engineer  2250000             INR         30428
## 128                   Data Scientist   700000             INR          9466
## 130               Lead Data Scientist  3000000             INR         40570
## 137                      ML Engineer  7000000             JPY         63711
## 138                      ML Engineer  8500000             JPY         77364
## 158 Applied Machine Learning Scientist   423000             USD        423000
## 178                   Data Scientist 30400000             CLP         40038
## 180                   Data Scientist   420000             INR          5679
## 181                 Big Data Engineer  1672000             INR         22611
## 198          Machine Learning Engineer  1799997             INR         24342
## 199               Data Science Manager  4000000             INR         54094
## 214                 Big Data Engineer   435000             INR          5882
## 223                   Data Scientist  2500000             INR         33808
```

```
## 226           Principal Data Scientist   416000             USD        416000
## 231                 Big Data Engineer  1200000             INR         16228
## 240                     Data Engineer  1600000             INR         21637
## 245                      AI Scientist  1335000             INR         18053
## 253            Principal Data Engineer   600000             USD        600000
## 254                     Data Scientist  2100000             INR         28399
## 263                     Data Scientist  1250000             INR         16904
## 264          Machine Learning Engineer  4900000             INR         66265
## 286               Data Science Manager  7000000             INR         94665
## 385          Head of Machine Learning  6000000             INR         79039
## 459              Business Data Analyst  1400000             INR         18442
## 460                     Data Scientist  2400000             INR         31615
## 464                     Data Scientist  1400000             INR         18442
## 483                      Data Engineer   324000             USD        324000
## 520             Applied Data Scientist   380000             USD        380000
## 524                Data Analytics Lead   405000             USD        405000
##     employee_residence remote_ratio company_location company_size
## 8                   HU           50               HU            L
## 12                  IN            0               IN            L
## 17                  JP          100               JP            S
## 19                  IN           50               IN            M
## 22                  IN          100               IN            L
## 26                  US          100               US            L
## 28                  MX            0               MX            S
## 34                  US            0               US            M
## 51                  IN            0               IN            S
## 64                  US          100               US            L
## 78                  IN           50               IN            M
## 93                  IN          100               IN            L
## 95                  IN           50               IN            L
## 98                  US          100               US            L
## 103                 HU           50               US            L
## 110                 IN          100               IN            L
## 128                 IN            0               IN            S
## 130                 IN           50               IN            L
## 137                 JP           50               JP            S
## 138                 JP           50               JP            S
## 158                 US           50               US            L
## 178                 CL          100               CL            L
## 180                 IN          100               US            S
## 181                 IN            0               IN            L
## 198                 IN          100               IN            L
## 199                 IN           50               US            L
## 214                 IN            0               CH            L
## 223                 IN            0               IN            M
## 226                 US          100               US            S
## 231                 IN          100               IN            L
## 240                 IN           50               IN            M
## 245                 IN          100               AS            S
## 253                 US          100               US            L
## 254                 IN          100               IN            M
## 263                 IN          100               IN            S
## 264                 IN            0               IN            L
## 286                 IN           50               IN            L
## 385                 IN           50               IN            L
## 459                 IN          100               IN            M
## 460                 IN          100               IN            L
## 464                 IN          100               IN            M
## 483                 US          100               US            M
## 520                 US          100               US            L
## 524                 US          100               US            L
## [1] "Outliers para columna salary_in_usd"
##  [1] 325000 450000 412000 450000 423000 416000 600000 324000 380000 405000
## [1] ""
##       X work_year experience_level employment_type
## 26   25      2020               EX              FT
## 34   33      2020               MI              FT
## 64   63      2020               SE              FT
## 98   97      2021               MI              FT
## 158 157      2021               MI              FT
## 226 225      2021               EX              CT
## 253 252      2021               EX              FT
## 483 482      2022               EX              FT
## 520 519      2022               SE              FT
## 524 523      2022               SE              FT
##                            job_title salary salary_currency salary_in_usd
## 26            Director of Data Science 325000             USD        325000
## 34                  Research Scientist 450000             USD        450000
## 64                      Data Scientist 412000             USD        412000
## 98               Financial Data Analyst 450000             USD        450000
## 158 Applied Machine Learning Scientist 423000             USD        423000
## 226           Principal Data Scientist 416000             USD        416000
## 253            Principal Data Engineer 600000             USD        600000
```

```
## 483                      Data Engineer 324000             USD        324000
## 520              Applied Data Scientist 380000             USD        380000
## 524                 Data Analytics Lead 405000             USD        405000
##     employee_residence remote_ratio company_location company_size
## 26                  US          100               US            L
## 34                  US            0               US            M
## 64                  US          100               US            L
## 98                  US          100               US            L
## 158                 US           50               US            L
## 226                 US          100               US            S
## 253                 US          100               US            L
## 483                 US          100               US            M
## 520                 US          100               US            L
## 524                 US          100               US            L
```

Encontrar datos duplicados

```
##  [1] X                 work_year         experience_level   employment_type
##  [5] job_title         salary            salary_currency    salary_in_usd
##  [9] employee_residence remote_ratio     company_location   company_size
## <0 rows> (or 0-length row.names)
```

```
## [1] "No hay datos duplicados en el dataframe"
```

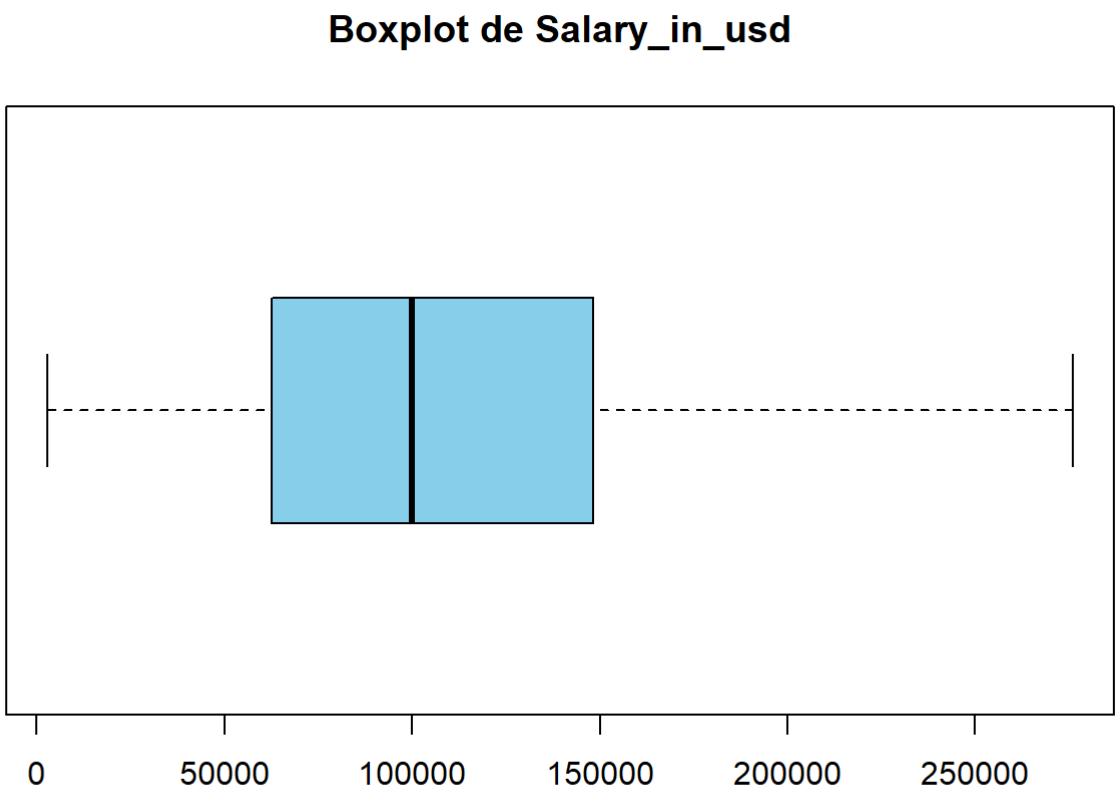
#c) Preparación de los Datos ## Selecciona el conjunto de datos a utilizar. Decide qué conjunto de datos se utilizará. Identifica variables objetivo. En caso necesario, explica por qué se incluyeron o excluyeron variables.

```
## 'data.frame':    607 obs. of  5 variables:
##  $ experience_level: chr  "MI" "SE" "SE" "MI" ...
##  $ employment_type : chr  "FT" "FT" "FT" "FT" ...
##  $ salary_in_usd   : int  79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...
##  $ company_location: chr  "DE" "JP" "GB" "HN" ...
##  $ company_size    : chr  "L" "S" "M" "S" ...
```

En caso de necesidad de recorte de datos (atípicos, faltantes, duplicados, etc), explica el motivo de la reducción.

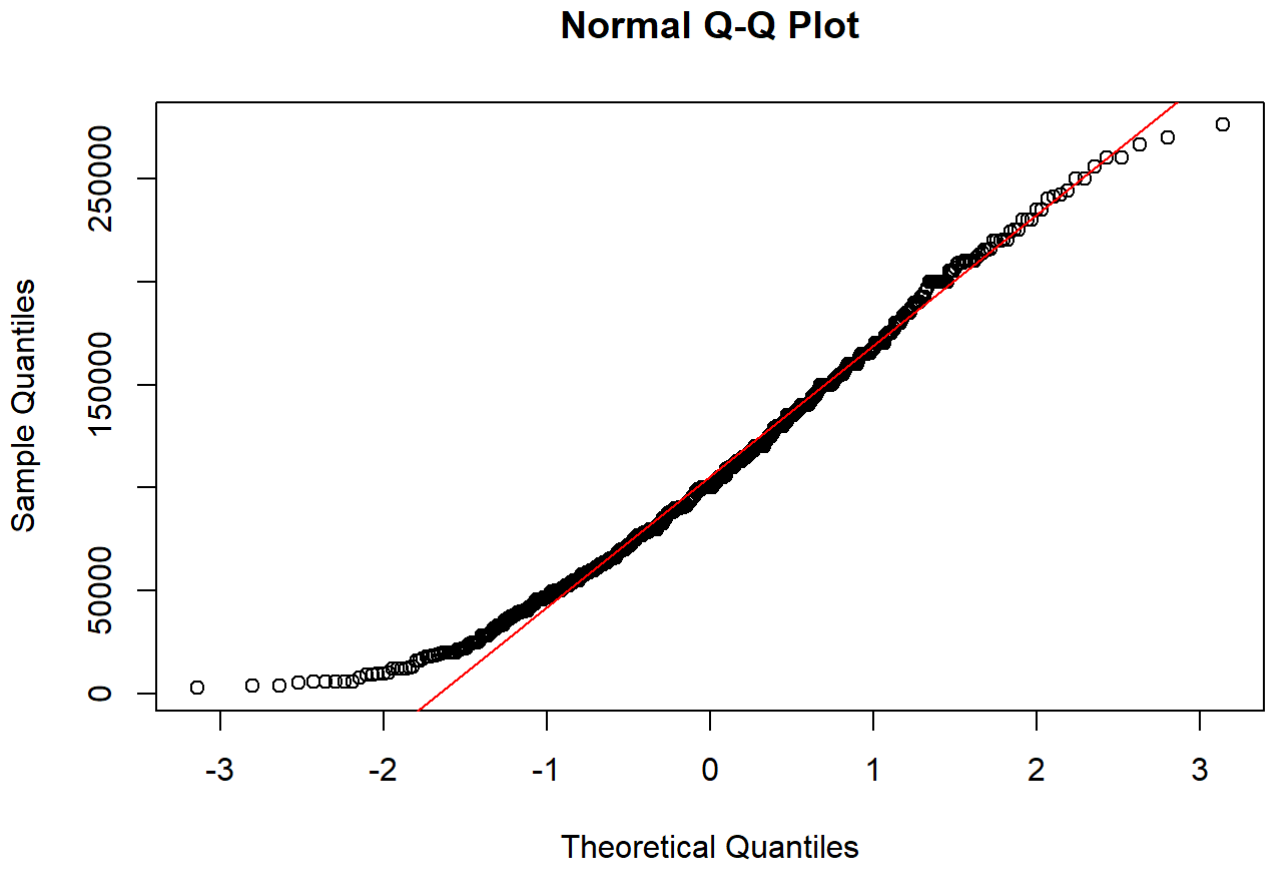
Quitar outliers

Se procede a quitar los valores outliers del dataframe en cuestión ya que está por fuera del 3er cuartíl y representan 10 registros de 670 de salarios excesivamente elevados en comparación



Transforma los datos en caso necesario. Revisa si es necesario discretizar los datos Revisa si es necesario escalar y normalizar los datos Construye atributos si es conveniente

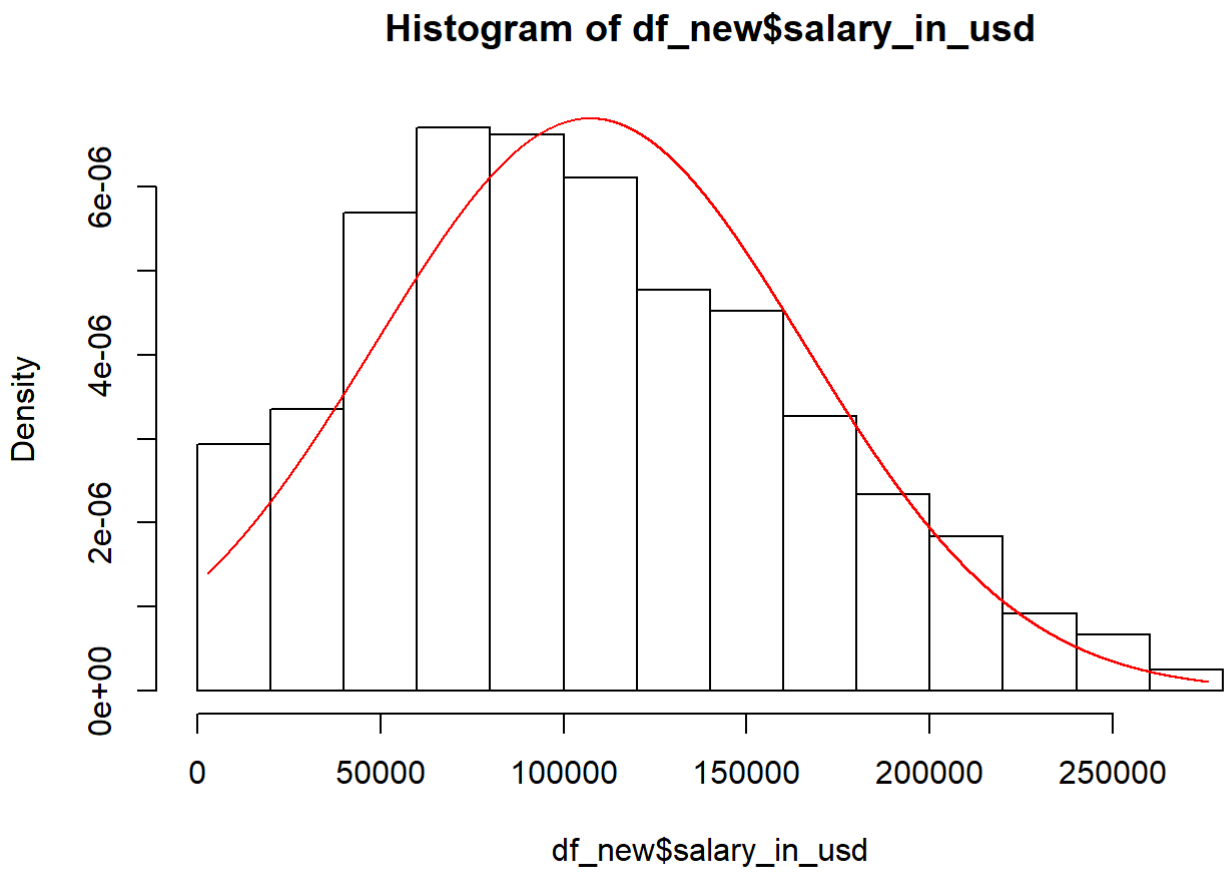
Explorar la normalidad de los datos



Interpretación Corresponde a

una distribución con colas delgadas (alta curtuosis, distribución Leptocúrtica).

Histograma de Distribución de la Data



Checar la cuortuosis y el sesgo

```
## Warning: package 'moments' was built under R version 4.1.3
```

```
##
## Attaching package: 'moments'
```

```
## The following object is masked from 'package:modeest':
##
##     skewness
```

```
## [1] 0.4185161
```

```
## [1] 2.601883
```

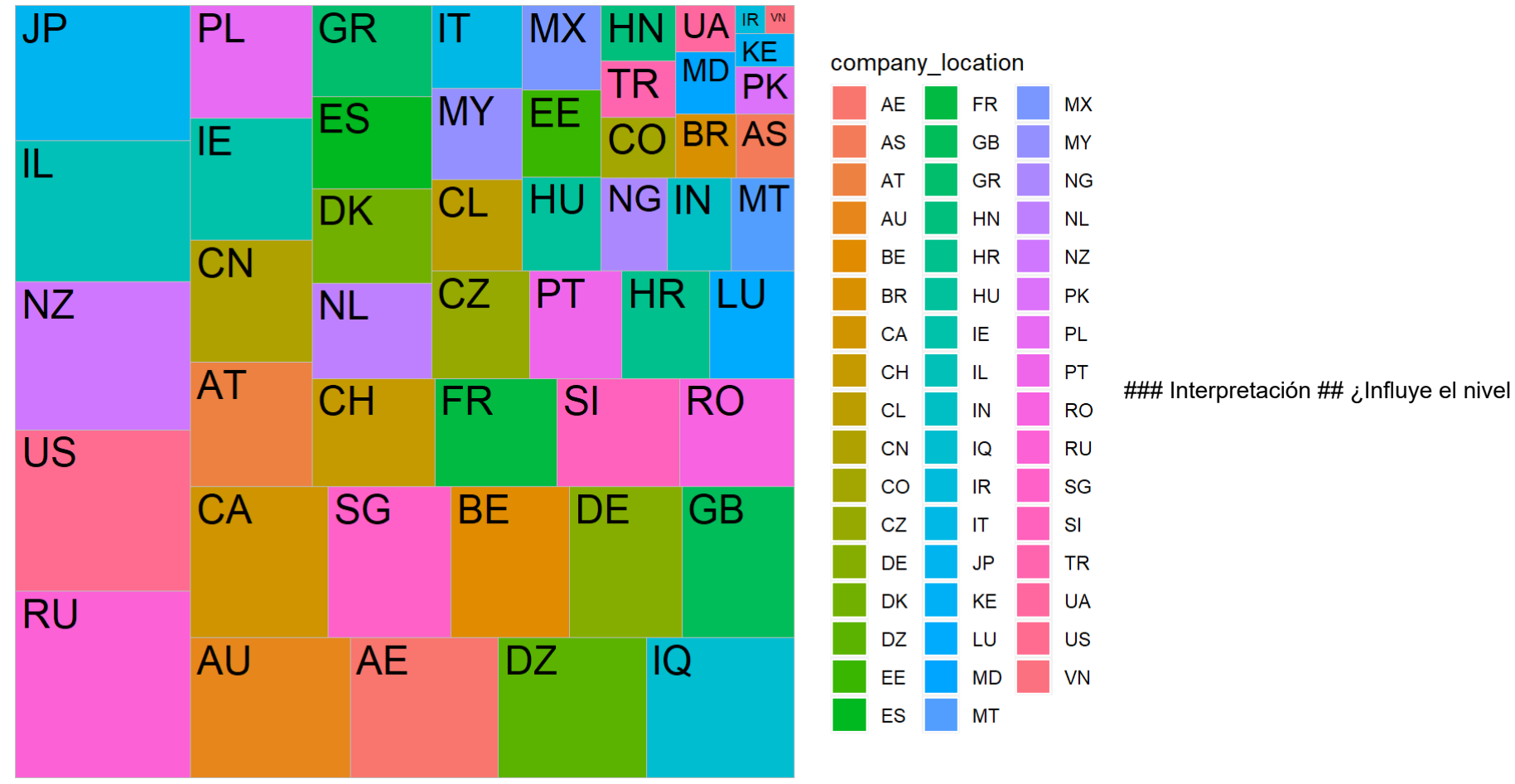
Análisis e Interpretación de la data

¿En qué países se ofrecen mejores salarios?

