Salarios

Nicolas Cardenas Valdez A01114959

2022-08-23

EL PROBLEMA

Identifica las condiciones 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. La informacion es muy variada con muchos datos atipicos por lo tanto la parte principal va a ser limpiar los datos y tratar de observar tendencias.

Resumen

Para este analisis buscamos analizar tendencias, tratar de agrupar datos, entre otras cosas. Nos basaremos principalmente en pruebas de ANOVA para encontrar las variables que mas tienen efecto sobre el salario. No necesariamente buscamos predecir el salario si no darnos una mejor idea. Como contamos con muchas variables categoricas ANOVA es nuestra mejor solucion para empezar a modelarlo.

DESCRIPCION DE LOS DATOS

Columna	Descripción
trabajo_año	El año en que se pagó el salario.
nivel de experiencia	El nivel de experiencia en el puesto durante el año con los siguientes valores posibles: EN Entry-level/Junior MI Mid-level/Intermediate SE Senior-level/Experto EX Executive-level/Director
Tipo de empleo	El tipo de empleo para el puesto: PT Tiempo parcial FT Tiempo completo CT Contrato FL Freelance
título profesional	Rol trabajado durante el año.
salario	El monto total del salario bruto pagado.
salario_moneda	La moneda del salario pagado como un código de moneda ISO 4217.
salario en usd	El salario en USD (tasa de cambio dividida por la tasa promedio de USD para el año respectivo a través de fxdata.foorilla.com).
residencia_empleado	El país de residencia principal del empleado durante el año laboral como código de país ISO 3166.
relación_remota	La cantidad total de trabajo realizado de forma remota, los valores posibles son los siguientes: 0 Sin trabajo remoto (menos del 20 %) 50 Parcialmente remoto 100 Totalmente remoto (más del 80 %)
Ubicación de la compañía	El país de la oficina principal del empleador o sucursal contratante como un código de país ISO 3166.

Columna	Descripción
tamaño de la empresa	Número promedio de personas que trabajaron para la empresa durante el año: S menos de 50 empleados (pequeño) M 50 a 250 empleados (mediano) L más de 250 empleados (grande)

EXPLORACION VARIABLES

Primero que nada determinaremos los datos que no son relevantes al analisis que queremos usar.

El primero que utilizaremos es WORK_YEAR, este es muy relevante ya que podemos ver que tanto han incrementado o declinado los salarios durante el tiempo.

```
## work_year
## 2020 2021 2022
##   72  217  318
```

Otro que es altamente relevante es el titulo o puesto.

```
## 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
```

Como podemos observar, hay muchos tipos de puestos, muchos en los cuales solo tenemos un solo dato, en el futuro buscaremos tratar de agrupar los puestos para tener una mejor idea por sector de la industria.

Tambien el nivel de que tan remoto es en conjunto con el tipo de empleo que es (FULL TIME, PART TIME, ETC.) Esto nos habla un poco mas sobre el trabajo ademas de solamente el puesto que tienen.

```
## employment_type
##  CT  FL  FT  PT
##   5   4 588  10
```

```
## remote_ratio
##   0  50 100
## 127  99 381
```

El mas evidente son las columnas de SALARIO + DIVISA, para esta informacion usaremos SALARY_IN_USD ya que nos brindara una unidad estandarizada entre todos los salarios. Si utilizaramos los de SALARIO + DIVISA tendriamos unidades en diferentes DIVISAS lo cual no seria bueno para nuestro analisis. Nos trae un estandar para el salario.

```
## [1] "Range  597141  /  Variance  5034932663.1761  /  STD  70957.2594113957"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   62726  101570  112298  150000  600000
```

La ubicacion de la empresa tambien puede ser relevante, no la descartaremos, sin embargo, no nos enforcaremos en esto porque para la mayoria de los paises solo contamos con uno o dos datos. Lo mismo va para employee

```
## 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
```

```
## 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
```

El tamano de la empresa tambien nos habla bien del salario. Esta tambien la tendremos que convertir a variable dummy.

```
## company_size
##   L   M   S
## 198 326  83
```

Problemas de datos: Primero checamos los NA en cada columna

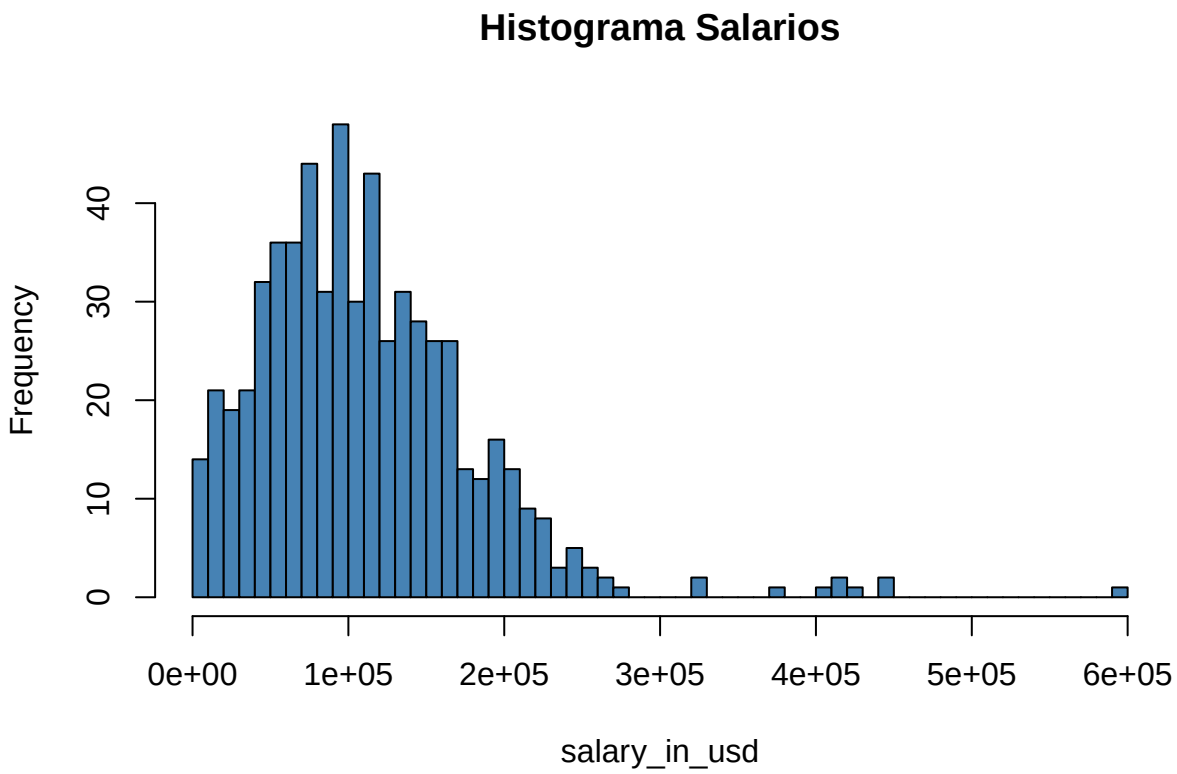
```
##                   X          work_year   experience_level    employment_type
##                   0                  0                  0                  0
##           job_title             salary    salary_currency      salary_in_usd
##                   0                  0                  0                  0
## employee_residence       remote_ratio   company_location       company_size
##                   0                  0                  0                  0
```

Vemos que no tenemos NAs y ya viendo las frecuencias de todos los valores en las columnas, vemos que no hay indiscripencias en lo datos y podemos proceder.

EXPLORACION

Primero que nada, buscaremos encontrar algunos datos generales sobre nuestra base de datos.

