TSUNAMIS - INTERNATIONAL

Emmanuel Rodríguez Silva

6/10/2023

1. TSUNAMIS AROUND THE WORLD - DATA PROVIDED BY THE NOAA (US)

This document contains some practice using datasets that were acquired in SQL (Structured Query Language).

This project have the following objectives:

1. Practice R programming
2. Combine the usage of SQL and R
3. Make better graphics

```
library(palmerpenguins)
library(ggplot2)
library(tidyverse)
library(dplyr)
library(lubridate)
library(here)     # referencing files easier
library(skimr)    # simplify data cleaning task (summarizing)
library(janitor)
```

```
tsunamis <- read.csv(file = "tsunamis.csv", header = T)
head(tsunamis)
```

```
##     country year water_ht deaths
## 1 INDONESIA 2004     50.9 167540
## 2 INDONESIA 1883     30.0  36000
## 3 SRI LANKA 2004     12.5  35322
## 4  PORTUGAL 1755     15.2  30000
## 5     INDIA 2004     17.3  16269
## 6     JAPAN 1771     85.4  13486
```

```
str(tsunamis)
```

```
## 'data.frame':    403 obs. of  4 variables:
##  $ country : chr  "INDONESIA" "INDONESIA" "SRI LANKA" "PORTUGAL" ...
##  $ year    : int  2004 1883 2004 1755 2004 1771 1765 2004 2004 1498 ...
##  $ water_ht: num  50.9 30 12.5 15.2 17.3 85.4 9 19.6 10 10 ...
##  $ deaths  : int  167540 36000 35322 30000 16269 13486 10000 8212 6051 5000 ...
```

```
str(tsunamis)
```

```
## 'data.frame':    403 obs. of  4 variables:
##  $ country : chr  "INDONESIA" "INDONESIA" "SRI LANKA" "PORTUGAL" ...
##  $ year    : int  2004 1883 2004 1755 2004 1771 1765 2004 2004 1498 ...
##  $ water_ht: num  50.9 30 12.5 15.2 17.3 85.4 9 19.6 10 10 ...
##  $ deaths  : int  167540 36000 35322 30000 16269 13486 10000 8212 6051 5000 ...
```

```
colnames(tsunamis)
```

```
## [1] "country"  "year"     "water_ht" "deaths"
```

Now, let's organize our dataset in order to create beautiful dataviz

```
 tsunamis %>% group_by(country) %>% arrange(country)
```