```
##    company_location salary_in_usd
## 44               RU     157500.00
## 49               US     136100.34
## 39               NZ     125000.00
## 25               IL     119059.00
## 30               JP     114127.33
## 4                AU     108042.67
## 1                AE     100000.00
## 15               DZ     100000.00
## 27               IQ     100000.00
## 7                CA      99823.73
## 45               SG      89294.00
## 5                BE      85699.00
## 13               DE      81887.21
## 19               GB      81583.04
## 3                AT      72920.75
## 10               CN      71665.50
## 24               IE      71444.00
## 41               PL      66082.50
## 8                CH      64114.00
## 18               FR      63970.67
## 46               SI      63831.00
## 43               RO      60000.00
## 38               NL      54945.75
## 14               DK      54386.33
## 17               ES      53060.14
## 20               GR      52293.09
## 12               CZ      50937.00
## 42               PT      47793.75
## 22               HR      45618.00
## 32               LU      43942.67
## 9                CL      40038.00
## 36               MY      40000.00
## 29               IT      36366.50
## 23               HU      35735.00
## 16               EE      32974.00
## 35               MX      32123.33
## 37               NG      30000.00
## 26               IN      28581.75
## 34               MT      28369.00
## 11               CO      21844.00
## 47               TR      20096.67
## 21               HN      20000.00
## 6                BR      18602.67
## 2                AS      18053.00
## 33               MD      18000.00
## 48               UA      13400.00
## 40               PK      13333.33
## 31               KE       9272.00
## 28               IR       4000.00
## 50               VN       4000.00
```

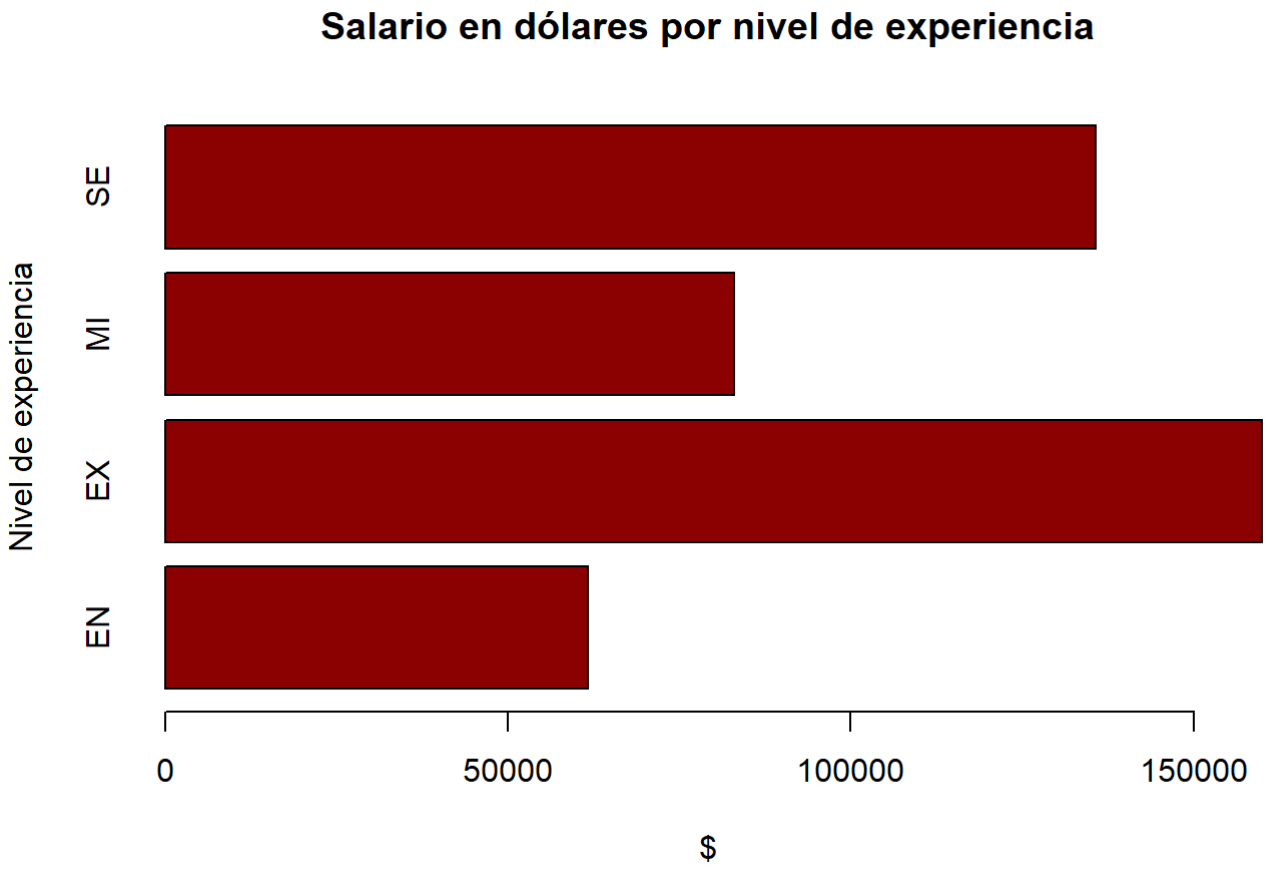
```
## Warning: package 'treemapify' was built under R version 4.1.3
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```



Interpretación ## ¿Influye el nivel de experiencia en el salario?

```
##   experience_level salary_in_usd
## 2               EX     159963.32
## 4               SE     135797.26
## 3               MI      82953.14
## 1               EN      61643.32
```



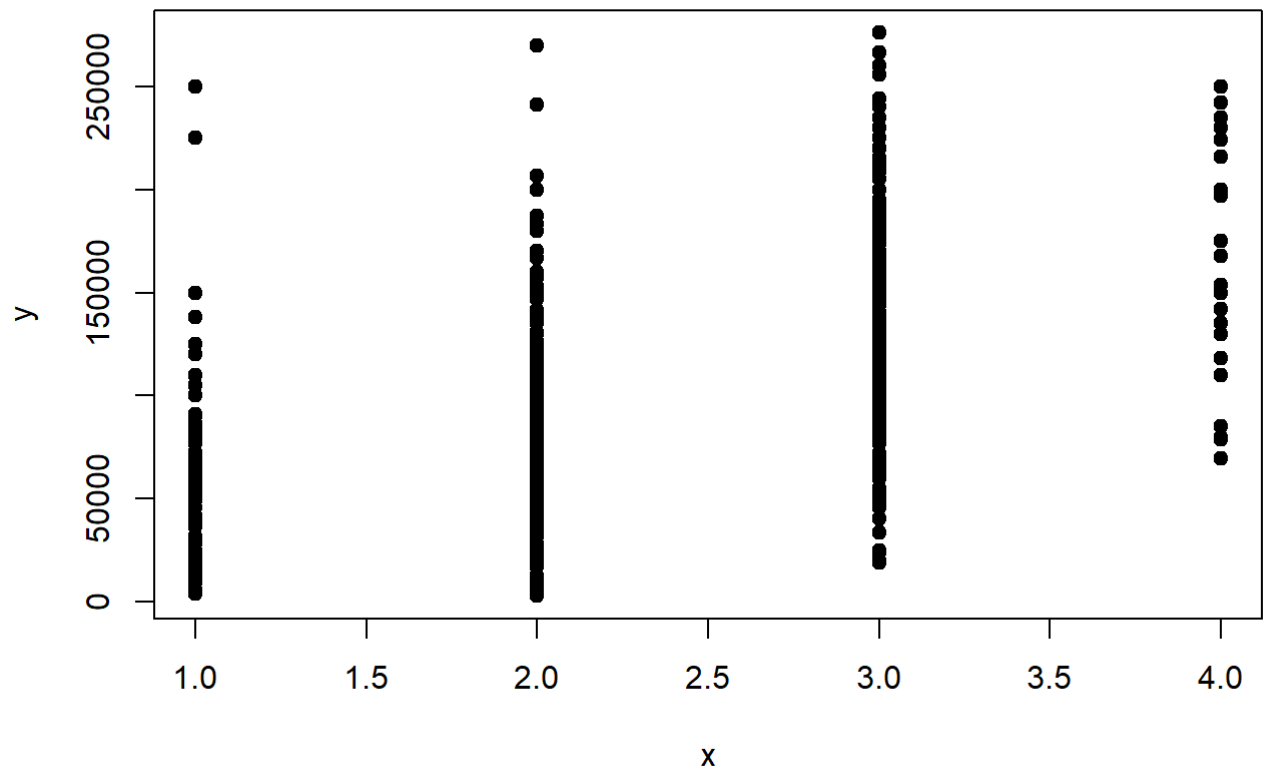
Interpretación

Se puede ver que el nivel de experiencia si influye en el Salario Promedio ya que los Expert EX Executive-level/Director tienen un salario promedio mayor a los demás niveles, seguido del Intermediate SE Senior-level y subsecuentemente los Junior MI Mid-level y por último los EN Entry-level.

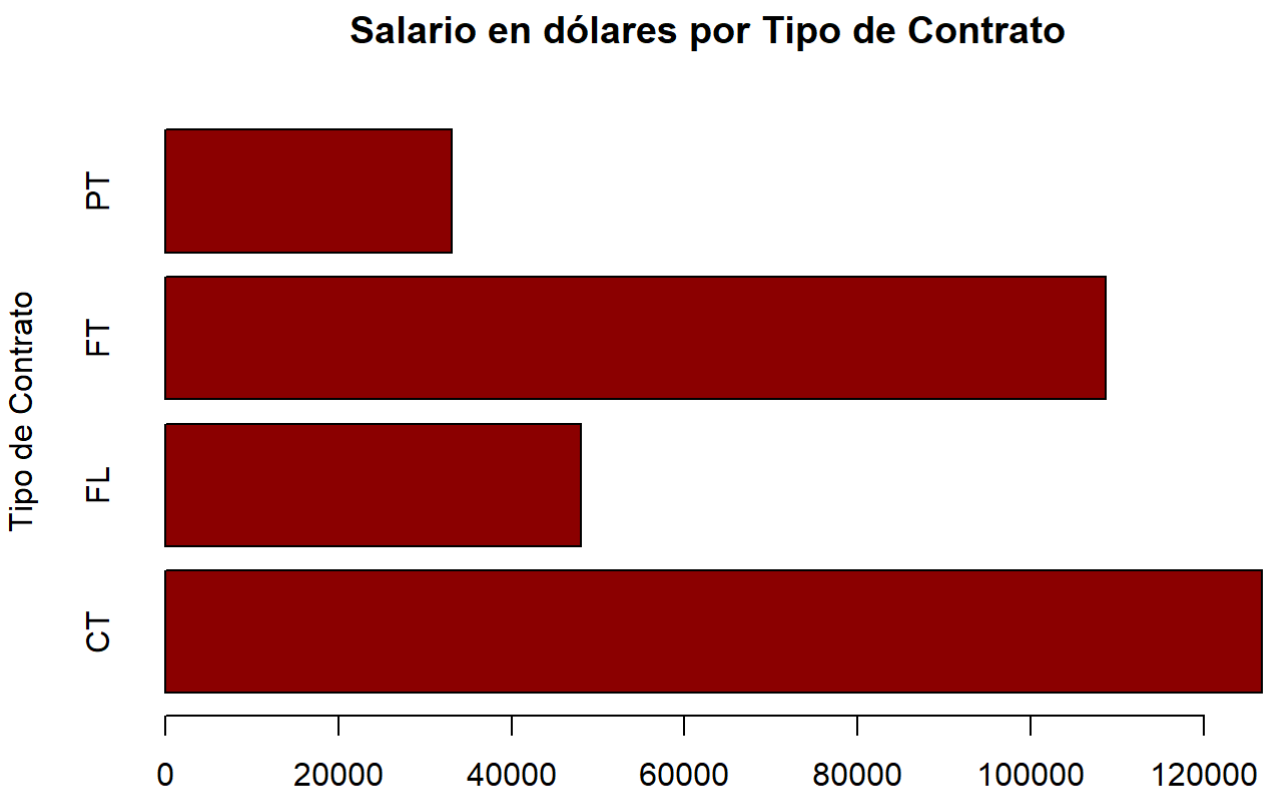
Añadir una columna numérica al dataset.

```
## 'data.frame':    597 obs. of  6 variables:
##  $ experience_level: chr  "MI" "SE" "SE" "MI" ...
##  $ employment_type : chr  "FT" "FT" "FT" "FT" ...
##  $ salary_in_usd   : int  79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...
##  $ company_location: chr  "DE" "JP" "GB" "HN" ...
##  $ company_size    : chr  "L" "S" "M" "S" ...
##  $ exp_num         : num  2 3 3 2 3 1 3 2 2 3 ...
```

Scatter Plot para evaluar la relación entre Experiencia-Salario




```
##   employment_type salary_in_usd
## 1              CT      126718.8
## 3              FT      108722.3
## 2              FL       48000.0
## 4              PT       33070.5
```



El tipo de contrato que ofrece

$

mejores salarios, de acuerdo al salario medio corresponde al de contrato ordinario (CT) ya que su salario promedio sobrepasa los $120’000 USD.

No obstante, se procederá a realizar un análisis más exhaustivo de ANOVA con 3 factores para determinar si el tipo de contrato, el tamaño de la compañía y/o el nivel de experiencia son factores determinantes para el salario de un trabajo data-oriented.Para ello, se requiere estudiar si existen diferencias significativas entre las medias de una variable continua (en este caso el salario promedio en usd) en diferentes niveles de una variable cualitativa (en este caso, hay 3 variables cualitativas a través de las cuales se van a evaluar la influencia así como su respectiva interacción entre si: nivel de experiencia del profesional, tamaño de la compañía y tipo de contrato) tras examinar el valor de las medias de cada grupo y su respectiva interacción entre sí.

Establece las hipótesis estadísticas.

Factores del Problema: * Nivel de Experiencia * Tamaño de la compañía. * Tipo de Contrato

Pregunta de Investigación:

¿Existe alguna influencia del nivel de experiencia, el tamaño de la compañía y/o el tipo de contrato en el salario de una profesión data-oriented?

Hipótesis estadísticas:

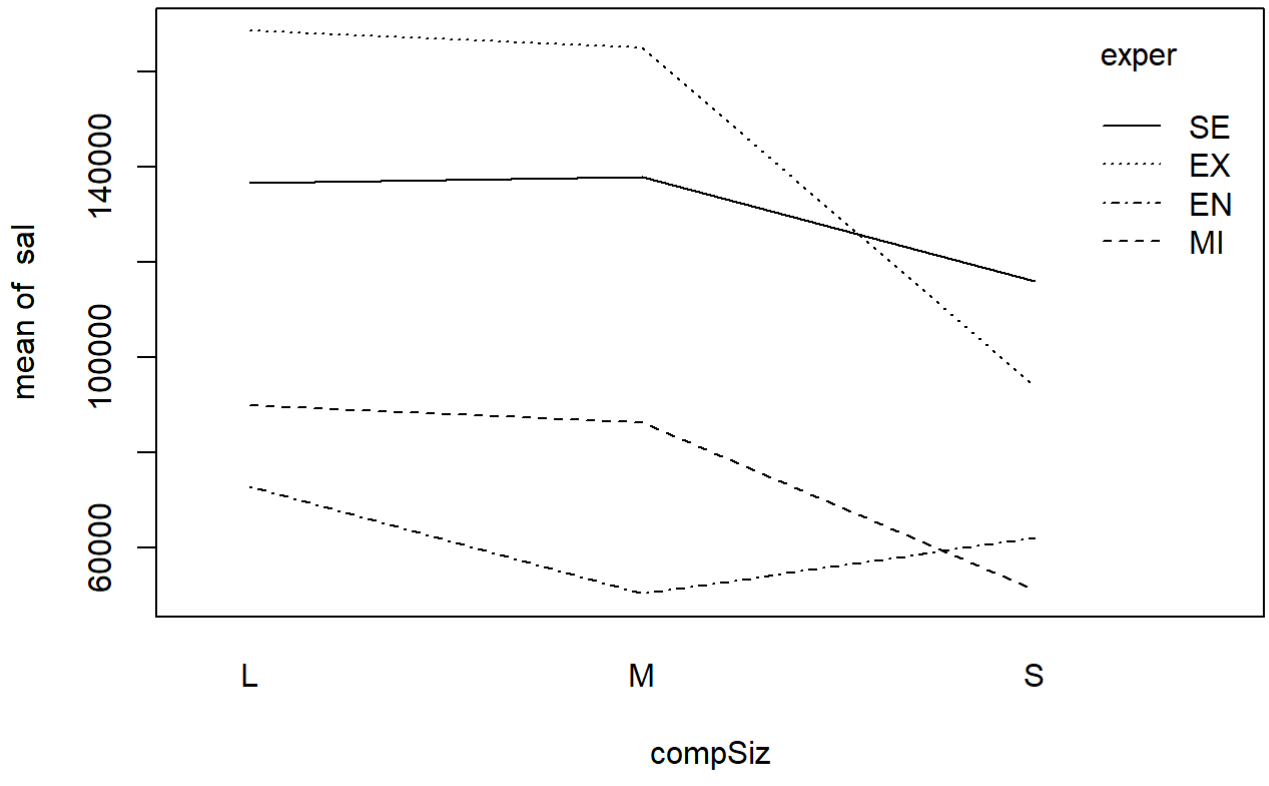
H0 = El nivel de experiencia no incide en el salario promedio de un data-oriented profesional. H1 = El tamaño de la compañía no incide en el salario promedio de un data-oriented profesional. H2 = El tipo de contrato no incide en el salario promedio de un data-oriented profesional. H3 = El nivel de experiencia y el tamaño de la compañía interactúan entre sí en el salario promedio de un data-oriented profesional. H4 = El tamaño de la

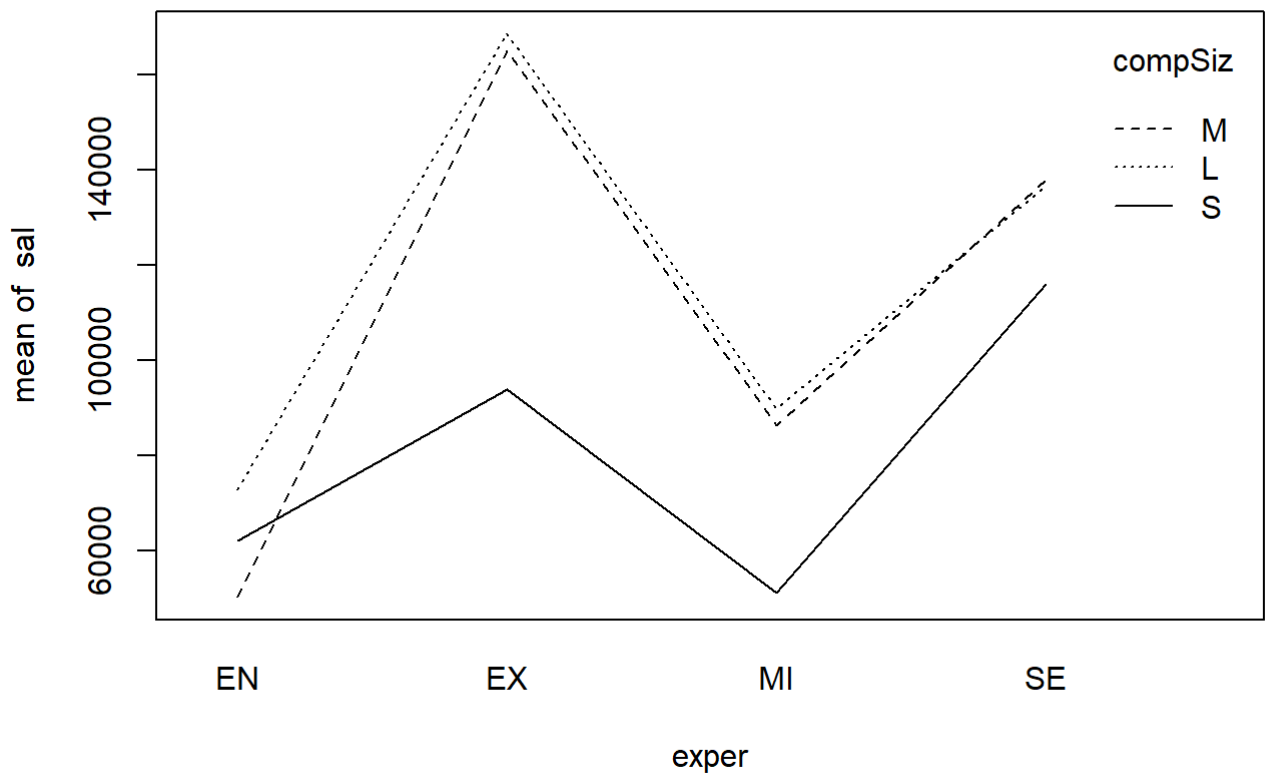
compañía y el tipo de contrato interactúan entre sí en el salario promedio de un data-oriented profesional. H5 = El nivel de experiencia y el tipo de contrato interactúan entre sí en el salario promedio de un data-oriented profesional.