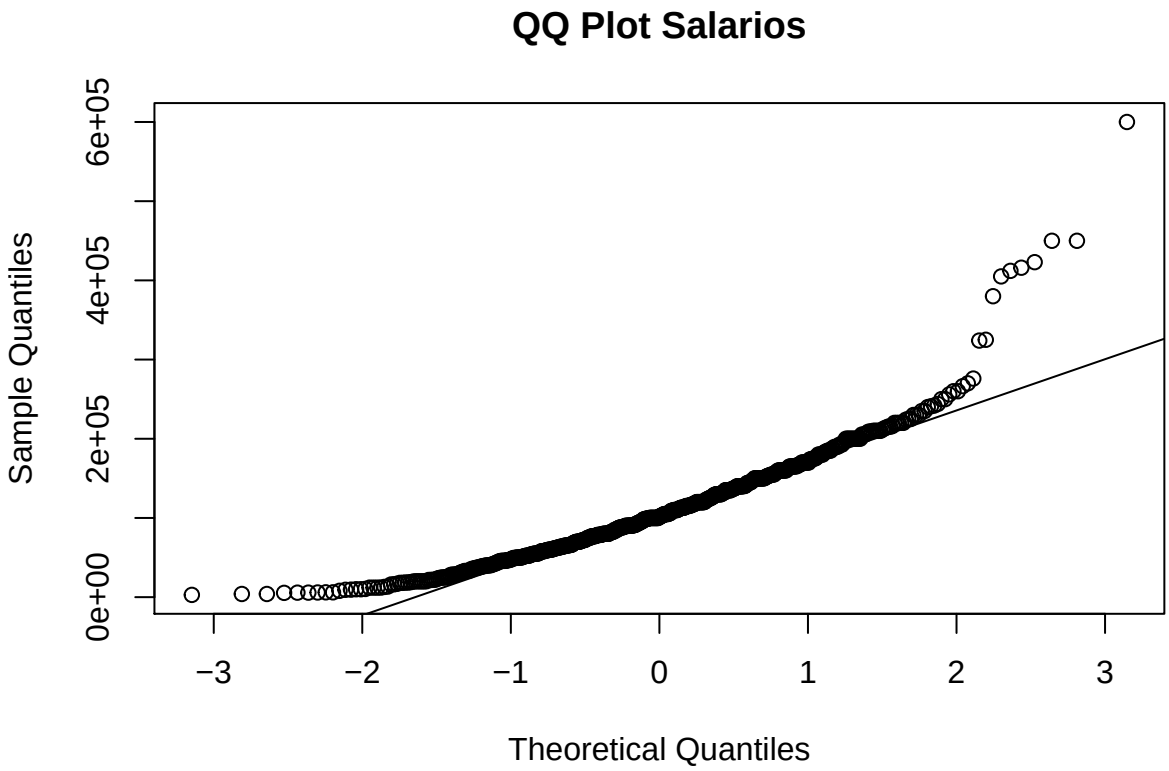
Haremos algunas pruebas de normalidad para los salarios (en USD), esto es solo con las datos que tenemos sin limpiar o filtrar por alguna categoria.



Histograma

El histograma (incluso si quitaramos los datos atipicos) muestra un sesgo hacia la derecha, lo cual no nos indica una distribucion normal

QQ Plot Salarios



Esto nos muestra la misma informacion que el histograma, que tiene un sesgo hacia la derecha y que no es una distribucion normal

SESGO

Ya que estamos hablando del sesgo, usaremos una libreria para encontrar el valor exacto.

```
## Warning: package 'e1071' was built under R version 4.0.5
```

```
## [1] 1.659312
```

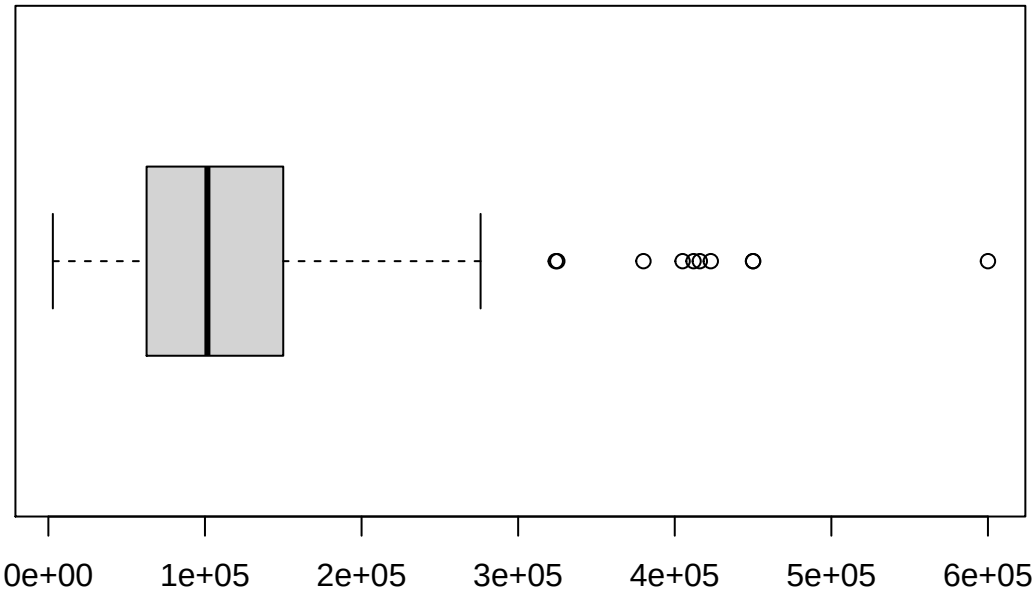
El sesgo es muy grande positivo, lo cual indica que se inclina a la derecha, que es lo que se muestra igualmente en nuestro QQplot e Histograma.

En conclusion todas nuestras pruebas indican que no es una distribucion normal. Sin embargo, debido a la cantidad de datos esto no es relevante. Lo mencionaremos a detalle en proximas secciones.

PREPARACION DE LOS DATOS

Prepararemos los datos alrededor de nuestra variable objetivo: los salarios (especificamente los salarios en usd, debido a lo que explicamos anteriormente). Ya que la mayoria de lo que nos interesa saber es como/donde provienen los salarios mas altos.

Como pudimos ver en el histograma, hay muchos datos atipicos. Para nuestro analisis, tenemos que determinar primero si queremos quitar los atipicos + extremos o solo los extremos, para esto hay que realizar un boxplot de los salarios y analizarlo

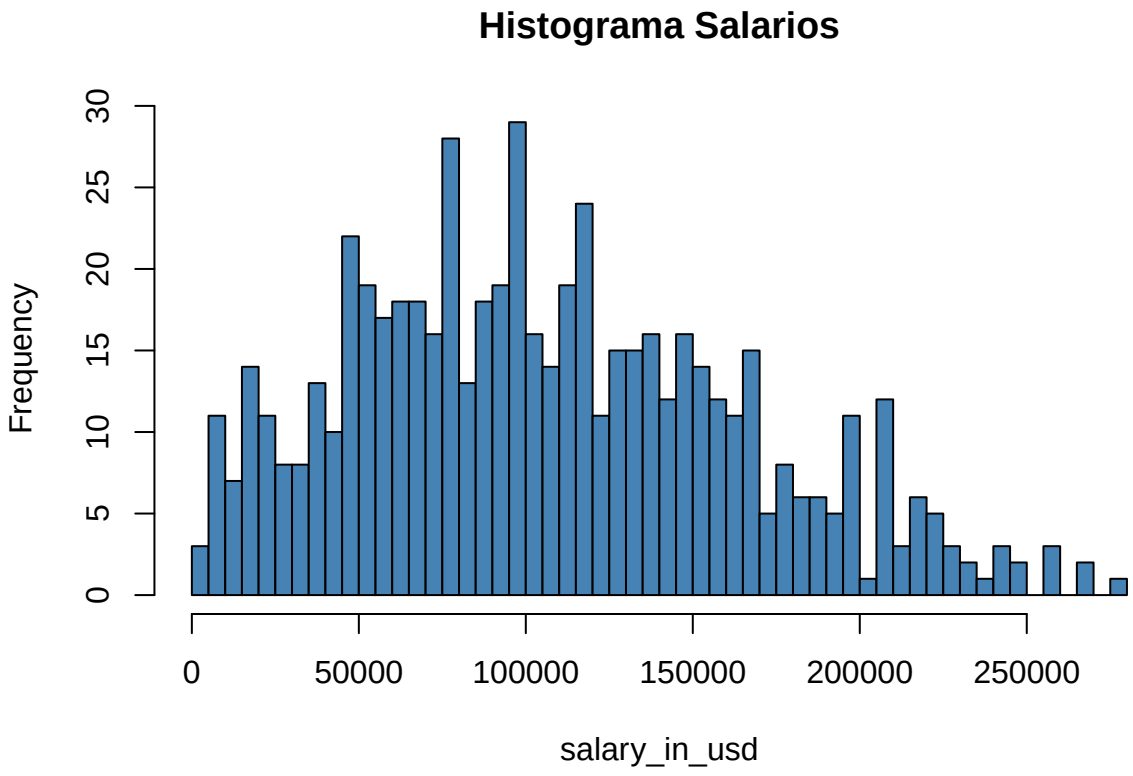


Viendo el boxplot vemos que seria buena idea quitar lo atipicos y extremos. Tiene mas sentido para tener un analisis propio que nos de una idea en lo mas comun. Hay muchos casos especiales donde hay salarios extremadamente inflados y los queremos sacar de nuestro analisis.