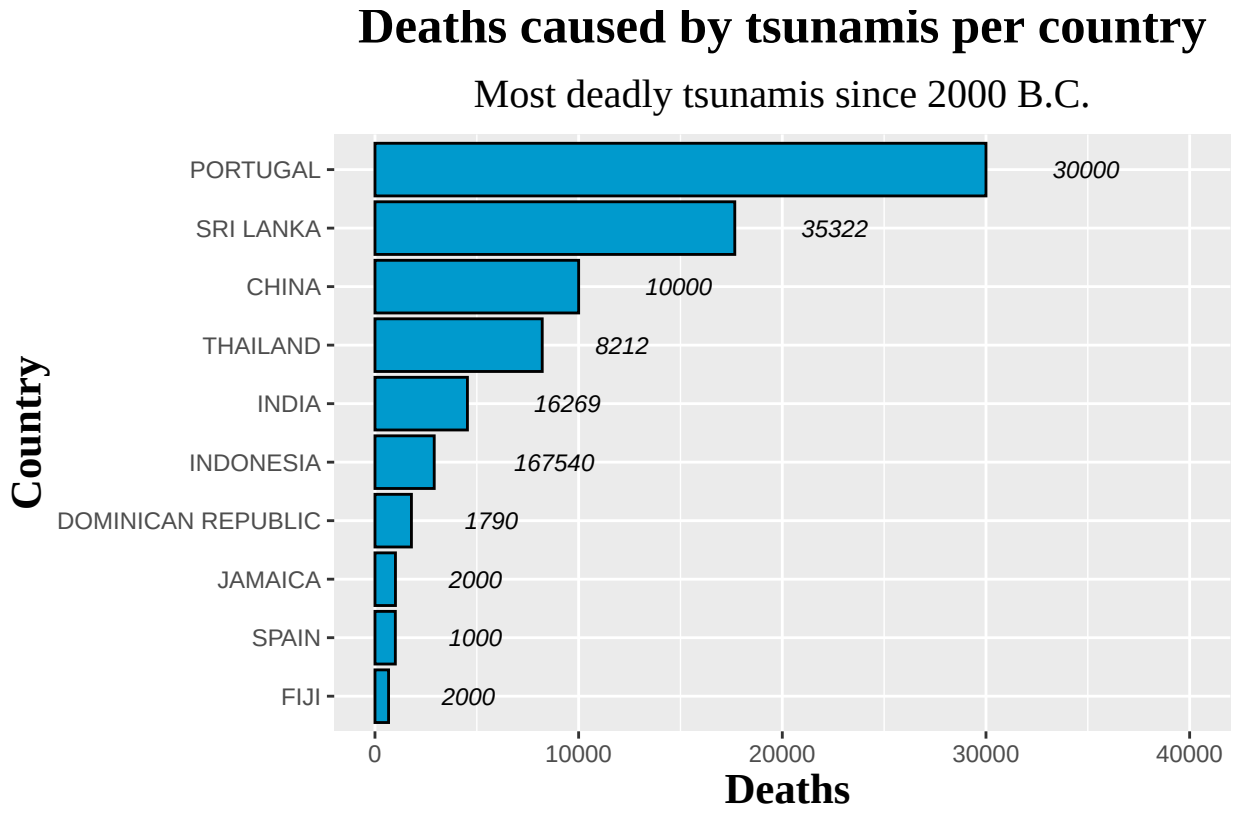
```
## # A tibble: 403 x 4
## # Groups:   country [50]
##    country  year water_ht deaths
##    <chr>   <int>    <dbl>  <int>
##  1 CANADA   1908     15       26
##  2 CANADA   1929      3        8
##  3 CANADA   1929     13        7
##  4 CANADA   1929     13        6
##  5 CANADA   1929     13        4
##  6 CANADA   1929     13        1
##  7 CANADA   1963      5.5      1
##  8 CHILE    1960      4      500
##  9 CHILE    1877     10      200
## 10 CHILE    1922      7      200
## # i 393 more rows
```

```
 tsunamis_impact <- tsunamis %>% group_by(country) %>%
  summarize(mean_deaths = round(mean(deaths)), max_deaths = max(deaths), min_deaths = min(deaths)) %>%
  arrange(desc(mean_deaths)) %>% head(n = 10)
```

Now that we have ordered and cleaned our data, let's do some visualizations

```
mi_col <- c("deepskyblue3")

ggplot(data = tsunamis_impact, aes(x = mean_deaths, y = reorder(country, mean_deaths))) +
  # I ll use colums to graph
  geom_col(fill = mi_col, col = "black")+
  # adding maximum values in order to show discrepancies
  geom_text(aes(label = max_deaths), hjust = -1, fontface = "italic", size = 3)+
  # ADDING LABELS
  labs(title = "Deaths caused by tsunamis per country",
       subtitle = "Most deadly tsunamis since 2000 B.C.",
       caption = "Source: The Global Historical Tsunami Database (NOAA, 2020)",
       y = "Country", x = " Deaths")+
  scale_x_continuous(limits = c(0, 40000.5), breaks = seq(0, 40000, 10000)) +
```

```
  theme_gray()+
  # CUSTOMIZING LABELS
  theme(plot.title = element_text(family = "Times", face = "bold", size = 19, hjust = 0.5, vjust = 2.5)
        plot.subtitle = element_text(family = "Times", size = 15, hjust = 0.5, vjust = 1.5),
        plot.caption = element_text(size = 13, vjust = -1),
        axis.title = element_text(family = "Times", face = "bold", size = 16))
```



As we can infer from the plot above is that mean does not represents well the deadly effect of tsunamis, because some countries had just one tsunami, and it caused many deaths

2. Student Mental Health Project (Kaggle)

This time I will explore data downloaded from Kaggle, I'm interested in investigating the mental health of university students

My objectives are:

1. Data cleaning and organize it
2. Exploring difference by sex, major, age
3. Elaborate some data viz

```
mental <- read.csv(file = "student_mental_health.csv", header = T)
head(mental)
```

```
##        Timestamp Choose.your.gender Age What.is.your.course.
## 1 8/7/2020 12:02             Female  18          Engineering
## 2 8/7/2020 12:04               Male  21    Islamic education
## 3 8/7/2020 12:05               Male  19                  BIT
## 4 8/7/2020 12:06             Female  22                Laws
## 5 8/7/2020 12:13               Male  23         Mathemathics
## 6 8/7/2020 12:31               Male  19          Engineering
##   Your.current.year.of.Study What.is.your.CGPA. Marital.status
## 1                     year 1        3.00 - 3.49             No
## 2                     year 2        3.00 - 3.49             No
## 3                     Year 1        3.00 - 3.49             No
## 4                     year 3        3.00 - 3.49            Yes
## 5                     year 4        3.00 - 3.49             No
## 6                     Year 2        3.50 - 4.00             No
##   Do.you.have.Depression. Do.you.have.Anxiety. Do.you.have.Panic.attack.
## 1                     Yes                   No                       Yes
## 2                      No                  Yes                        No
## 3                     Yes                  Yes                       Yes
## 4                     Yes                   No                        No
## 5                      No                   No                        No
## 6                      No                   No                       Yes
##   Did.you.seek.any.specialist.for.a.treatment.
## 1                                           No
## 2                                           No
## 3                                           No
## 4                                           No
## 5                                           No
## 6                                           No
```

I want to explore the differences between sex

```
colnames(mental)
```

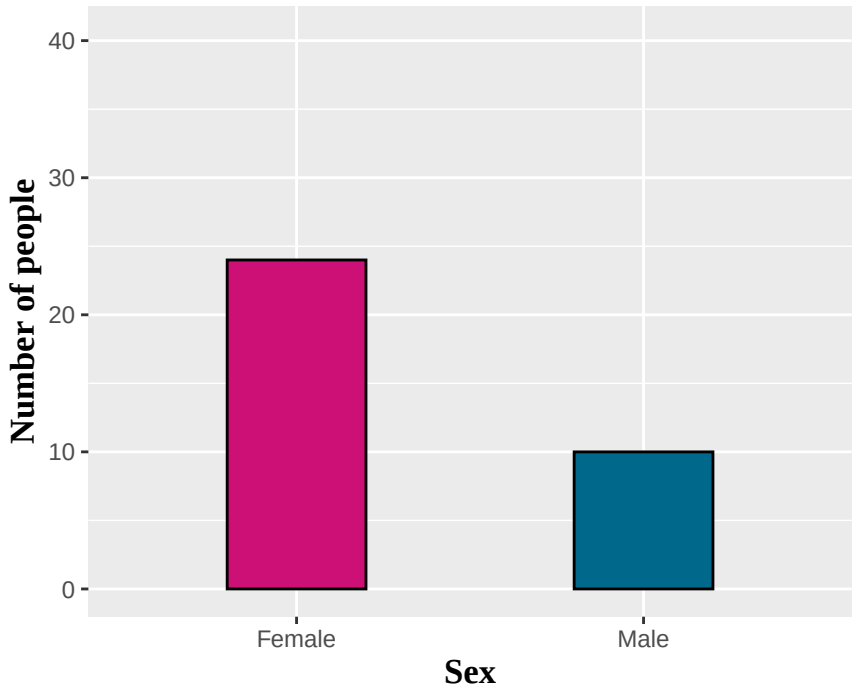
```
##  [1] "Timestamp"
##  [2] "Choose.your.gender"
##  [3] "Age"
##  [4] "What.is.your.course."
##  [5] "Your.current.year.of.Study"
##  [6] "What.is.your.CGPA."
##  [7] "Marital.status"
##  [8] "Do.you.have.Depression."
##  [9] "Do.you.have.Anxiety."
## [10] "Do.you.have.Panic.attack."
## [11] "Did.you.seek.any.specialist.for.a.treatment."
```

```
mental_2 <- mental %>% mutate(sex = Choose.your.gender, anxiety = Do.you.have.Anxiety., depression = Do
  group_by(sex) %>% summarize(how_many_have_anxiety = sum(anxiety == "Yes"), how_many_depressed = sum(d

colnames(mental_2)
```

```
## [1] "sex"                   "how_many_have_anxiety" "how_many_depressed"
```

```
#  group_by(sex)
```

```
my_colors <- c("deeppink3", "deepskyblue4")

ggplot(data = mental_2, aes(x = sex, y = how_many_have_anxiety)) +
  geom_col(fill = my_colors, col = "black", width = 0.4)+
  scale_y_continuous(limits = c(0, 40.5), breaks = seq(0, 40, 10))+
  labs(title = "Anxiety in University Students",
       subtitle = "Differences in mental health between sex",
       caption = "Data set was collected by a survey conducted by Google forms",
       x = "Sex", y = "Number of people")+
  theme(plot.title = element_text(family = "Times", face = "bold", size = 19, hjust = 0.5, vjust = 2.5)
        plot.subtitle = element_text(family = "Times", size = 15, hjust = 0.5, vjust = 1.5),
        plot.caption = element_text(family = "Times", size = 11),
        axis.title = element_text(family = "Times", size = 13, face = "bold"))+
  theme(aspect.ratio = 0.8)
```




```
mental_3 <- mental %>% mutate(sex = Choose.your.gender, anxiety = Do.you.have.Anxiety., depression = Do

mental_4 <- mental_3 %>% distinct(course)
# there is nothing interesting about courses, we need more data
```