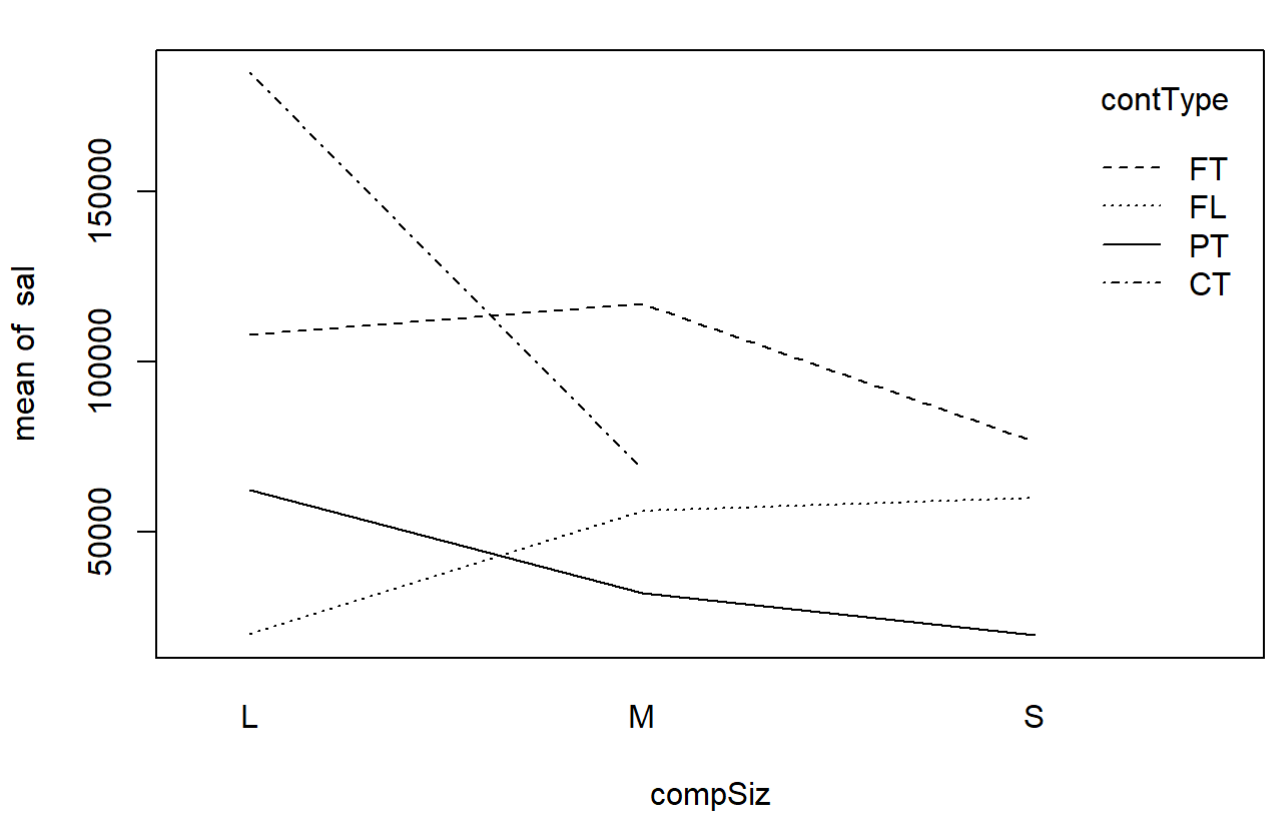
Preparación de los datos

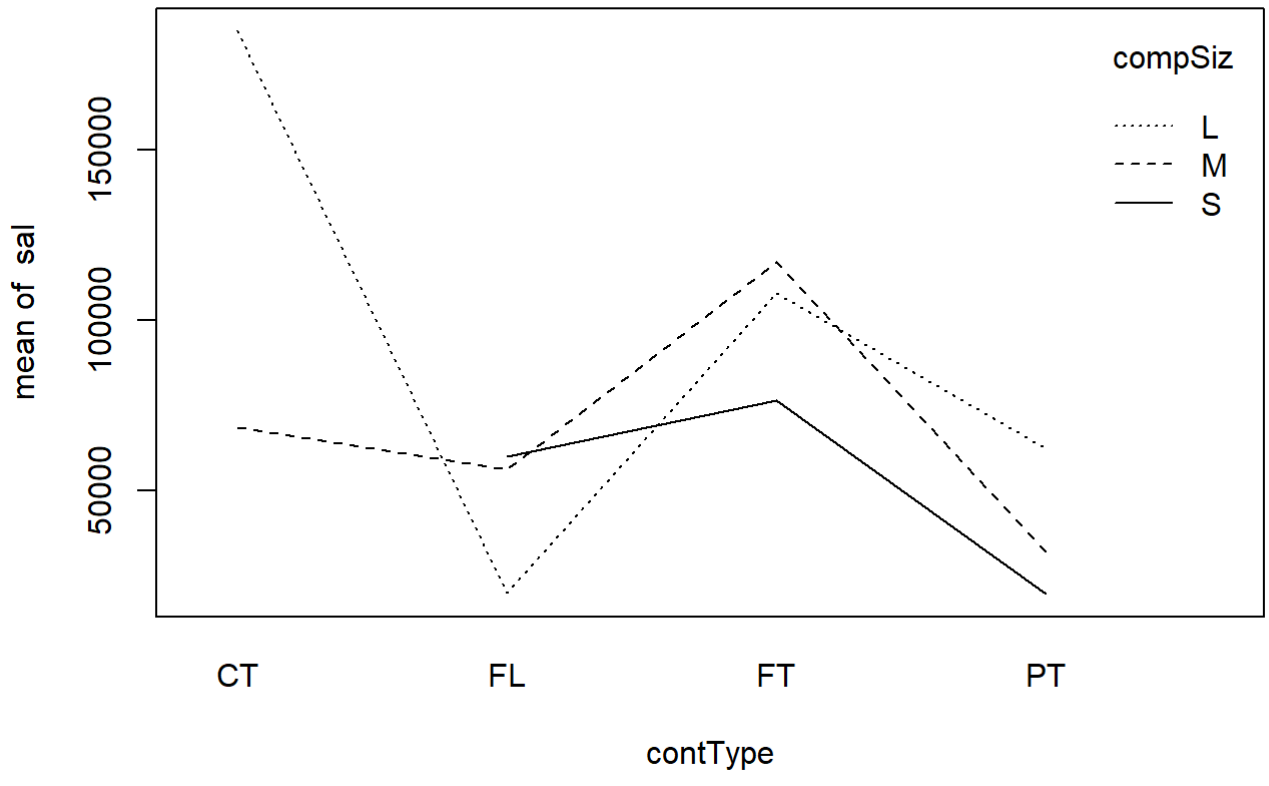
Exploración de los datos

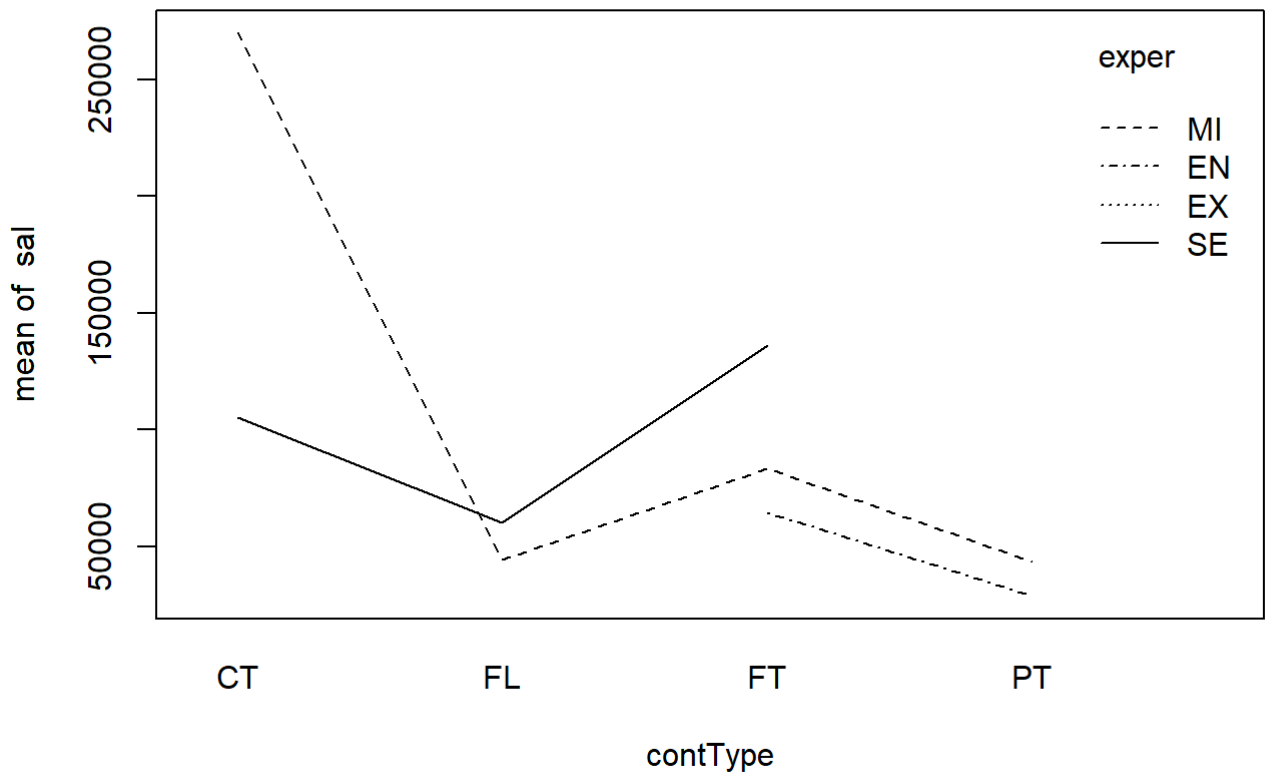
Haz la gráfica de interacción de dos factores.

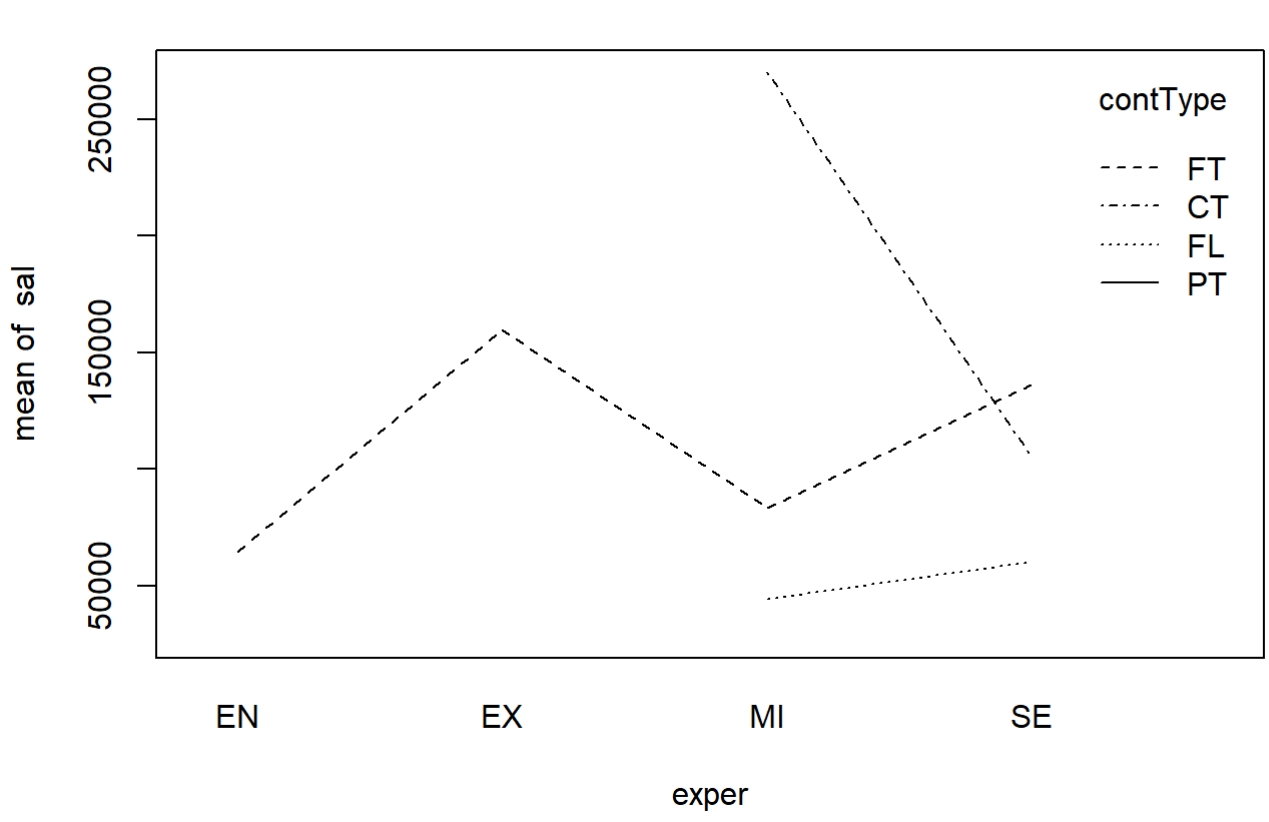

mean of sal
140000
100000
60000
L
M
S
compSiz
exper
SE
EX
EN
MI


mean of sal
140000
100000
60000
EN
EX
MI
SE
exper
compSiz
M
L
S


mean of sal
150000
100000
50000
L
M
S
compSiz
contType
FT
FL
PT
CT







A partir de las gráficas de interacción

se pudo rescatar lo siguiente: * Independientemente del tamaño de la compañía, los profesionales más experimentados (EX) son los que tienen mayores salarios promedios, seguidos de los intermedios (Intermediate Senior-level - SE), nivel medio Junio (Junior Mid-Level - MID) y finalmente los de primer ingreso (EN). No obstante, en las empresas de tamaño pequeño (S) si existe una pequeña excepción a este patrón ya que se puede ver que los profesionales de primer ingreso ganan más que los de nivel intermedio. Por ende, se puede ver un efecto de interacción mínimo/limitado entre las variables (nivel de experiencia y tamaño de la compañía). También se puede ver claramente que independientemente

del nivel de expertise del profesionista, los profesionales que trabajan en compañías de tamaño grande, tienen mayor salario promedio que en compañías de otros tamaños.No obstante, si existe un pequeño efecto de interacción para las compañías de tamaño mediano y pequeño ya que los profesionales de nivel intermedio en compañías pequeñas ganan más que aquellos trabajando en compañías medianas.

- Debido a los cruces en los gráficos, existe un efecto de interacción entre el tipo de contrato y el tamaño de la compañía, especialmente por el ejemplo de los Contracts (CT), donde al parecer los profesionistas de este esquema ganan mucho más en las compañías de tamaño pequeño y grande que en las compañas medianas. Sin embargo, el esquema de Full-Time (FT) es más remunerado que los contratos (CT) en las compañías de tamaño mediano (M). Además, los profesionistas bajo el esquema de Free-Lancer (FL) suelen ganar más que los profesionistas Part-Time (PT) en las compañías de tamaño Mediano y Pequeño. Por ende, si existe un efecto de interacción muy definido entre ambas variables (tipo de contrato y tamaño de la compañía). En efecto, también existe un cruce importante ya que en las compañías pequeñas (S) los profesionistas que están bajo esquema de contrato (CT), son mejor remunerados que en los demás tamaños de empresas.
- Con respecto a la interacción entre el tipo de contrato y el nivel de expertise del profesional, cabe recalcar que existe un efecto de interacción mínimo entre ambas variables ya que para los profesionistas de nivel intermedio que están bajo el esquema de contract (CT) son menos remunerados que los profesionistas de nivel experto (EX), no obstante, en el esquema de Full-Time (FT), los intermediate Senior-Level son mejor remunerados que los expertos.

Por ende, únicamente la interacción tamaño de la compañía-tipo de contrato es significativa para obtener un mejor salario como profesionista de datos.

Escribe tus conclusiones parciales

Por consecuencia, se rechazan las hipótesis H3 y H5 ya que la interacción es mínima para tamaño de empresa-Nivel de Experiencia y Tipo de Contrato-Nivel de Experiencia, por lo que es necesario reducir el modelo a las primeras hipótesis:

H0 = El nivel de experiencia no incide en el salario promedio de un data-oriented profesional. H1 = El tamaño de la compañía no incide en el salario promedio de un data-oriented profesional. H2 = El tipo de contrato no incide en el salario promedio de un data-oriented profesional. H4 = El tamaño de la compañía y el tipo de contrato interactúan entre sí en el salario promedio de un data-oriented profesional.

Aplicación del Modelo

```
##                   Df    Sum Sq   Mean Sq F value   Pr(>F)
## compSiz            2 1.141e+11 5.706e+10  24.280 7.44e-11 ***
## contType           3 5.569e+10 1.856e+10   7.899 3.59e-05 ***
## exper              3 4.824e+11 1.608e+11  68.422  < 2e-16 ***
## compSiz:contType   5 2.105e+10 4.210e+09   1.791    0.113
## Residuals        583 1.370e+12 2.350e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Warning: package 'lsr' was built under R version 4.1.3
```

```
##                      eta.sq eta.sq.part
## compSiz          0.01445772  0.02110692
## contType         0.01177593  0.01725933
## exper            0.23626261  0.26055112
## compSiz:contType 0.01029996  0.01512881
```

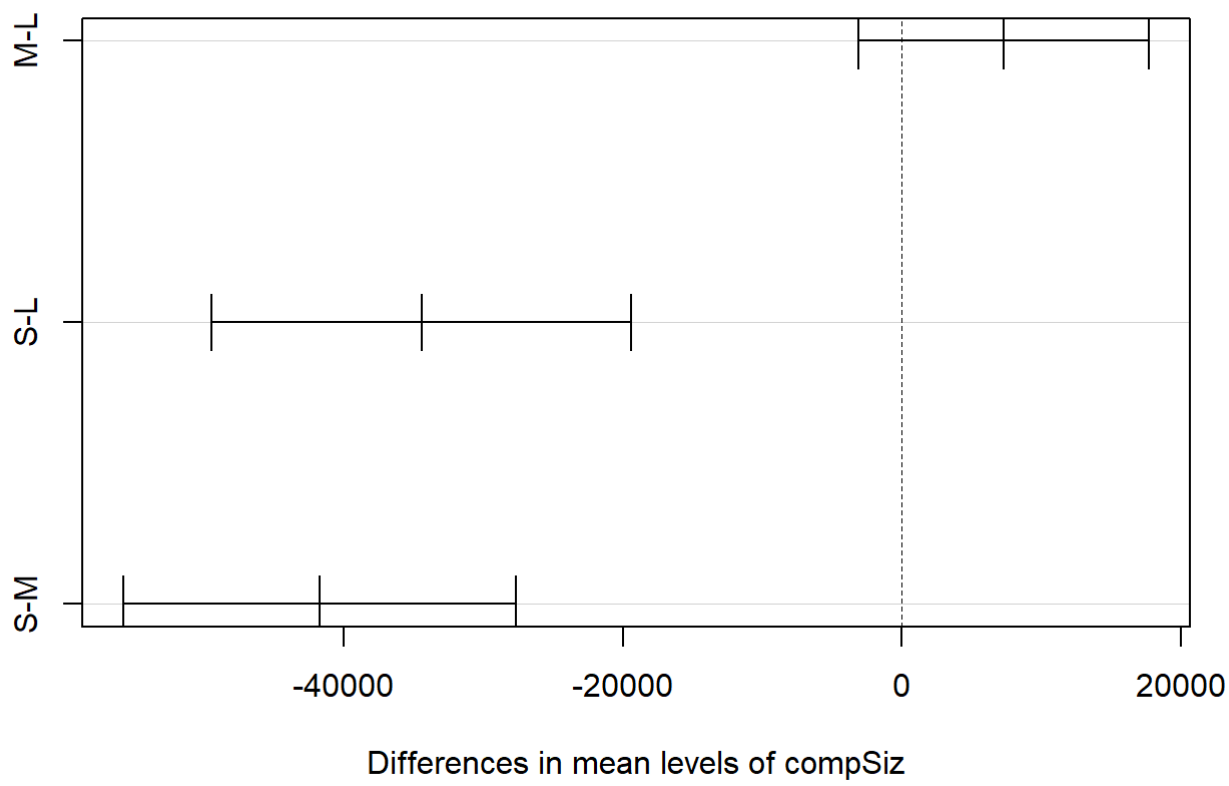
Interpreta el resultado desde la perspectiva estadística y en el contexto del problema.

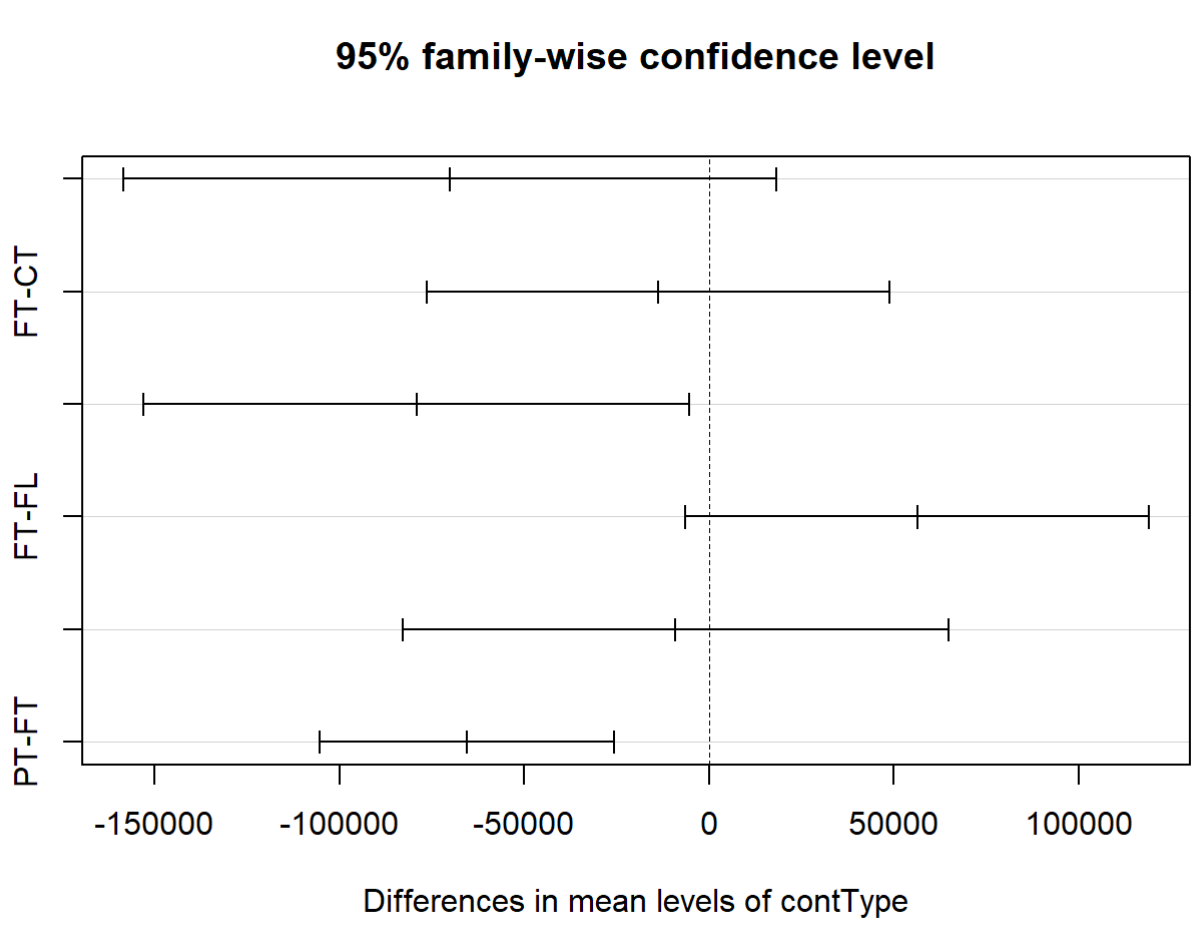
A partir del análisis de ANOVA, se puede ver que el valor F (Fisher-Snedecor) de la interacción entre compSiz-contType es muy cercano a uno además de que su valor-p es superior al valor de alpha (0.05) por lo que esta interacción no es estadísticamente significativa para determinar el salario promedio de los profesionistas de Datos. Esto significa que la intervarianza y la intravarianza de esta interacción son muy cercanas entre sí, por lo que no hay diferencia en el efecto. Al contrario, se puede ver que el valor F es muy elevado para el factor de expertise (superior 65) con un valor p menor a 0.05, lo que indica que este factor si es estadísticamente significativo para determinar el salario de un profesionista de datos, lo que indica que tanto la intervarianza como la intravarianza están muy separadas entre sí y que consecuentemente tienen distintos valores de medias entre cada agrupción de acuerdo al nivel de experiencia de cada profesionista. Adicionalmente, se puede ver que el factor de compSiz y contType también tienen valores de F superiores a 1 y que su valor p es menor a 0.05 lo cuál reafirma su significancia, respectivamente en ese orden.