Primero determinaremos los rangos intercuartiles

Ya con los limites que tenemos, actualizaremos la matriz con la nueva informacion

VOLVER A CHECAR DISTRIBUCION DE X



El histograma nos muestra que sigue sin ser una distribucion normal pero como tenemos muchos datos no nos importa la distribucion de x ya que $\bar{x}$ sera distribuida normalmente

Asi quedo mucho mejor, nos da un mejor "scope" en nuestra informacion. Los datos atipicos nomas alterarian nuestro analisis.

DUMMY VARIABLES

Ya con nuestros datos atipicos limpios, debemos cambiar las columnas restantes a dummy variables. Con R este es un proceso sencillo. Para las dummys en realidad necesitamos n-1 columnas pero para hacerlo mas facil de leer y entender usaremos diferentes columnas para cada dato.

Para nuestra finalidad, como nuestra variable objetivo es el salario, podemos agrupar por estas variables sin tener que hacer las dummies, sin embargo, si en un futuro deseamos hacer algun tipo de modelo de regresion lineal o algo parecido, seran necesitadas las dummy variables.

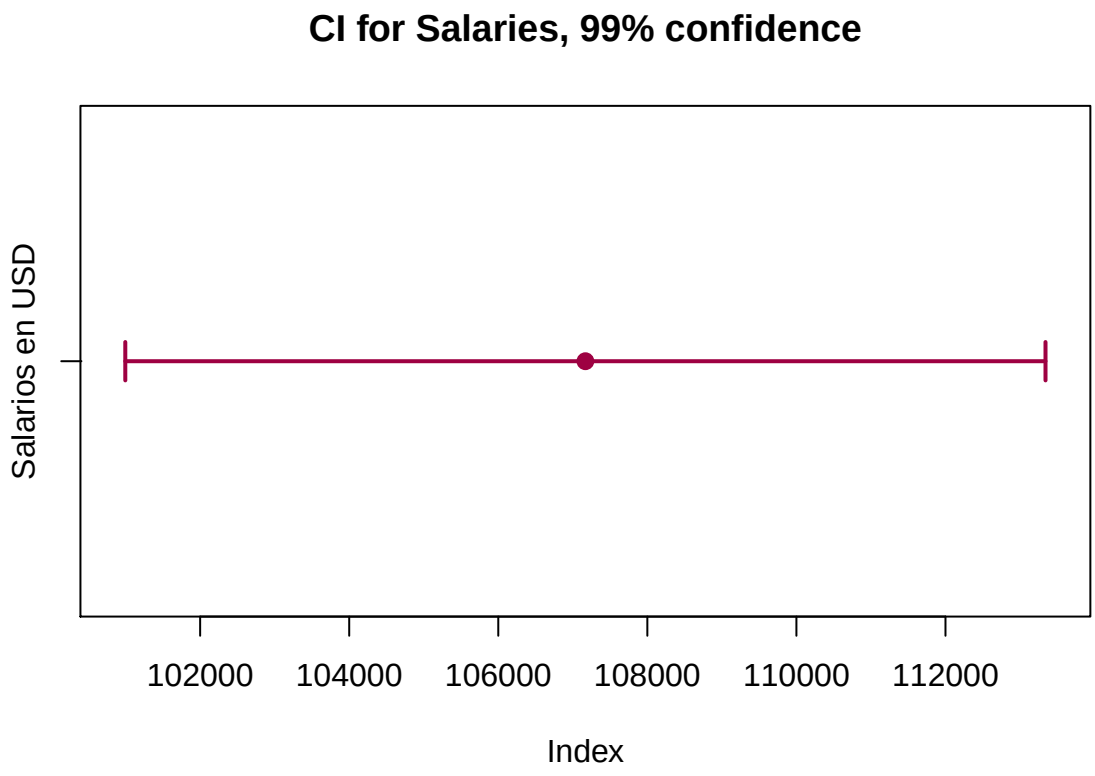
INTERVALOS DE CONFIANZA

A continuacion haremos algunas funciones para calcular los intervalos de confianza de acuerdo a nuestra funcion

```
## Warning: package 'RColorBrewer' was built under R version 4.0.5
```

Primero con nuestros salarios sin categorizar y un alpha de 0.01 (99% de confianza) que es la que estaremos utilizando para todos nuestros calculos. Estamos usando un nivel de confianza mas alto que el normal (95%) ya que este es

```
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  100995"
## [1] "Average:  107168"
## [1] "Upper Limit:  113341"
```



es un buen "fit" ya que podemos estar 99% seguros que la media de la poblacion esta entre y lo cual es un rango muy pequeno.

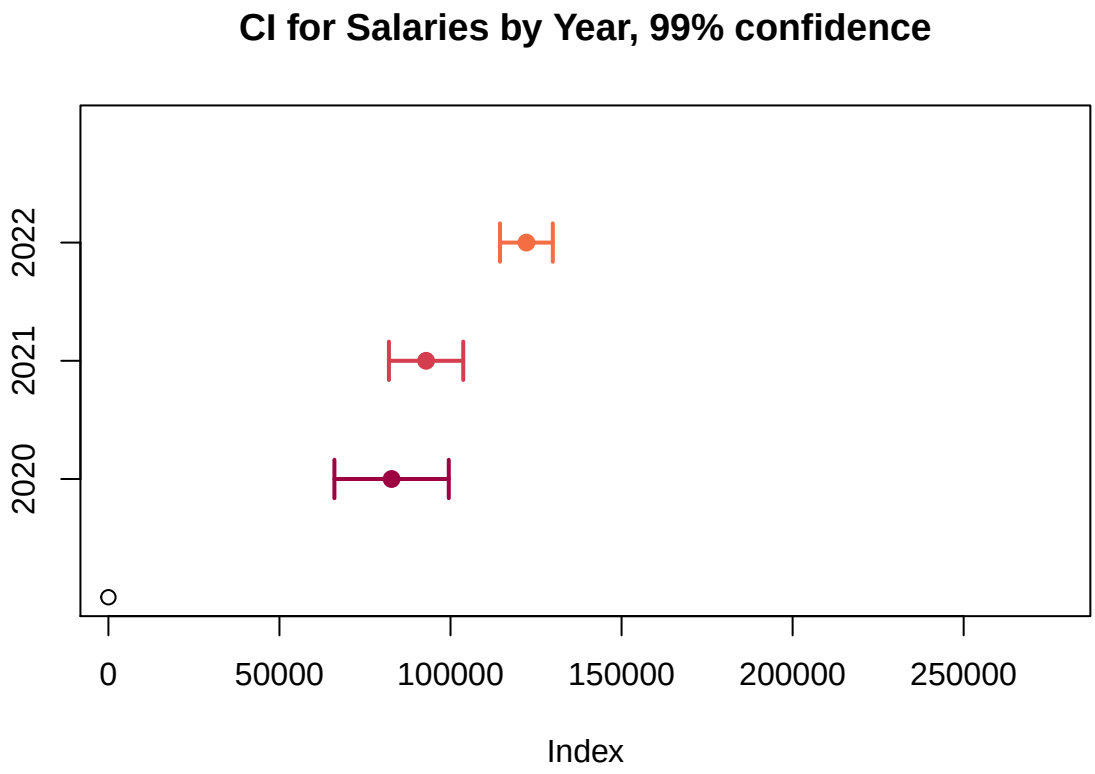
ANALISIS DE DATOS

Buscaremos tener un vistazo a los salarios de acuerdo con diferentes variables categoricas.