Let's construct a linear regression model based on information regarding experience and salary

I have just 2 main porpuses:

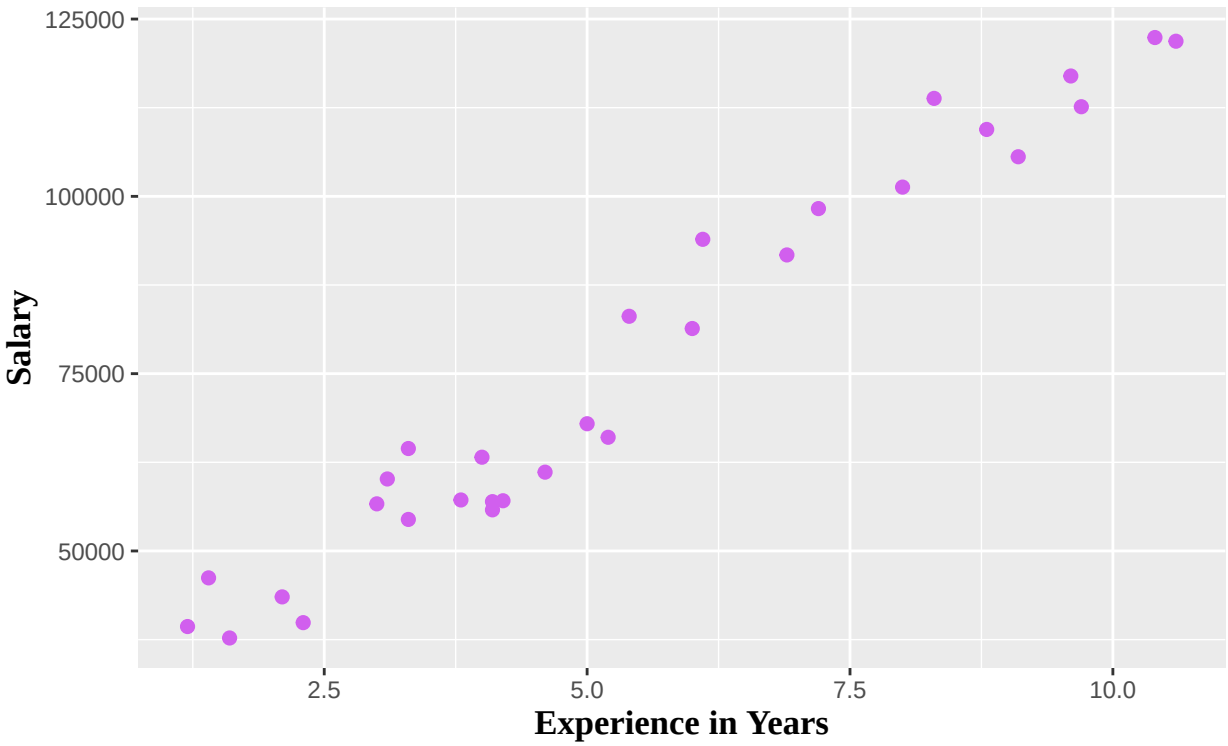
to recall the process of modelling Linear Regression models and, to interpret the results

```
salary <- read.csv(file = "Salary_dataset.csv", header = T)
head(salary, n = 8)
```

```
##   YearsExperience Salary
## 1             1.2  39344
## 2             1.4  46206
## 3             1.6  37732
## 4             2.1  43526
## 5             2.3  39892
## 6             3.0  56643
## 7             3.1  60151
## 8             3.3  54446
```

