M2.851 - Tipología y ciclo de vida de los datos (aula 1)

PRÁCTICA-2

Autor: Jesus Navajas Briones
Fecha : Diciembre 2017

Índice de contenido

A - Fase previa: Elección dataset

Buscando posibles datasets a utilizar para la práctica he encontrado los resultados sobre una encuesta de DataScience realizada en 2017 en kaggle y creo puede ser interesante para esta practica; dado que aparte de ser datos de una encuesta -creo más complejos de tratar que de otras fuentes: sensores, mercados...- el propio tema me resulta de interés por razones obvias.

En un primer análisis:

- Se dispone de los resultados de 16.716 encuestas
- Existen 290 atributos distintos:
  - Existen atributos numéricos (edad, salarios..) , categóricos y de texto libre. Con respuesta simple o múltiple.
  - Se dividen en dos ficheros según sean de respuesta de texto libre (FREE) o de selección/tipadas (SELECT). No existe relación e incluso los de respuesta libre han sido aleatorizados/anonimizados (permutados atributos por filas) .
  - algunos solo son respondidos por subconjuntos según otras respuestas/categorias (si están trabajando...) y otras por todos (All).

Contar - Column	FICHERO		
Asked	FREE	SELECT	**Total Resultado**
All	27	43	**70**
CodingWorker	24	137	**161**
CodingWorker-NC	1	4	**5**
Learners	6	35	**41**
Non-switcher	1		**1**
Non-worker		2	**2**
OnlineLearners	1	1	**2**
Worker	1	1	**2**
Worker1	1	5	**6**
Total Resultado	**62**	**228**	**290**

Como creo que la finalidad de la práctica es enfrentarte a un conjunto de datos heterogéneo, realizar un data profiling (incluyendo pruebas estadisticas) y plantear transformaciones útiles para el tratamiento posterior; mi idea es la siguiente (teniendo en cuenta que el conjunto de atributos es muy grande):

1. Hacer un análisis rápido de todas las columnas: semanticamente, formatos y distribución de valores (no incluyendo las de texto libre de principio).
2. Seleccionar un conjunto relevante para la descripción de la población encuestada de data scientists/encuestados (edad, genero, país, nivel de formación, salario...). Una vez reducido el conjunto de columnas/atributos realizar:
   - las transformaciones y pruebas estadísticas
   - representación de los resultados a partir de tablas y gráficas.
3. Seleccionar un conjunto de atributos y ver si es posible construir un modelo de regresión del sueldo recibido a partir de los mismos.
4. Seleccionar un conjunto relevante para la descripción de métodos, tecnologías analizar las relaciones entre las mismas a nivel de la población. Presentando:

- las transformaciones y pruebas estadísticas
- representación de los resultados a partir de tablas y gráficas.

Respecto al problema/pregunta a responder no existe como tal, más que un estudio los datos en dos contextos planteados. Creemos no es la finalidad de esta práctica la construcción de un modelo basado en métodos de data mining, si bien presentaremos test estadísticos y trasformaciones que podrían servir para dicho procesamiento (ruego nos indiquen si debemos/podemos usar métodos de clustering o descubrimiento de reglas de asociación para el análisis).

B - Estudio global y determinación de preguntas

B.1. Estudio dataset:análisis datos en bruto

Tal y como hemos planteado en esta primera fase realizaremos una primer análisis del dataset restringiendonos a las preguntas de seleccióny datos numéricos (multipleChoiceResponses.csv) no incluyendo aquelas de texto libre. Haremos propiamente un análisis de los propios datos como tales y como están informados.

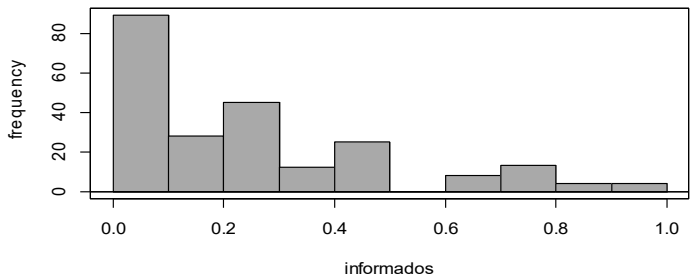
Procedemos a cargarlo en R y realizar un summary y una serie de operaciones para contabilizar los distintos datos vacios, distintos...:

```
#------------------------CARGA DE DATOS---------------------
datos.origen <-
read.csv("C:/Users/jesus/Documents/Estudios/UOC/1CUAT/TIPOLOYCICLO/PRA2/kaggle-survey-2017/multipleChoiceResponses.csv")
#------------------------ESTUDIO INICIAL---------------------
datos.summary<-summary(datos.origen)
write.table(datos.summary,
"C:/Users/jesus/Documents/Estudios/UOC/1CUAT/TIPOLOYCICLO/PRA2/columnas1.txt", sep="\t")
datos.distintos<-lapply(datos.origen, function(x) length(unique(x)))
write.table(datos.distintos,
"C:/Users/jesus/Documents/Estudios/UOC/1CUAT/TIPOLOYCICLO/PRA2/columnas2.txt", sep="\t")
datos.vacios<-lapply(datos.origen, function(x) c(sum(is.na(x)),sum((x==''), na.rm = TRUE)))
write.table(datos.vacios,
"C:/Users/jesus/Documents/Estudios/UOC/1CUAT/TIPOLOYCICLO/PRA2/columnas3.txt", sep="\t")
datos.isfactor<-unlist(lapply(datos.origen,is.factor),use.names = FALSE)
datos.distintos<-unlist(lapply(datos.origen, function(x) length(unique(x))),use.names = FALSE)
datos.vacios2<-unlist(lapply(datos.vacios, function(x) sum(x)),use.names = FALSE)
datos.informados<-(16716-datos.vacios2)/16716
datos.meta<-
data.frame(nombres=names(datos.origen),esfactor=as.factor(datos.isfactor),distintos=datos.distintos,informados=datos.informados)
```

Estamos tratando con un dataset con las siguientes características:

- Se tratan de 228 columnas.
- De los cuales 214 son factores y 14 numéricas.
- Respecto al nivel de datos informados en las columnas:

```
with(datos.meta, Hist(informados, scale="frequency", breaks="Sturges",col="darkgray"))
hist<-with(datos.meta, Hist(informados, scale="frequency",
breaks="Sturges",col="darkgray",plot=FALSE))
rbind(hist$breaks,hist$counts)
```

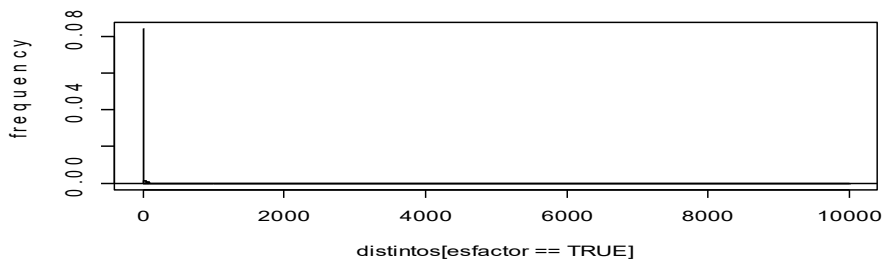


```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
[1,]    0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8   0.9     1
[2,]   89 28.0 45.0 12.0 25.0  0.0  8.0 13.0  4.0   4.0
```

Donde apreciamos disponemos de muchas columnas (89+28=117) informadas menos del 20%, por lo que habrá que ser cuidadoso en la selección de los campos para poder trabajar con los suficientes valores (téngase en cuenta que muchas preguntas solo van dirigidas a estudiantes, trabajadores...).

- Si realizamos un histograma de los valores distintos de las columnas que son factores (no numéricos, de selección):

```
with(datos.meta, Hist(distintos[esfactor==TRUE], scale="frequency",breaks =  c(0,10,50,100,1000,10000), col="darkgray"))
hist<-with(datos.meta, Hist(distintos[esfactor==TRUE], scale="frequency",breaks = c(0,10,50,100,1000,10000), col="darkgray",plot=FALSE))
rbind(hist$breaks,hist$counts)
```




```
     [,1] [,2] [,3] [,4] [,5]  [,6]
[1,]    0   10   50  100 1000 10000
[2,]  181   12    6    5   11
```

Apreciamos que la mayoria son selecciones por debajo de 10 elementos distintos , lo cual resulta muy aconsejable de aplicar para estudios de varianza, posibles regresiones...