Para esto crearemos una funcion que nos de las estadisticas a partir de un subset de los datos:

SUMMARY DE SALARIO POR ANIO

Encontraremos la media y la desviacion estandar por ano para los salarios

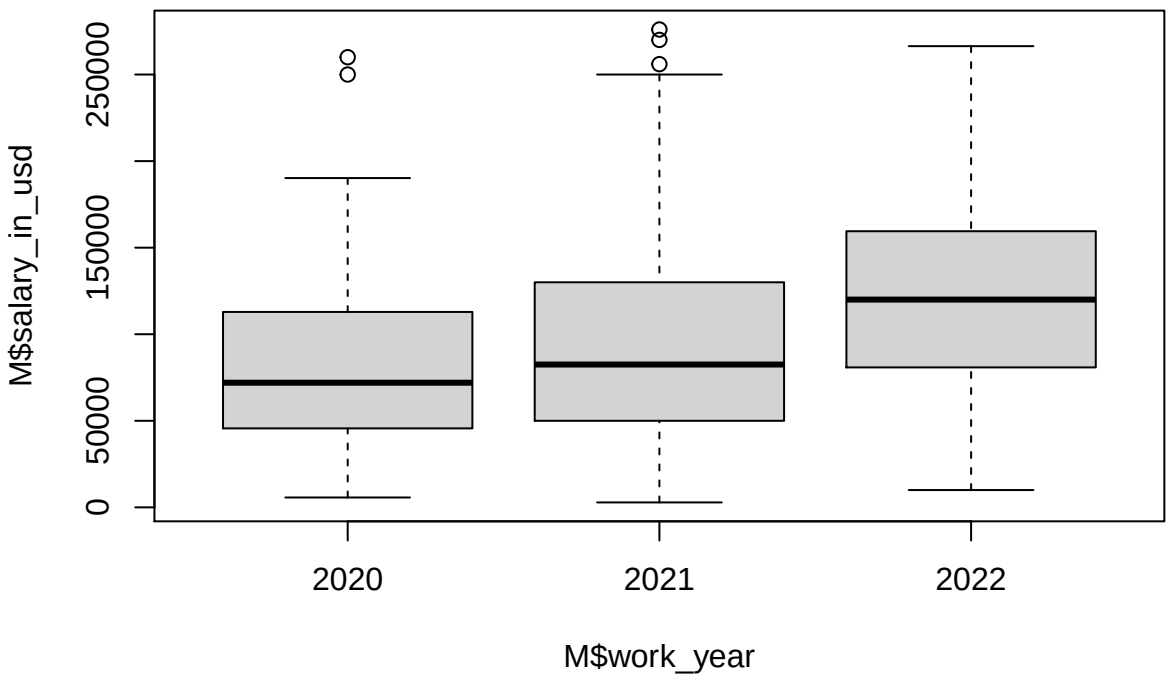



```
## [1] "Year  2020 :"
## [1] "N datos:  69  /  Range  254293  /  Variance  2903846799.54518  /  STD  53887.3528719419"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5707   45618   72000   82776  112872  260000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  66065"
## [1] "Average:  82775"
## [1] "Upper Limit:  99485"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  66065"
## [1] "Average:  82775"
## [1] "Upper Limit:  99485"
## [1] "Year  2021 :"
## [1] "N datos:  213  /  Range  273141  /  Variance  3786098734.23771  /  STD  61531.2825661688"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   50000   82500   92860  130000  276000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  82000"
## [1] "Average:  92860"
```

```
## [1] "Upper Limit:  103720"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  82000"
## [1] "Average:  92860"
## [1] "Upper Limit:  103720"
## [1] "Year  2022 :"
## [1] "N datos:  315  /  Range  256400  /  Variance  2827090363.81383  /  STD  53170.3899159469"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10000   80833  120000  122187  159500  266400
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  114470"
## [1] "Average:  122187"
## [1] "Upper Limit:  129903"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  114470"
## [1] "Average:  122187"
## [1] "Upper Limit:  129903"
```

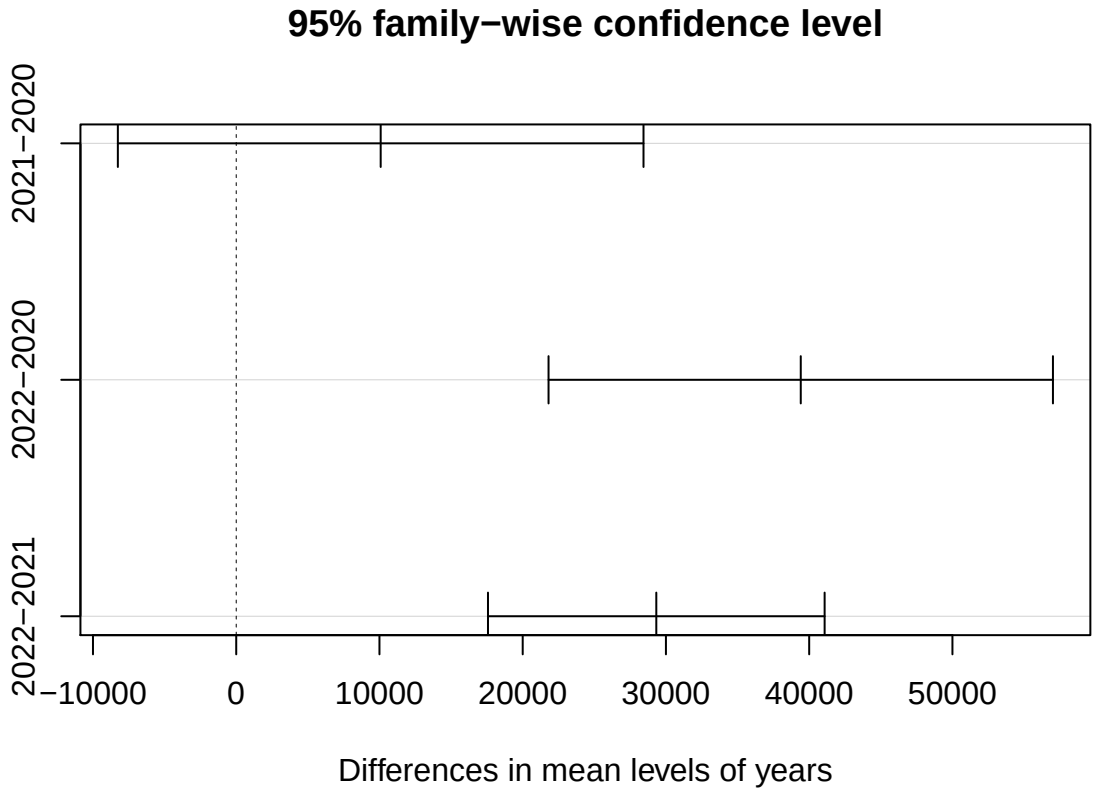
Lo que podemos ver a partir de esta informacion es que el salario ha incrementaddo durante los anios. Entre 2020 y 2021 es casi cierto que no incrementaron pero de 2021 a 2022 es donde estamos 99% seguros que si incremento. La desvicacion estandar sigue siendo aproximadamente lo mismo. La cantidad de datos es la misma con todos los anios entonces nos indica que tenemos aproximadamente el mismo nivel de precision en nuestros datos. Esto nos muestra que la demanada para analistas de datos esta subiendo, este incremento no solo es igual a la inflacion si no mas incluso durante en periodos de pandemia lo cual es una buena indicacion.

