# Reporte: los salarios de los científicos de datos

Antonio David Gutiérrez Páez, Arnold Torres Maldonado

2023-05-03

## Resumen (Abstract)

Los objetivos de esta investigación son conocer los paises donde los cientificos de datos tienen más ingresos económicos por el trabajo que desempeñan y predecir si a futuro en estos paises se seguirá percibiendo ese nivel de ingresos. Estas incógnitas se resuelven usando el modelo de minería de datos CRISP-DM para crear un modelo de regresión lineal, el cual arroja que los X paises donde las profesiones mencionadas son más pagadas son X,Y y Z; también se da a conocer que estos paises seguirán siendo en los que mayor salario percibe un científico de datos.

## Introducción

La ciencia de datos es una disciplina que ha estado a la alza recientemente con los avances tecnológicos y con la cantidad de datos que se producen todos los días, siendo actualmente una disciplina indispensable para el análisis y tratamiento de datos en masa. Debido a esto, la relación entre el salario percibido para esta disciplina contra el año en el que una persona se dedicaba a esto va en subida, siendo el último año uno de los mejores pagados para esta profesión, esto también debido a la alta demanda de procesamiento de datos y la poca disponibilidad de personas profesionales que se dediquen a esta disciplina. Por estas razones se plantea generar un modelo que explique la relación entre variables como los años de experiencia y el país donde el trabajador reside, para que a partir de este se puedan generar conclusiones que sirvan para que una persona que quiera dedicarse a esta rama de las ciencias de la computación tome desiciones informadas.

## Marco teorico

La ciencia de datos es un campo multidisciplinario que combina conceptos y técnicas de matemáticas, estadísticas, informática y dominios específicos para abordar problemas relacionados con el manejo y análisis de datos. Implica la recopilación, limpieza, procesamiento, análisis y visualización de datos con el objetivo de obtener información valiosa y útil para la toma de decisiones. (InLab FIB, s.f.). Así pues, los científicos de datos son profesionales especializados en la recopilación, análisis y interpretación de grandes volúmenes de datos complejos extraen conocimientos para la toma de decisiones informadas, pueden descubrir patrones, identificar tendencias y realizar predicciones. Algunas de las razones por las que estos cobraron tanta importancia son:

- La toma de decisiones basadas en datos: Ayudan a las organizaciones a tomar decisiones fundamentadas en información cuantitativa y basada en evidencia, en lugar de depender únicamente de intuiciones o suposiciones.
- La mejora de la eficiencia y la productividad: Al analizar grandes cantidades de datos, los científicos de datos pueden identificar patrones y tendencias que pueden ayudar a mejorar la eficiencia operativa, optimizar procesos y aumentar la productividad.
- La innovación y ventaja competitiva: La ciencia de datos permite descubrir ideas innovadoras y soluciones que pueden impulsar la innovación y la ventaja competitiva de una organización. Al aprovechar los datos, los científicos de datos pueden identificar nuevas oportunidades de negocio, mejorar productos y servicios existentes, e incluso crear nuevos modelos de negocio.

La disciplina de científico de datos tiene sus raíces en campos como las estadísticas, la informática y la inteligencia artificial. A medida que la tecnología y la cantidad de datos han aumentado exponencialmente, la necesidad de expertos en ciencia de datos se ha vuelto cada vez más importante. La disciplina ha evolucionado rápidamente en las últimas décadas, impulsada por avances en el procesamiento de datos, el aprendizaje automático y la inteligencia artificial. Se espera que esta disciplina siga creciendo y desempeñe un papel fundamental en la sociedad y los negocios. Con el avance de la tecnología y la creciente disponibilidad de datos, se prevé que la demanda de científicos de datos continúe en aumento. Además, se espera que se desarrollen nuevas técnicas y enfoques en la ciencia de datos, lo que permitirá un análisis más sofisticado y una toma de decisiones más precisa y basada en datos (Manyika et al., 2011).