B.2. Estudio dataset: análisis con semántica

Si tenemos en cuenta la semantica de los campos y valores con mayor ocurrencia:

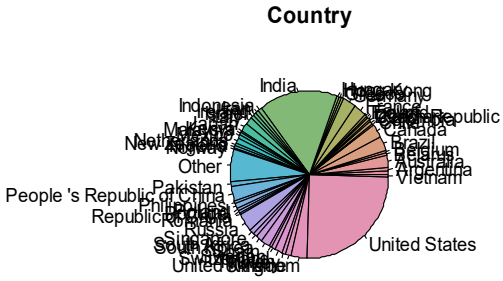
- Los 14 campos numéricos pueden clasificarse en los siguientes grupos:
  - Datos numéricos puros: edad (Age) y el sueldo(compensation Amount).
  - 6 son el porcentaje dedicado a cada tipo de aprendizaje (LearningCategory...)
  - 6 son el porcentaje dedicado a cada tipo de trbajo (Time...)
- De los datos 214 campos no numericos (factores), 165 pueden entenderse como agrupaciones de numéricos (discretos ordenados) , al tratarse de:
  - Niveles de estudios.
  - Preguntas con respuestas limitadas a:
    - Nunca, poco, normalmente, habitualmente, siempre.
    - Años : <1, entre 1 y 2 ….

      ….

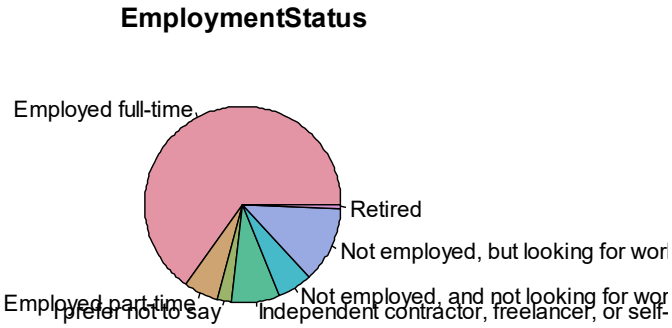
*Esto es importante para el tratamiento, ya que no simplemente es una clasificación (donde

aplicariamos valores dummy 0/1), sino que el orden tiene relevancia y en otros casos incluso podría usarse media del rango como valor (para una regresión, por ejemplo).

- Si nos referimos a paises de origen de la encuesta EEUU(25%) e India(16%) pueden tomarse como subconjuntos mas homogéneos siendo un conjunto importante. De España existen 320 registro y de un pais similar como Francia 442 (por si quisieramos hacer comparativas).



- Respecto a si su situacion laboral (EmploymentStatus) el 65% representa está Employed full-time.



*La información de todas las columnas se presenta como un fichero adicional de hoja de calculo denominado **ESTUDIO_COLUMNAS.ods**, donde indicamos tanto valores de que se trata cada columna (NUMERICO o factor) , numero de valores vacios, datos del summary (valores mas presentes o estadisticos principales si es numérico:Min,max,media,mediana y quantiles); asi como el analisis semantico realizado posteriormente(si es un factor ordenado...).*

B.3. Preguntas/problemas a plantear

A partir de este primer análisis de los datos nos vamos plantear los siguientes problemas/preguntas:

1. Estudio de la encuesta como tal por países: determinando si la edad media y número de encuestas por millón de habitantes por pais sigue distribuciones normales. Así como la comparativa con España y su correlación con el nivel de vida del país.
2. Intentar construir un modelo de regresión (y predicción) del sueldo a partir de un conjunto

de variables explicativas.

Nos hubiera gustado haber plantado un tercer estudio donde se hubieran aplicado métodos de contrate en tablas (tipo Chi cuadrado) para las diferentes tecnologías...., pero por falta de tiempo no ha podido ser.

C - Pregunta I: Caracterización de la encuesta por paises

C.1. Descripción del problema/pregunta

La idea de esta primera parte es realizar un análisis de dos datos principales: Edad media y nº de respuestas por millón de habitantes por país y su relación respecto a la encuesta en España.

Procederemos de la siguiente forma:

1. Veremos en primer lugar si dichos valores pueden entenderse como una distribución normal
2. Comprobaremos si el valor en la encuesta española respecto a los demás paises, viendo si podemos decir es un pais medio.
3. Analizaremos si el nivel de producto interior bruto es un factor determinante en estos valores.

C.2. Obtención de datos y limpieza de datos

La obtención de datos se ha realizado a partir de los datos de la encuesta unido con una segunda fuente de datos obtenida del Banco mundial (mírese referencias). Se han filtrado los paises de la encuesta y se han obtenido los datos:

- Country:Nombre pais a cruzar
- Population: Poblacion total
- ClasPerGDP: Redondeo de renta per capita dividida por 10000$ (permitirá clasificar paises)

Obtendremos la media de edad del campo Age de la fuente principal (eliminando los no informados N/A) y contaremos numero de encuestas por pais(no_rows). Procederemos a cruzar dichos datos con la tabla de paises (eliminando pais vacio y Others al no cruzarse).

```
aux1 <- read.csv("C:/Users/jesus/Documents/Estudios/UOC/1CUAT/TIPOLOYCICLO/PRA2/datos/datos_paises.csv",na.strings = '#N/D')
aux1<-aux1[c('Country','Population','ClasPerGDP')]

#Cruce de datos con paises y calculo porcentajes por millon de habitantes
aux2<-datos.origen %>%  group_by(Country) %>% summarise(no_rows = length(Country),Age=mean(Age,na.rm = TRUE))
datos.paises <- merge(aux1,aux2,by="Country")
datos.paises$perRespuestas=datos.paises$no_rows/(datos.paises$Population/1000000)
summary(datos.paises)
```

```
      Country      no_rows            Age          Population
 Argentina: 1   Min.   :  51.00   Min.   :25.71   Min.   :   4692700
 Australia: 1   1st Qu.:  73.75   1st Qu.:29.96   1st Qu.:  10383866
 Belarus  : 1   Median : 119.50   Median :32.01   Median :  44426038
 Belgium  : 1   Mean   : 319.13   Mean   :32.19   Mean   : 147738391
 Brazil   : 1   3rd Qu.: 259.75   3rd Qu.:34.74   3rd Qu.: 100665442
 Canada   : 1   Max.   :4197.00   Max.   :40.06   Max.   :1378665000
 (Other)  :46                                     NA's   :6
   ClasPerGDP    perRespuestas
 Min.   :0.000   Min.   : 0.0486
 1st Qu.:1.000   1st Qu.: 2.1189
 Median :2.000   Median : 4.9334
 Mean   :2.739   Mean   : 6.7625
 3rd Qu.:5.000   3rd Qu.: 9.8480
 Max.   :9.000   Max.   :32.8145
 NA's   :6       NA's   :6
```

Disponemos en total de 52 paises de los cuales 6 no se dispone de su poblacion, por lo que no se aplicarán a la prueba de respuestas por población.