ANOVA ANIO




```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## years         2 1.557e+11 7.786e+10    24.5 5.98e-11 ***
## Residuals   594 1.888e+12 3.178e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Esto nos indica que si efectivamente si existe efecto para esta variable



Esto nos indica que el efecto entre 2020 y 2021 es igual pero entre 2022 es diferente a estas. Mostrando que incremento, al igual a nuestro otro analisis

SALARIO POR TAMANO DE LA EMPRESA

Creo que un analisis interesante podria ser el salario a partir del tamano de la empresa.

Hacemos lo mismo, pero ahora con respecto al tamano de la empresa


```
## [1] "Size  L :"
## [1] "N datos:  191  /  Range  260518  /  Variance  3828442890.25048  /  STD  61874.4122416567"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5882   59551   96282  107933  150500  276000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  96400"
## [1] "Average:  107932"
## [1] "Upper Limit:  119464"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  96400"
## [1] "Average:  107932"
## [1] "Upper Limit:  119464"
## [1] "Size  M :"
## [1] "N datos:  324  /  Range  262400  /  Variance  3057722116.02748  /  STD  55296.6736434253"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4000   77921  112900  115238  150820  266400
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  107325"
## [1] "Average:  115238"
```

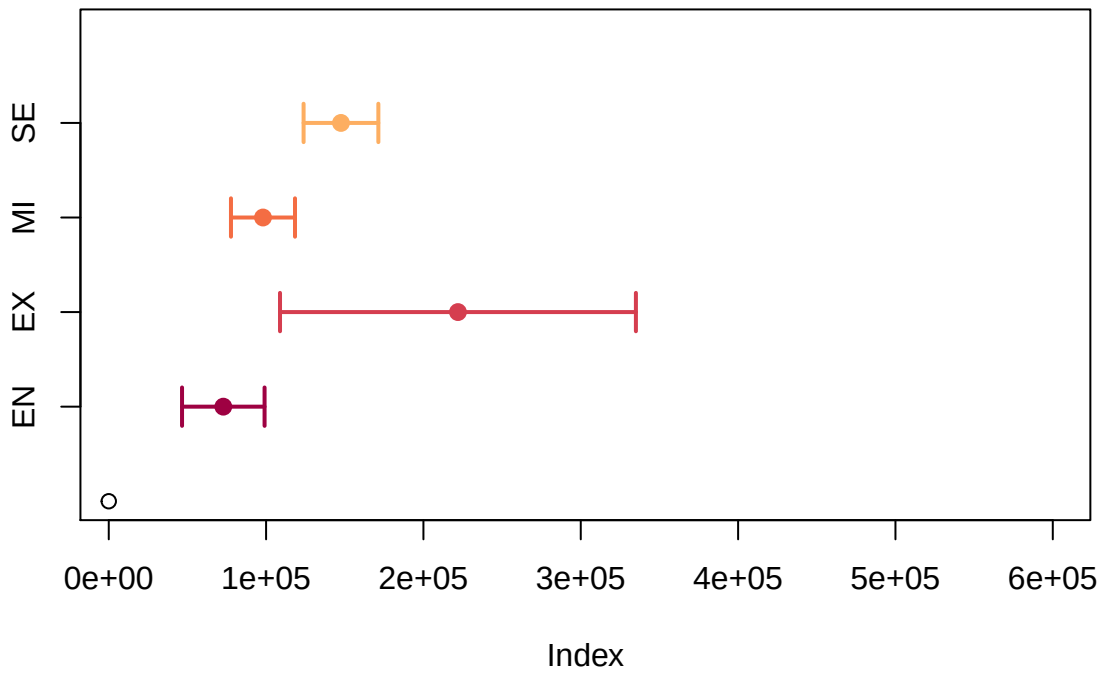
```
## [1] "Upper Limit:  123151"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  107325"
## [1] "Average:  115238"
## [1] "Upper Limit:  123151"
## [1] "Size  S :"
## [1] "N datos:  82  /  Range  263541  /  Variance  2646384389.19904  /  STD  51443.0208016504"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   41816   65000   73506   98937  260000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  58873"
## [1] "Average:  73506"
## [1] "Upper Limit:  88139"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  58873"
## [1] "Average:  73506"
## [1] "Upper Limit:  88139"
```

Como podemos observar, la media por tamano de empresa resulta ser interesante ya que la mediana fue la que resulto en lo mas alto (con la media de la muestra). Sin embargo podemos estar muy seguros que la media de la poblacion en cuanto a tamano de la empresa es la misma para empresas grandes y empresas medianas. Lo que podemos tener asegurados es que ambas de estas tienen media mas grande que la pequena. Creo que esto se puede deber a que en empresas grandes necesitan a muchas personas para poder funcionar correctamente lo cual puede disminuir el salario promedio, sin embargo, tambien hay muchas posiciones ejecutivas que pueden aumentar el salario. Para esto analizaremos el promedio de salarios en empresas grandes por nivel de experiencia para tener una mejor idea.

Primero que nada para ajustar con estos ejecutivos, hay que utilizar nuestro “dataset” con todos los datos (incluyendo atipicos)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5882   61042  100000  119243  154600  600000
```

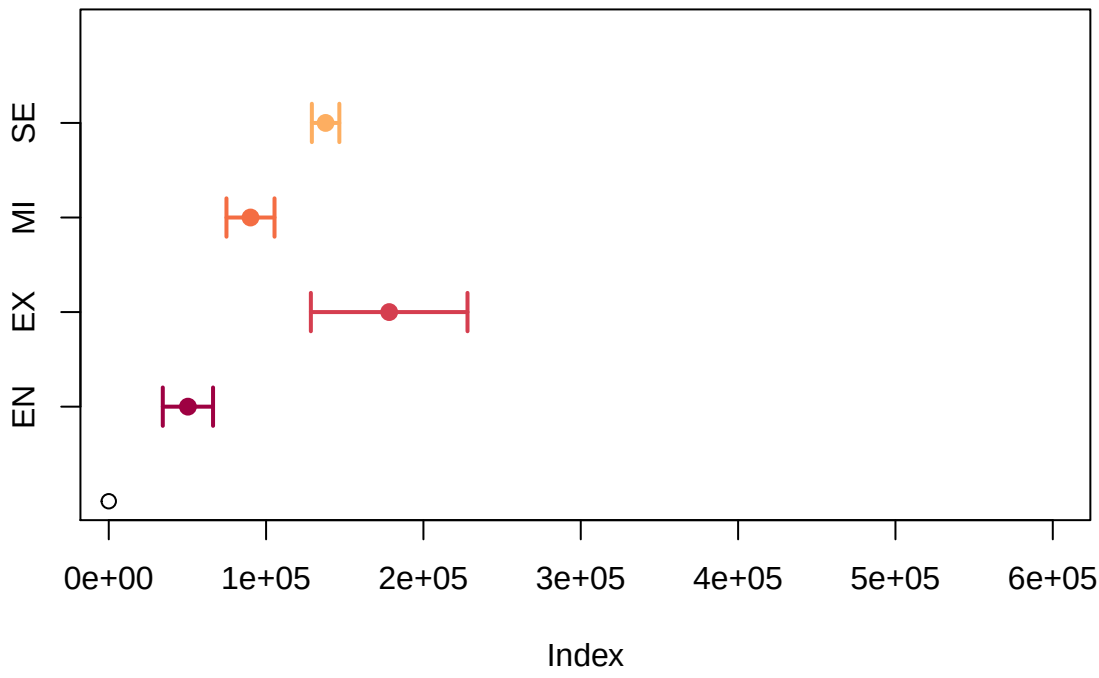
Como podemos ver ya con los datos atipicos incluidos, el promedio es mucho mas alto que las empresas medianas y pequenas lo cual indica que mueve mucho la media incluir todos estos puestos ejecutivos. Ahora analizaremos esta misma informacion (company_size = L) pero divido por nivel de experiencia.




```
## [1] "Nivel de Experiencia en Empresa Grande:  EN :"
## [1] "N datos:  29  /  Range  260518  /  Variance  3005732490.47537  /  STD  54824.5610148898"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5882   37300   63831   72813   91000  250000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  46589"
## [1] "Average:  72813"
## [1] "Upper Limit:  99036"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  46589"
## [1] "Average:  72813"
## [1] "Upper Limit:  99036"
## [1] "Nivel de Experiencia en Empresa Grande:  EX :"
## [1] "N datos:  11  /  Range  187361  /  Variance  21206337968.3636  /  STD  145623.960831876"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   79039  145923  196979  221942  242500  600000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  108844"
## [1] "Average:  221942"
```

```
## [1] "Upper Limit:  335039"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  108844"
## [1] "Average:  221942"
## [1] "Upper Limit:  335039"
## [1] "Nivel de Experiencia en Empresa Grande:  MI :"
## [1] "N datos:  86  /  Range  260328  /  Variance  5365379346.75404  /  STD  73248.7497965258"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6072   51178   86000   98030  116436  450000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  77684"
## [1] "Average:  98030"
## [1] "Upper Limit:  118375"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  77684"
## [1] "Average:  98030"
## [1] "Upper Limit:  118375"
## [1] "Nivel de Experiencia en Empresa Grande:  SE :"
## [1] "N datos:  72  /  Range  246229  /  Variance  6126121151.47868  /  STD  78269.5416588003"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20171   94640  147000  147591  185000  412000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  123831"
## [1] "Average:  147591"
## [1] "Upper Limit:  171350"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  123831"
## [1] "Average:  147591"
## [1] "Upper Limit:  171350"
```