## Materiales y Métodos

Para la investigación se usó el lenguaje de programación R, R es un lenguaje de programación estadística ampliamente utilizado para análisis de datos, modelado estadístico y visualización. Es una plataforma gratuita y de código abierto que ofrece una amplia gama de herramientas y paquetes especializados para el procesamiento y análisis de datos. RStudio, por otro lado, es un entorno de desarrollo integrado (IDE) diseñado específicamente para trabajar con R. Proporciona una interfaz gráfica fácil de usar que facilita la escritura, ejecución y depuración de código en R. RStudio también ofrece características adicionales como paneles de visualización, administración de proyectos y colaboración en línea, lo que lo convierte en una herramienta popular entre los usuarios de R. Para desarrollar el modelo de regresión lineal hicimos uso de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM por sus siglas), esta es una metodología para la minería de datos que se utiliza comúnmente en el análisis de datos empresariales. La metodología consta de seis fases:

1. Comprensión del problema: esta fase inicia por comprender los requisitos y objetivos del proyecto para establecer un contexto.
2. Entendimiento de los datos: consiste en explorar y analizar los datos disponibles para obtener una comprensión detallada de su contenido, calidad y estructura.
3. Preparación de los datos: fase donde se realizan tareas de limpieza, transformación y selección de datos para garantizar la calidad y adecuación de los mismos para el análisis.
4. Modelado: periodo de aplicar técnicas de minería de datos y construir modelos predictivos o descriptivos utilizando algoritmos apropiados.
5. Evaluación: fase para poner a prueba el rendimiento y la eficacia de los modelos construidos mediante métricas y pruebas para verificar su validez y precisión.
6. Despliegue: donde se implementan los resultados y las conclusiones del proyecto en el entorno.

Por otra parte, la regresión lineal es una técnica conocida de modelado estadístico que se utiliza para analizar la relación entre dos variables continuas. Se utiliza para predecir el valor de una variable dependiente a partir del valor de una o más variables independientes, esta asume que la relación entre las variables es lineal, lo que significa que el cambio en la variable independiente tiene un cambio proporcional en la variable dependiente. Una forma de aplicar la regresión lineal es mediante el método de mínimos cuadrados, este es un método matemático utilizado para encontrar una línea de regresión que mejor se ajusta a los datos. El objetivo del método es minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por la línea de regresión. El conjunto de datos que se analizó proviene de el sitio web ai-jobs.net, el cual es un lugar donde usuarios publican ofertas de trabajo relacionadas con inteligencia artificial, machine learning y ciencia de datos. Tiene un apartado dedicado para que personas que ya trabajan usando esas tecnologías publiquen de manera anónima detalles sobre su trabajo tales como su salario, experiencia o lugar de residencia; esto con el objetivo de proporcionar una guía para los principiantes que quieren dedicarse a lo mismo. vamos a normalizar datos?

### Comprensión del problema

Se empezó por analizar los salarios que se perciben a través de los años y sus desviaciones estándar, después lo hicimos con el promedio de los profesionistas de data mining.

### Entendimiento de los datos

Tenemos un conjunto de datos que contiene 11 variables:

- Año trabajado: desde 2020 hasta 2023.
- Nivel de experiencia: Ingeniero Senior (SE), Gerente/Instructor (MI), Nivel Principiante (EN) y Experto (EX).
- Tipo de empleo: Tiempo completo (FT), independiente (FL) y por contrato (CT).
- Título del trabajo.
- Salario.
- Divisa del salario.
- Salario en dólares.
- Lugar de residencia.
- Trabajo remoto (en porcentaje).
- Ubicación de la compañia.
- Tamaño de la compañia: pequeña (S), mediana (M) y grande (L).

Debido a las diferencias de divisas utilizaremos la variable del salario en dólares para normalizar los datos recibidos, año trabajado y el nivel de experiencia, por lo que en primera instancia se hicieron dos data frames. El primero con las variables salario en dólares y año trabajado; y el segundo con el salario en dólares y el nivel de experiencia.