Country	no_rows	Age	Population	ClasPerGDP	perRespuestas
1 Argentina	92	34,82	43.847.430	1	2,10
2 Australia	421	36,39	24.127.159	6	17,45
3 Belarus	54	27,59	9.507.120	1	5,68
4 Belgium	91	34,70	11.348.159	5	8,02
5 Brazil	465	31,86	207.652.865	1	2,24
6 Canada	440	35,77	36.286.425	5	12,13
7 Chile	51	31,76	17.909.754	2	2,85
8 Colombia	113	31,49	48.653.419	1	2,32
9 Czech Republic	53	31,92	10.561.633	2	5,02
10 Denmark	78	34,33	5.731.118	6	13,61
11 Egypt	66	30,17	NA	NA	NA
12 Finland	67	33,22	5.495.096	5	12,19
13 France	442	32,73	66.896.109	4	6,61
14 Germany	460	33,80	82.667.685	5	5,56
15 Greece	81	31,83	10.746.740	2	7,54
16 Hong Kong	65	31,68	NA	NA	NA
17 Hungary	66	34,02	9.817.958	1	6,72
18 India	2704	27,57	1.324.171.354	0	2,04
19 Indonesia	131	26,42	261.115.456	0	0,50
20 Iran	112	28,47	NA	NA	NA
21 Ireland	94	36,01	4.773.095	7	19,69
22 Israel	105	36,42	8.547.100	3	12,28
23 Italy	238	35,69	60.600.590	3	3,93
24 Japan	277	34,71	126.994.511	5	2,18
25 Kenya	59	27,63	48.461.567	0	1,22
26 Malaysia	79	29,86	31.187.265	1	2,53
27 Mexico	126	33,61	127.540.423	1	0,99
28 Netherlands	205	36,98	17.018.408	5	12,05
29 New Zealand	74	40,06	4.692.700	4	15,77
30 Nigeria	73	30,32	185.989.640	0	0,39
31 Norway	53	34,46	5.232.929	9	10,13
32 Other	1023	31,49	NA	NA	NA
33 Pakistan	161	26,88	193.203.476	0	0,83
34 People 's Republic of China	471	27,02	1.378.665.000	1	0,34
35 Philippines	84	28,07	103.320.222	0	0,81
36 Poland	184	30,80	37.948.016	2	4,85
37 Portugal	93	34,42	10.324.611	2	9,01
38 Republic of China	67	27,15	1.378.665.000	1	0,05
39 Romania	59	32,79	19.705.301	1	2,99
40 Russia	578	29,98	144.342.396	1	4,00
41 Singapore	184	31,49	5.607.283	5	32,81
42 South Africa	127	32,09	55.908.865	1	2,27
43 South Korea	194	32,21	NA	NA	NA
44 Spain	320	36,73	46.443.959	3	6,89
45 Sweden	89	35,19	9.903.122	6	8,99
46 Switzerland	129	35,22	8.372.098	8	15,41
47 Taiwan	254	30,46	NA	NA	NA
48 Turkey	144	29,92	79.512.426	1	1,81
49 Ukraine	196	29,07	45.004.645	0	4,36
50 United Kingdom	535	35,81	65.637.239	4	8,15
51 United States	4197	35,22	323.127.513	5	12,99
52 Vietnam	71	25,71	92.701.100	0	0,77

C.3. Analisis de datos y Representación de resultados

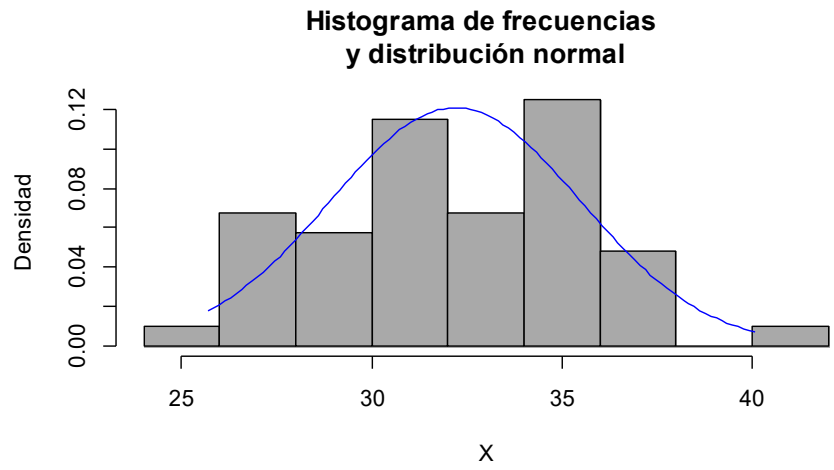
C.3.I. Test de normalidad

En primer lugar procedermos a realizar el test de Shapiro-Wilks sobre los campos de media de edad y nº de contestaciones por millon de habitantes , el cual establece como hipotesis nula que la

población está distribuida normalmente.

Creamos una funcion que muestre histograma y normal y procedemos a ejecutarlo sobre la edad:

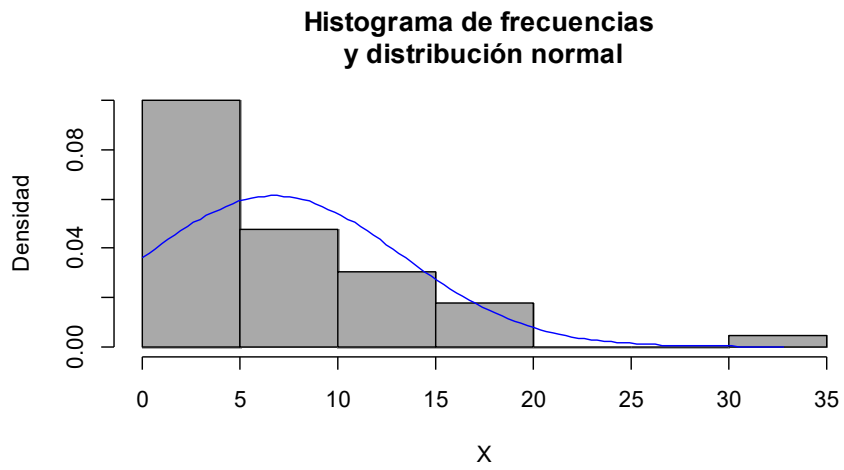
```
#FUNCION AUXILIAR
mifun.plotn <- function(x,main="Histograma de frecuencias \ny distribución normal",
                xlab="X",ylab="Densidad") {
  min <- min(x)
  max <- max(x)
  media <- mean(x)
  dt <- sd(x)

  hist(x,freq=F,main=main,xlab=xlab,ylab=ylab,col="darkgray",breaks="Sturges")
  curve(dnorm(x,media,dt), min, max,add = T,col="blue")
}
#Normalidad edad por pais
aux<-datos.paises$Age[!is.na(datos.paises$Age)]
mifun.plotn(aux)
shapiro.test(aux)
```




```
        Shapiro-Wilk normality test

data:  aux
W = 0.97334, p-value = 0.2914
```

Y despues sobre el porcentaje de respuestas por millón:

```
#Normalidad respuestas por pais
aux<-datos.paises$perRespuestas[!is.na(datos.paises$perRespuestas)]
mifun.plotn(aux)
shapiro.test(aux)
```




```
        Shapiro-Wilk normality test

data:  aux
W = 0.84108, p-value = 0.00001816
```

Los valores de p-value nos hacen rechazar claramente la hipotesis en el segundo caso y aceptarla en el primero(podríamos aplicar un 99.9% o más y lo cumpliría) .

C.3.II. Comparativa encuesta española

Para ver si España representa un pais medio en estos valores comprobaremos si el valor obtenido en España puede ser la media de la muestra.

Para ello realizaremos en el caso de la edad un t-test con un 95% de confianza donde la hipotesis (dado suponemos es normal la distribución a partir de lo estimado anteriormente) :

```
paisesSinEspana<-datos.paises[datos.paises$Country!='Spain',]
aux<-paisesSinEspana$Age[!is.na(paisesSinEspana$Age)]
mesp<-datos.paises[datos.paises$Country=='Spain','Age']
mesp
t.test(aux,mu=mesp)
```

```
[1] 36.72698

One Sample t-test

data:  aux
t = -10.104, df = 50, p-value = 1.134e-13
alternative hypothesis: true mean is not equal to 36.72698
95 percent confidence interval:
 31.18498 33.02298
sample estimates:
mean of x
 32.10398
```

Donde vemos es exterior al intervalo y posee un valor de p-test muy reducido, por lo que podemos deducir que no es probable la meia de la muestra coincida con la Española.