Como podemos ver nuestra hipotesis es correcta, el salario de los ejecutivos esta incluso fuera de nuestra muestra, la mayoria son datos atipicos que no se mostrarian con nuestra limpieza de datos. Por lo que podemos ver aqui es que EN y MI son igual, pero Ejecutivo y Senior no solo son mas altos si no que mucho mas altos. Y observamos mucha variacion en cuanto al intervalo de ejecutivo. Tambien podemos deducir que la unica razon por la que las empresas grandes tienen el promedio mas grande debido a la posicion de los ejecutivos y seniors. Y solo para tener la idea completa, tambien analizaremos mas a profundidad por nivel de experiencia en empresas medianas

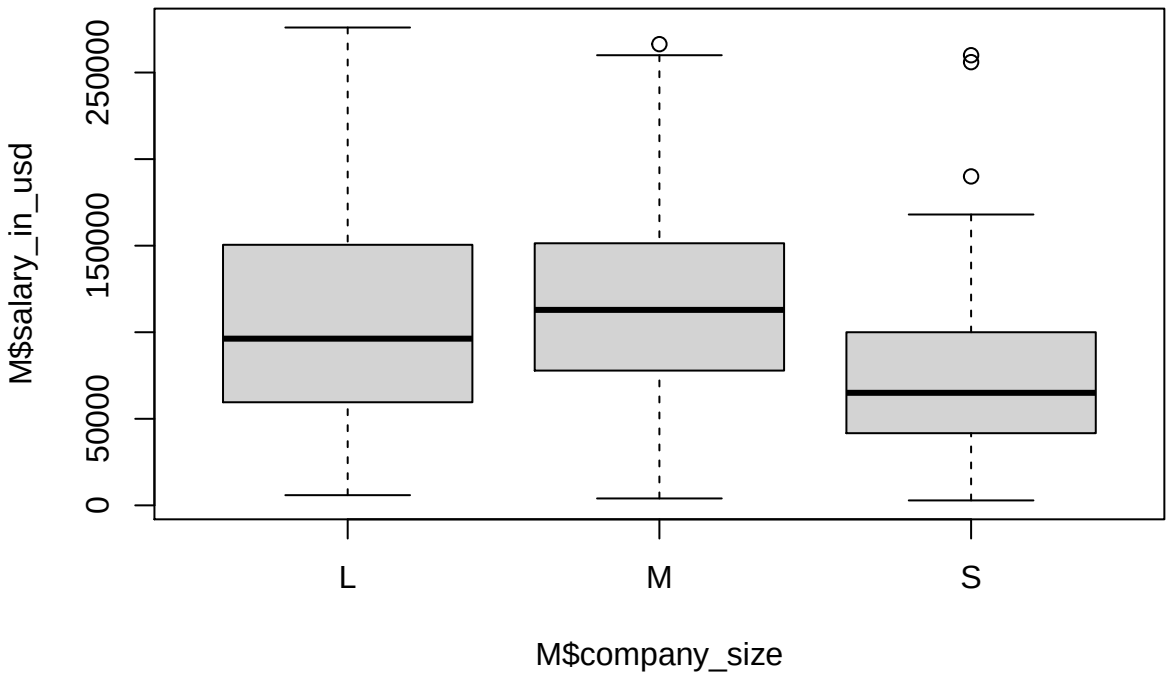



```
## [1] "Nivel de Experiencia en Empresa Grande:  EN :"
## [1] "N datos:  30  /  Range  262400  /  Variance  1154328446.92414  /  STD  33975.40944454"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4000   21689   49823   50322   69250  125000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  34343"
## [1] "Average:  50321"
## [1] "Upper Limit:  66299"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  34343"
## [1] "Average:  50321"
## [1] "Upper Limit:  66299"
## [1] "Nivel de Experiencia en Empresa Grande:  EX :"
## [1] "N datos:  12  /  Range  181400  /  Variance  4481301725.84091  /  STD  66942.52554125"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   85000  130020  171438  178242  218000  324000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  128464"
## [1] "Average:  178241"
```

```
## [1] "Upper Limit:  228018"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  128464"
## [1] "Average:  178241"
## [1] "Upper Limit:  228018"
## [1] "Nivel de Experiencia en Empresa Grande:  MI :"
## [1] "N datos:  98  /  Range  262400  /  Variance  3442414150.42626  /  STD  58672.0900465141"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4000   52351   78659   90091  117217  450000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  74824"
## [1] "Average:  90091"
## [1] "Upper Limit:  105357"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  74824"
## [1] "Average:  90091"
## [1] "Upper Limit:  105357"
## [1] "Nivel de Experiencia en Empresa Grande:  SE :"
## [1] "N datos:  186  /  Range  247493  /  Variance  2139271433.06356  /  STD  46252.2586806694"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18907  105000  135500  137816  165400  266400
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  129079"
## [1] "Average:  137815"
## [1] "Upper Limit:  146551"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  129079"
## [1] "Average:  137815"
## [1] "Upper Limit:  146551"
```

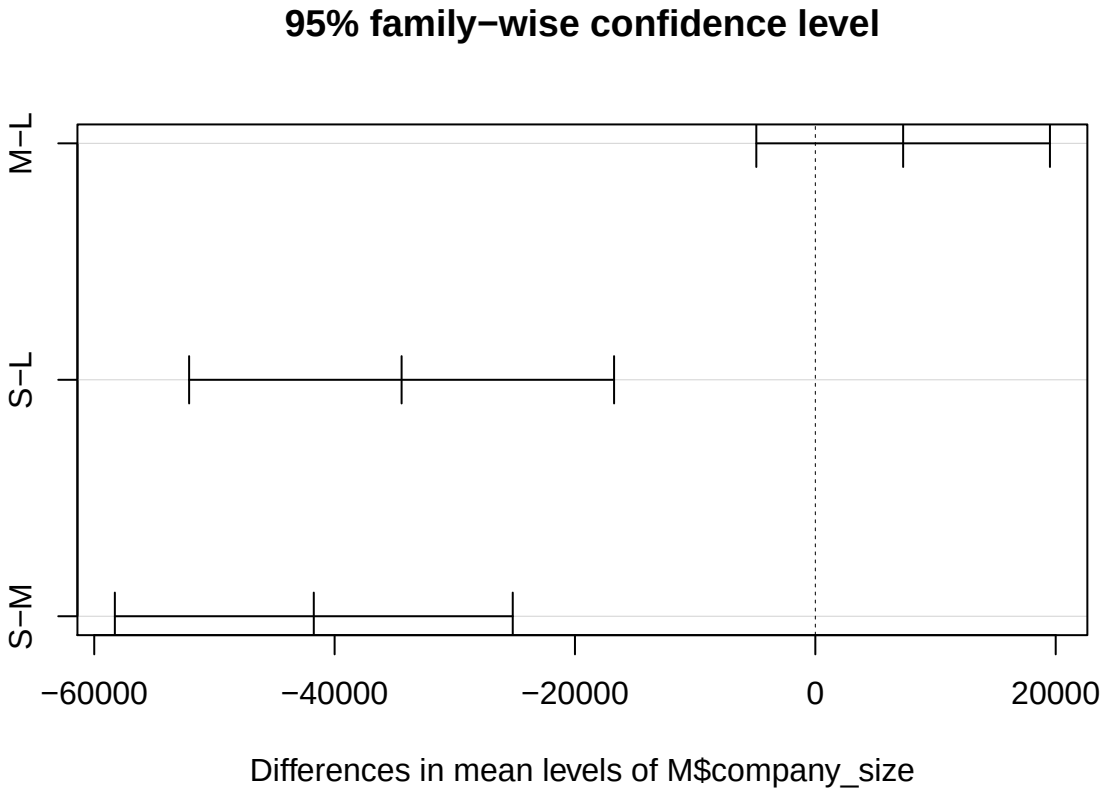
Al observar estos ultimos datos nos damos cuenta que la media de cada nivel de puesto en una empresa mediana es por el nivel de experiencia, excepto a nivel senior/ejectivo que es donde son iguales. Tambien observamos una variacion mas pequena que las empresas