Let's observe a scatterplot to observe if there is an association between these two variables

```
ggplot(data = salary, aes(x = YearsExperience, y = Salary))+
  geom_point(cex = 2, col = "mediumorchid2")+
  labs(title = "Salary vs. Experience",
       caption = "Data provided by Allena Venkata from her Kaggle  account (2023)",
       x = "Experience in Years")+
  theme_grey()+
  theme(plot.title = element_text(family = "Times", face = "bold", size = 19, hjust = 0.5, vjust = 2.5)
        plot.caption = element_text(family = "Times", size = 11),
        axis.title = element_text(family = "Times", size = 13, face = "bold"))
```



Data provided by Allena Venkata from her Kaggle account (2023)

Now, I have seen some evidence about the relationship between salary and experience, I need some statistical proves, I'll apply test in order to discover the type of distribution, if data is normal, I ll apply parametric statistics, otherwise, we should use non-parametric statistics.

```
shapiro.test(salary$Salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salary$Salary
## W = 0.91032, p-value = 0.01516
```

```
shapiro.test(salary$YearsExperience)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salary$YearsExperience
## W = 0.94206, p-value = 0.1034
```

According to the normality test, only the experience has a normal distribution, because P is greater than **0.05**

So, let's try to predict the salary based on experience, but first of all, I need to make sure there is significance relationship between them

```
cor.test(salary$Salary, salary$YearsExperience)
```

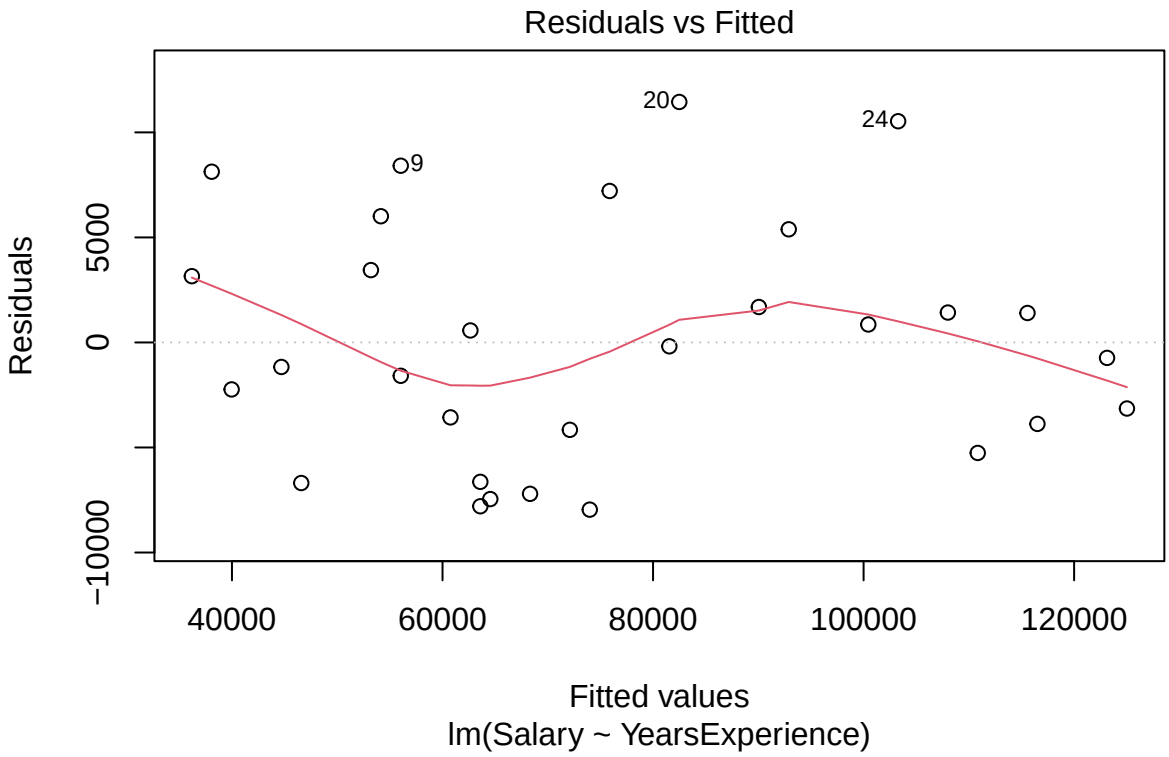
```
##
##  Pearson's product-moment correlation
##
## data:  salary$Salary and salary$YearsExperience
## t = 24.95, df = 28, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9542949 0.9897078
## sample estimates:
##       cor
## 0.9782416
```