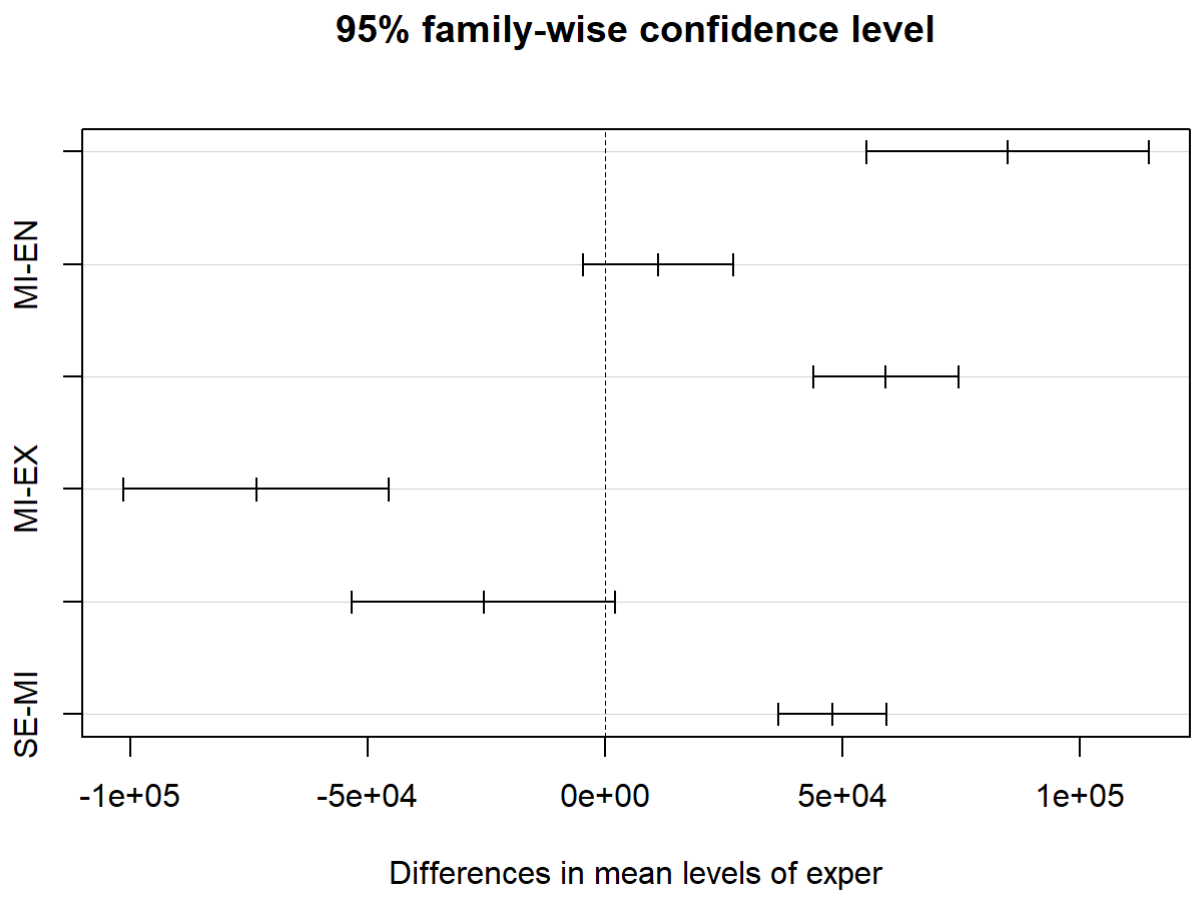
En lo que respecta al análisis de varianza, se puede ver claramente que el factor de expertise indica una mayor proporción de varianza que puede ser explicada por el modelo en comparación a los demás factores que presentan un nivel menor al 2%, en comparación a expertise que tiene una varianza del 24% lo que se traduce en un efecto pequeño. ## Realiza la prueba de comparaciones múltiples de Tukey. Grafica los intervalos de confianza de Tukey.

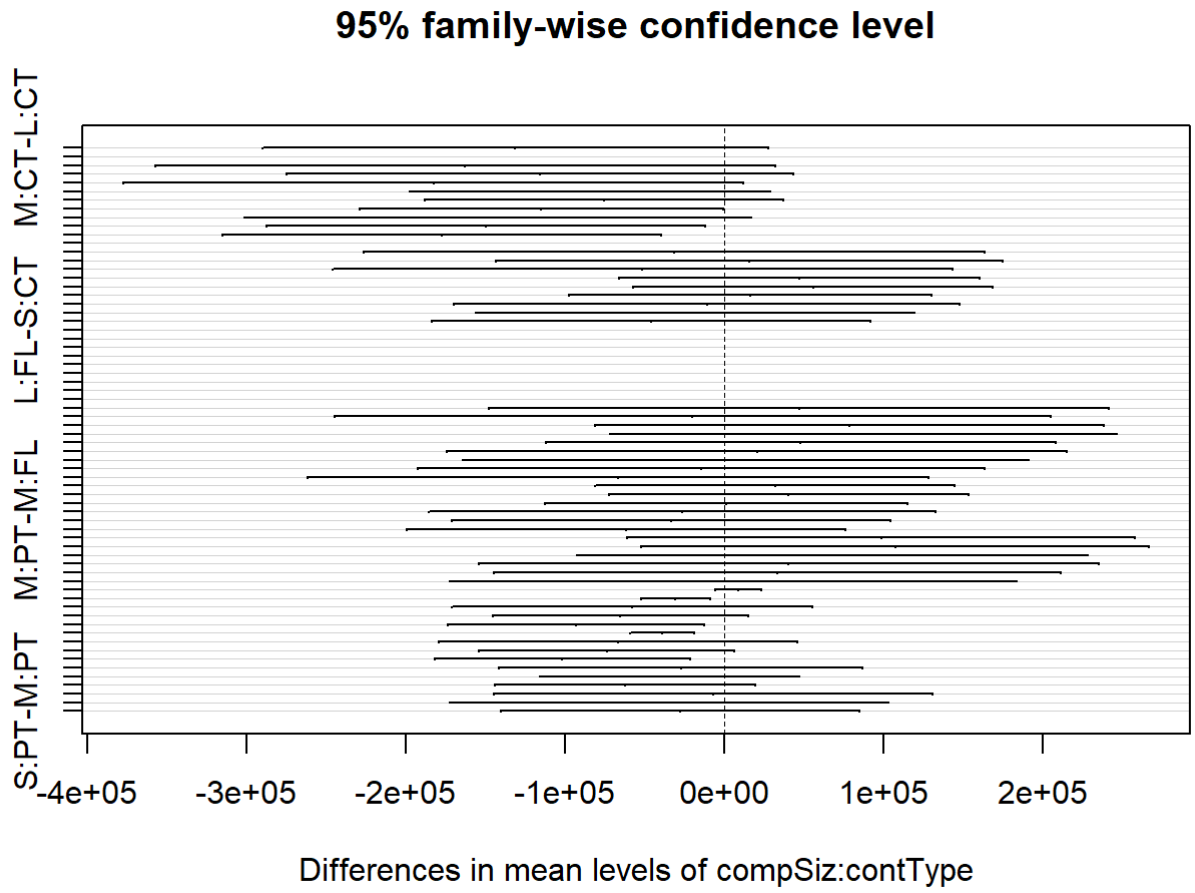
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sal ~ compSiz * contType + exper)
##
## $compSiz
##           diff        lwr       upr     p adj
## M-L   7305.687  -3086.113  17697.49 0.2248614
## S-L -34426.285 -49465.808 -19386.76 0.0000003
## S-M -41731.972 -55813.830 -27650.11 0.0000000
##
## $contType
##             diff         lwr        upr     p adj
## FL-CT -70112.179 -158434.297  18209.940 0.1727275
## FT-CT -13752.515  -76421.040  48916.010 0.9422850
## PT-CT -79147.167 -153042.753  -5251.581 0.0303319
## FT-FL  56359.664   -6308.861 119028.189 0.0953094
## PT-FL  -9034.989  -82930.574  64860.597 0.9891687
## PT-FT -65394.653 -105233.140 -25556.165 0.0001595
##
## $exper
##            diff         lwr       upr     p adj
## EX-EN  84696.75   54923.374 114470.13 0.0000000
## MI-EN  11141.24   -4720.164  27002.64 0.2697175
## SE-EN  59030.56   43746.116  74315.01 0.0000000
## MI-EX -73555.51 -101545.809 -45565.22 0.0000000
## SE-EX -25666.19  -53333.624   2001.24 0.0800865
## SE-MI  47889.32   36460.554  59318.09 0.0000000
##
## $`compSiz:contType`
##                    diff         lwr        upr     p adj
## M:CT-L:CT -131322.5702 -290405.769  27760.629 0.2244520
## S:CT-L:CT           NA          NA         NA        NA
## L:FL-L:CT -162825.5767 -357661.908  32010.755 0.2084109
## M:FL-L:CT -115943.9639 -275027.163  43139.235 0.4134012
## S:FL-L:CT -182806.6358 -377642.968  12029.696 0.0898625
## L:FT-L:CT  -84310.6061 -197402.576  28781.364 0.3765703
## M:FT-L:CT  -75636.0333 -188480.258  37208.191 0.5507061
## S:FT-L:CT -115000.7023 -228941.039  -1060.365 0.0455474
## L:PT-L:CT -142437.6966 -301520.895  16645.502 0.1304449
## M:PT-L:CT -149607.4477 -287377.539 -11837.356 0.0201742
## S:PT-L:CT -177447.5417 -315217.633 -39677.450 0.0016349
## S:CT-M:CT           NA          NA         NA        NA
## L:FL-M:CT  -31503.0065 -226339.338 163333.325 0.9999954
## M:FL-M:CT   15378.6063 -143704.592 174461.805 1.0000000
## S:FL-M:CT  -51484.0657 -246320.397 143352.266 0.9993753
## L:FT-M:CT   47011.9641  -66080.006 160103.934 0.9697229
## M:FT-M:CT   55686.5369  -57157.687 168530.761 0.9016261
## S:FT-M:CT   16321.8679  -97618.469 130262.205 0.9999987
## L:PT-M:CT  -11115.1264 -170198.325 147968.072 1.0000000
## M:PT-M:CT  -18284.8775 -156054.969 119485.214 0.9999994
## S:PT-M:CT  -46124.9716 -183895.063  91645.120 0.9946907
## L:FL-S:CT           NA          NA         NA        NA
## M:FL-S:CT           NA          NA         NA        NA
## S:FL-S:CT           NA          NA         NA        NA
## L:FT-S:CT           NA          NA         NA        NA
## M:FT-S:CT           NA          NA         NA        NA
## S:FT-S:CT           NA          NA         NA        NA
## L:PT-S:CT           NA          NA         NA        NA
## M:PT-S:CT           NA          NA         NA        NA
## S:PT-S:CT           NA          NA         NA        NA
## M:FL-L:FL   46881.6127 -147954.719 241717.944 0.9997452
## S:FL-L:FL  -19981.0592 -244958.676 204996.558 1.0000000
## L:FT-L:FL   78514.9706  -80995.298 238025.239 0.9031844
## M:FT-L:FL   87189.5434  -72145.170 246524.257 0.8200682
## S:FT-L:FL   47824.8744 -112288.000 207937.749 0.9980591
## L:PT-L:FL   20387.8801 -174448.452 215224.212 1.0000000
## M:PT-L:FL   13218.1290 -164642.294 191078.552 1.0000000
## S:PT-L:FL  -14621.9651 -192482.388 163238.458 1.0000000
## S:FL-M:FL  -66862.6719 -261699.004 127973.660 0.9934452
## L:FT-M:FL   31633.3578  -81458.612 144725.328 0.9989361
## M:FT-M:FL   40307.9307  -72536.294 153152.155 0.9908318
## S:FT-M:FL     943.2617 -112997.075 114883.599 1.0000000
## L:PT-M:FL  -26493.7327 -185576.931 132589.466 0.9999937
## M:PT-M:FL  -33663.4838 -171433.575 104106.608 0.9997043
## S:PT-M:FL  -61503.5778 -199273.669  76266.514 0.9493538
## L:FT-S:FL   98496.0298  -61014.239 258006.298 0.6752916
## M:FT-S:FL  107170.6026  -52164.111 266505.316 0.5450894
## S:FT-S:FL   67805.9336  -92306.941 227918.808 0.9652730
## L:PT-S:FL   40368.9392 -154467.392 235205.271 0.9999418
## M:PT-S:FL   33199.1882 -144661.235 211059.611 0.9999797
## S:PT-S:FL    5359.0941 -172501.329 183219.517 1.0000000
```

```
## M:FT-L:FT    8674.5728   -6027.418  23376.564 0.7359955
## S:FT-L:FT  -30690.0962  -52247.693  -9132.499 0.0002320
## L:PT-L:FT  -58127.0905 -171219.061  54964.880 0.8739261
## M:PT-L:FT  -65296.8416 -145689.177  15095.494 0.2464964
## S:PT-L:FT  -93136.9356 -173529.271 -12744.600 0.0086694
## S:FT-M:FT  -39364.6690  -59582.366 -19146.972 0.0000000
## L:PT-M:FT  -66801.6634 -179645.888  46042.561 0.7318246
## M:PT-M:FT  -73971.4144 -154014.858   6072.030 0.1019158
## S:PT-M:FT -101811.5085 -181854.953 -21768.064 0.0020321
## L:PT-S:FT  -27436.9943 -141377.331  86503.343 0.9997433
## M:PT-S:FT  -34606.7454 -116188.204  46974.713 0.9648443
## S:PT-S:FT  -62446.8395 -144028.298  19134.619 0.3345564
## M:PT-L:PT   -7169.7511 -144939.842 130600.340 1.0000000
## S:PT-L:PT  -35009.8451 -172779.936 102760.246 0.9995690
## S:PT-M:PT  -27840.0940 -140328.903  84648.715 0.9996655
```


95% family-wise confidence level
M-L
S-L
S-M
-40000
-20000
0
20000
Differences in mean levels of compSiz


95% family-wise confidence level
FT-CT
FT-FL
PT-FT
-150000
-100000
-50000
0
50000
100000
Differences in mean levels of contType


95% family-wise confidence level
MI-EN
MI-EX
SE-MI
-1e+05
-5e+04
0e+00
5e+04
1e+05
Differences in mean levels of exper



Interpreta el resultado desde la perspectiva estadística y en el contexto del problema. Tras observar la gráfica de turkey para el factor de compSiz, se puede ver que las diferencias entre medias en las que el intervalo de confianza que engloba los límites inferior y superior no contienen el valor de 0 y que son estadísticamente significativas fueron la de los grupos S-M y S-L, mientras que para el grupo M-L no fueron. Por lo que se puede inferir, que el salario promedio para las copañías grandes (L) y medianas (M) es equivalente.

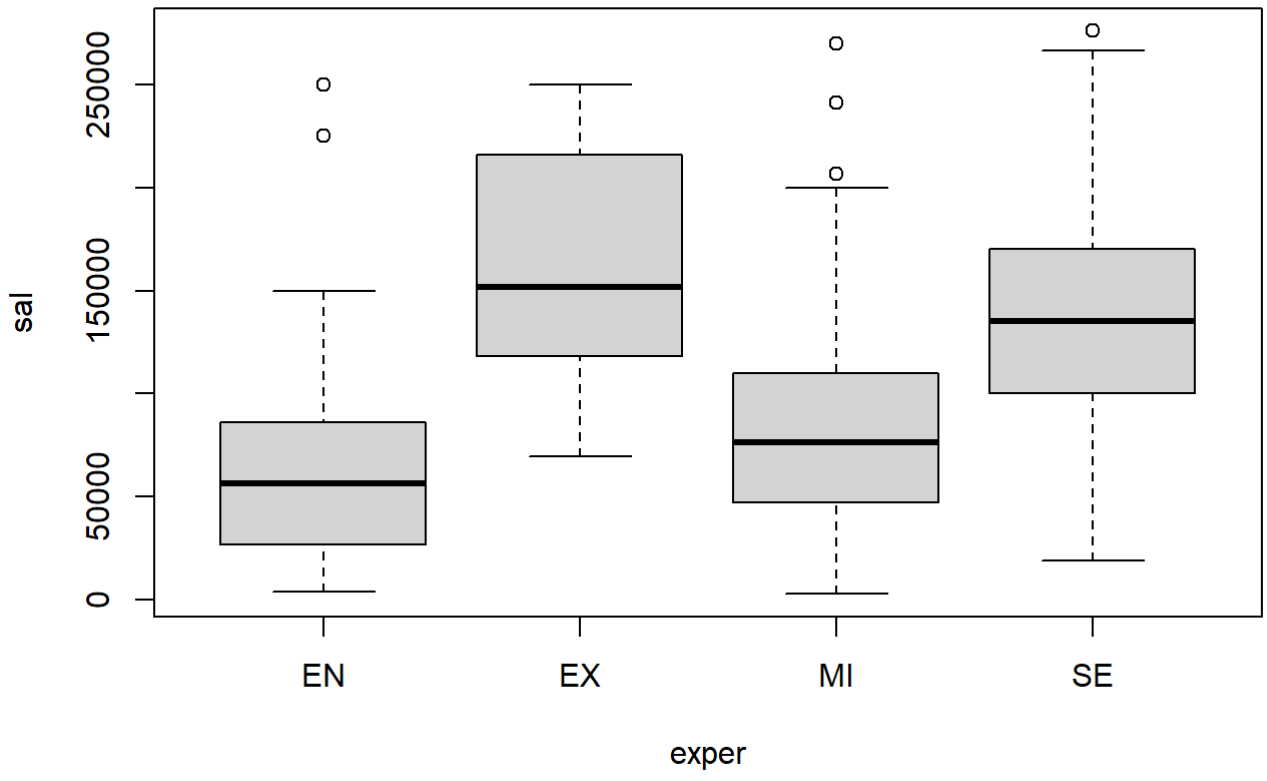
Por otro lado, la gráfia de turkey para el factor de contType indica que para casi todos los grupos por tipo de contrato, las medias no son estadísticamente significativas ya que todas pasan por 0 a excepción del grupo Part-Time (PT) - Full-Time (FT). Esto significa que el facto de tipo de contrato no es significativo para el nivel de salario promedio de un profesionista de datos.

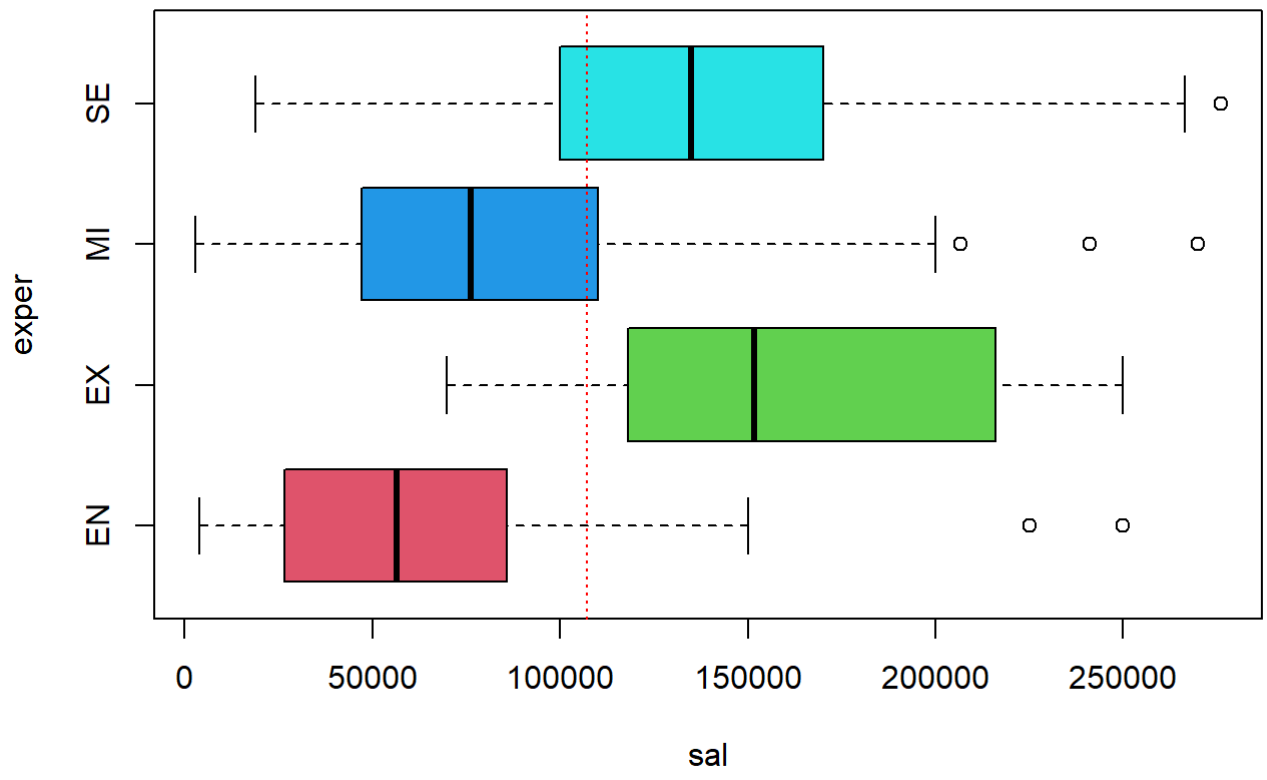
FInalmente, el gráfico de turkey para el tercer factor de nivel de expertise del profesional, muestra que para los grupos de MI-EN y SE-MI, la diferencia entre las dos medias no es estadísticamente significativa ya que ambas incluyen el valor de 0 en sus intervalos de confianza. Por lo que, el único grupo que tiene una diferencia significativa en sus medias es el de MI(Junior Mid-Level) - EX (Expert Level/Director).

Para observar mejor los efectos de los factores principales, se calcula la media por nivel y se grafica por nivel. También se calcula la media general.

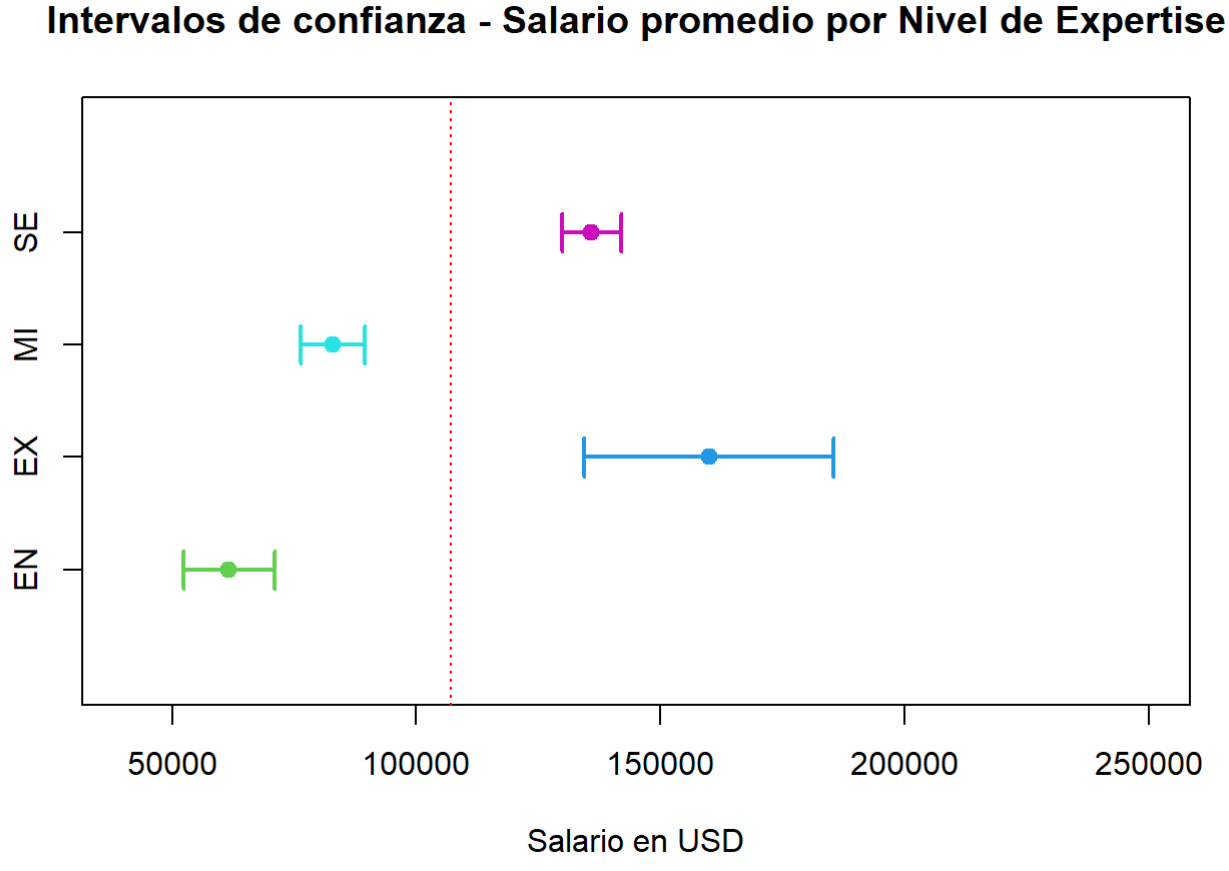
```
##        EN        EX        MI        SE
##  61643.32 159963.32  82953.14 135797.26
```