ANOVA TAMANO EMPRESA




```
##                 Df    Sum Sq   Mean Sq F value   Pr(>F)
## M$company_size   2 1.141e+11 5.706e+10   17.57 3.87e-08 ***
## Residuals      594 1.929e+12 3.248e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

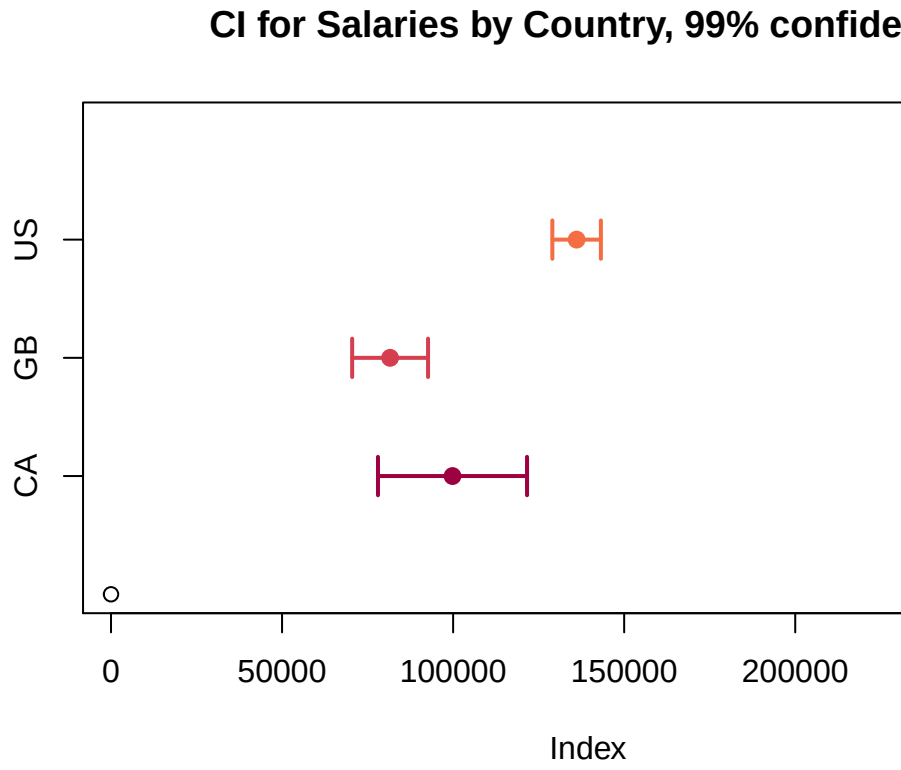
Esto nos indica que si efectivamente si existe efecto para esta variable



Esto nos indica que la diferencia entre el efecto pequena y (mediana y grande) es la misma y que el efecto de mediana y grande es igual.

SALARIO POR PAIS

Ahora analizaremos los salarios en termino de pais, pero como hay muchos paises con pocos datos, utilizare-



mos un filtro para solo mostrar los paises

```
## [1] "Pais  VN :"
## [1] "N datos:  30  /  Range  214400  /  Variance  2146385650.68506  /  STD  46329.1015527504"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   52000   69730   81896   99824  117916  225000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  78036"
## [1] "Average:  99823"
## [1] "Upper Limit:  121611"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  78036"
## [1] "Average:  99823"
## [1] "Upper Limit:  121611"
## [1] "Pais  VN :"
## [1] "N datos:  47  /  Range  229100  /  Variance  871667234.172063  /  STD  29524.0111463883"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37300   57575   78526   81583  103931  183228
##
```

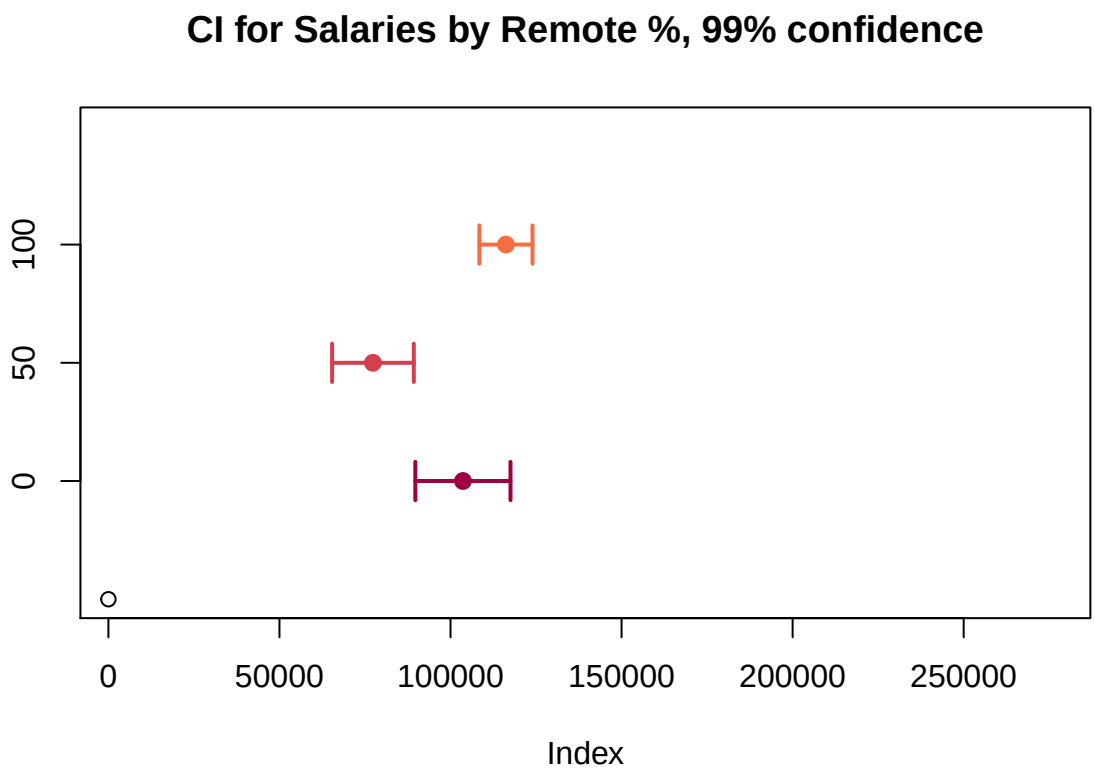
```
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  70490"
## [1] "Average:  81583"
## [1] "Upper Limit:  92675"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  70490"
## [1] "Average:  81583"
## [1] "Upper Limit:  92675"
## [1] "Pais  VN :"
## [1] "N datos:  345  /  Range  260721  /  Variance  2618096759.77803  /  STD  51167.340753434"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5679  100000  135000  136100  167000  276000
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  129004"
## [1] "Average:  136100"
## [1] "Upper Limit:  143196"
##
##
##
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  129004"
## [1] "Average:  136100"
## [1] "Upper Limit:  143196"
```

Como podemos ver hay muy pocos datos en la mayoria de los paises, y nos quedaron solo unos pocos pero estos dan buena indicacion de cual de estos paises es mejor para trabajar, la opcion mas clara es Estados Unidos ya que todo el intervalo esta arriba de los demas. CA y GB estan muy cercanos y podemos decir que son similares. Pudieramos tomar paises que tengan menos de 30 datos pero tendriamos que checar la distribucion y encontrar el modelo, lo que realizamos pero no queda en el "scope" de este analisis. (La distribucion es BETA).

No se hare ANOVA para los paises ya que es muy pequena la muestra y no queremos analizar todos los paises porque eso no seria efectivo.

SALARIO POR MODALIDAD

Por ultimo, a mi en lo personal me interesa como cambian los salarios dependiendo de que tan remoto es. Ahora con lo de la pandemia y hacia donde se esta moviendo el mundo (trabajos cada vez mas tecnologicos), se estan abriendo muchos trabajos remotos, por lo cual me interesaria saber si hay diferencia y que tanta, asi que haremos el mismo analisis pero para el tipo de modalidad.

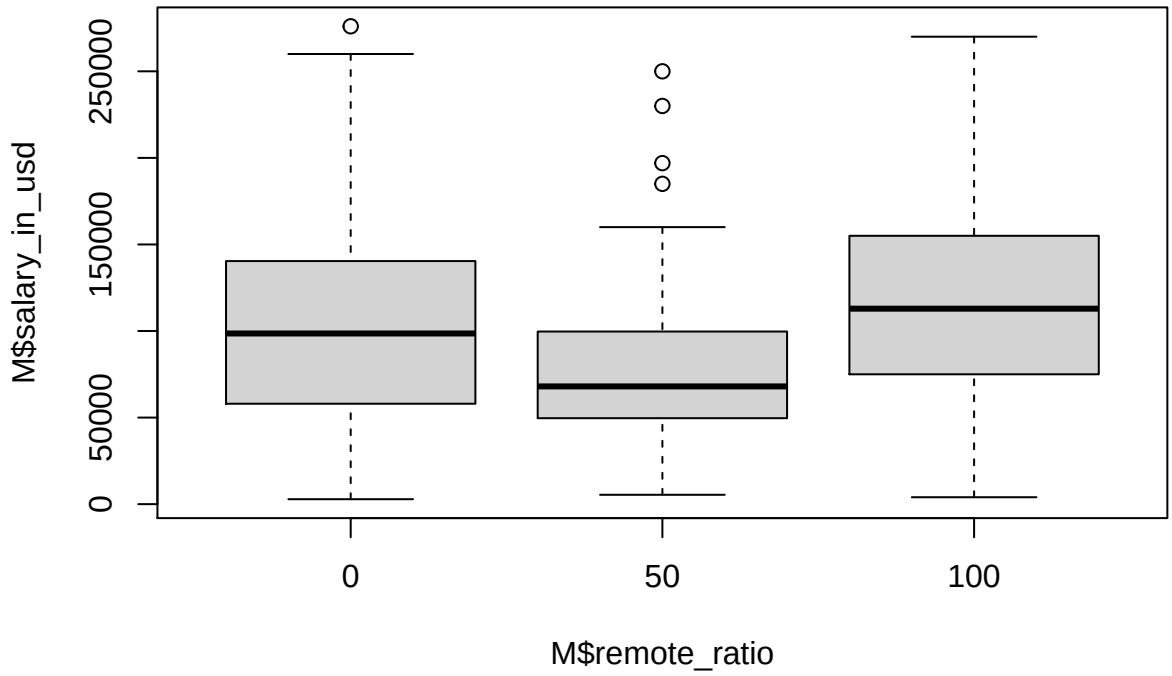