### Preparación de los datos

Instalaremos y utilizarémos el paquete de "stringr" para poder utilizar análisis de cadenas de texto dentro de el dataSet. Inicialmente, en el conjunto de datos la variable del tamaño de la compañia, la experiencia y el país vienen representados con letras, por lo que tenemos que buscar estas letras y después cambiarlos por valores numéricos de la siguiente manera utilizando un ciclo

```
salarios <- read.csv("~/Downloads/ds_salaries.csv")

library("stringr")
options(scipen=999)

##### as.factor()
### relacion entre el nivel de experiencia, salario, tamaño de la empresa y el pais
###

experiencia <- salarios$experience_level
experiencia
experienciaNueva <- list()

tamano <- salarios$company_size
tamano
tamanoNuevo <- list()

lugar <- salarios$company_location
lugar
lugarNuevo <- list()

salarioUsd <- salarios$salary_in_usd
salarioUsd

# Cambiar valores de exp a numeros
for (x in 1 : length(experiencia)) {
  if (str_detect(experiencia[[x]], "SE")) experienciaNueva[x] <- 1 #1 para SE
```

```
   if(str_detect(experiencia[[x]],"MI")) experienciaNueva[x] <- 2 #2 para MI
   if(str_detect(experiencia[[x]],"EN")) experienciaNueva[x] <- 3 #3 para EN
   if(str_detect(experiencia[[x]],"EX")) experienciaNueva[x] <- 4 #4 para EX

}

#Cambiar valores de tamaño de empresa a numeros
for (x in 1 : length(tamano)) {
  if (str_detect(tamano[[x]], "M")) tamanoNuevo[x] <- 2 #2 para mediano
  if(str_detect(tamano[[x]],"L")) tamanoNuevo[x] <- 3 #3 para grande
  if(str_detect(tamano[[x]],"S")) tamanoNuevo[x] <- 1 #1 para pequeño
}
#Cambiar valores de pais a números (ignorar comentarios en esta parte jaja)
for (x in 1 : length(lugar)) {
  if (str_detect(lugar[[x]], "US")) lugarNuevo[x] <- 1 #1 para SE
  if(str_detect(lugar[[x]],"GB")) lugarNuevo[x] <- 2 #2 para MI
  if(str_detect(lugar[[x]],"CA")) lugarNuevo[x] <- 3 #3 para EN
  if(str_detect(lugar[[x]],"ES")) lugarNuevo[x] <- 4 #4 para EX
  if (str_detect(lugar[[x]], "IN")) lugarNuevo[x] <- 5 #1 para SE
  if(str_detect(lugar[[x]],"DE")) lugarNuevo[x] <- 6 #2 para MI
  if(str_detect(lugar[[x]],"FR")) lugarNuevo[x] <- 7 #3 para EN
  if(str_detect(lugar[[x]],"BR")) lugarNuevo[x] <- 8 #4 para EX
  if (str_detect(lugar[[x]], "PT")) lugarNuevo[x] <- 9 #1 para SE
  if(str_detect(lugar[[x]],"AU")) lugarNuevo[x] <- 10 #2 para MI
  if(str_detect(lugar[[x]],"GR")) lugarNuevo[x] <- 11 #3 para EN
  if(str_detect(lugar[[x]],"NL")) lugarNuevo[x] <- 12 #4 para EX
  if (str_detect(lugar[[x]], "MX")) lugarNuevo[x] <- 13 #1 para SE
  if(str_detect(lugar[[x]],"IE")) lugarNuevo[x] <- 14 #2 para MI
  if(str_detect(lugar[[x]],"SG")) lugarNuevo[x] <- 15 #3 para EN
  if(str_detect(lugar[[x]],"JP")) lugarNuevo[x] <- 16 #4 para EX
  if (str_detect(lugar[[x]], "AT")) lugarNuevo[x] <- 17 #1 para SE
  if(str_detect(lugar[[x]],"PL")) lugarNuevo[x] <- 18 #2 para MI
  if(str_detect(lugar[[x]],"CH")) lugarNuevo[x] <- 19 #3 para EN
  if(str_detect(lugar[[x]],"NG")) lugarNuevo[x] <- 20 #4 para EX
  if (str_detect(lugar[[x]], "TR")) lugarNuevo[x] <- 21 #1 para SE
  if(str_detect(lugar[[x]],"LV")) lugarNuevo[x] <- 22 #2 para MI
  if(str_detect(lugar[[x]],"PR")) lugarNuevo[x] <- 23 #3 para EN
  if(str_detect(lugar[[x]],"IT")) lugarNuevo[x] <- 24 #4 para EX