```
## [1] 107168.9
```





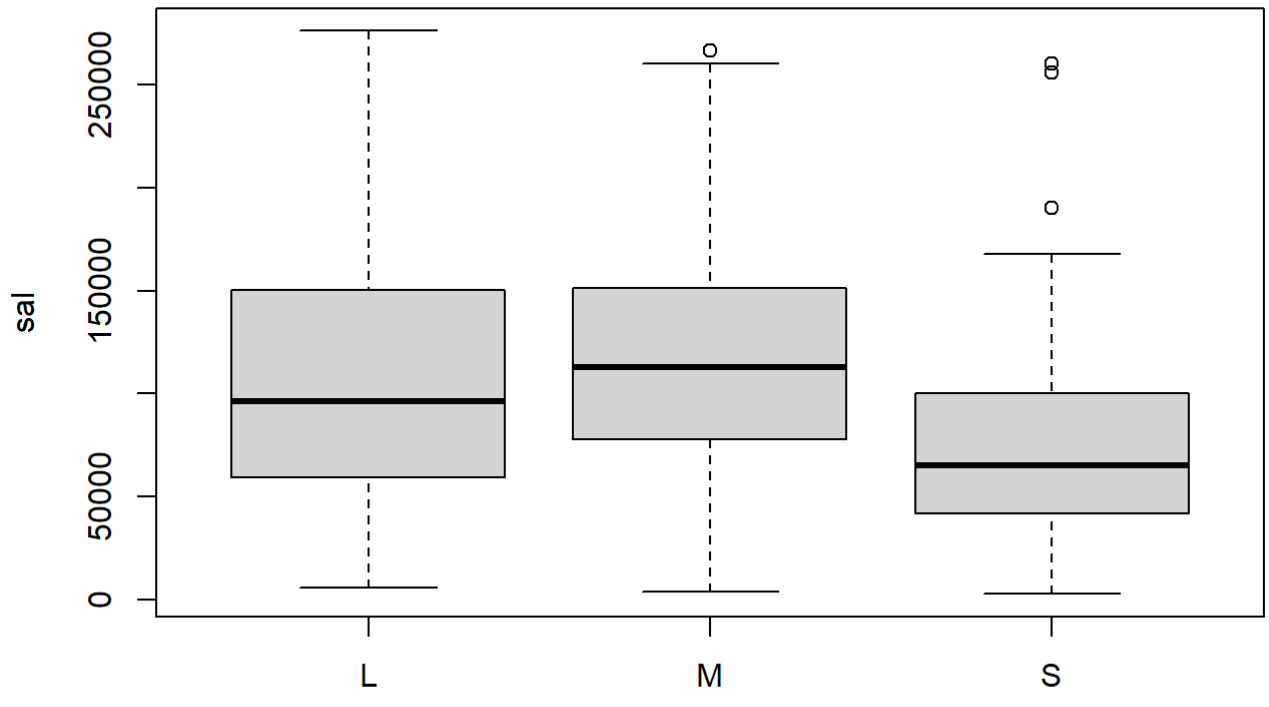
Intervalos de confianza - Salario promedio por Nivel de Expertise

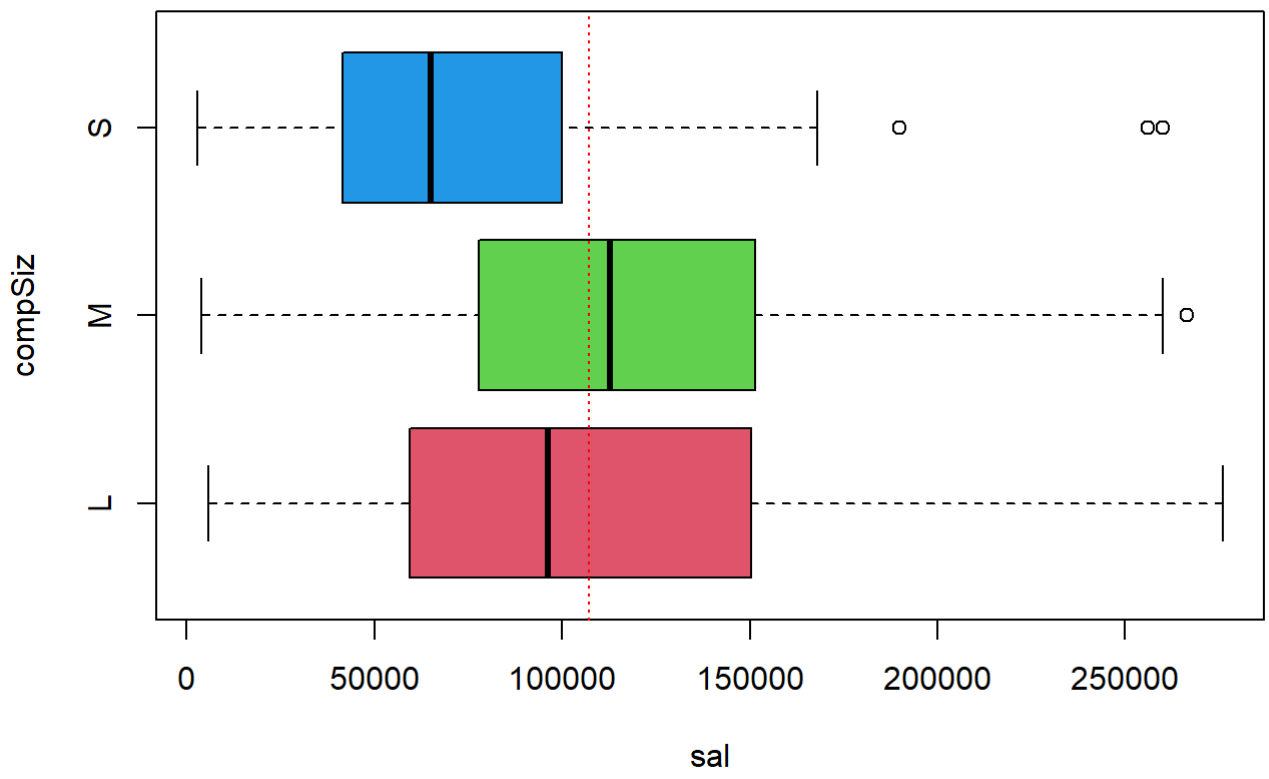


Interpreta el resultado desde la perspectiva estadística y en el contexto del problema. A través de los intervalos de confianza se puede ver que para todos los intervalos de acuerdo al nivel de expertise del profesionista, el rango de valores que abarcan son muy variados ya que cada nivel tiene un rango de valores distinto y no se translapan unos con otros, por lo que se puede inferir que el Nivel de expertise es un factor decisivo para determinar el salario promedio de un profesionista.

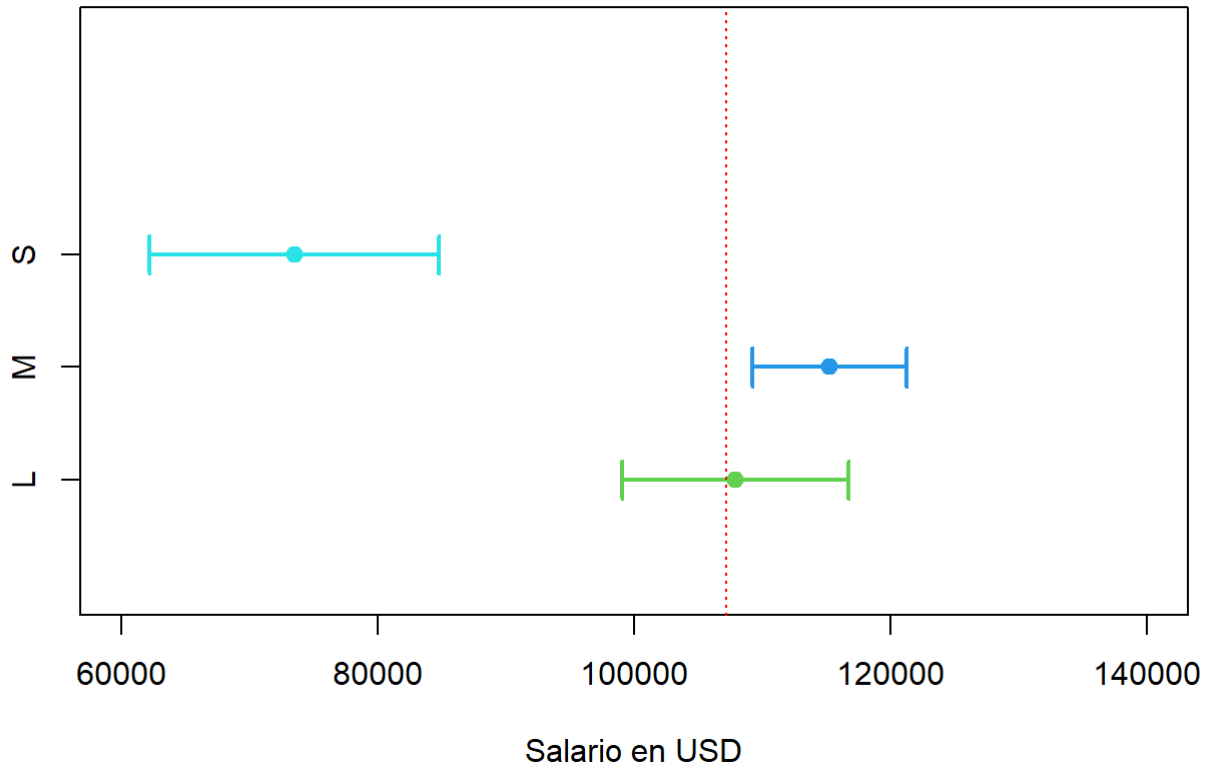
```
##         L         M         S
## 107932.53 115238.22  73506.24
```

```
## [1] 107168.9
```





Intervalos de confianza - Salario promedio por Tamaño de la Empresa

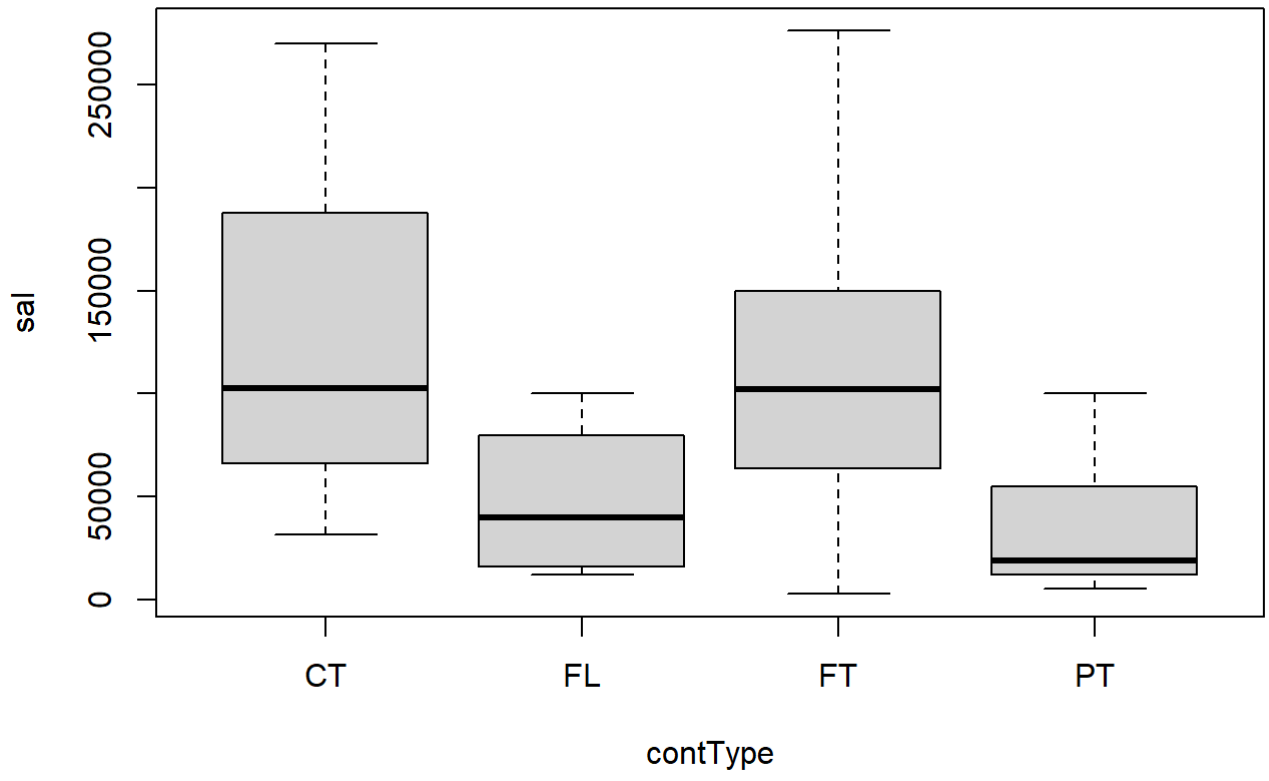


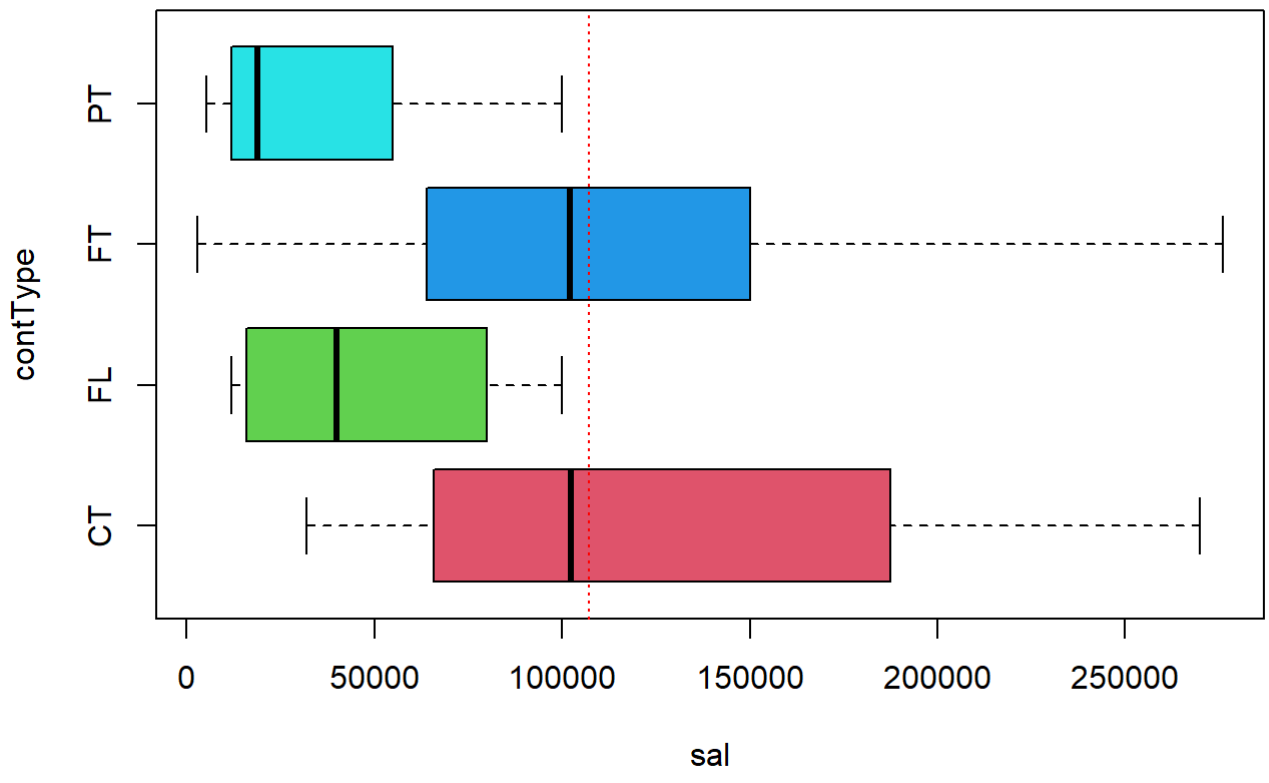
Interpreta el resultado desde la perspectiva estadística y en el contexto del problema. A través de los intervalos de confianza se puede ver que para todos los intervalos de acuerdo al tamaño de la empresa, el rango de valores que abarca cada tipo de empresa está muy definido para las empresas de tamaño chico (S), sin embargo, para las empresas de tamaño mediano (M) y de tamaño grande (L), estos intervalos se translapan por lo que abría que realizar

un análisis más profundo para verificar si comparten la misma media, por lo que se puede decir que el tamaño de la empresa es un factor decisivo para determinar el salario promedio de un profesionista en una distinción de empresas grandes o chicas. Pero, aparentemente entre empresas medianas y grandes parece ser que no hay una distincción muy marcada en la que ganen un mayor salario.

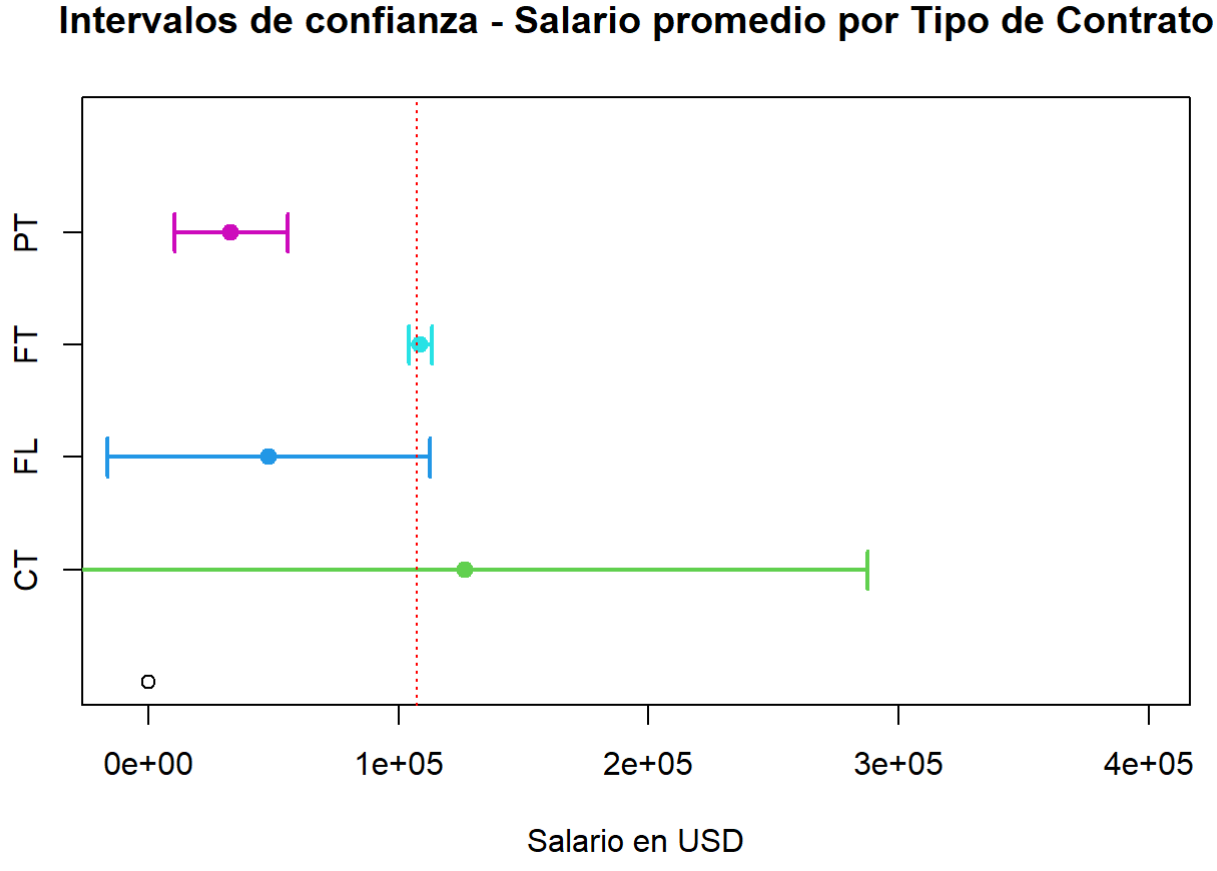
```
##       CT       FL       FT       PT
## 126718.8  48000.0 108722.3  33070.5
```