```
## [1] "% Remoto  0 : "
##
## EN EX MI SE
## 14  3 55 54
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  89723"
## [1] "Average:  103627"
## [1] "Upper Limit:  117530"
## [1] "% Remoto  50 : "
##
## EN EX MI SE
## 25  5 41 27
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  65396"
## [1] "Average:  77331"
## [1] "Upper Limit:  89266"
## [1] "% Remoto  100 : "
##
##  EN  EX  MI  SE
##  49  14 114 196
## [1] "Intervalo de  99 % de confianza:"
## [1] "Lower Limit:  108438"
## [1] "Average:  116204"
## [1] "Upper Limit:  123970"
```

Esta grafica creo que es la mas interesante de todas ya que podemos ver que la diferencia entre trabajos

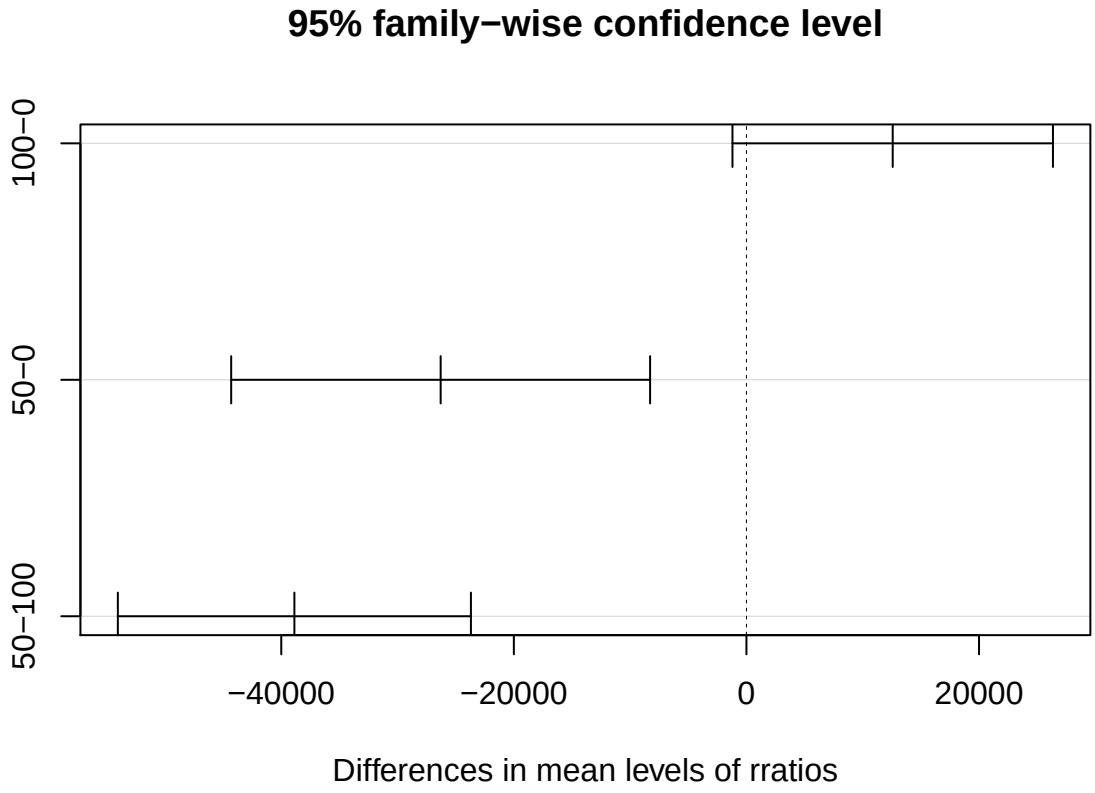
hibridos (50%) remoto, y los (100%) remotas es no solo grande pero tambien definitivamente (99% de confianza) mas alta si trabajas de manera completamente remota. Sin embargo, hay que considerar que no tenemos los datos atipicos, los cuales nos subirian la media de los que debido a que muchos de estos trabajos de ejecutivos sorprendentemente son remotos actualmente. Como podemos ver en la tabla de frecuencias el salario creo que no se debe al trabajo remoto si no mas a la posicion. Porque podemos ver que los puestos con mas experiencia actualmente son completamente remotos. Lo que si podemos confirmar es que los que son completamente presenciales ganan mas que los hibridos y igual a los remotos. Lo cual nos indica que definitivamente hibrido es el peor modelo, pero debido a la distribucion de puestos dentro de cada modalidad.

ANOVA MODALIDAD




```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## rratios       2 1.193e+11 5.964e+10   18.41 1.75e-08 ***
## Residuals   594 1.924e+12 3.239e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Esto nos indica que si efectivamente si existe efecto para esta variable



Esto nos indica que 0 y 100% remoto tienen el mismo efecto, y hibrido es el que se diferencia de esto. Esto respalda nuestro analisis previo.

CONCLUSION

Los salarios para analistas de datos son altos incluso a niveles de experiencia bajos, y tambien concluimos que entre empresas medianas y grandes la unica diferencia real de salarios es en cuanto a la gran magnitud de varianza entre los salarios de los ejecutivos. El puesto, o nivel de experiencia, sin duda es el factor mas grande en cuanto al salario. En cuanto a la modalidad no pudimos tener un analisis completo debido a esto mismo, que las distribuciones de los puestos eran lo que mas influenciaban en el promedio de la modalidad. Al igual, hay bastante diferencia entre paises en cuanto a salarios pero necesitamos mas informacion para aclarar eso mas concisamente, lo que queda claro es que Estados Unidos es definitivamente mejor pais para trabajar incluso que CA y GB. En general la demanda para analistas de datos esta incrementando ya que cada vez hay salarios mas altos a lo largo de los anios (2020/2021 -> 2022).

REPOSITORIO

Archivos Utilizados para este analisis: https://drive.google.com/drive/folders/16Y6_cbbXaWo_AuxQGE0QRvrL4VdnlmkS?usp=sharing Dentro de la carpeta “Salarios”

REFERENCIAS

Data Science Job Salaries. (2022). Retrieved 15 September 2022, from https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries