Data Salaries Expectations

Franco Quintanilla

2022-08-24

Importamos los datos

```
df = read.csv("/Users/francoquintanilla/Documents/R/ds_salaries.csv")
head(df)
```

```
##   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
```

Variables cuantitativas

Definimos las variables cuantitativas como una variable estadística que puede ser medible, por lo que definimos las siguientes variables:

```
work_year = df$work_year
salary = df$salary
salary_USA = df$salary_in_usd
remote_ratio = df$remote_ratio
```

Medidas de tendencia central y Medidas de dispersión.

Para las medidas de tendencia de cada una de estas variables, vamos a sacar el promedio, la mediana, y la moda. Para las medidas de dispersión vamos a calcular la varianza y la desviación estándar

Para la moda, vamos a definir la siguiente función, ya que R no cuenta con la función de la moda que a nosotros nos interesa.

```
moda <- function(x)
  {
   return(as.numeric(names(which.max(table(x)))))
  }
```

Work Year

```
# Mean
p_work_year = mean(work_year)
cat("El promedio de 'Work Year' es de:", p_work_year, "\n")

## El promedio de 'Work Year' es de: 2021.405

# Median
med_work_year = median(work_year)
cat("La mediana de 'Work Year' es de:", med_work_year, "\n")

## La mediana de 'Work Year' es de: 2022

# Mode
moda_work_year = moda(work_year)
cat("La moda de 'Work Year' es de:", moda_work_year, "\n")

## La moda de 'Work Year' es de: 2022

# Variance
var_work_year = var(work_year)
cat("La varianza de 'Work Year' es de:", var_work_year, "\n")

## La varianza de 'Work Year' es de: 0.4790481

# Standard deviation
desv_work_year = sd(work_year)
cat("La desviación estandar de 'Work Year' es de:", desv_work_year, "\n")

## La desviación estandar de 'Work Year' es de: 0.692133
```

Salary

```
# Mean
p_salary = mean(salary)
cat("El promedio de 'Salary' es de:", p_salary, "\n")

## El promedio de 'Salary' es de: 324000.1

# Median
med_salary = median(salary)
cat("La mediana de 'Salary' es de:", med_salary, "\n")

## La mediana de 'Salary' es de: 115000

# Mode
moda_salary = moda(salary)
cat("La moda de 'Salary' es de:", moda_salary, "\n")

## La moda de 'Salary' es de: 80000

# Variance
var_salary = var(salary)
cat("La varianza de 'Salary' es de:", var_salary, "\n")

## La varianza de 'Salary' es de: 2.38504e+12

# Standard deviation
desv_salary = sd(salary)
cat("La desviación estandar de 'Salary' es de:", desv_salary, "\n")

## La desviación estandar de 'Salary' es de: 1544357
```

Salary in USD

```
# Mean
p_salary_in_usd = mean(salary_USA)
cat("El promedio de 'Salary in USD' es de:", p_salary_in_usd, "\n")

## El promedio de 'Salary in USD' es de: 112297.9

# Median
med_salary_in_usd = median(salary_USA)
cat("La mediana de 'Salary in USD' es de:", med_salary_in_usd, "\n")

## La mediana de 'Salary in USD' es de: 101570

# Mode
moda_salary_in_usd = moda(salary_USA)
cat("La moda de 'Salary in USD' es de:", moda_salary_in_usd, "\n")

## La moda de 'Salary in USD' es de: 1e+05

# Variance
var_salary_in_usd = var(salary_USA)
cat("La varianza de 'Salary in USD' es de:", var_salary_in_usd, "\n")
```

```
## La varianza de 'Salary in USD' es de: 5034932663

# Standard deviation
desv_salary_in_usd = sd(salary_USA)
cat("La desviación estandar de 'Salary in USD' es de:",
desv_salary_in_usd, "\n")

## La desviación estandar de 'Salary in USD' es de: 70957.26
```

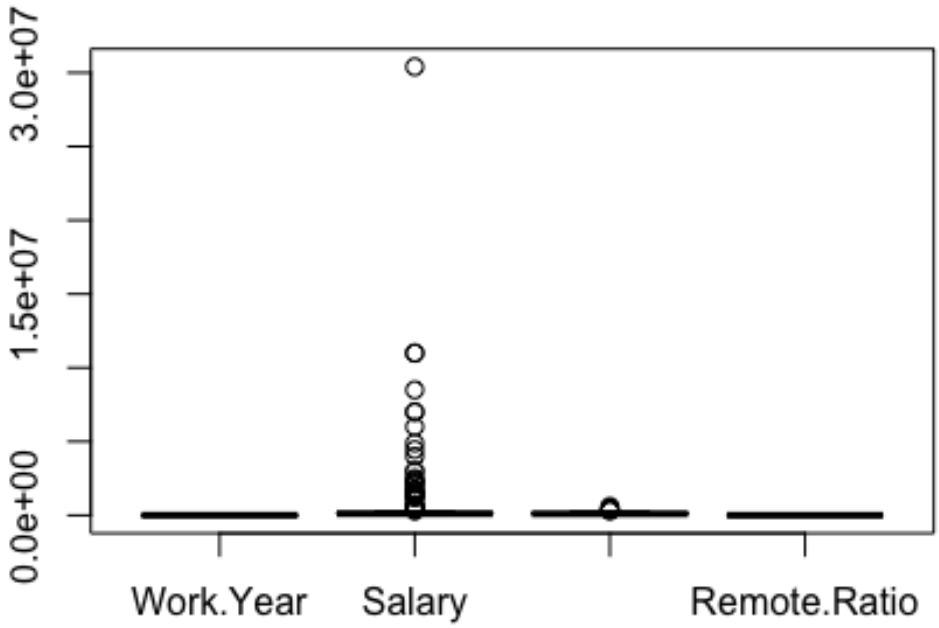
Remote Ratio

```
# Mean
p_remote_ratio = mean(remote_ratio)
cat("El promedio de 'Salary' es de:", p_remote_ratio, "\n")

## El promedio de 'Salary' es de: 70.92257

# Median
med_remote_ratio = median(remote_ratio)
cat("La mediana de 'Salary' es de:", med_remote_ratio, "\n")

## La mediana de 'Salary' es de: 100

# Mode
moda_remote_ratio = moda(remote_ratio)
cat("La moda de 'Salary' es de:", moda_remote_ratio, "\n")

## La moda de 'Salary' es de: 100

# Variance
var_remote_ratio = var(remote_ratio)
cat("La varianza de 'Salary' es de:", var_remote_ratio, "\n")

## La varianza de 'Salary' es de: 1657.233

# Standard deviation
desv_remote_ratio = sd(remote_ratio)
cat("La desviación estandar de 'Salary' es de:", desv_remote_ratio, "\n")

## La desviación estandar de 'Salary' es de: 40.70913
```

Visualización de las medidas de tendencia central, y de dispersión.

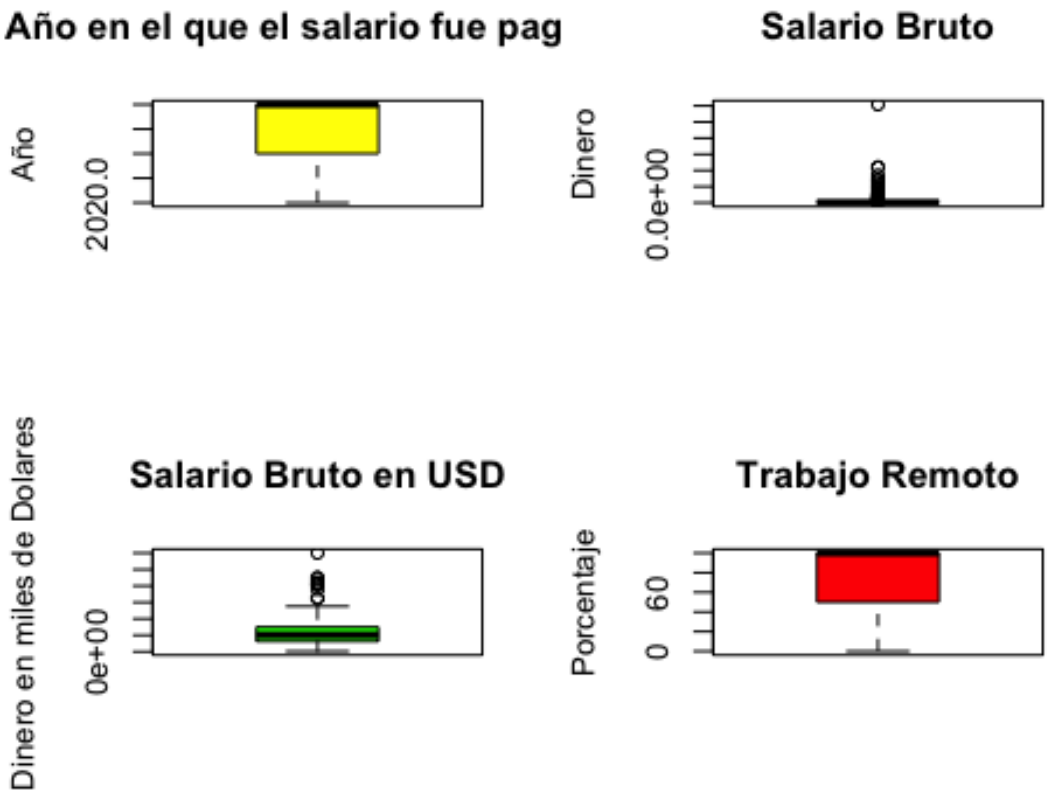
Boxplot

```
df_boxplot = data.frame("Work Year"=work_year, "Salary"=salary, "Salary
in USD"=salary_USA, "Remote Ratio"=remote_ratio)
boxplot(df_boxplot)
```



Con este gráfico, nos damos cuenta de que no podemos graficar las 4 variables cualitativas por sus diferentes magnitudes, por lo que vamos a graficar cada una por separado.

```
par(mfrow=c(2,2))

boxplot(work_year, main="Año en el que el salario fue pagado",
ylab="Año", col="yellow")
boxplot(salary, main="Salario Bruto", ylab="Dinero", col="blue")
boxplot(salary_USA, main="Salario Bruto en USD", ylab="Dinero en miles de
Dolares", col="green")
boxplot(remote_ratio, main="Trabajo Remoto", ylab="Porcentaje",
col="red")
```



Se ve la gran diferencia de los datos, y ahora si podemos visualizar mejor el "Box plot".

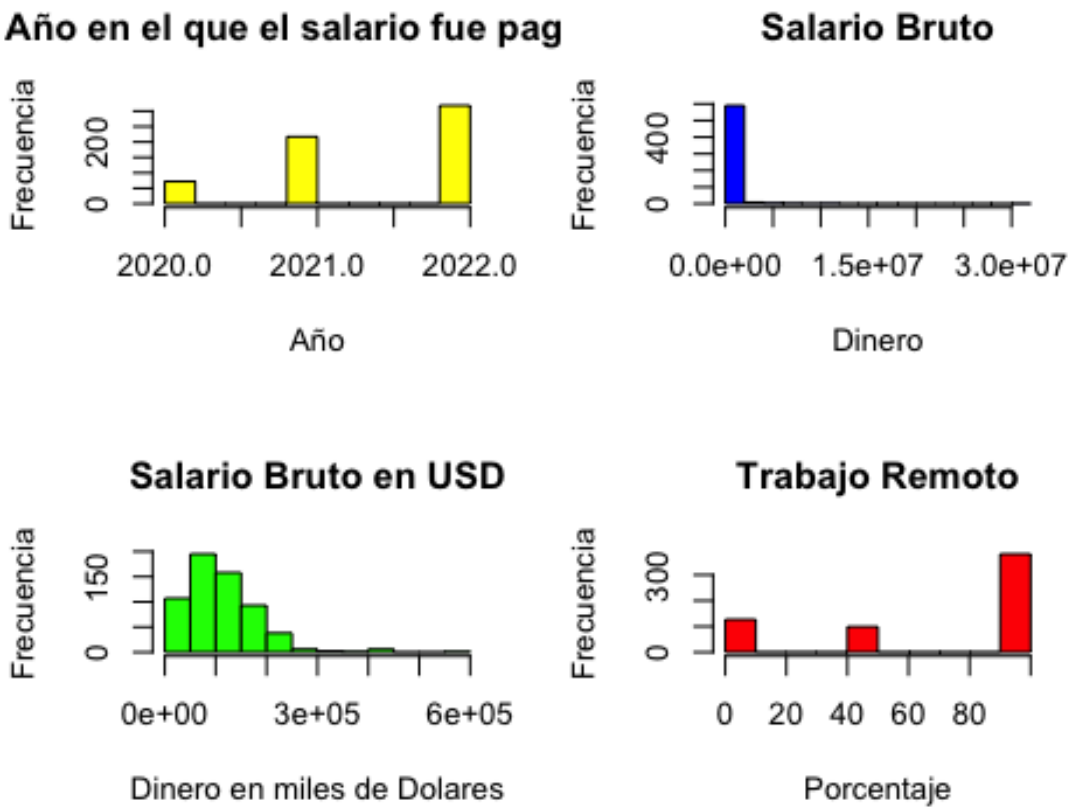
Viendo los plots anteriores nos damos cuenta de que tanto "Salary in USD" como "Total gross salary" tienen outliers. En "Total gross salary" se puede observar como prácticamente todos los datos son distintos y por eso no cuenta con un "box" que represente los intercuartiles ni la media, sino que se apreciaría como una línea, y por eso mismo la mayoría de los datos son atípicos. Esto lo tomaremos en cuenta a la hora de la limpieza de los datos.

Histograma

El histograma es una herramienta que nos ayuda a visualizar la distribución de nuestros datos. En este caso vamos a realizar el histograma con nuestras variables cualitativas.

```
par(mfrow=c(2,2))

hist(work_year, main="Año en el que el salario fue pagado", xlab="Año",
ylab="Frecuencia", col="yellow")
hist(salary, main="Salario Bruto", xlab="Dinero", ylab="Frecuencia",
col="blue")
hist(salary_USA, main="Salario Bruto en USD", xlab="Dinero en miles de
Dolares", ylab="Frecuencia", col="green")
```

```
hist(remote_ratio, main="Trabajo Remoto", xlab="Porcentaje",
ylab="Frecuencia", col="red")
```



En esta visualización de histograma podemos ver que tanto “Year the salary was paid”, cómo “Work done remotely” son datos que solo ocurren 3 veces en ciertos momentos, es decir, en el primer caso es por que esa variable representa el año en el cual el salario fue pagado, y en el segundo es por que representa lo siguiente: 0 = Sin trabajo a distancia (menos del 20%), 50 = Parcialmente a distancia, 100 = Totalmente a distancia (más del 80%).

En el caso de “Total gross salary” parecería que tenemos una distribución delta, ya que es un dato mayoritario en el dataset y no sigue, o al menos no en una cantidad remarcable. Por último, en el caso de “Salary in USD” podemos ver que sigue una posible distribución normal sesgada a la derecha, lo que nos dice que los salarios en su mayor frecuencia no son tan grandes como uno esperaría, sino, que se encuentra con un promedio de $ $ 112,297.9$

Variables cualitativas

Definimos las variables cualitativas como una variable estadística que describe las cualidades, no numérica, por lo que definimos las siguientes variables:

```
experiencia = table(df$experience_level)
tipo_empleo = table(df$employment_type)
titulo = table(df$job_title)
lugar_trabajador = table(df$employee_residence)
tamaño_empresa = table(df$company_size)
lugar_empresa = table(df$company_location)
```

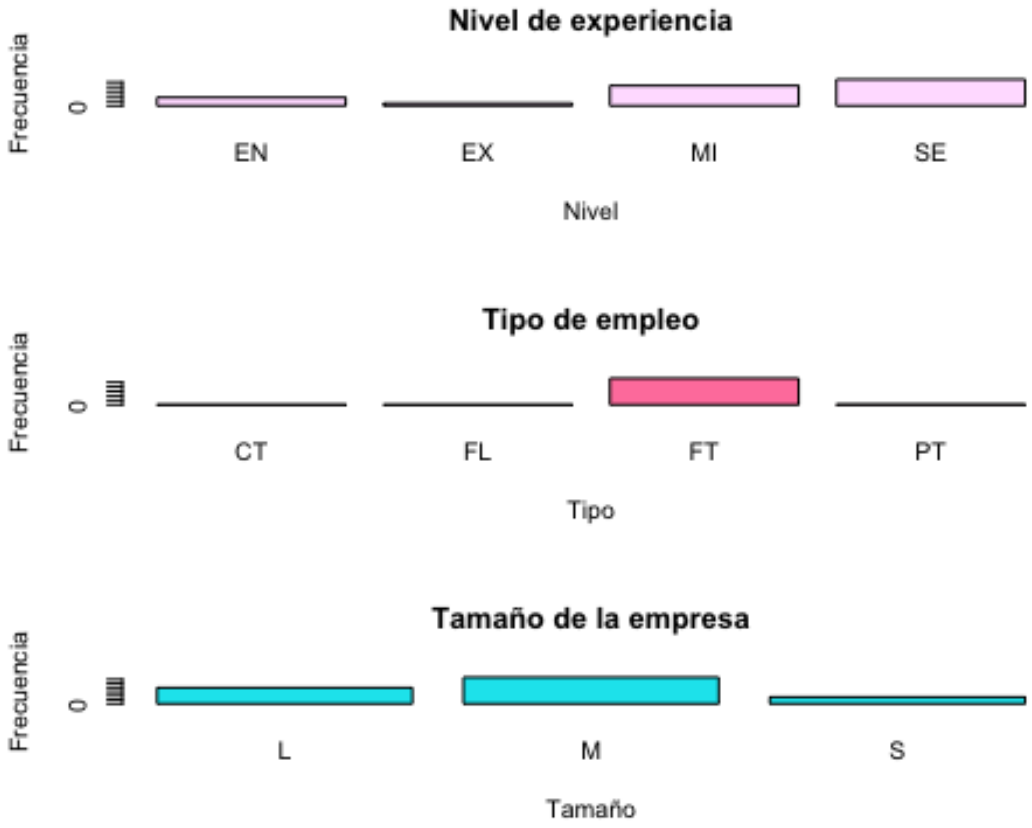
Para la moda, vamos a definir la siguiente función para las variables categóricas

```
moda <- function(x)
  {
   uniqv <- unique(x)
   uniqv[which.max(tabulate(match(x, uniqv)))]
   }
```

Distribución de frecuencia

Bar Plot

```
par(mfrow=c(3,1))

barplot(experiencia, main="Nivel de experiencia", col="thistle1",
xlab="Nivel", ylab="Frecuencia")
barplot(tipo_empleo, main="Tipo de empleo", col="palevioletred1",
xlab="Tipo", ylab="Frecuencia")
barplot(tamaño_empresa, main="Tamaño de la empresa", col="turquoise2",
xlab="Tamaño", ylab="Frecuencia")
```

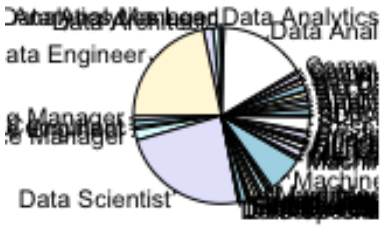


En estos primeros gráficos, es decir los de barra, podemos ver una distribución muy arreglada de los datos con su frecuencia, esto se debe a que estas columnas contienen pocas categorías por cada variable.

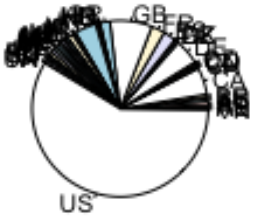
Pie Chart

```
par(mfrow=c(1,3))

pie(titulo, radius=1, main="Titulo del trabajo")
pie(lugar_empresa, radius=1, main="Lugar de la empresa")
pie(lugar_trabajador, radius=1, main="Lugar del trabajador")
```

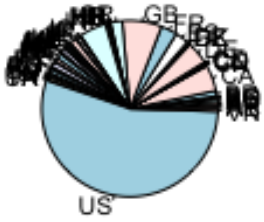
Titulo del trabajo



Lugar de la empresa



Lugar del trabajador



En estos gráficos de pastel, podemos observar como son demasiadas variables categóricas que ni podemos distinguir cual es cual. Por lo que después de este análisis más visual de los datos, se optó por limpiar nuestro set de datos.

Moda

```
moda_experiencia = moda(df$experience_level)
cat("La moda del 'Nivel de experiencia' es:", moda_experiencia, "\n")

## La moda del 'Nivel de experiencia' es: SE

moda_tipo_empleo = moda(df$employment_type)
cat("La moda del 'Tipo de empleo' es:", moda_tipo_empleo, "\n")

## La moda del 'Tipo de empleo' es: FT

moda_titulo = moda(df$job_title)
cat("La moda del 'Titulo de trabajo' es:", moda_titulo, "\n")

## La moda del 'Titulo de trabajo' es: Data Scientist

moda_lugar_trabajador = moda(df$employee_residence)
cat("La moda del 'Lugar del trabajador' es:", moda_lugar_trabajador,
"\n")
```

```
## La moda del 'Lugar del trabajador' es: US

moda_tamaño_empresa = moda(df$company_size)
cat("La moda del 'Tamaño de la empresa' es:", moda_tamaño_empresa, "\n")

## La moda del 'Tamaño de la empresa' es: M

moda_lugar_empresa = moda(df$company_location)
cat("La moda del 'Lugar de la compañia' es:", moda_lugar_empresa, "\n")

## La moda del 'Lugar de la compañia' es: US
```

Investigación y limpieza de datos

Con los datos anteriores, podemos enfocarnos solamente en los datos que nos interesan.

Investigando un poco, nos dimos cuenta que las variables que más afectan son "job_title" y "company_size". Ya que con estas características, las demás variables pueden ser hasta un poco repetitivas.

Unos ejemplos son "salary" y "salary_currency", ya que contamos con una característica que nos da el condensado de ambas variables en dólares, y así nos evitamos distintas monedas, cambios, etc. Por otra parte está el "experience_level" que va de la mano con "job_title" ya que si una persona tiene más experiencia, entonces su título de trabajo va a cambiar a manager o director. También podemos hacer un análisis de cómo fue cambiando el sueldo en los distintos "work_year" para poder visualizar una tendencia, si es que existe.

Dicho esto, procedemos a hacer una limpia de nuestro data frame y nos quedamos con las variables de nuestro interés.

```
df_clean = subset(df, select = c(job_title, salary_in_usd, company_size,
work_year))
head(df_clean)

##                    job_title salary_in_usd company_size work_year
## 1             Data Scientist         79833            L      2020
## 2 Machine Learning Scientist        260000            S      2020
## 3          Big Data Engineer        109024            M      2020
## 4       Product Data Analyst         20000            S      2020
## 5  Machine Learning Engineer        150000            L      2020
## 6               Data Analyst         72000            L      2020
```

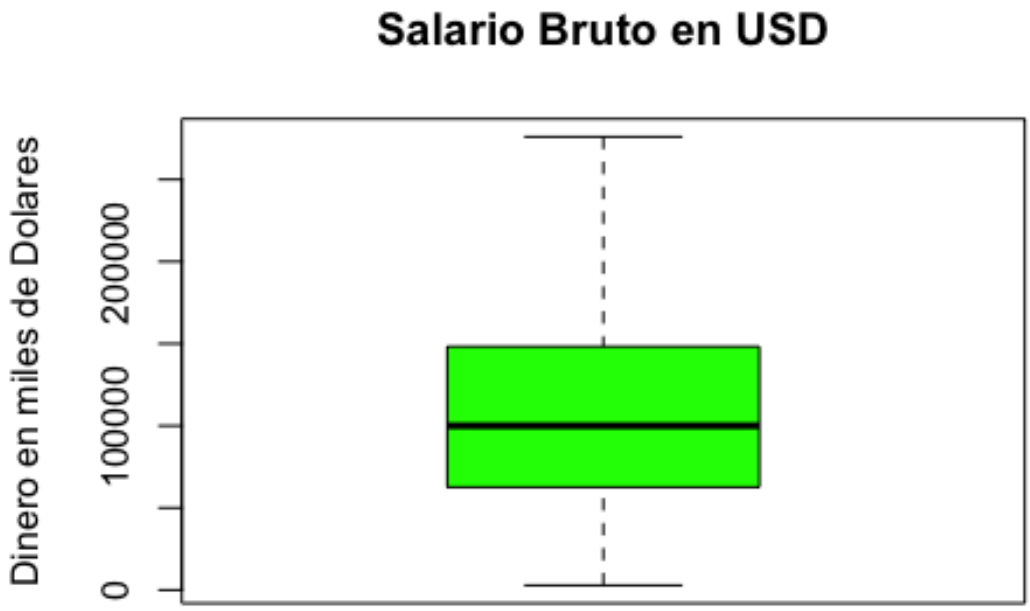
Como nos podemos acordar por los boxplots de antes, la variable "salary_in_usd" tenía algunos outliers que nos podrían hacer ruido en nuestra graficación e interpretación de los datos, por lo que optamos por quitar esos datos atípicos.

```
q1 = quantile(df_clean$salary_in_usd, 0.25)  # Cuantil 1
q3 = quantile(df_clean$salary_in_usd, 0.75) # Cuantil 3
```

```
ri = q3 - q1 # Rango intercuartilico
q_lim = q3 + 1.5*ri  # Datos atipicos superiores

df_clean = subset(df_clean, df_clean$salary_in_usd < q_lim)
```

Verificamos si ya no tenemos más outliers en “salary_in_usd”.

```
boxplot(df_clean$salary_in_usd, main="Salario Bruto en USD", ylab="Dinero
en miles de Dolares", col="green")
```



Con este nuevo data frame podemos separar los datos por año, para así graficar cada año, por lo que primero lo separamos.

```
df_2020 = subset(df_clean, df_clean$work_year == 2020)
df_2021 = subset(df_clean, df_clean$work_year == 2021)
df_2022 = subset(df_clean, df_clean$work_year == 2022)
```

Con los datos limpios y divididos por año, podemos graficar para ver una visualización más concreta de su comportamiento. Para esto vamos a utilizar la librería de GGPLOT, la cual simplifica el trabajo gráfico.

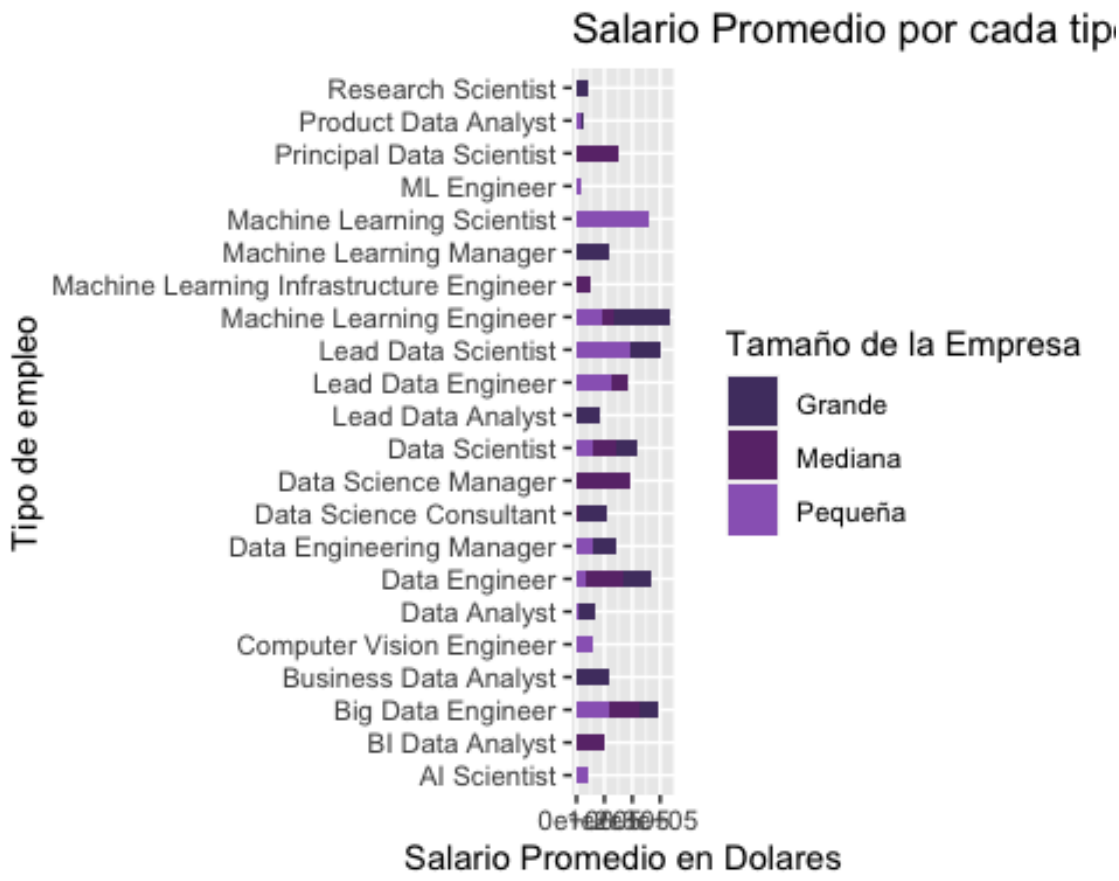
```
library(ggplot2)
```

Iniciada la librería, podemos graficar los datos de la mejor manera para su entendimiento simple.Primero vamos a definir todas las gráficas para después poder llamarlas.

```
gp_2020 = ggplot(data=df_2020) + geom_bar(aes(x = salary_in_usd, y =
job_title, fill = company_size), fun = "mean", stat = "summary",
width=0.5) + scale_fill_manual(values = c("#513D70", "#6B3479",
"#9A67C0"), labels = c("Grande", "Mediana", "Pequeña")) + guides(fill =
guide_legend(title = "Tamaño de la Empresa")) + labs(title="Salario
Promedio por cada tipo de Empleo 2020", x = "Salario Promedio en
Dolares", y = "Tipo de empleo")

gp_2021 = ggplot(data=df_2021) + geom_bar(aes(x = salary_in_usd, y =
job_title, fill = company_size), fun = "mean", stat = "summary",
width=0.5) + scale_fill_manual(values = c("#2F4F4F", "#87CEFA",
"#98F5FF"), labels = c("Grande", "Mediana", "Pequeña")) + guides(fill =
guide_legend(title = "Tamaño de la Empresa")) + labs(title="Salario
Promedio por cada tipo de Empleo 2021", x = "Salario Promedio en
Dolares", y = "Tipo de empleo")

gp_2022 = ggplot(data=df_2022) + geom_bar(aes(x = salary_in_usd, y =
job_title, fill = company_size), fun = "mean", stat = "summary",
width=0.5) + scale_fill_manual(values = c("#008B45", "#548B54",
"#98FB98"), labels = c("Grande", "Mediana", "Pequeña")) + guides(fill =
guide_legend(title = "Tamaño de la Empresa")) + labs(title="Salario
Promedio por cada tipo de Empleo 2022", x = "Salario Promedio en
Dolares", y = "Tipo de empleo")
```

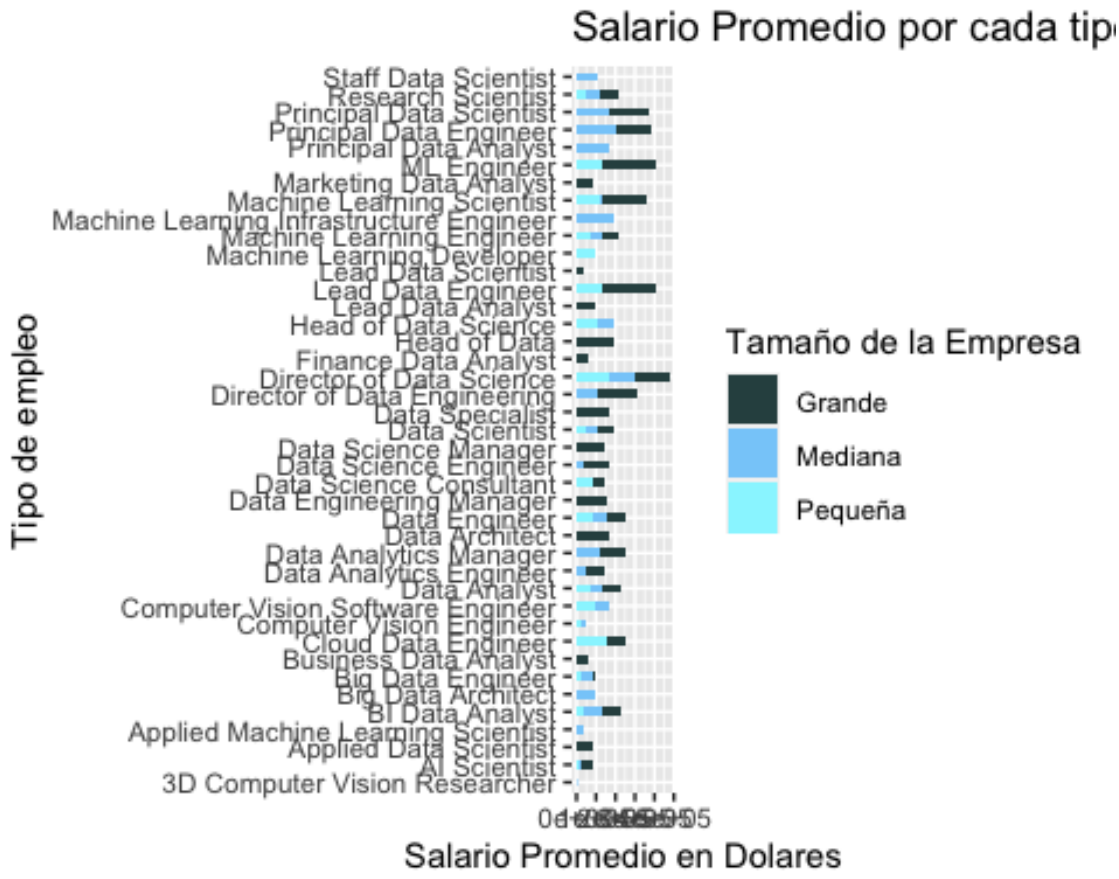
Gráfica del 2020

```
gp_2020
```



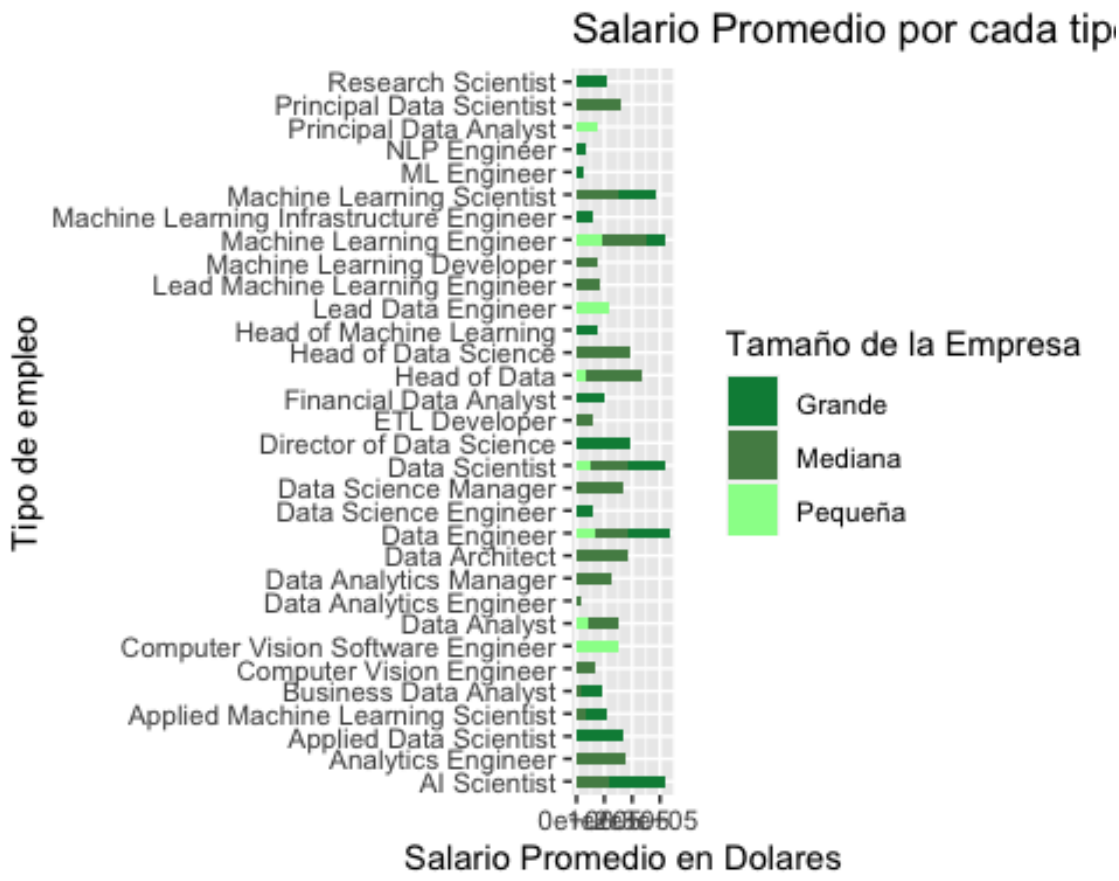
Gráfica del 2021

```
gp_2021
```



Gráfica del 2022

```
gp_2022
```



Con estas gráficas, nosotros podemos ver que del 2020 al 2022 se han tanto creado como eliminado distintos tipos de empleos, lo cual nos ayuda a tomar una decisión más informada sobre la expectativa de salario como un científico de datos.

A partir de este resultado, vamos a tomar la decisión de irnos con los datos más nuevos, es decir con los datos del 2022, esto más que nada por la creación y eliminación de nuevos y viejos empleos, el cual personalmente, es un factor que influye en la toma de decisiones.

Dicho esto, con nuestro dataframe del 2022, procedemos a sacar las variables y gráficas más significativas.

Análisis de los datos del 2022

Media y Desviación estándar

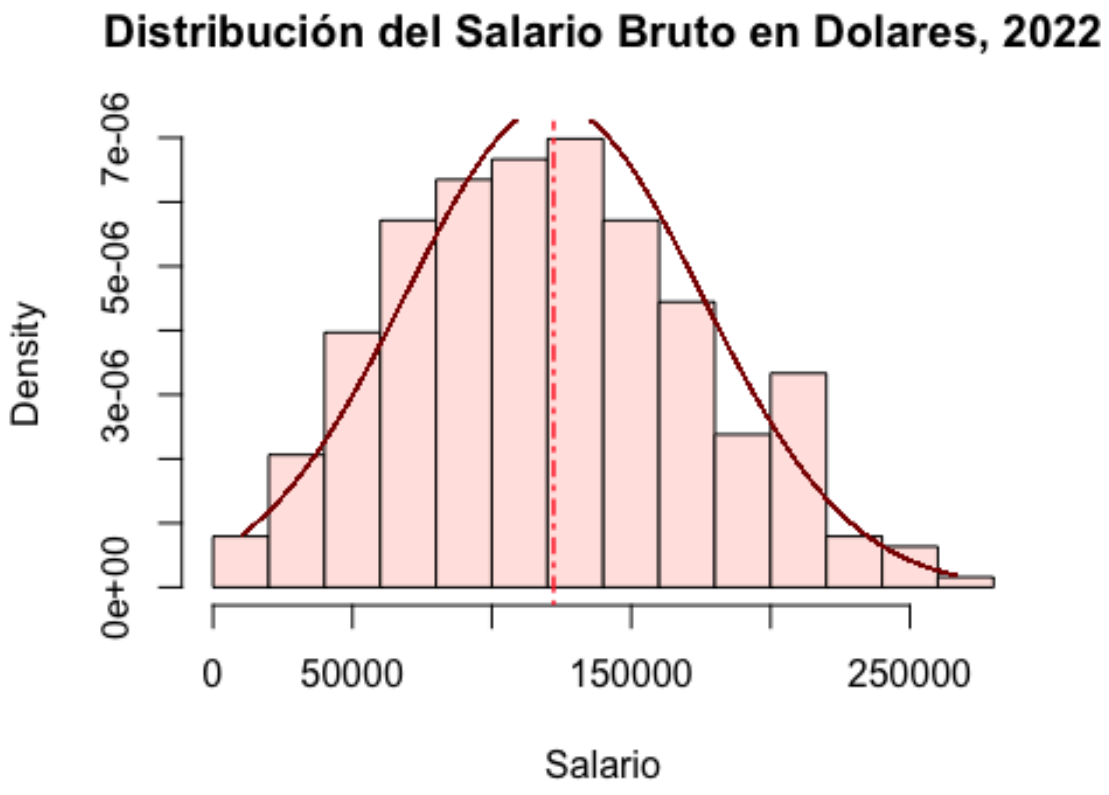
```
mu = mean(df_2022$salary_in_usd)
cat("La media de 'Salary in USD' en el 2022 es de:", mu, "\n")

## La media de 'Salary in USD' en el 2022 es de: 122187.3
```

```r
dv = sd(df_2022$salary_in_usd)
cat("La desviación estandar de 'Salary in USD' en el 2022 es de:", dv,
"\n")

## La desviación estandar de 'Salary in USD' en el 2022 es de: 53170.39
```

Ahora, vamos a graficar si histograma para visualizar su distribución normal.

Histograma

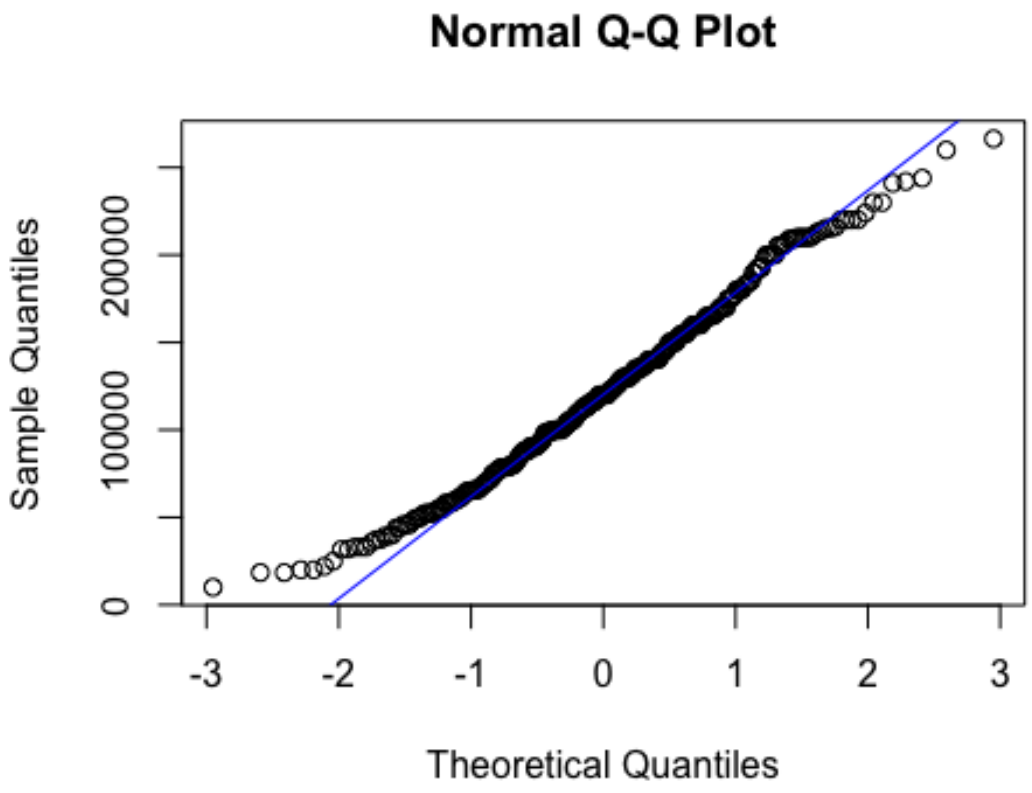
```r
hist(df_2022$salary_in_usd, prob=TRUE, col="mistyrose1",
main="Distribución del Salario Bruto en Dolares, 2022", xlab="Salario")
x = seq(min(df_2022$salary_in_usd), max(df_2022$salary_in_usd), 0.1)
y = dnorm(x, mean(df_2022$salary_in_usd), sd(df_2022$salary_in_usd))
lines(x, y, col="darkred")
abline(v=mean(df_2022$salary_in_usd), col="firebrick1", lty = 4, lwd=1.5)
```



Vemos que los datos siguen la tendencia de una distribución normal, aunque con la ayuda de Q-Q plot podremos inferir mejor por que no se adapta de la mejor manera, además de el histograma, observamos que la línea punteada es el valor de la media y la línea roja oscura es la representación de la distribución de los datos del salario en USD en el 2022.

Gráfica de Normalidad (Q-Q Plot)

```
qqnorm(df_2022$salary_in_usd)
qqline(df_2022$salary_in_usd, col="blue")
```



Gracias a la gráfica de Q-Q Plot podemos observar que nuestros datos de "salary_in_usd" en el 2022 se comporta como una distribución con colas delgadas es decir, alta curtosis y una distribución leptocúrtica.

Intervalos de confianza

Salary in USD

Por último, con los datos anteriores, vamos a sacar un estimado con la distancia z, para poder saber que tan alejados estamos de la media en base a un salario diferente. Esto lo vamos a lograr con la ayuda de la librería *BSDA*.

```
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange

z.test(df_2022$salary_in_usd, conf.level=0.95,
sigma.x=sd(df_2022$salary_in_usd), mu=mu)

##
##  One-sample z-Test
##
## data:  df_2022$salary_in_usd
## z = 0, p-value = 1
## alternative hypothesis: true mean is not equal to 122187.3
## 95 percent confidence interval:
##  116315.6 128059.0
## sample estimates:
## mean of x
##  122187.3
```

Lo que a nosotros nos interesa del z test son los intervalos de confianza, que en este caso son Confidence Intervals in Salary_USD = [116,315.6 128,059.0]

Lo que quiere decir que el rango de salario de un "científico de datos" ronda entre los 116,315.6 y 128,059.0, con lo cual podemos hacer una interpretación más educada.

Salario en USD vs Tamaño de empresa

Análisis

```
m = tapply(df_2022$salary_in_usd, df_2022$company_size, mean)
cat("Media de los salarios en los distintos tamaños de empresa:", m,
"\n")

## Media de los salarios en los distintos tamaños de empresa: 119249.1
124962.9 77046.54

s = tapply(df_2022$salary_in_usd, df_2022$company_size, sd)
cat("Desviación estandar de los salarios en los distintos tamaños de
empresa:", s, "\n")

## Desviación estandar de los salarios en los distintos tamaños de
empresa: 59162.4 51835.73 38510.28

n = tapply(df_2022$salary_in_usd, df_2022$company_size, length)
cat("Tamaño de la muestra de los salarios en los distintos tamaños de
empresa:", n, "\n")

## Tamaño de la muestra de los salarios en los distintos tamaños de
empresa: 44 258 13

sm = s/sqrt(n)
E = abs(qt(0.025,n-1))*sm
```

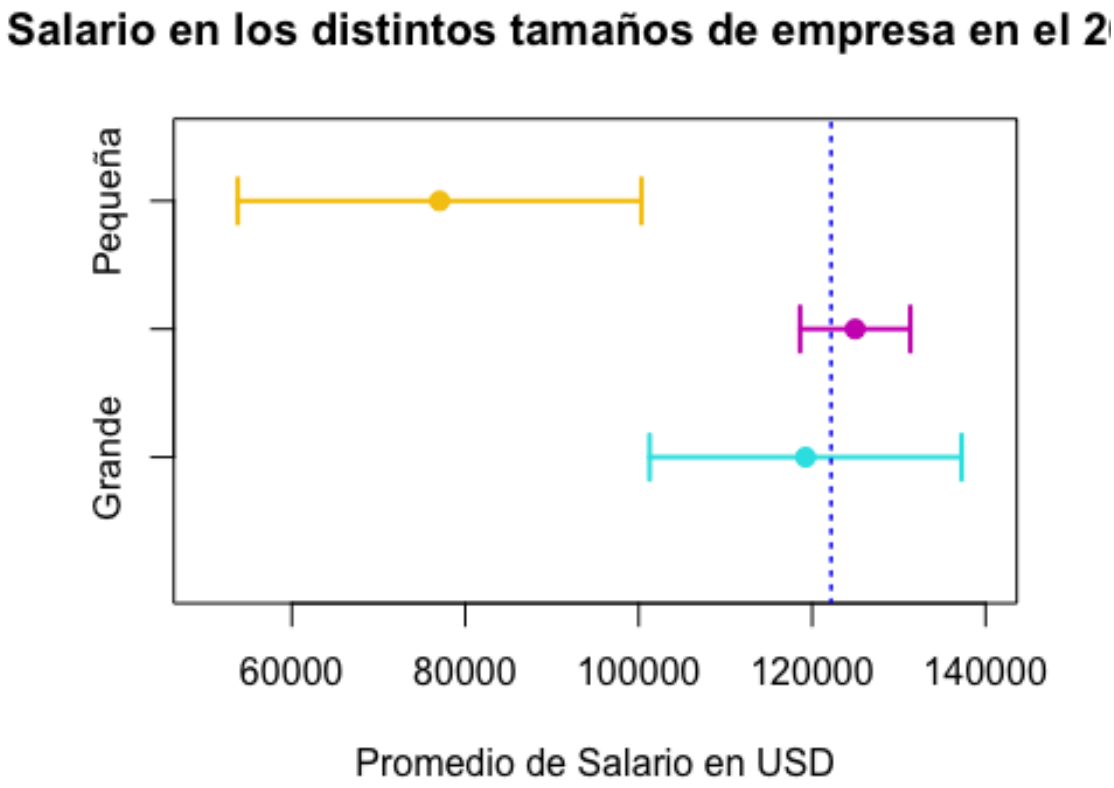
```r
In = m - E
cat("Intervalos de confianza inferiores de los salarios en los distintos tamaños de empresa:", In, "\n")
```

```
## Intervalos de confianza inferiores de los salarios en los distintos tamaños de empresa: 101262.1 118607.9 53775.01
```

```r
Sup = m + E
cat("Intervalos de confianza superiores de los salarios en los distintos tamaños de empresa:", Sup, "\n")
```

```
## Intervalos de confianza superiores de los salarios en los distintos tamaños de empresa: 137236.1 131317.9 100318.1
```

```r
plot(0, ylim=c(0,3.5), xlim=c(50000, 140000), yaxt="n", ylab="", xlab="Promedio de Salario en USD", main="Salario en los distintos tamaños de empresa en el 2022")
axis(2, at=c(1:3), labels=c("Grande", "Mediana", "Pequeña"))

for(i in 1:3)
  {
arrows(In[i], i, Sup[i], i, angle=90, code=3, length = 0.1, lwd = 2, col=i+4)
points(m[i], i, pch=19, cex=1.1, col=i+4)
}

abline(v=mean(df_2022$salary_in_usd), lty=3, col="blue", lwd=1.5)
```



Salario en USD vs Año Laboral

Análisis

```
m = tapply(df_clean$salary_in_usd, df_clean$work_year, mean)
cat("Media de los salarios en los distintos años laborales:", m, "\n")
```

```
## Media de los salarios en los distintos años laborales: 82775.88
92860.44 122187.3
```

```
s = tapply(df_clean$salary_in_usd, df_clean$work_year, sd)
cat("Desviación estandar de los salarios en los distintos años
laborales:", s, "\n")
```

```
## Desviación estandar de los salarios en los distintos años laborales:
53887.35 61531.28 53170.39
```

```
n = tapply(df_clean$salary_in_usd, df_clean$work_year, length)
cat("Tamaño de la muestra de los salarios en los distintos años
laborales:", n, "\n")
```

```
## Tamaño de la muestra de los salarios en los distintos años laborales:
69 213 315
```

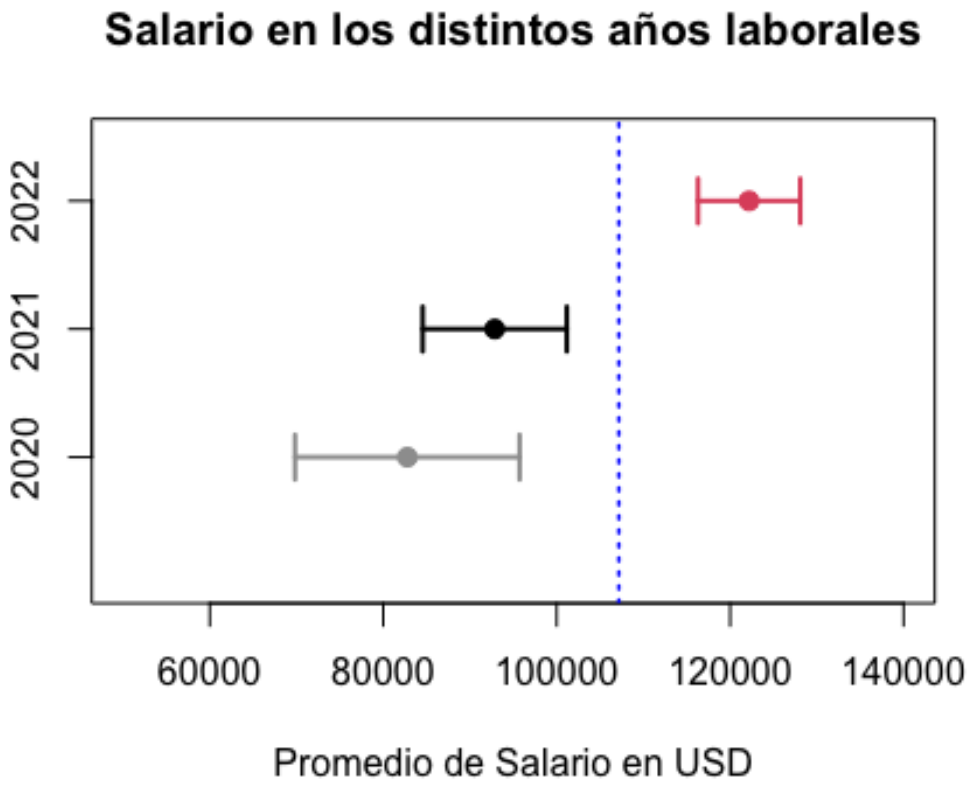
```r
sm = s/sqrt(n)
E = abs(qt(0.025,n-1))*sm

In = m - E
cat("Intervalos de confianza inferiores de los salarios en los distintos años laborales:", In, "\n")
```

```
## Intervalos de confianza inferiores de los salarios en los distintos años laborales: 69830.73 84549.68 116292.9
```

```r
Sup = m + E
cat("Intervalos de confianza superiores de los salarios en los distintos años laborales:", Sup, "\n")
```

```
## Intervalos de confianza superiores de los salarios en los distintos años laborales: 95721.04 101171.2 128081.7
```

```r
plot(0, ylim=c(0,3.5), xlim=c(50000, 140000), yaxt="n", ylab="", xlab="Promedio de Salario en USD", main="Salario en los distintos años laborales")
axis(2, at=c(1:3), labels=c("2020", "2021", "2022"))

for(i in 1:3)
  {
arrows(In[i], i, Sup[i], i, angle=90, code=3, length = 0.1, lwd = 2, col=i+7)
points(m[i], i, pch=19, cex=1.1, col=i+7)
}

abline(v=mean(df_clean$salary_in_usd), lty=3, col="blue", lwd=1.5)
```



Gracias a todo este análisis estadístico, podemos responder las siguientes preguntas.

Preguntas:

- ¿Cuál es el salario al que puede aspirar un analista de datos?

El rango de salarios de un científico de datos puede oscilar entre los 116,315.6 y 128,059.0, en el año 2022, en base a nuestras estadísticas con un intervalo de confianza del 95%, además se puede observar visualmente en la gráfica de intervalos de confianza de "Salario en los distintos años laborales".

- ¿Se han incrementado los salarios a lo largo del tiempo?

La forma más sencilla pero a la vez no tan confiable, es con los promedios de salario en los diferentes años, en donde después de limpiar nuestro dataset, esperamos que se comporten de una manera *Normal*, si tomamos esto, vemos que Promedio de Salary_USD $= \begin{bmatrix} 2020 & 2021 & 2022 \\ 82775.88 & 92860.44 & 122187.30 \end{bmatrix}$, en donde el promedio con mayor valor monetario es el del año 2022, y podemos inferir que el salario aumentó a lo largo del tiempo.

Por otra parte, viendo nuevamente el gráfico de los intervalos de confianza de "Salario en los distintos años laborales", podemos observar que efectivamente los salarios han

incrementado a lo largo del tiempo. Si tomamos como observación la gráfica pasada, podemos observar que los intervalos de confianza sobre los salarios en los años 2020 y 2021 se traslapan, por lo que podemos asumir que uno contiene a otro y que son iguales. A diferencia del año 2022, en donde podemos observar que su intervalo de confianza es mayor a la media poblacional de los salarios en dólares de los años anteriores, donde además se ve que está alejado de los demás intervalos de confianza.

- ¿Influye el tamaño de la compañía en el salario que ofrece en 2022?

Viendo la gráfica de "Salario en los distintos tamaños de empresa en el 2022", podemos afirmar que si, y que no, esto se debe por que la gráfica de los intervalos de confianza de "Salario en los distintos tamaños de empresa en el 2022" muestra que para una empresa pequeña tiene un intervalo de confianza que ocupa lugar en los salarios menores. Por otra parte, Las compañías medianas y grandes se encuentran traslapan en los intervalos de confianza, por lo que podemos asumir que son las mismas ya que una contiene a la otra.

Dicho esto, podemos afirmar que el tamaño de una empresa pequeña en 2022 si influye en el salario que ofrece, no obstante, cuando hablamos de una empresa mediana y una empresa grande, podemos afirmar con el 95% de confianza que no varía, y podríamos decir que son la misma.