  if (str_detect(lugar[[x]], "DK")) lugarNuevo[x] <- 25 #1 para SE
  if(str_detect(lugar[[x]],"SI")) lugarNuevo[x] <- 26 #2 para MI
  if(str_detect(lugar[[x]],"CO")) lugarNuevo[x] <- 27 #3 para EN
  if(str_detect(lugar[[x]],"BE")) lugarNuevo[x] <- 28 #4 para EX
  if (str_detect(lugar[[x]], "PK")) lugarNuevo[x] <- 29 #1 para SE
  if(str_detect(lugar[[x]],"UA")) lugarNuevo[x] <- 30 #2 para MI
  if(str_detect(lugar[[x]],"AR")) lugarNuevo[x] <- 31 #3 para EN ME QUEDE AQUI
  if(str_detect(lugar[[x]],"LU")) lugarNuevo[x] <- 32 #4 para EX
  if (str_detect(lugar[[x]], "AS")) lugarNuevo[x] <- 33 #1 para SE
  if(str_detect(lugar[[x]],"CZ")) lugarNuevo[x] <- 34 #2 para MI
  if(str_detect(lugar[[x]],"TH")) lugarNuevo[x] <- 35 #3 para EN
  if(str_detect(lugar[[x]],"FI")) lugarNuevo[x] <- 36 #4 para EX
  if (str_detect(lugar[[x]], "RU")) lugarNuevo[x] <- 37 #1 para SE
  if(str_detect(lugar[[x]],"HR")) lugarNuevo[x] <- 38 #2 para MI
  if(str_detect(lugar[[x]],"AE")) lugarNuevo[x] <- 39 #3 para EN
```

```
  if(str_detect(lugar[[x]],"LT")) lugarNuevo[x] <- 40 #4 para EX
  if (str_detect(lugar[[x]], "GH")) lugarNuevo[x] <- 41 #1 para SE
  if(str_detect(lugar[[x]],"IL")) lugarNuevo[x] <- 42 #2 para MI
  if(str_detect(lugar[[x]],"RO")) lugarNuevo[x] <- 43 #3 para EN
  if(str_detect(lugar[[x]],"KE")) lugarNuevo[x] <- 44 #4 para EX
  if (str_detect(lugar[[x]], "EE")) lugarNuevo[x] <- 45 #1 para SE
  if(str_detect(lugar[[x]],"CF")) lugarNuevo[x] <- 46 #2 para MI
  if(str_detect(lugar[[x]],"SE")) lugarNuevo[x] <- 47 #3 para EN
  if(str_detect(lugar[[x]],"HU")) lugarNuevo[x] <- 48 #4 para EX
  if (str_detect(lugar[[x]], "ID")) lugarNuevo[x] <- 49 #1 para SE
  if(str_detect(lugar[[x]],"BS")) lugarNuevo[x] <- 50 #2 para MI
  if(str_detect(lugar[[x]],"MK")) lugarNuevo[x] <- 51 #3 para EN
  if(str_detect(lugar[[x]],"HN")) lugarNuevo[x] <- 52 #4 para EX
  if (str_detect(lugar[[x]], "MT")) lugarNuevo[x] <- 53 #1 para SE
  if(str_detect(lugar[[x]],"AM")) lugarNuevo[x] <- 54 #2 para MI
  if(str_detect(lugar[[x]],"AL")) lugarNuevo[x] <- 55 #3 para EN
  if(str_detect(lugar[[x]],"SK")) lugarNuevo[x] <- 56 #4 para EX
  if (str_detect(lugar[[x]], "MY")) lugarNuevo[x] <- 57 #1 para SE
  if(str_detect(lugar[[x]],"MD")) lugarNuevo[x] <- 58 #2 para MI
  if(str_detect(lugar[[x]],"DZ")) lugarNuevo[x] <- 59 #3 para EN
  if(str_detect(lugar[[x]],"IR")) lugarNuevo[x] <- 60 #4 para EX
  if (str_detect(lugar[[x]], "BO")) lugarNuevo[x] <- 61 #1 para SE
  if(str_detect(lugar[[x]],"CR")) lugarNuevo[x] <- 62 #2 para MI
  if(str_detect(lugar[[x]],"NZ")) lugarNuevo[x] <- 63 #3 para EN
  if(str_detect(lugar[[x]],"BA")) lugarNuevo[x] <- 64 #4 para EX
  if (str_detect(lugar[[x]], "PH")) lugarNuevo[x] <- 65 #1 para SE
  if(str_detect(lugar[[x]],"HK")) lugarNuevo[x] <- 66 #2 para MI
  if(str_detect(lugar[[x]],"EG")) lugarNuevo[x] <- 67 #3 para EN
  if(str_detect(lugar[[x]],"MA")) lugarNuevo[x] <- 68 #4 para EX
  if (str_detect(lugar[[x]], "CL")) lugarNuevo[x] <- 69 #1 para SE
  if(str_detect(lugar[[x]],"VN")) lugarNuevo[x] <- 70 #2 para MI
  if(str_detect(lugar[[x]],"CN")) lugarNuevo[x] <- 71 #3 para EN
  if(str_detect(lugar[[x]],"IQ")) lugarNuevo[x] <- 72 #4 para EX
}
```

Ya que necesitamos guardar en algún lugar los valores, primero definimos una lista para guardar los valores nuevos después del ciclo. Para poder utilizar estos valores, tendremos que hacer el cambio de una lista a valores numéricos, por lo que procedemos a utilizar "as.numeric" para convertir nuestras listas en valores numéricos

```
experienciaNueva <- as.numeric(experienciaNueva)
tamanoNuevo <- as.numeric(tamanoNuevo)
lugarNuevo <- as.numeric(lugarNuevo)
```

### Modelado

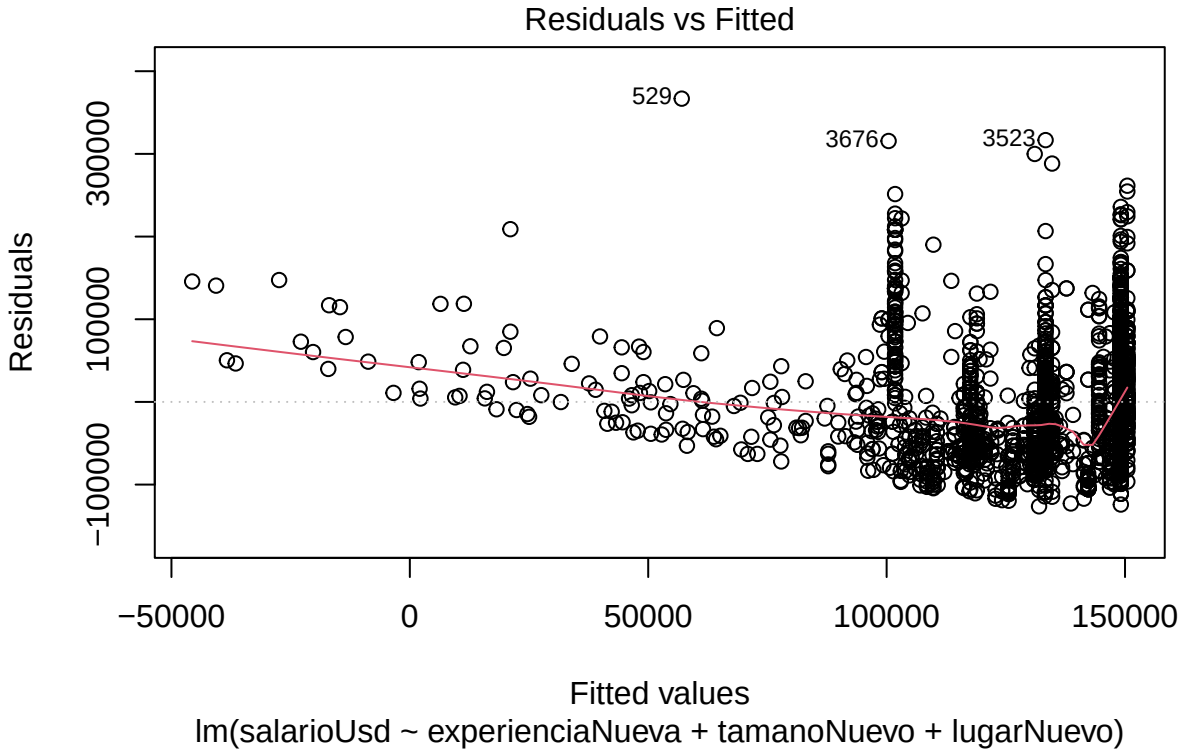
Una vez listas nuestras variables, generamos un modelo de regresión lineal con estas.

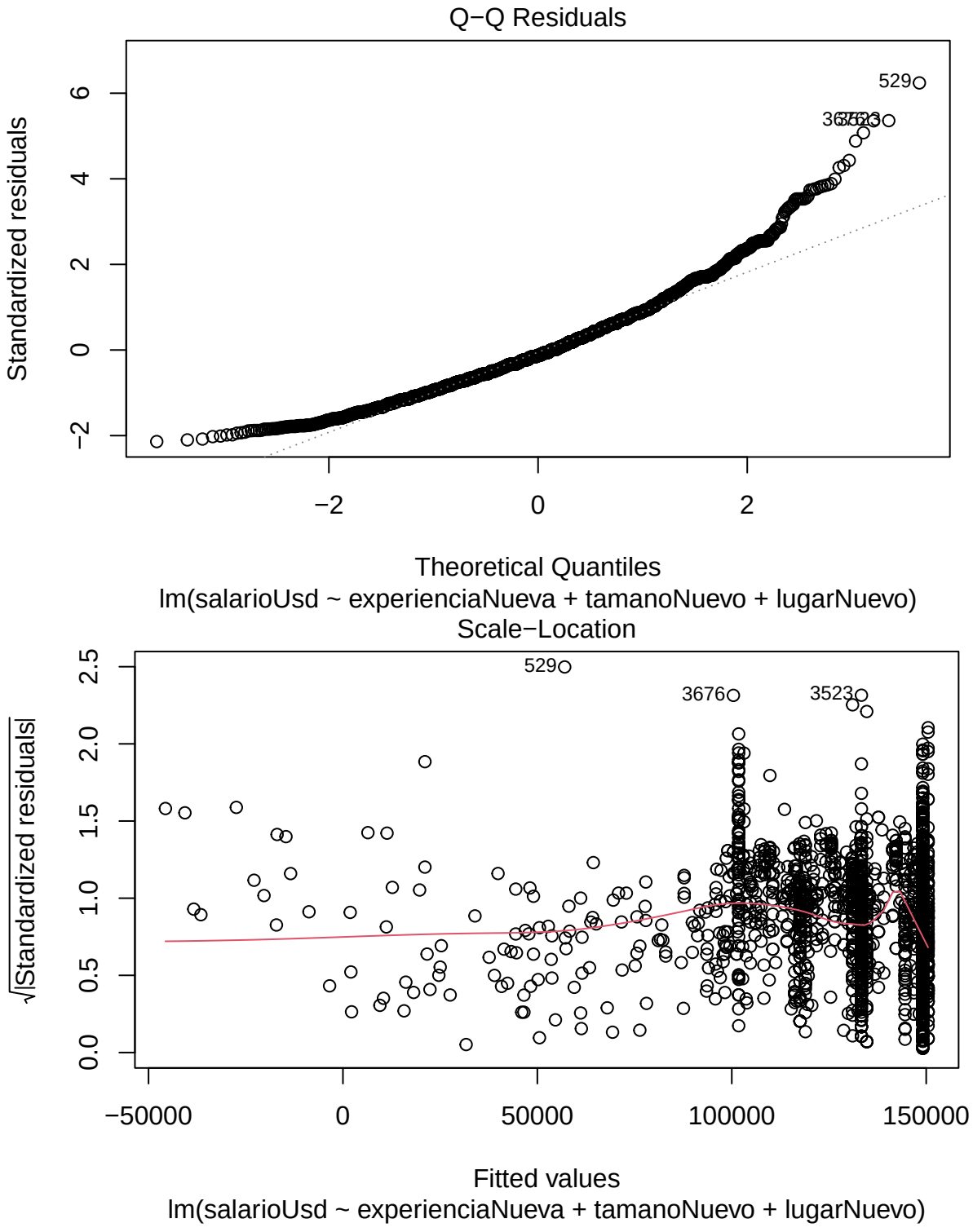
```
modeloLineal <- lm(salarioUsd~experienciaNueva+tamanoNuevo+lugarNuevo)
summary(modeloLineal)
```