Por otra parte realizaremos el test de Wilconox sobre la muestra de preguntas por millon de habitantes para ver si el valor en España podría ser la media del total:

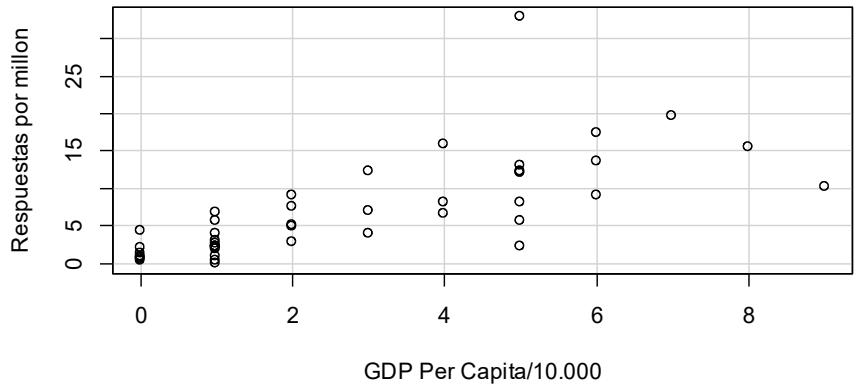
```
#TEST: Preguntas por millon Spain caracteristic
paisesSinEspana<-datos.paises[datos.paises$Country!='Spain',]
aux<-paisesSinEspana$perRespuestas[!is.na(paisesSinEspana$perRespuestas)]
mesp<-datos.paises[datos.paises$Country=='Spain','perRespuestas']
mesp
wilcox.test(aux, mu=mesp)
```

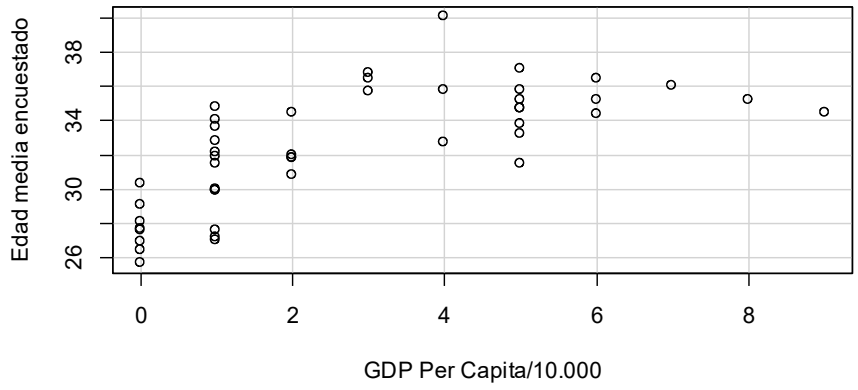
```
 [1] 6.890024


        Wilcoxon signed rank test

data:  aux
V = 447, p-value = 0.4331
alternative hypothesis: true location is not equal to 6.890024
```

Donde si se cumple la hipotesis.

C.3.III. Correlacion con Renta per capita

Primero realizaremos un analisis gráfico representando en unos graficos de dispersión la renta per capita/10000 (eje X) y las respuestas y edad media en la otra (eje Y):





Donde si parece apreciarse cierta correlación.

Procederemos ahora a clasificar los paises en aquellos con renta >=30000$ y menor y realizaremos un test de si la media es similar(test de dos muestras), primero para la edad (test-t):

```
#PREPARACION TEST de dos muestras
datos.paises$discRper=as.factor(datos.paises$ClasPerGDP>=3)

#TEST:Dos vias edad por pais
aux<-datos.paises[!is.na(datos.paises$Age),]
aux<-aux[!is.na(aux$discRper),]

t.test(aux$Age~aux$discRper)
var.test(aux$Age~aux$discRper)
```

```
 t.test(aux$Age~aux$discRper)

        Welch Two Sample t-test

data:  aux$Age by aux$discRper
t = -7.6694, df = 43.27, p-value = 0.000000001345
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.403824 -3.737584
sample estimates:
mean in group FALSE  mean in group TRUE
           30.17683            35.24754

> var.test(aux$Age~aux$discRper)

        F test to compare two variances

data:  aux$Age by aux$discRper
F = 2.2377, num df = 25, denom df = 19, p-value = 0.07585
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9167856 5.1877986
sample estimates:
ratio of variances
          2.237662
```

Donde podemos rechazar que la media es igual (con %5). Ademas verificamos si la varianza es similar (tampoco) , aunque el t-test se había aplicado sin dicha suposición (var.equal=T).

Procederemos ha realizar el test no parametrico similar sobre el numero de encuestas por millón:

```
#TEST:Dos vias preguntas por millon por pais
aux<-datos.paises[!is.na(datos.paises$perRespuestas),]
aux<-aux[!is.na(aux$discRper),]

wilcox.test(aux$perRespuestas~aux$discRper)
```

```
        Wilcoxon rank sum test

data:  aux$perRespuestas by aux$discRper
W = 34, p-value = 0.00000002314
alternative hypothesis: true location shift is not equal to 0
```

Donde podemos rechazar la igualdad de la media en las muestras.

Ahora bien, estos test nos permiten rechazar la igualdad de las muestras (a nivel de medias), pero para si existe una correlacion obteniendodremos el coeficiente de correlación (de Pearson para la muestra suponemos normal y de Spearman como no parametrico).

```
#CORRELACION edad
```

```
aux<-datos.paises[!is.na(datos.paises$Age),]
aux<-aux[!is.na(aux$discRper),]
cor(aux$Age,aux$ClasPerGDP)

#CORRELACION respuestas
aux<-datos.paises[!is.na(datos.paises$perRespuestas),]
aux<-aux[!is.na(aux$discRper),]
cor.test(aux$perRespuesta,aux$ClasPerGDP,method='spearman')
```

```
 [1] 0.7023939


Warning in cor.test.default(aux$perRespuesta, aux$ClasPerGDP, method = "spearman") :
  Cannot compute exact p-value with ties

        Spearman's rank correlation rho

data:  aux$perRespuesta and aux$ClasPerGDP
S = 2779.3, p-value = 1.175e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.828599
```

Donde apreciamos valores nada cercanos a 0.

Por ultimo realizaremos una regresión completa y miraremos la bondad del ajuste:

```
#REGRESION edad
aux<-datos.paises[!is.na(datos.paises$Age),]
aux<-aux[!is.na(aux$discRper),]
summary(lm(aux$Age~aux$ClasPerGDP))
```

```
 Call:
lm(formula = aux$Age ~ aux$ClasPerGDP)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1490 -1.5391 -0.1801  1.3846  6.4212

Coefficients:
               Estimate Std. Error t value     Pr(>|t|)
(Intercept)     29.6563     0.5535  53.580      < 2e-16 ***
aux$ClasPerGDP   0.9949     0.1520   6.546 0.0000000531 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.474 on 44 degrees of freedom
Multiple R-squared:  0.4934,  Adjusted R-squared:  0.4818
F-statistic: 42.85 on 1 and 44 DF,  p-value: 0.00000005313
```

```
#REGRESION respuestas
aux<-datos.paises[!is.na(datos.paises$perRespuestas),]
aux<-aux[!is.na(aux$discRper),]
summary(lm(aux$perRespuestas~aux$ClasPerGDP))
```

```
Call:
lm(formula = aux$perRespuestas ~ aux$ClasPerGDP)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9698 -1.5732 -0.6224  1.0250 21.6635

Coefficients:
               Estimate Std. Error t value     Pr(>|t|)
(Intercept)     1.4456     1.0179   1.420        0.163
aux$ClasPerGDP  1.9411     0.2795   6.944 0.0000000138 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.549 on 44 degrees of freedom
Multiple R-squared:  0.5229,  Adjusted R-squared:  0.5121
F-statistic: 48.23 on 1 and 44 DF,  p-value: 0.00000001379
```

Donde apreciamos:

- residuos simetricamente distribuido,
- valores de significacion altos de los coeficientes de regresión (p-values bajos y estrellas)
- un coeficiente de correrlación cercano al 0.5 ,
- y un alto grado de confianza al rechazarse la hipotesis de ser 0 el coeficiente mediante el test F de Fischer (p-values bajos)

  en ambos casos.

C.4. Conclusiones.Resolución del problema.