```
## [1] 107168.9
```





Intervalos de confianza - Salario promedio por Tipo de Contrato



Interpreta el resultado desde la perspectiva estadística y en el contexto del problema. A través de los intervalos de confianza se puede ver que para todos los intervalos de acuerdo al tipo de contrato, el rango de valores que abarcan son muy variados se translapan, por lo que se puede decir que el tipo de contrato no es un factor decisivo para determinar el salario promedio de un profesionista.El intervalo más grande corresponde al esquema de Contrato (CT) y

este abarca a los demás intervalos, siendo estos mucho más acotados. Si bien parece ser que existe una diferencia en la media entre el esquema de contrato (CT) con respecto a part-time (PT) y con respecto a Free-Lancer (FL), la media entre Contract (CT) y Full-Time (FT) están bastante cercanas.

Escribe tus conclusiones parciales

El efecto del Tipo de Contacto no es significativo ya que los valores de la media para los distintos tipos de contrato son muy similares, por lo que se descarta la hipótesis H2. En adición, se comprobó que la interacción entre el tamaño de la compañía y el tipo de contrato no es estadísticamente significativa para el salario promedio de un data-drive professional. Entonces, se descarta la hipótesis h4. Se reduce el modelo únicamente a las siguientes hipótesis: H0 = El nivel de experiencia no incide en el salario promedio de un data-oriented profesional. H1 = El tamaño de la compañía no incide en el salario promedio de un data-oriented profesional.

Realiza el segundo modelo de ANOVA

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## exper         3 5.939e+11 1.980e+11  82.661  < 2e-16 ***
## compSiz       2 3.432e+10 1.716e+10   7.165 0.000842 ***
## Residuals   591 1.415e+12 2.395e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##        EN        EX        MI        SE
##  61643.32 159963.32  82953.14 135797.26
```

```
## [1] 107168.9
```

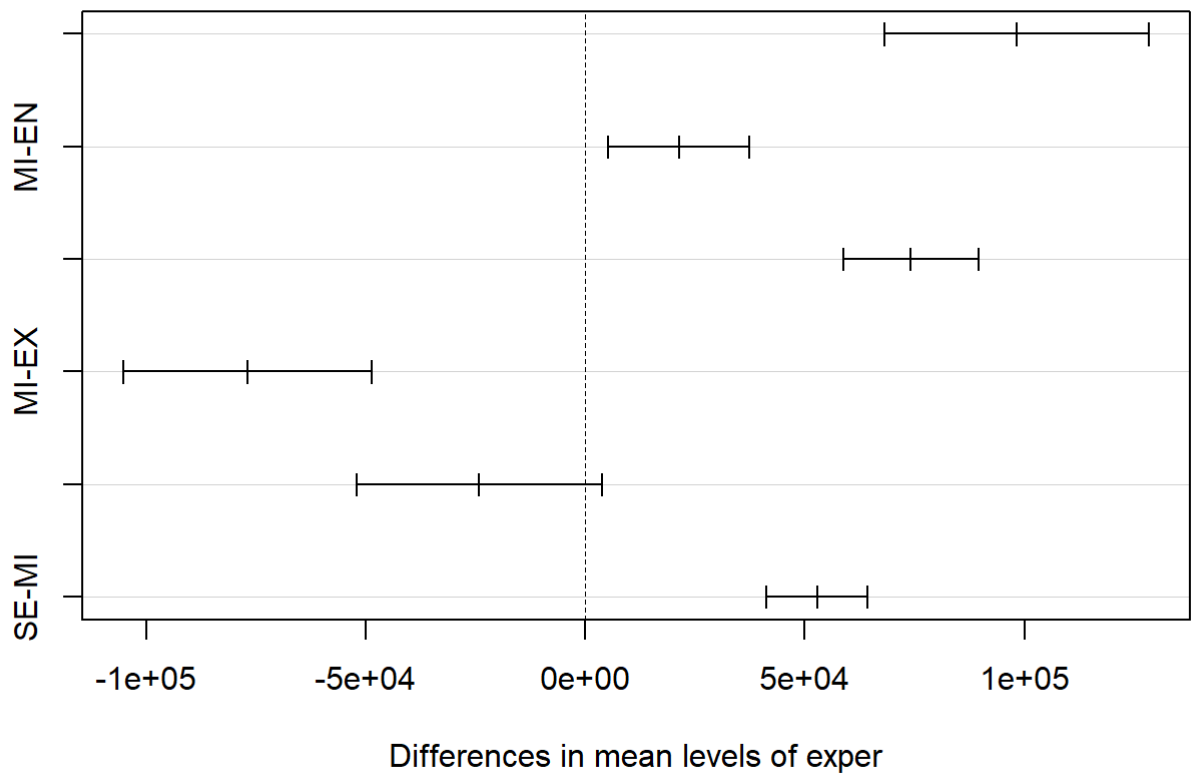
```
##             eta.sq eta.sq.part
## exper   0.25155787  0.26643807
## compSiz 0.01679447  0.02367459
```

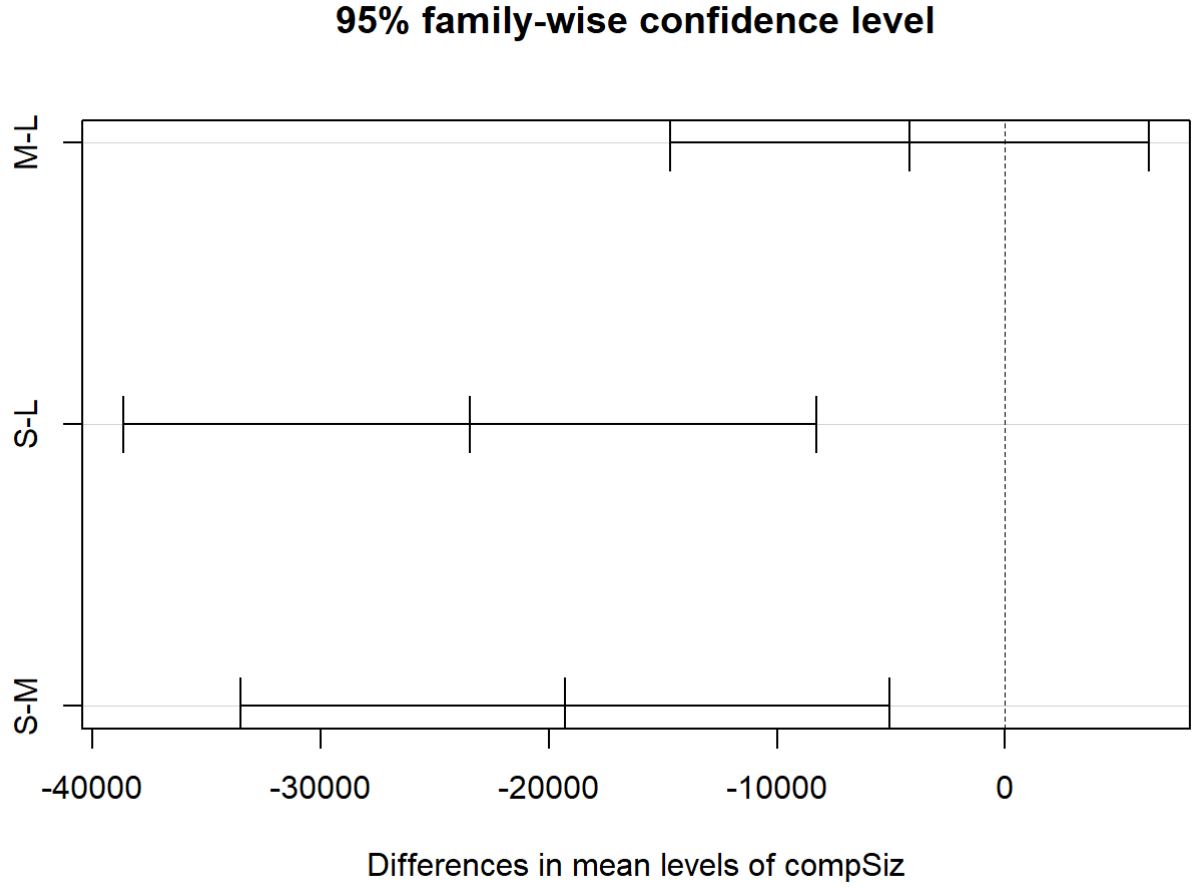
Interpreta el resultado desde la perspectiva estadística y en el contexto del problema.

Nuevamente, en este segundo análisis de ANOVA, se puede apreciar que el facto de expertise es estadísticamente muy significativo para predecir el salario de un profesionista de datos debido a su elevado valor de F que es inclusive más alto que el valor F del primer modelo, y el valor p sigue siendo menor a alpha. También se puede ver que el tamaño de la compañía es estadísticamente significativo para la variable a predecir ya que es superior por 6 puntos a 1 y su valor p es menor que alpha, pero en menor proporción al factor de expertise. Esto también se puede comprobar a partir del análisis de varianza donde el efecto de la variable de expertise es mucho mayor al efecto de compSize, por lo que se podría decir que el factor de expertise tiene una asociación más fuerte con la variable de salarios que el tamaño de la compañía.