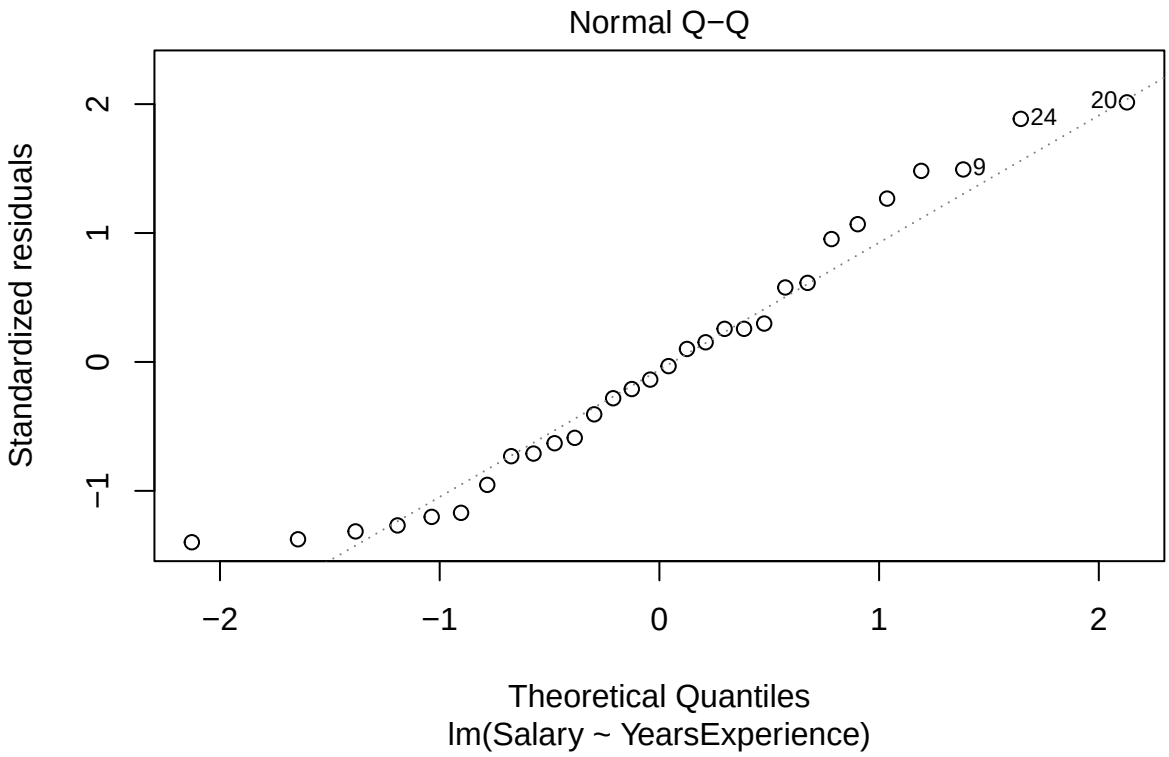
After applying the **pearson's correlation coefficient**, I noticed that exist a strong positive correlation between our variables, having said that, I will proceed with the elaboration of the linear model.

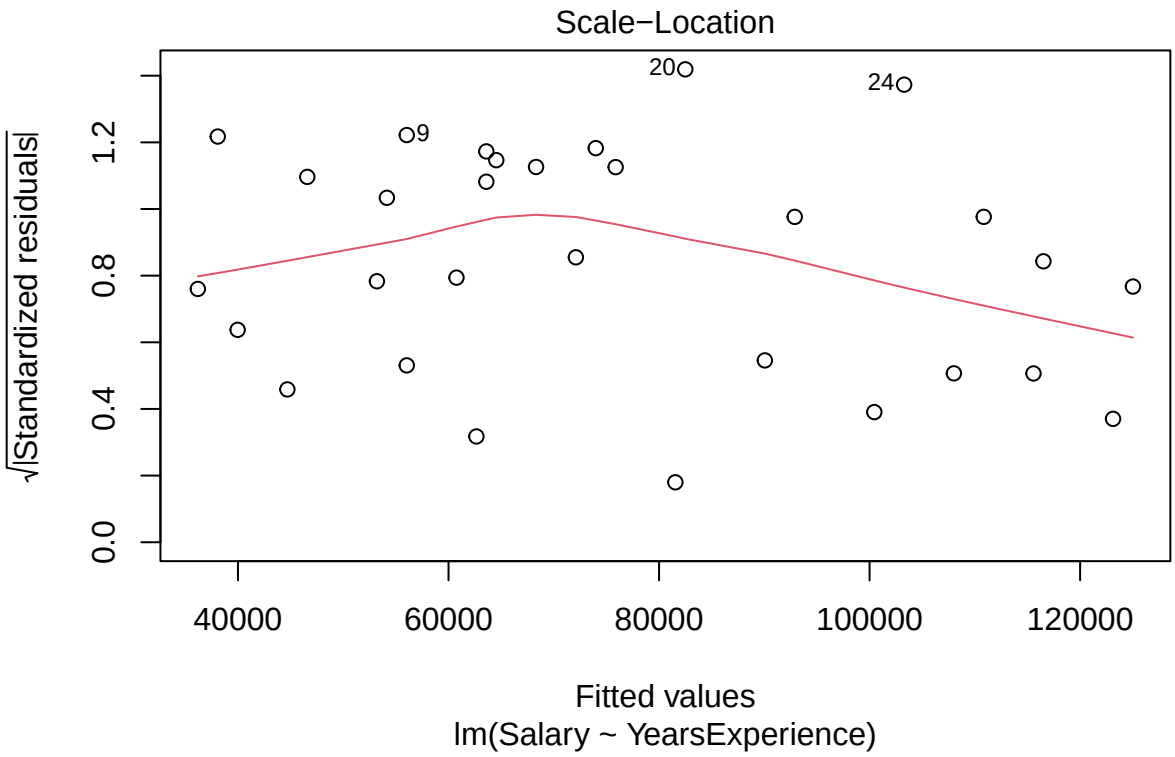
```
model_salary <- lm(Salary ~ YearsExperience, data = salary)
summary(model_salary)
```

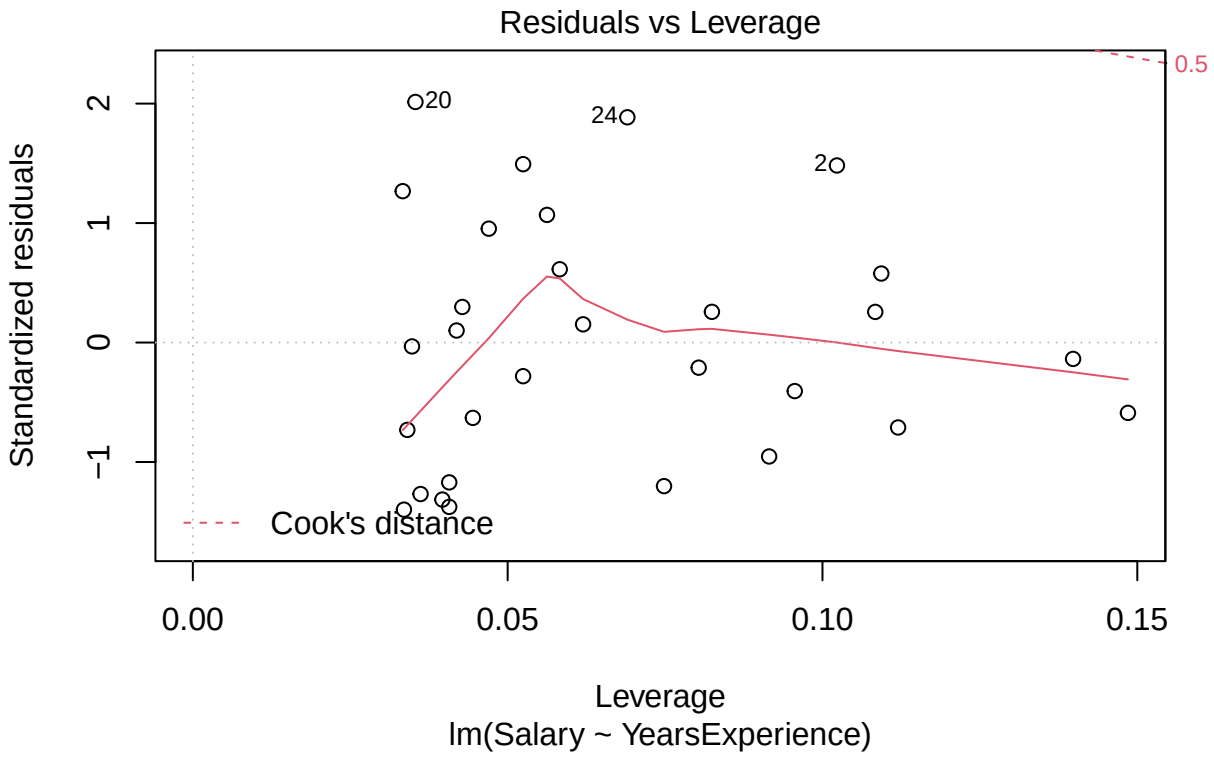
```
##
## Call:
## lm(formula = Salary ~ YearsExperience, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7958.0 -4088.5  -459.9  3372.6 11448.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      24848.2     2306.7   10.77 1.82e-11 ***
## YearsExperience   9450.0      378.8   24.95  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5788 on 28 degrees of freedom
## Multiple R-squared:  0.957,  Adjusted R-squared:  0.9554
## F-statistic: 622.5 on 1 and 28 DF,  p-value: < 2.2e-16
```

```
plot(model_salary)
```


Residuals vs Fitted
20
24
9
Residuals
5000
0
−10000
40000
60000
80000
100000
120000
Fitted values
lm(Salary ~ YearsExperience)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Salary ~ YearsExperience)
24
20
9
−2
−1
0
1
2


Scale−Location
20
24
9
√|Standardized residuals|
0.0
0.4
0.8
1.2
40000
60000
80000
100000
120000
Fitted values
lm(Salary ~ YearsExperience)



The previous model indicates that, for one (unit = year)increment in experience, the salary expected will increase **$9,450.00** Knowing that, We can predict a salary based on how many years of experience a person has.

```
# manually predicting salary according with coefficients model

print("y =24848.2 + 9450.0 * 10")
```

```
## [1] "y =24848.2 + 9450.0 * 10"
```

```
# calculating salary for a person with 10 years of experience

salary_predicted = 24848.2 + 9450.0 * 10

salary_predicted
```

```
## [1] 119348.2
```

```
value <- data.frame(YearsExperience = 10)


predict(model_salary, value)
```

```
##        1
## 119347.8
```

```
ggplot(data = salary, aes(x = YearsExperience, y = Salary)) +
  geom_point() +
  geom_smooth(method='lm')+
  geom_text(x = 8.5, y = 48000, label = "y = 24848.2 + 9450.0 * experience + e \n  y = b0 + b1*x + e")+
  labs(title = "Linear Regression (Predicting Salary)",
       caption = "Source: Salary Dataset on Kaggle by Allena Venkata (2023)")+
  theme_grey()+
  theme(plot.title = element_text(family = "Times", face = "bold", size = 19, hjust = 0.5, vjust = 2.5),
        plot.caption = element_text(family = "Times", size = 11),
        axis.title = element_text(family = "Times", face = "bold", size = 15))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```