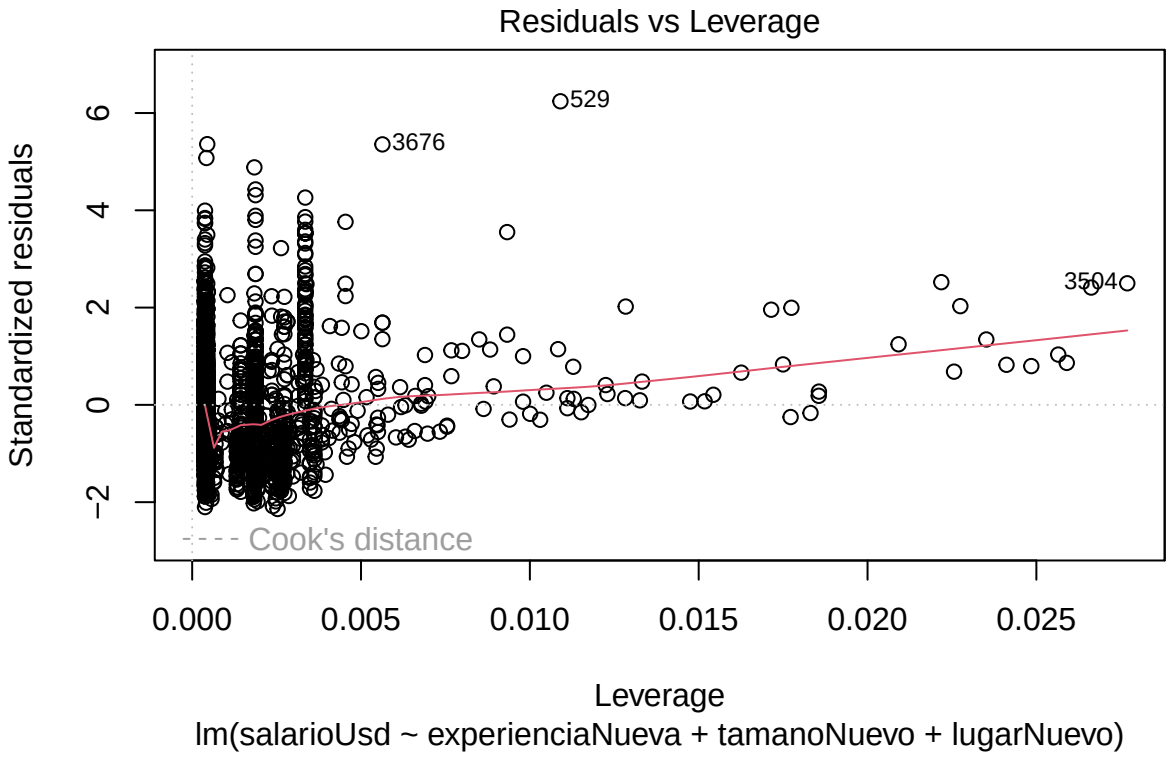
```
##
## Call:
## lm(formula = salarioUsd ~ experienciaNueva + tamanoNuevo + lugarNuevo)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -126294  -40286   -7595   34270  366805
##
## Coefficients:
##                   Estimate Std. Error t value            Pr(>|t|)
## (Intercept)      164447.9     5441.1  30.223 <0.0000000000000002 ***
## experienciaNueva -15780.5     1269.4 -12.431 <0.0000000000000002 ***
## tamanoNuevo        1366.0     2464.1   0.554               0.579
## lugarNuevo        -2279.4      139.2 -16.379 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59100 on 3751 degrees of freedom
## Multiple R-squared:  0.1223, Adjusted R-squared:  0.1216
## F-statistic: 174.3 on 3 and 3751 DF,  p-value: < 0.00000000000000022
```

Generamos una gráfica para que se viera más clara la relación:




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(salarioUsd ~ experienciaNueva + tamanoNuevo + lugarNuevo)
Scale−Location
√|Standardized residuals|
Fitted values
lm(salarioUsd ~ experienciaNueva + tamanoNuevo + lugarNuevo)




Y después hicimos una predicción.

```
prediccion <- predict(modeloLineal)
prediccion
```

## Conclusiones

Cuando realizamos nuestro modelo de regresión lineal tenemos que nuestro Adjusted R-squared es de 0.1216 lo que equivale a una confiabilidad de los datos de 12.16% aproximadamente, por lo que nuestra predicción del futuro acerca de esta profesión con sus salarios respecto a su nivel de experiencia, lugar de trabajo y tamaño de la empresa es muy poca confiable.

## Referencias


1. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium.
2. InLab FIB. (s.f.). ¿Qué es un Data Scientist? Recuperado de https://inlab.fib.upc.edu/es/blog/que-es-un-data-scientist
3. Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. Harvard Business Review, 90(10), 70-76.
4. 1. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.
5. Chaki, A. (2023). Data Science Salaries 2023. [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023