Realiza la prueba de comparaciones múltiples de Tukey. Grafica los intervalos de confianza de Tukey.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sal ~ exper + compSiz)
##
## $exper
##            diff         lwr        upr     p adj
## EX-EN  98320.00   68267.140 128372.860 0.0000000
## MI-EN  21309.82    5299.531  37320.118 0.0036121
## SE-EN  74153.95   58726.026  89581.865 0.0000000
## MI-EX -77010.18 -105263.213 -48757.138 0.0000000
## SE-EX -24166.05  -52093.198   3761.089 0.1165192
## SE-MI  52844.12   41308.072  64380.170 0.0000000
##
## $compSiz
##          diff       lwr       upr     p adj
## M-L  -4188.65 -14678.04  6300.743 0.6162754
## S-L -23462.36 -38643.12 -8281.593 0.0008929
## S-M -19273.71 -33487.81 -5059.603 0.0043269
```





Interpreta el resultado desde la perspectiva estadística y en el contexto del problema. Tras observar la gráfica de turkey para el factor de exper, se puede ver que para todas las interacciones entre las agrupaciones, existen diferencias entre medias muy marcadas en las que el intervalo de confianza que engloba los límites inferior y superior no contienen el valor de 0 y que, por ende, son estadísticamente significativas. Por lo que se puede inferir, que el salario promedio para cada nivel de expertise es totalmente diferente y es estadísticamente significativo.

Por otro lado, la gráfica de compSiz muestra que las diferencias entre medias en las que el intervalo de confianza que engloba los límites inferior y superior no contienen el valor de 0 y que son estadísticamente significativas fueron la de los grupos S-M y S-L, mientras que para el grupo M-L no fueron. Por lo que se puede inferir, que el salario promedio para las copañías grandes (L) y medianas (M) es muy cercano.

Escribe tus conclusiones parciales

A través del análisis de ANOVA y de varianza, se comprueba estadísticamente que tanto el nivel de experiencia como el tamaño de la compañía son factores decisivos e influyentes sobre el salario promedio de un data-oriented professional debido a su efecto estadístico y que sus medias internas son diferentes entre sí, de modo que se procede a rechazar las siguientes hipótesis: H0 = El nivel de experiencia no incide en el salario promedio de un data-oriented professional. H1 = El tamaño de la compañía no incide en el salario promedio de un data-oriented professional.

Con respecto a los intervalos de confianza con la prueba de comparaciones múltiples de Tukey a un 95% de confianza, se puede observar que todos los intervalos comprenden un rango de valores menores a 0, lo que indica: * Las compañías M pagan igual que las compañías L. * Las compañías L pagan más que las compañías S. * M1 es mejor que M2

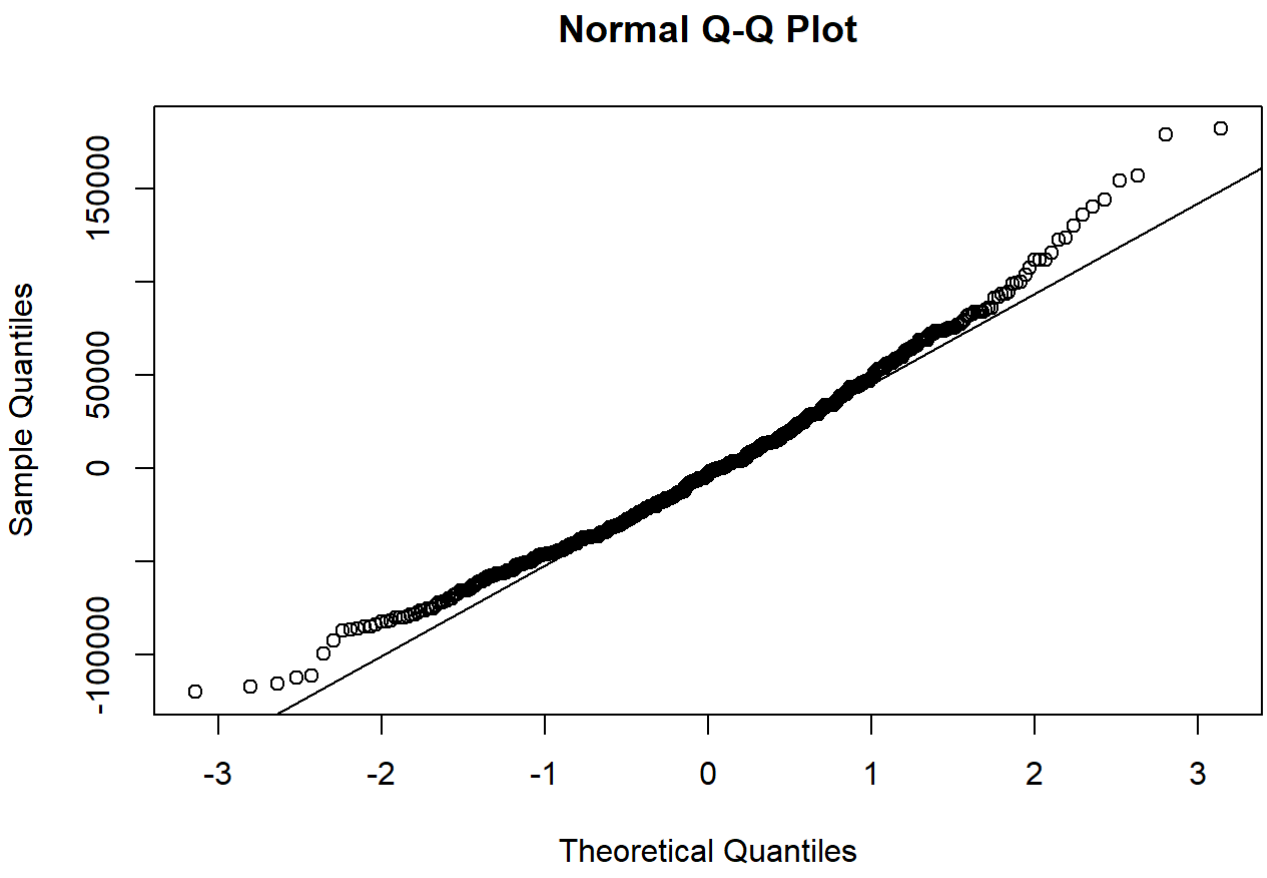
Por ende, M = L > S

- Los profesionales expertos (EX) son los mayormente remunerados.
- Los profesionales intermedios (SE) son mejor remunerados que los MID y EN pero menor remunerados que los expertos (EX).
- Los profesionales medios (MI) son mejor remunerados que los EN pero menores a los SE y los EX.
- Los profesionales entry-level (EN) son los menormente remunerados.

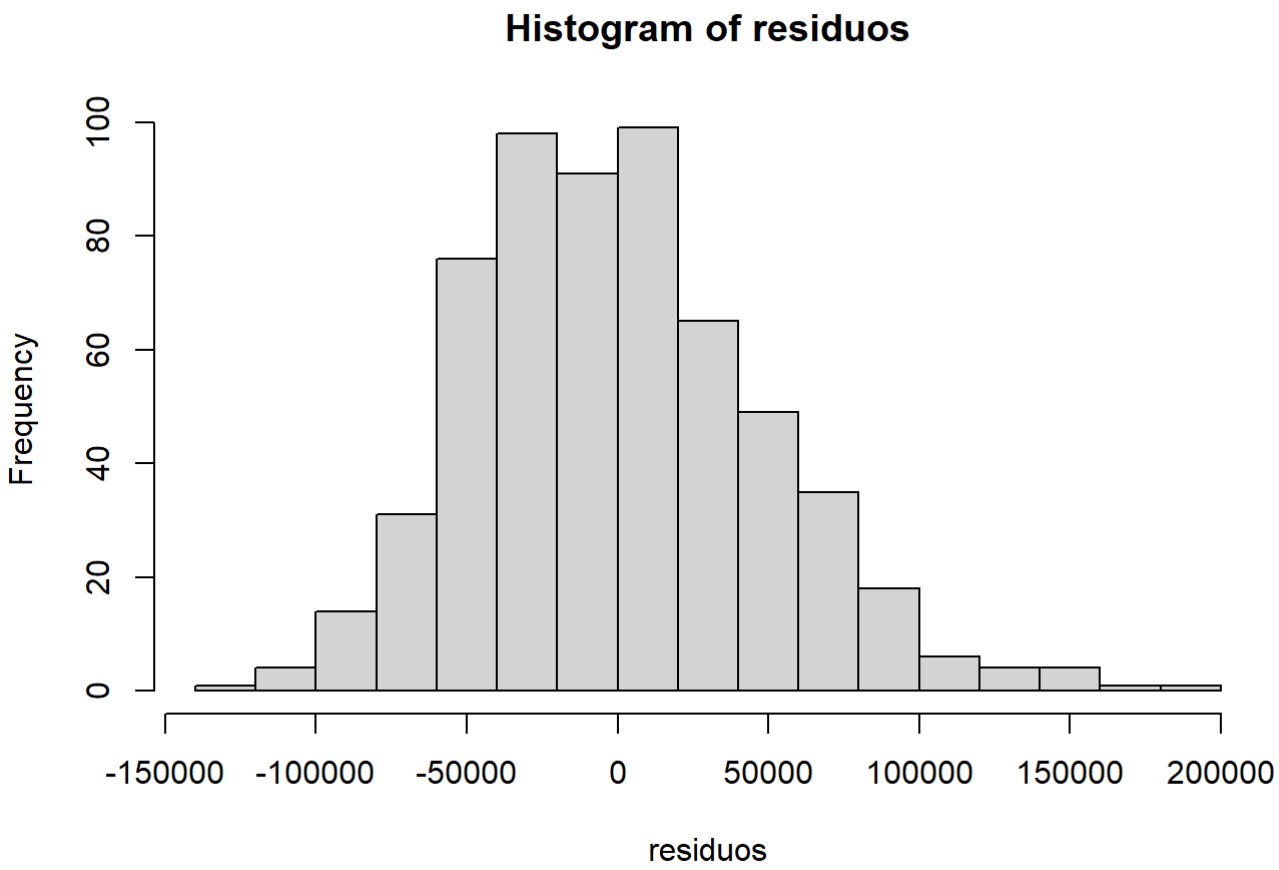
Por ende, EX > SE > MI >EN

En conclusión, se ha comprobado que el factor de nivel de expertise y de tamaño de la compañía sí son determinantes e influyen directamente en el salario promedio de los profesionales de datos. Dicho de otra forma, existe un efecto significativo del nivel de expertise y el tamaño de la compañía sobre el salario promedio al que puede aspirar un profesional de datos, independientemente del tipo de contrato que tengan. #

Comprobación de la validez del modelo. ## Normalidad

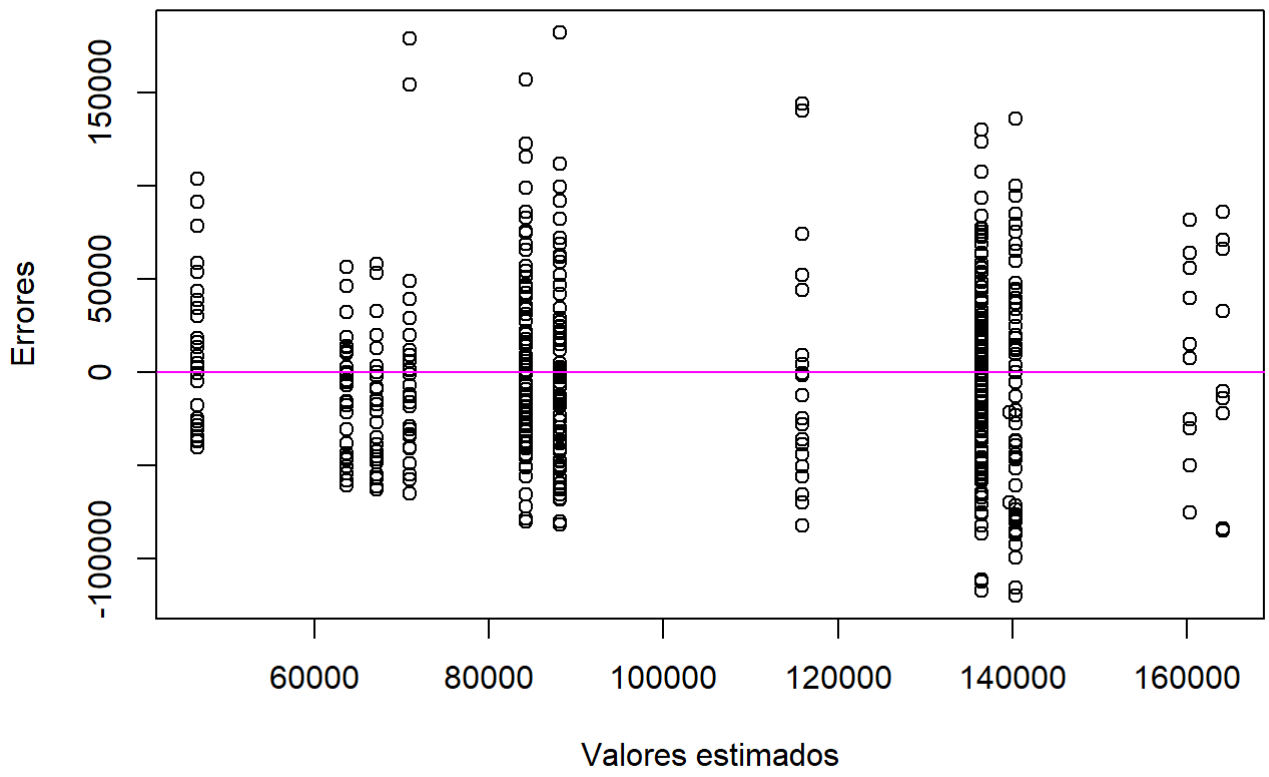


El comportamiento de esta gráfica de probabildad normal presenta una distribución con colas suaves la cual posee una alta curtosis en forma de una distribución Leptocúrtica. Los residuos en efecto, se comportan como una distribución normal ya que se ajustan casi perfectamente a una línea recta y tienen una tendencia creciente. De igual manera, se procedió a graficar un histograma de frecuencias para observar la distribución de la data.



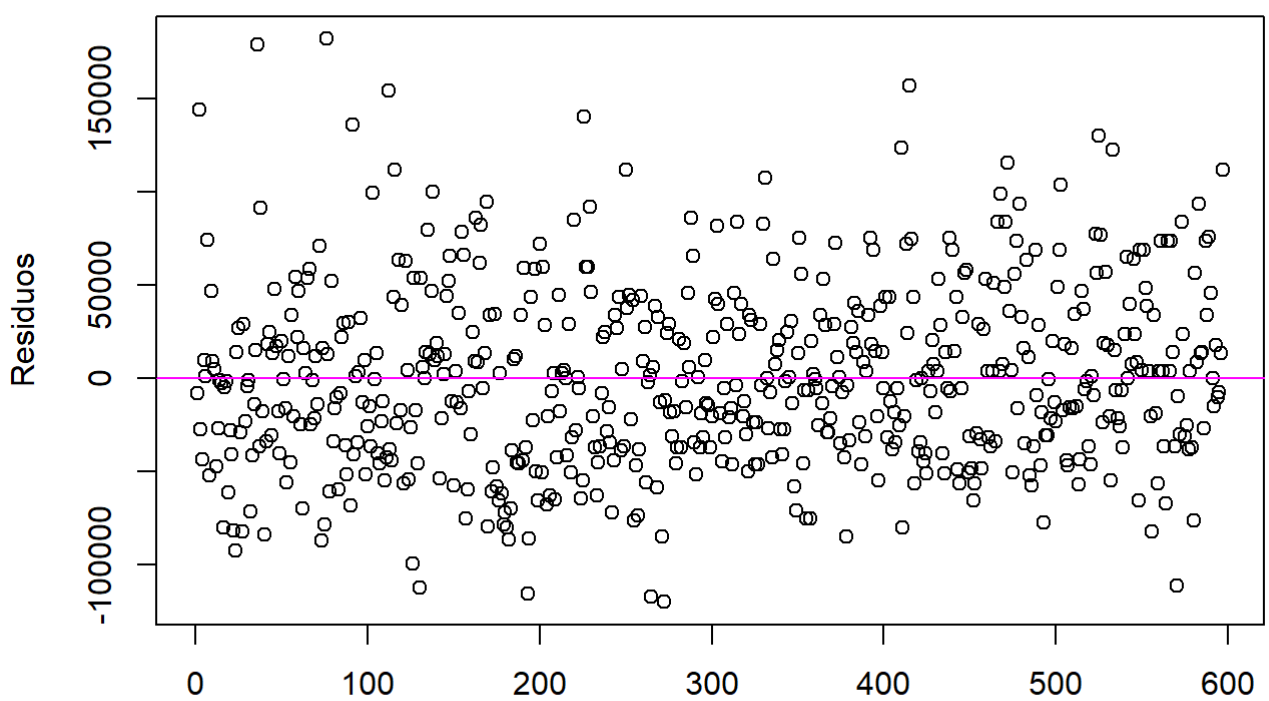
En este histograma, se puede verificar claramente que la mayor agrupación de los datos está en el centro de la distribución y la menor proporción de los datos se encuentra en los extremos, por ende, se asemeja casi perfectamente a una distribución normal ya que no presenta una simetría perfecta con repecto a la media de la distribución.

Homocedasticidad

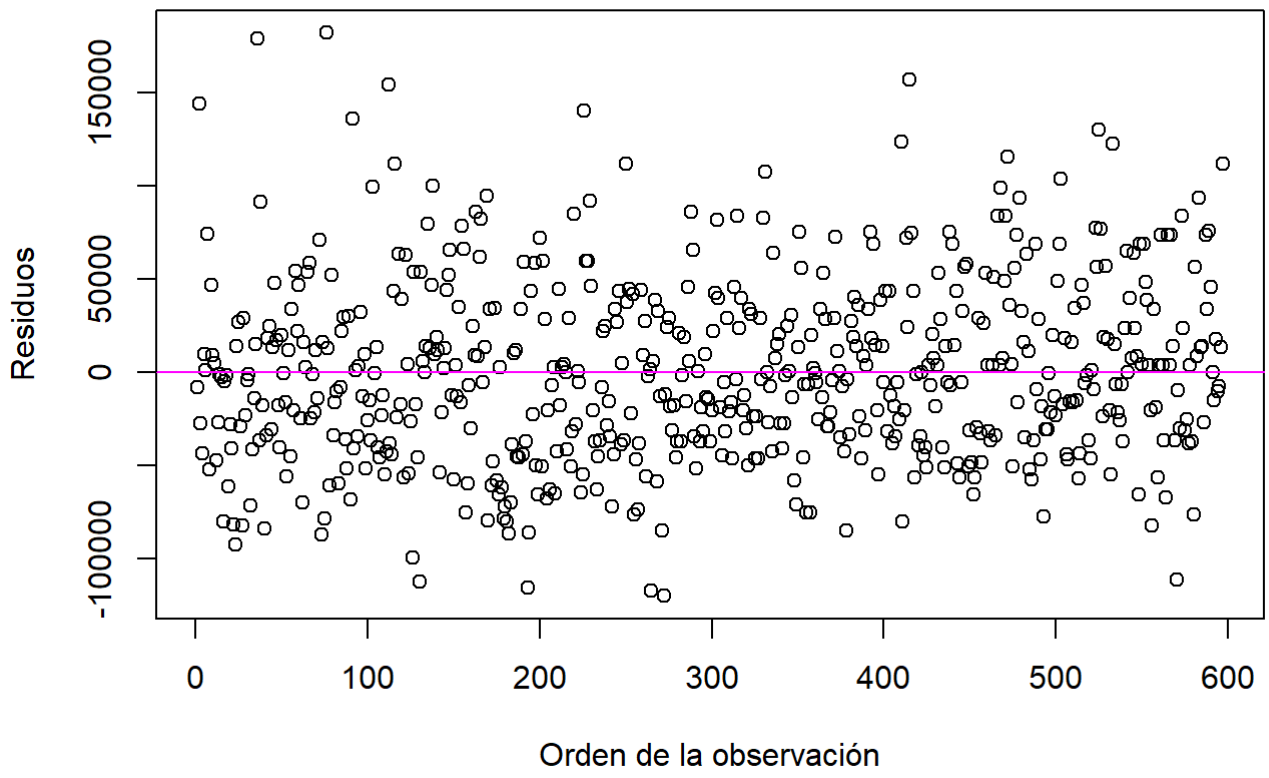


Con respecto a la homocedasticidad, se puede ver a través de la gráfica que los residuos presentan una dispersión constante ya que cada grupo tiene la misma distancia con respecto al error, y su variabilidad es constante para cada valor de x. Adicionalmente, se aprecia que la media de los errores fue de 0, por lo que sí tienen una distribución Normal.

Independencia

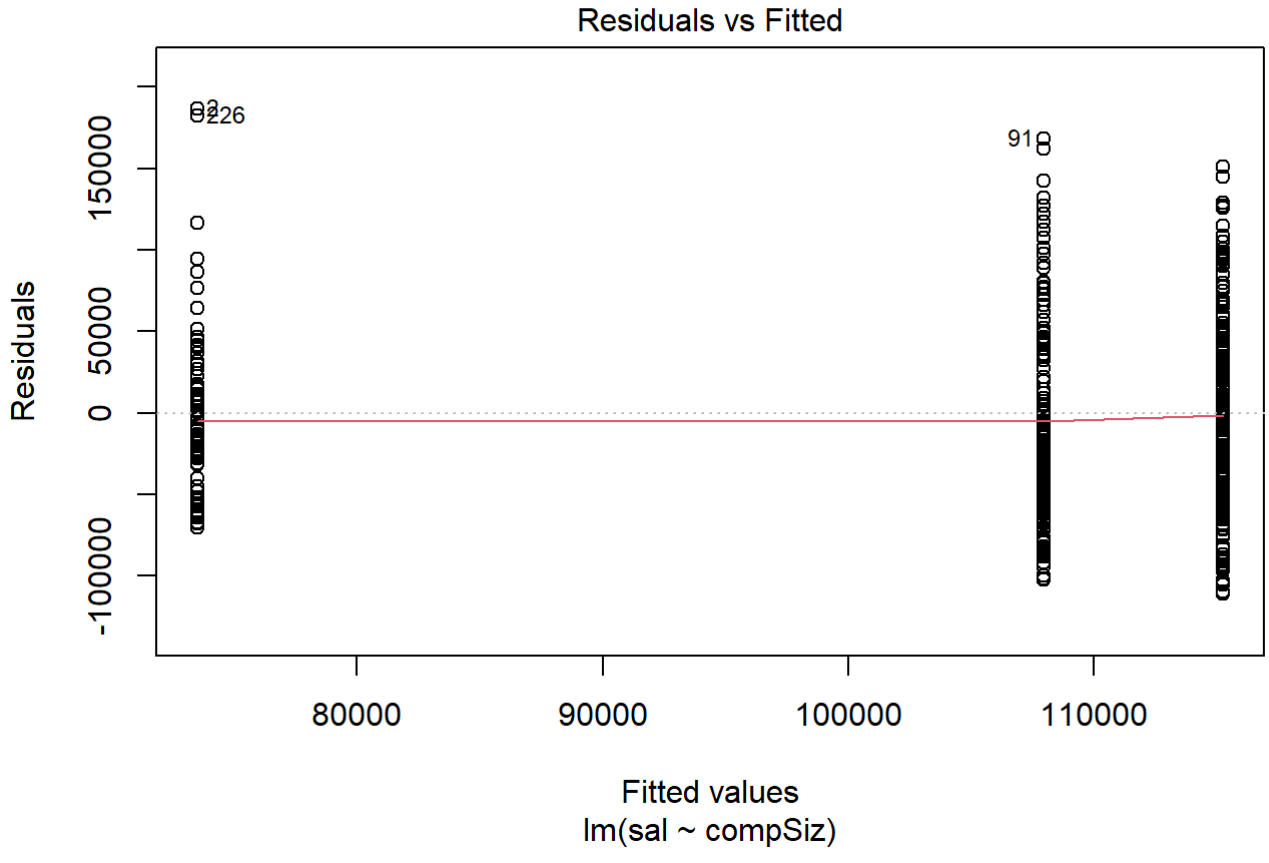


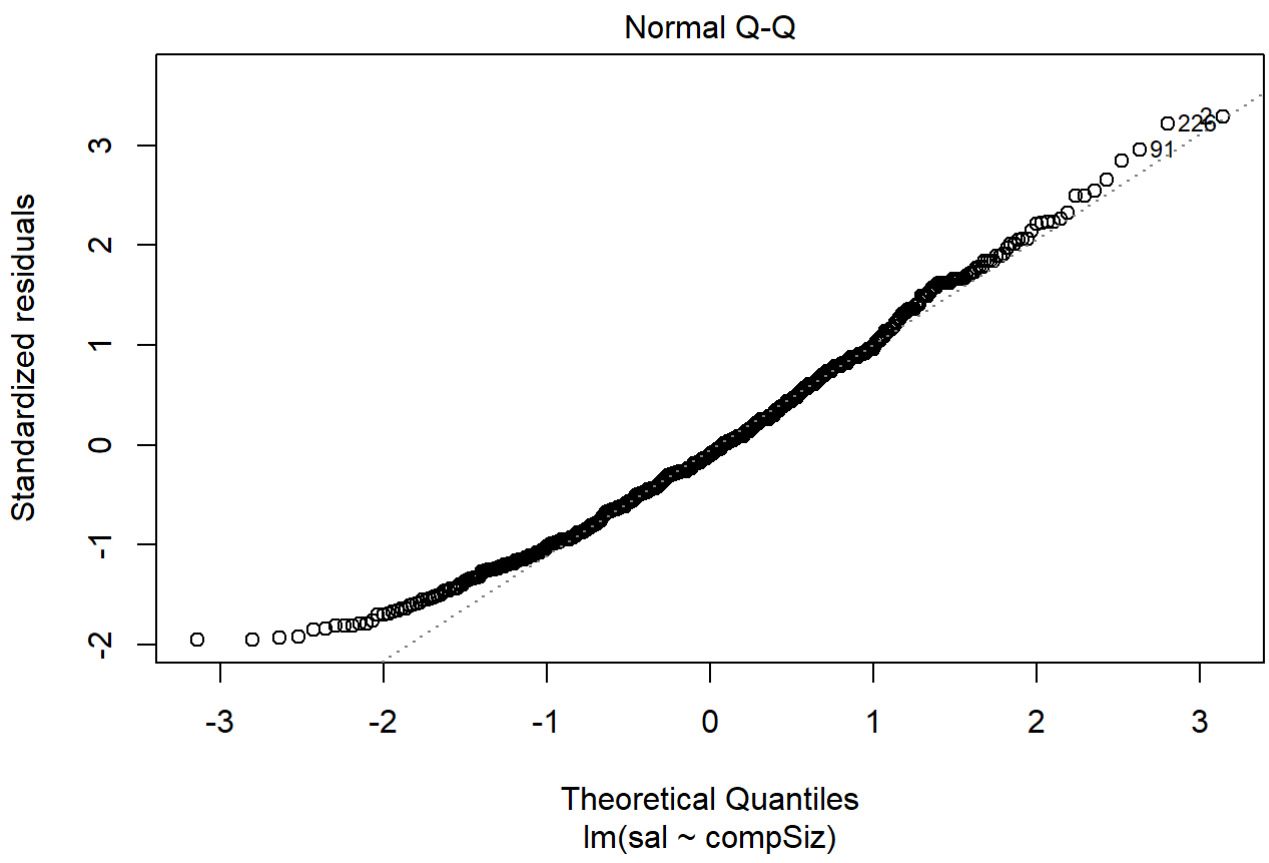
En efecto, se puede ver que no existe una tendencia clara en el comportamiento de los residuos, por lo que se puede decir que son independientes del orden de las observaciones. Es decir, no existe una correlación determinada entre ellos, lo que asegura su relación de independencia.

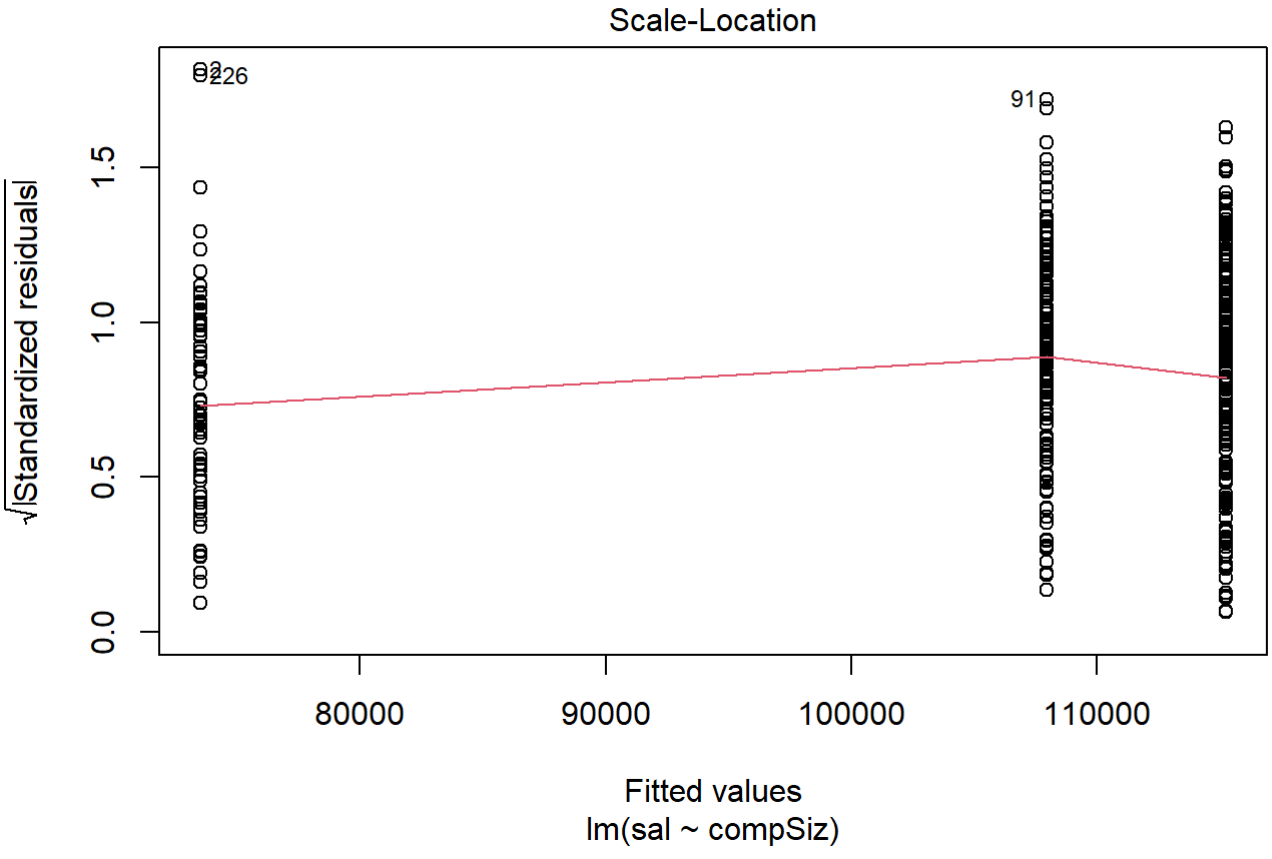


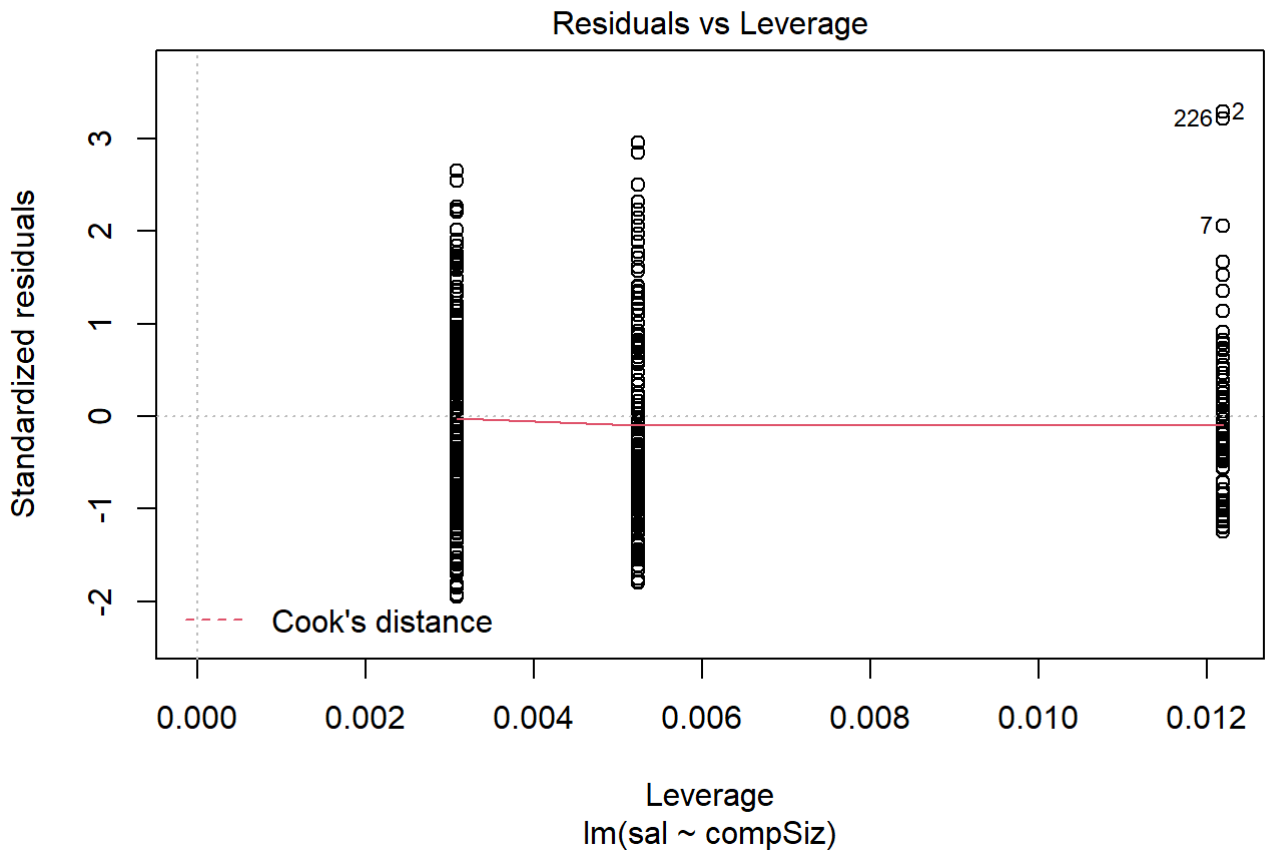
En efecto, se puede ver que no existe una tendencia clara en el comportamiento de los residuos, por lo que se puede decir que son independientes del orden de las observaciones. Es decir, no existe una correlación determinada entre ellos, lo que asegura su relación de independencia.

Relación lineal entre las variables (coeficiente de determinación).


Residuals vs Fitted
Residuals
150000
50000
0
-100000
2
226
91
80000
90000
100000
110000
Fitted values
lm(sal ~ compSiz)

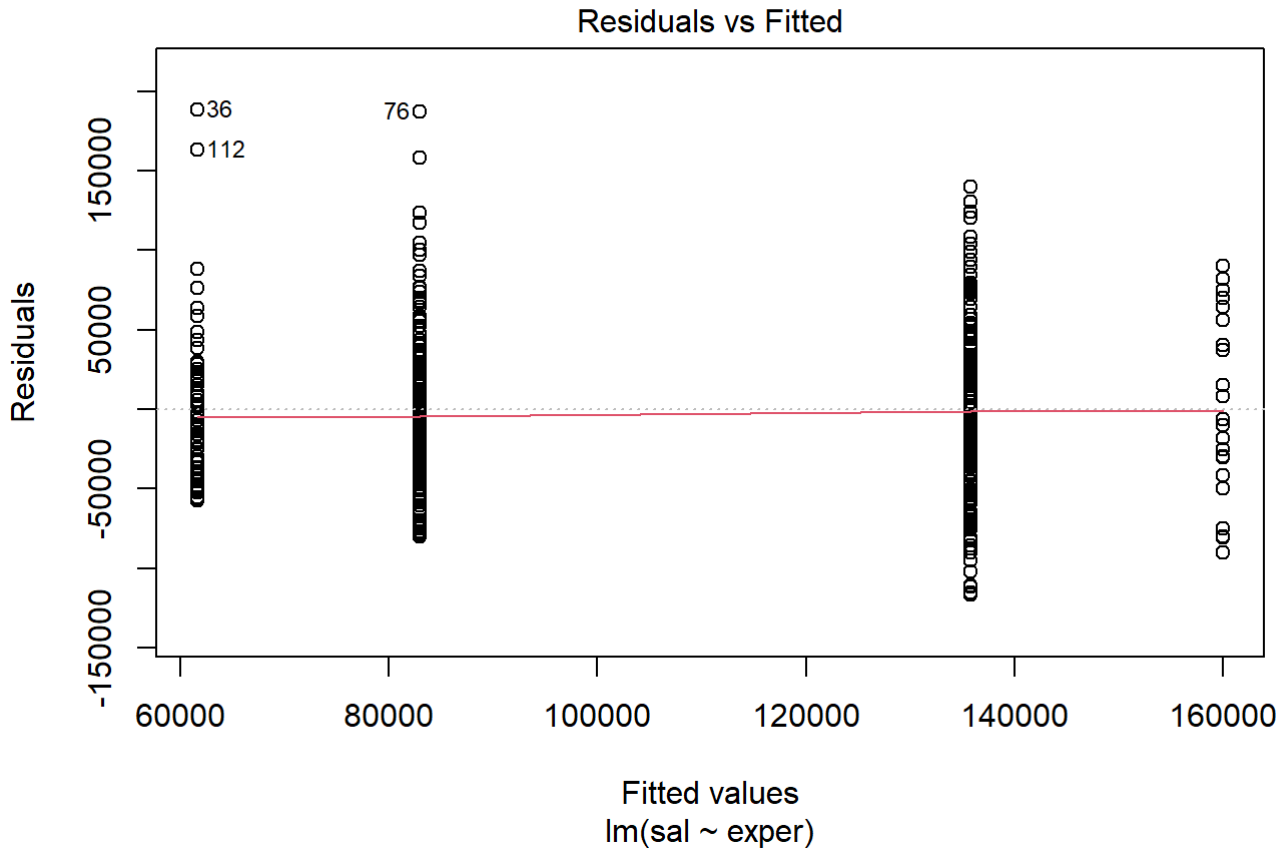

Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
2
226
91
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(sal ~ compSiz)

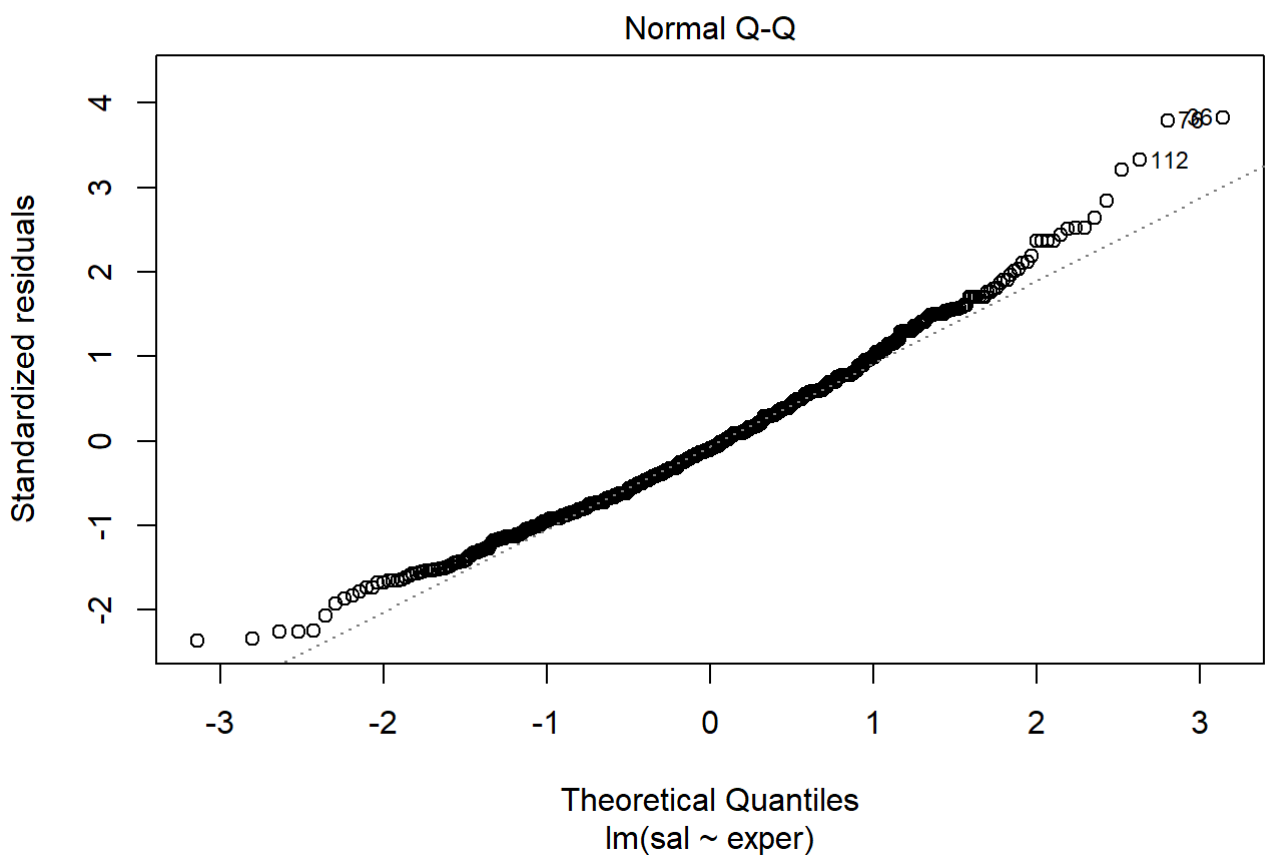

Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
2
226
91
80000
90000
100000
110000
Fitted values
lm(sal ~ compSiz)

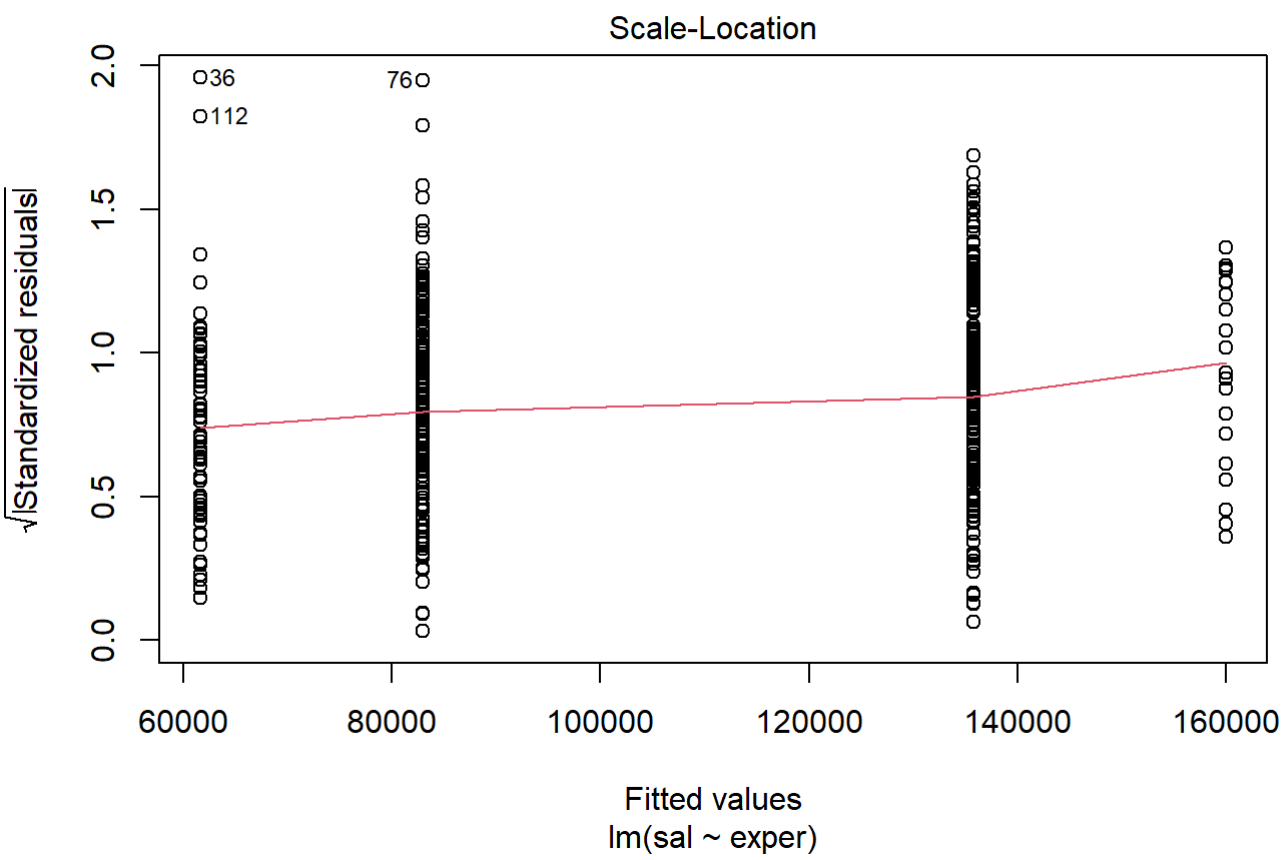


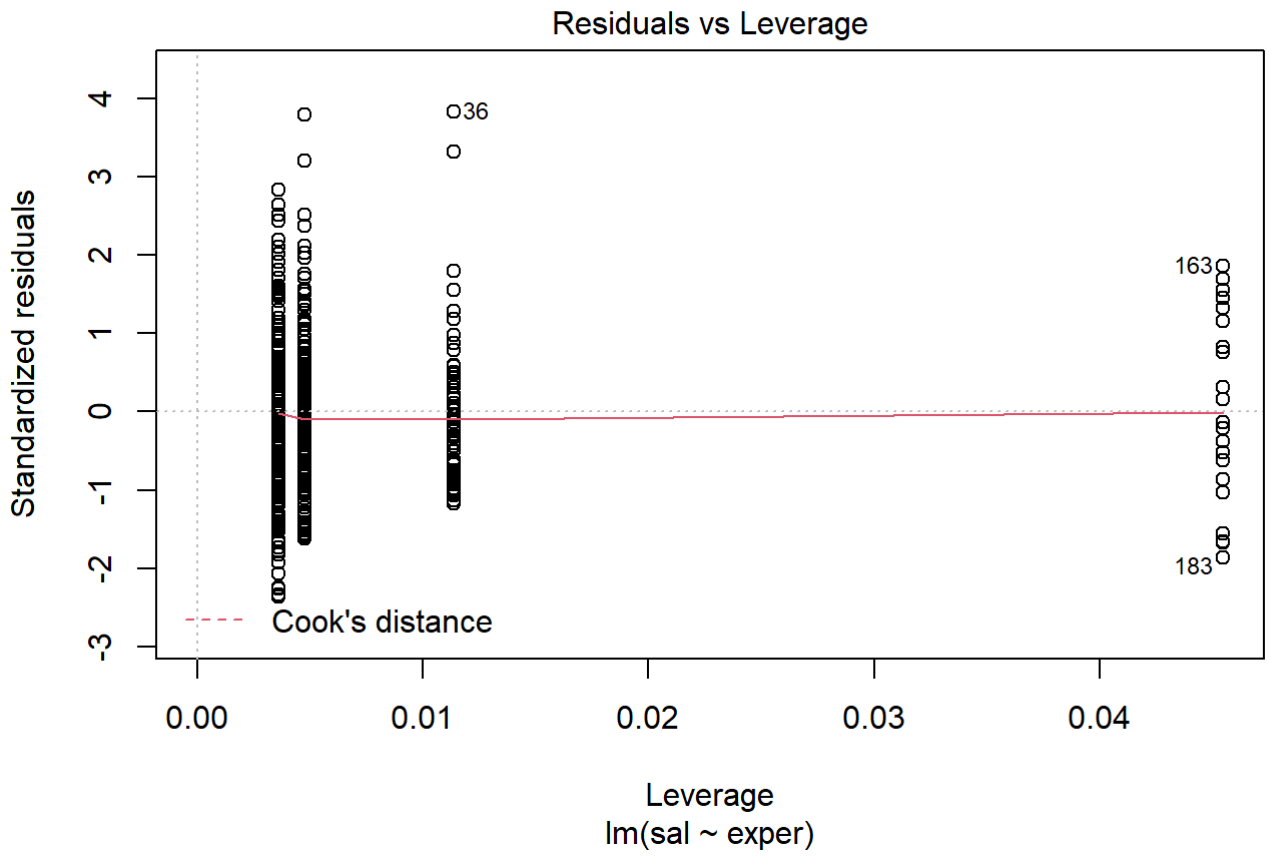

```
## [1] "El coeficiente de Determinación es:  0.0236800706538239"
```

A partir de las gráficas anteriores, se puede observar que existe una correlación entre el tamaño de la compañía y el nivel de salario promedio de un profesionista de datos por el comportamiento creciente de los residuos estandarizados en relación a los cuartiles teóricos. De igual manera, el cálculo de coeficiente de correlación indica una correlación positiva entre ambas variables de 2.37%.


Residuals vs Fitted
Residuals
150000
50000
-50000
-150000
36
112
76
60000
80000
100000
120000
140000
160000
Fitted values
lm(sal ~ exper)


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
76
66
112
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(sal ~ exper)


Scale-Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
36
112
76
60000
80000
100000
120000
140000
160000
Fitted values
lm(sal ~ exper)




```
## [1] "El coeficiente de Determinación es:  0.295634426800737"
```

A partir de las gráficas anteriores, se puede observar que existe una correlación entre el tamaño de la compañía y el nivel de salario promedio de un profesionista de datos por el comportamiento creciente de los residuos estandarizados en relación a los cuartiles teóricos. De igual manera, el cálculo de coeficiente de correlación indica una correlación positiva entre ambas variables de 29.56%.

Conclusión final en el contexto del problema.

En síntesis, cabe recalcar que los efecto perteneciente del factor de tamaño de compañía y el nivel de expertise fueron significativos para la determinación del salario promedio de un profesinal de datos, independientemente del tipo de contrato que tengan. En segunda instancia, se halló que los 4 niveles de experiencia en cuestión producen efectos diferentes en el salario promedio que puede llegar a tener un profesional de datos. Por un lado, el nivel de expertise más alto (EX) resultó ser el más remunerado en base a la media del salario promedio en dólares de la muestra ya que su salario fue significativamente mayor al de la media general. Por otro lado, el nivel de experiencia de Entre-Level resultó ser el menos remunerado ya que su media fue significativamente menor a la media general.

En efecto, se encontró que los profesionistas que trabajan en compañías grandes (L) y medianas (M) reciben en promedio casi el mismo salario y este a su vez, es mayor que el salario de los profesionistas que laboran en compañías pequeñas (S), independientemente del tipo de contrato que tengan.

En otro aspecto, el tipo de contrato no presentó un efecto concreto ya que es indiferente ante el salario promedio de los profesionistas de datos debido a que sus valores de medias fueron muy cercanos entre sí. Finalmente, se demostró que el primer método incrementó el rendimiento de los estudiantes con respecto a la media general, por lo que se puede decir que es el mejor método en términos de eficiencia que aporta al rendimiento de los estudiantes.

Por otra parte, se ha encontrado que el modelo propuesto es capaz de explicar (en conjunto nivel de expertise y tipo de compañía) el 32% de la variación, de modo que el tamaño de la compañía y el nivel de expertise si son factores determinante sobre el salario promedio de los profesionales de datos. No obstante, claramente la combinación de estos factores no es capaz de explicar la mitad de la muestra, por lo que existen otros factores externos (que no se estén considerando dentro del análisis) que también tiene la mayor incidencia en el salario promedio de los profesionistas de datos en el porcentaje de variación restante (exactamente el 68% de la muestra).

A partir de la interpretación de los gráficos Q-Q y los residuos vs. el valor esperado, los datos sí cumplen con las características de normalidad e independencia lo cuál sustenta la validación del modelo propuesto.

Por ende, se rechaza las primera hipótesis de H0 y H1.