Respecto a las preguntas planteadas podemos deducir con los grados de confianza establecidos (De principio 95%):

1. La edad media de los encuestados por pais sigue una distribución normal y no así él numero de encuestados por millón.
2. España no es un pais medio en edad de encuestados pero si podría serlo en numero de encuestas por millón.
3. El nivel de renta del pais es relevante para los dos valores, existiendo una clara correlación positiva.

D - Pregunta II: Regresion de Sueldo

D.1. Descripción del problema/pregunta

Se plantea realizar una regresión del sueldo de un científico de datos a partir de los datos de la encuesta.

Para ello se hace una preselección de posibles parámetros se estiman relevantes:

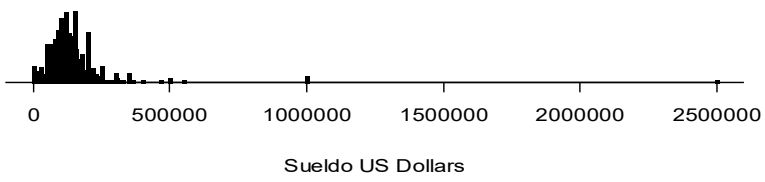
		TIPO				VALORES
NOMBRE COL	DESCRIPCION	TIPO	IMPORTADA COMO	TIPO_ADICIONAL	TIPO_ADICIONAL_2	DISTINTOS
GenderSelect	Genero	FACTOR	FACTOR			5
Country	Pais	FACTOR	FACTOR			53
Age	Edad	NUMERO	NUMERO			85
CodeWriter	Escribe codigo	FACTOR	FACTOR	SI/NO		3
LearningDataScienceTime	Años aprendiendo Ciencia de datos	FACTOR	FACTOR	RANGO	YEARS	7
FormalEducation	Nivel de educación	FACTOR	FACTOR	RANGO	EDUCATIONA	8
EmployerSize	Tamaño empresa	FACTOR	FACTOR	RANGO	TAMAÑOEMP	11
EmployerMLTime	Tiempo en empresa	FACTOR	FACTOR	RANGO	YEAR2	7
AlgorithmUnderstandingLevel	Nivel cnocimiento de algoritmos	FACTOR	FACTOR	RANGO	UNDESTAL	7
CompensationAmount	Sueldo	NUMERO	FACTOR			907
CompensationCurrency	Moneda	FACTOR	FACTOR			88

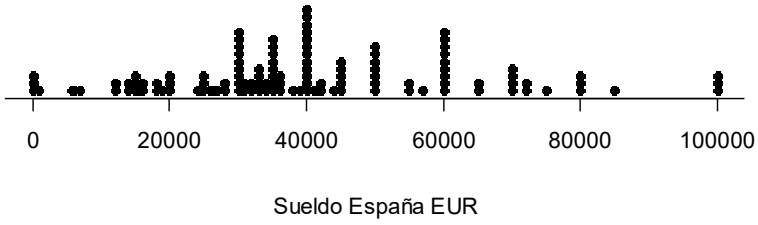
D.2. Obtención de datos y limpieza de datos

D.2.I. Filtrado y tratamiento de sueldo

Para el estudio de los salarios y sus posibles factores restringiremos los datos a España y EEUU para tener unas muestras más homogéneas (cada una la trataremos por separado). A su vez dentro de estos solo tomaremos los que estén en moneda local o esta no haya sido informada(supondremos es la local EUR/DOLLAR). Nos quedan dos dataset de 133 (España) y 1520 (US) observaciones.

Analizando gráficamente:





Apreciamos existen:

- valores muy bajos en ciertos casos, lo cual nos lleva a deducir se refieren a miles. Tomaremos los valores inferios a 500 como miles y lo corregiremos.
- Valores muy altos en el modelo americano, y aunque pueda ser cierto restringiremos nuestra muestra entre 10.000 y 500.0000 (redondeanndo ya a miles para ser mas comodo su tratamiento).

Nos quedan finalmente 128(España) y 1495 (US) observaciones.

```
aux<- read.csv("C:/Users/jesus/Documents/Estudios/UOC/1CUAT/TIPOLOYCICLO/PRA2/datos/kaggle-survey-2017/multipleChoiceResponses.csv")
datos.origen<-
aux[,c('GenderSelect','Country','Age','EmploymentStatus','CodeWriter','CurrentJobTitleSelect','CurrentEmployerType','LearningDataScienceTime','FormalEducation','ParentsEducation','EmployerIndustry','EmployerSize','EmployerMLTime','JobFunctionSelect','AlgorithmUnderstandingLevel','CompensationAmount','CompensationCurrency','SalaryChange','JobSatisfaction')]


#------------------------PROCESODATOS-----------------------------
#Seleccionamos con cantidad informada y convertmos a numero
aux<-datos.origen[!is.na(datos.origen$CompensationAmount),]
aux$CompensationAmount<-as.numeric(as.character(aux$CompensationAmount))
#Filtramos por pais y moneda
datos.US<-subset(aux, (Country=='United States') & (CompensationCurrency=='USD' | CompensationCurrency=='' ))
#Filtramos por pais y moneda
datos.SP<-subset(aux, (Country=='Spain') & (CompensationCurrency=='EUR' | CompensationCurrency=='' ))


#Aplicamos rango a sueldo y redondeamos a miles
datos.SP<-mutate(datos.SP,CompensationAmount = round(ifelse(CompensationAmount < 500 , CompensationAmount , CompensationAmount/1000)))
datos.SP<-subset(datos.SP, CompensationAmount>10 & CompensationAmount < 200 )

datos.US<-mutate(datos.US,CompensationAmount = round(ifelse(CompensationAmount < 500 , CompensationAmount , CompensationAmount/1000)))
datos.US<-subset(datos.US, CompensationAmount>10 & CompensationAmount < 500 )
```

D.2.II. Variables explicativas

En un primer análisis:

- Eliminamos Code Writer dado que todos son YES.
- Eliminamos LearningDataScienceTime dado que no esta informado (solo estudiantes).

Transformamos el resto de campos para su análisis:

- GenderSelect se convierte en dos variable binarias(dummy) isMale y isFemale.
- FormalEducation, se establecen los siguientes valores:
  - "I prefer not to answer"=0,
  - "I did not complete any formal education past high school"=1,
  - "Some college/university study without earning a bachelor's degree"=2,
  - "Bachelor's degree"=3,
  - "Master's degree"=4,
  - "Professional degree"=5,
  - "Doctoral degree"=6
- LemployerSize,se establecen los siguientes valores:
  - vacio,"I prefer not to answer"=0,,"I don't know"=0,
  - "Fewer than 10 employees"=5,
  - "10 to 19 employees"=15,
  - "20 to 99 employees"=60,
  - "100 to 499 employees"=250,
  - "500 to 999 employees"=750,
  - "1,000 to 4,999 employees"=2500,
  - "5,000 to 9,999 employees"=7500,
  - "10,000 or more employees"=15000)
- EmployerMLTime,se establecen los siguientes valores:
  - "Don't know"=0,
  - "Less than one year"=0.5,
  - "1-2 years"=1.5,
  - "3-5 years"=4,
  - "6-10 years"=8,
  - "More than 10 years"=12
- AlgorithmUnderstandingLevel,se establecen los siguientes valores:
  - "Enough to run the code / standard library"=1,
  - "Enough to explain the algorithm to someone non-technical"=2,
  - "Enough to code it again from scratch, albeit it may run slowly"=3
  - "Enough to code it from scratch and it will run blazingly fast and be super efficient"=4
  - "Enough to refine and innovate on the algorithm"=5,

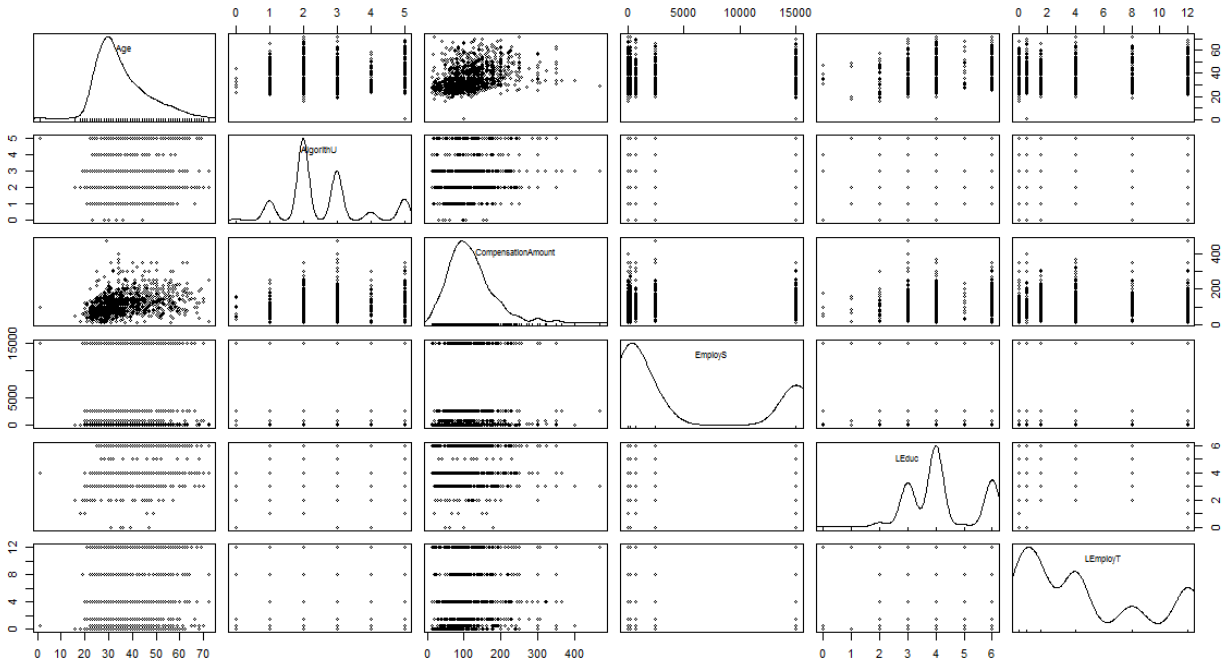
Construimos un nuevo dataset con solo estos valores y la variable objetivo:

```
#Construimos factores accesorios
#dataFrame auxiliares auxiliares

datos.mergeFormalEducation=data.frame(FormalEducation=c("","I prefer not to answer","I did not
complete any formal education past high school","Some college/university study without earning a
bachelor's degree","Bachelor's degree","Master's degree","Professional degree","Doctoral
degree"),LEduc=c(0,0,1,2,3,4,5,6))

datos.EmployerSize=data.frame(EmployerSize=c("","I prefer not to answer","I don't know","Fewer than 10
employees","10 to 19 employees","20 to 99 employees","100 to 499 employees","500 to 999
employees","1,000 to 4,999 employees","5,000 to 9,999 employees"=7500,"10,000 or more
employees"),EmployS=c(0,0,0,5,15,60,250,750,2500,7500,15000))

datos.EmployerMLTime=data.frame(EmployerMLTime=c("","Don't know","Less than one year","1-2
years","3-5 years","6-10 years","More than 10 years"),LEmployT=c(0,0,0.5,1.5,4,8,12))

datos.AlgorithmUnderstandingLevel=data.frame(AlgorithmUnderstandingLevel=c("","Enough to run the
code / standard library","Enough to explain the algorithm to someone non-technical","Enough to code it
again from scratch, albeit it may run slowly","Enough to code it from scratch and it will run blazingly fast
and be super efficient","Enough to refine and innovate on the algorithm"),AlgorithU=c(0,1,2,3,4,5))


#DatosSP
aux<-datos.SP
aux<-mutate(aux,isMale=ifelse(GenderSelect=='Male',1,0))
aux<-mutate(aux,isFemale=ifelse(GenderSelect=='Female',1,0))
aux<-mutate(aux,isFemale=ifelse(GenderSelect=='Female',1,0))
aux<- merge(aux,datos.mergeFormalEducation,by="FormalEducation")
aux<- merge(aux,datos.EmployerSize,by="EmployerSize")
aux<- merge(aux,datos.EmployerMLTime,by="EmployerMLTime")
aux<- merge(aux,datos.AlgorithmUnderstandingLevel,by="AlgorithmUnderstandingLevel")
datos.SP2<-
aux[,c('CompensationAmount','Age','isMale','isFemale','LEduc','EmployS','LEmployT','AlgorithU')]

aux<-datos.US
aux<-mutate(aux,isMale=ifelse(GenderSelect=='Male',1,0))
aux<-mutate(aux,isFemale=ifelse(GenderSelect=='Female',1,0))
aux<-mutate(aux,isFemale=ifelse(GenderSelect=='Female',1,0))
aux<- merge(aux,datos.mergeFormalEducation,by="FormalEducation")
aux<- merge(aux,datos.EmployerSize,by="EmployerSize")
aux<- merge(aux,datos.EmployerMLTime,by="EmployerMLTime")
aux<- merge(aux,datos.AlgorithmUnderstandingLevel,by="AlgorithmUnderstandingLevel")
datos.US2<-
aux[,c('CompensationAmount','Age','isMale','isFemale','LEduc','EmployS','LEmployT','AlgorithU')]
```

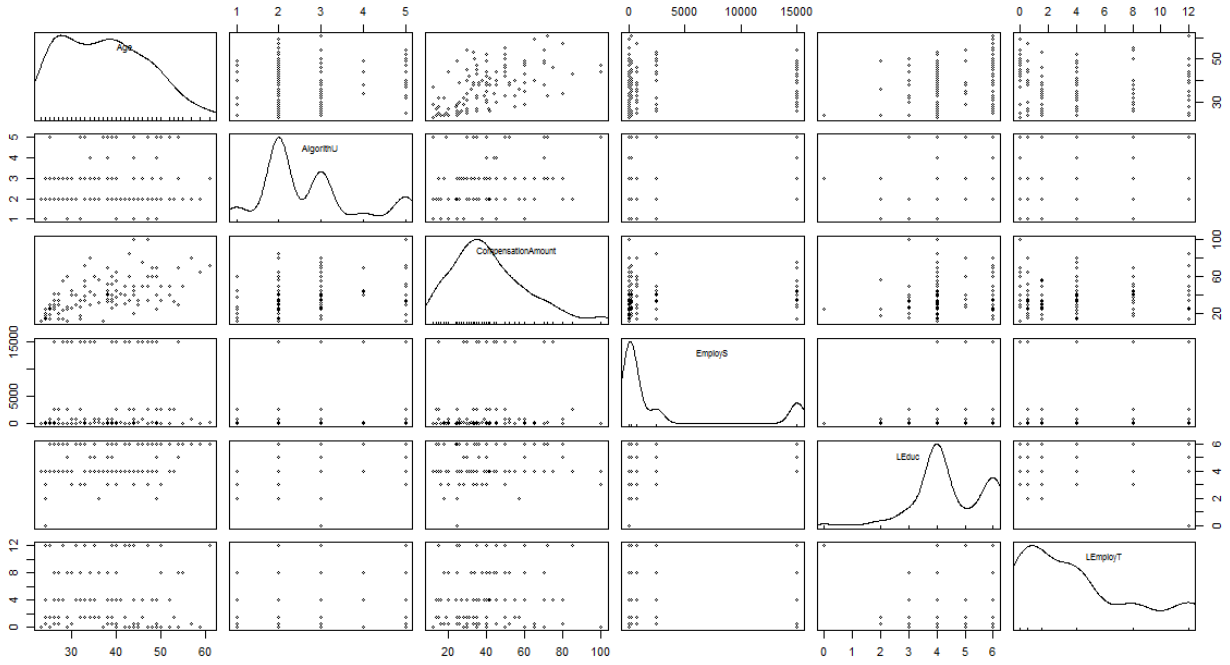
Al descartar tomar valores no informados(N/A) nos quedan 115 (España) y 1283(US) observaciones.

D.3. Análisis de datos

Empezaremos con un análisis grafico para ver si hay correlaciones claras, mostrando los graficos de dispersión cruzadas(entro todos los elementos excepto los binarios isMAle y isFemale):



(Primero US y luego España)



Donde solo apreciamos claramente una correlación entre la edad y el sueldo.

Si obtenemos matrices de correlaciones:

```
cor(datos.US2,use ="complete.obs")
cor(datos.SP2,use ="complete.obs")
```

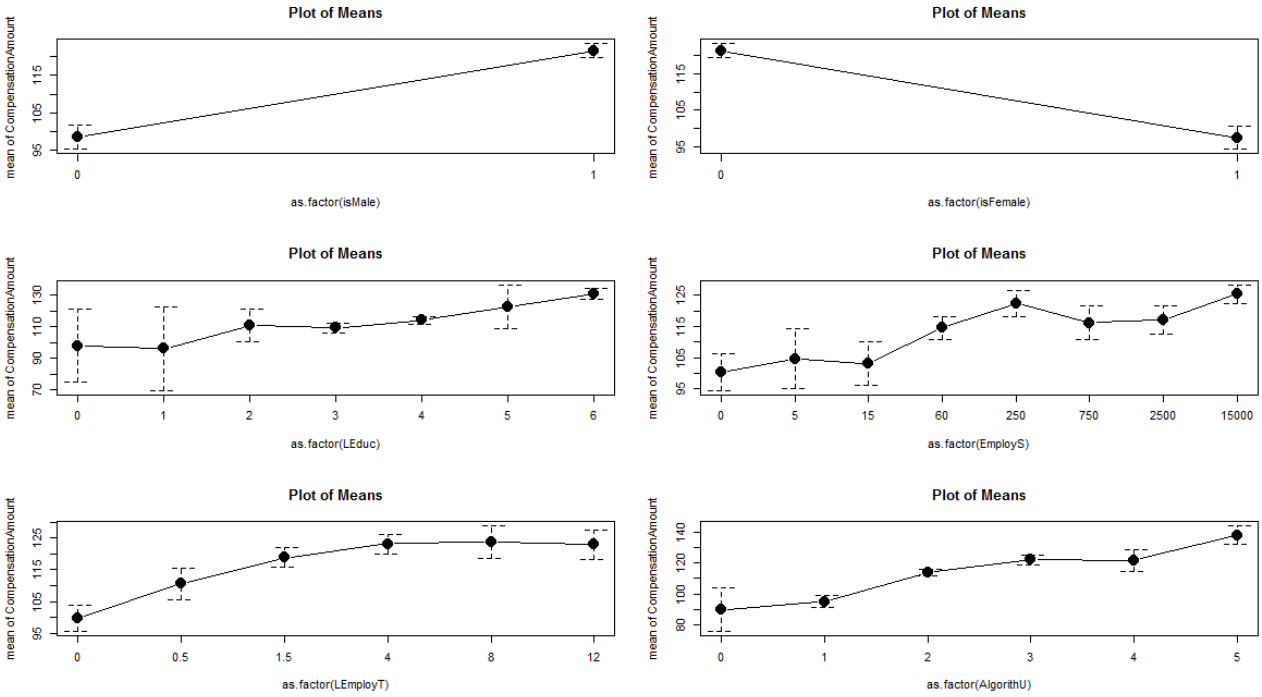
	CompensationAmount	Age	isMale	isFemale	LEduc	EmployS	LEmployT	AlgorithU
CompensationAmount	1.00000000	0.35689629	0.15236993	-0.154929492	0.139304099	0.093103775	0.09780895	0.17513085
Age	0.35689629	1.00000000	0.08990744	-0.099495167	0.214158868	-0.026986585	0.03922917	0.07749600
isMale	0.15236993	0.08990744	1.00000000	-0.963897890	-0.045466637	0.018832874	0.01317768	0.07259557
isFemale	-0.15492949	-0.09949517	-0.96389789	1.000000000	0.051480773	-0.009469622	-0.01580201	-0.07376762
LEduc	0.13930410	0.21415887	-0.04546664	0.051480773	1.000000000	0.003653028	0.16450758	0.22945777
EmployS	0.09310378	-0.02698658	0.01883287	-0.009469622	0.003653028	1.000000000	0.26592751	-0.01354126
LEmployT	0.09780895	0.03922917	0.01317768	-0.015802008	0.164507581	0.265927510	1.00000000	0.12737655
AlgorithU	0.17513085	0.07749600	0.07259557	-0.073767617	0.229457775	-0.013541261	0.12737655	1.00000000

	CompensationAmount	Age	isMale	isFemale	LEduc	EmployS	LEmployT	AlgorithU
CompensationAmount	1.00000000	0.555700676	0.11290196	-0.12102180	0.09739695	0.09613456	0.118630240	0.28554674
Age	0.55570068	1.000000000	0.08810669	-0.07120702	0.25917011	0.14051546	-0.007818808	0.15645451
isMale	0.11290196	0.088106690	1.00000000	-0.96680293	0.07939587	0.12745856	-0.231234383	0.06879938
isFemale	-0.12102180	-0.071207025	-0.96680293	1.00000000	-0.07088408	-0.18500927	0.211840552	-0.07766656
LEduc	0.09739695	0.259170108	0.07939587	-0.07088408	1.00000000	0.09435812	0.095841030	0.14450631
EmployS	0.09613456	0.140515459	0.12745856	-0.18500927	0.09435812	1.00000000	0.106469642	0.11485578
LEmployT	0.11863024	-0.007818808	-0.23123438	0.21184055	0.09584103	0.10646964	1.000000000	0.09440820
AlgorithU	0.28554674	0.156454511	0.06879938	-0.07766656	0.14450631	0.11485578	0.094408196	1.00000000

Vemos claramente la existente entre isMale y isFemale(se habia planteado en dos porque había no definidos), así como la ya detectada entre la edad y el sueldo, siendo remarcables:

- Nivel educativo y edad, los científicos de datos mayores tienen niveles educativos altos.
- Como el tamaño de la empresa esta relacionado con otros factores como edad o genero en España.
- Como el tiempo de permanencia en la empresa esta relacionado con el genero en España.

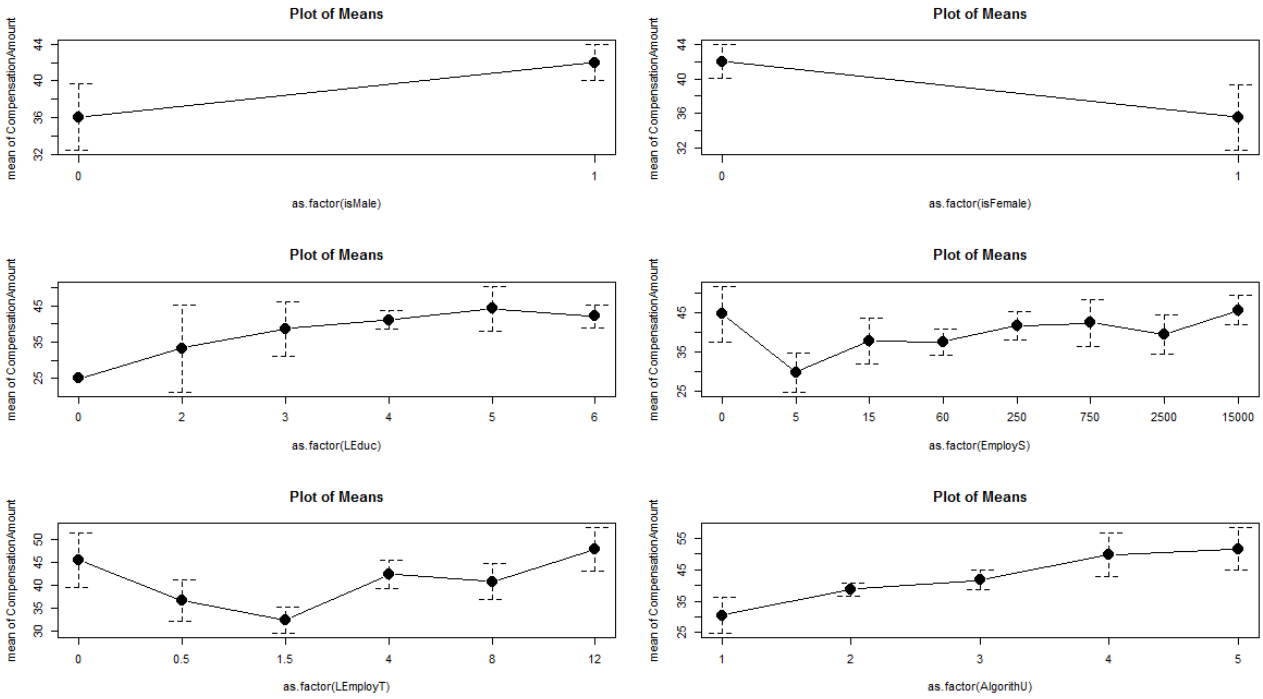
Si mostramos por medias y varianzas del sueldo respecto a los factores:



Donde apreciamos en el la encuesta americana se produce los siguiente:

1. Un claro sesgo por sexo, siendo inferior y mucho más variable en mujeres
2. Un crecimiento con el nivel educativo siendo mucho más variable en niveles inferiores.
3. Un crecimiento con el tamaño de la empresa.
4. Un crecimiento por el tiempo en la empresa, siendo más marcado a partir hasta el 4 año.
5. Un crecimiento por el conocimiento de algoritmos.

Mientras que en la española:



1. Un sesgo comparable por sexo.
2. Un crecimiento similar por nivel educativo, con altas varianzas a niveles bajos.
3. Se produce un efecto por tamaño de empresa, donde en pequeñas (autónomos seguramente) parte de un nivel suprior.
4. Se aprecia claramente el efecto de la crisis de hace unos años, donde los sueldos son inferiores para aquellos contratados hace 2 años.
5. Un crecimiento similar por conocimiento de algoritmos.

D.4. Representación de los resultados:Modelo de regresion

Para la construcción del modelo optimo (eficaz) sería requisito la homocedasticidad del sueldo, pero ya hemos visto en las graficas que la varianza depende de los valores de los factores en muchos casos.

En segundo lugar el requisito de normalidad no se cumple a nivel global (ya que no lo cumple la muestra), pero habría que inspeccionar para cada uno de las combinaciones de valores tomadas.

De todas formas, y aun conociendo las posibles limitaciones del modelo, procederemos a calcular los valores del modelo e intentar un diagnostico (selección de valores mediante el procedimiento step basado en el Akaike's Information Criterion).

D.4.I. Modelo USA

Construimos un modelo a partir de todos los valores y analizamos sus parametros:

```
modelo.US <- lm(CompensationAmount ~ Age + isMale + EmployS + AlgorithU + LEmployT + LEduc  ,
data=datos.US2)
summary(modelo.US)
anova(modelo.US)
```

```
 Call:
lm(formula = CompensationAmount ~ Age + isMale + EmployS + AlgorithU +
    LEmployT + LEduc, data = datos.US2)

Residuals:
    Min      1Q  Median      3Q     Max
-169.83  -33.93   -6.53   25.29  360.07
```

```
Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)  3.2721769  7.9068984   0.414   0.679063
Age          1.9099564  0.1522016  12.549    < 2e-16 ***
isMale      17.2312499  3.9341044   4.380 0.00001285 ***
EmployS      0.0008381  0.0002421   3.462   0.000553 ***
AlgorithU    6.7612824  1.3826104   4.890 0.00000114 ***
LEmployT     0.5026183  0.3741008   1.344   0.179339
LEduc        1.8750221  1.3433362   1.396   0.163021
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.36 on 1268 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.1749,  Adjusted R-squared:  0.171
F-statistic: 44.79 on 6 and 1268 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Response: CompensationAmount
           Df  Sum Sq Mean Sq  F value        Pr(>F)
Age         1  578523  578523 195.7422     < 2.2e-16 ***
isMale      1   66247   66247  22.4145 0.00000244407 ***
EmployS     1   45616   45616  15.4340 0.00009004143 ***
AlgorithU   1   91003   91003  30.7907 0.00000003496 ***
LEmployT    1    7124    7124   2.4103        0.1208
LEduc       1    5758    5758   1.9482        0.1630
Residuals 1268 3747622    2956
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Donde apreciamos lo siguiente:

- Los residuos no aparecen centrados en 0 (no son normales centrados en 0).
- El punto de interceptación estimado se situa cercano a los 34000 dolares.
- Los valores de los parametros estimados son positivos como era de esperar a partir de lo visto en la grafica, siendo algunos significativos ((Pr(>|t|) bajo) para Edad(Age), genero(isMale), tamaño empresa(EmployS) y conocimiento algoritmos(AlgorithU) ,
- El error es muy grande al (54) y los coeficientes de Regresion R bajos (deberían ser por encima de 0.6 para empezar a valorar un ajuste bueno).
- Viendose el tamaño del error medio claramente en el analisis ANOVA.

Si procedemos a una optimización del modelo con step/AIC:

```
modelo.resulUS<-step(modelo.US, direction="both")
summary(modelo.resulUS)
anova(modelo.resulUS)
```

```
 Start:  AIC=10196.06
CompensationAmount ~ Age + isMale + EmployS + AlgorithU + LEmployT +
    LEduc

            Df Sum of Sq     RSS   AIC
- LEmployT   1      5335 3752957 10196
- LEduc      1      5758 3753380 10196
<none>                   3747622 10196
- EmployS    1     35433 3783054 10206
- isMale     1     56699 3804321 10213
- AlgorithU  1     70680 3818301 10218
- Age        1    465420 4213041 10343

Step:  AIC=10195.88
CompensationAmount ~ Age + isMale + EmployS + AlgorithU + LEduc

            Df Sum of Sq     RSS   AIC
<none>                   3752957 10196
+ LEmployT   1      5335 3747622 10196
- LEduc      1      7547 3760503 10196
- EmployS    1     46739 3799695 10210
- isMale     1     56945 3809901 10213
```

```
- AlgorithU  1     75362 3828318 10219
- Age        1    466310 4219267 10343
> summary(modelo.resulUS)

Call:
lm(formula = CompensationAmount ~ Age + isMale + EmployS + AlgorithU +
    LEduc, data = datos.US2)

Residuals:
    Min      1Q  Median      3Q     Max
-167.67  -33.70   -6.02   25.45  364.35

Coefficients:
             Estimate Std. Error t value    Pr(>|t|)
(Intercept)  3.4127349  7.9087139   0.432       0.666
Age          1.9117118  0.1522443  12.557     < 2e-16 ***
isMale      17.2680500  3.9352567   4.388 0.000012387 ***
EmployS      0.0009264  0.0002330   3.975 0.000074221 ***
AlgorithU    6.9467535  1.3761378   5.048 0.000000511 ***
LEduc        2.1257664  1.3307307   1.597       0.110
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.38 on 1269 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.1737,  Adjusted R-squared:  0.1704
F-statistic: 53.35 on 5 and 1269 DF,  p-value: < 2.2e-16

> anova(modelo.resulUS)
Analysis of Variance Table

Response: CompensationAmount
            Df  Sum Sq Mean Sq  F value       Pr(>F)
Age          1  578523  578523 195.6181    < 2.2e-16 ***
isMale       1   66247   66247  22.4003 0.0000024617 ***
EmployS      1   45616   45616  15.4243 0.0000904995 ***
AlgorithU    1   91003   91003  30.7711 0.0000000353 ***
LEduc        1    7547    7547   2.5518       0.1104
Residuals 1269 3752957    2957
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Deduce eliminar el atributo tiempo trabajando (EmployT), obteniendo un modelo similar a nivel de bondad de ajuste.

D.4.II. Modelo España

Como en el caso anterior, construimos un modelo a partir de todos los valores y analizamos sus parametros:

```
modelo.SP <- lm(CompensationAmount ~ Age + isMale + EmployS + AlgorithU + LEmployT + LEduc  ,
data=datos.SP2)
summary(modelo.SP)
anova(modelo.SP)
```

```
Call:
lm(formula = CompensationAmount ~ Age + isMale + EmployS + AlgorithU +
    LEmployT + LEduc, data = datos.SP2)

Residuals:
    Min      1Q  Median      3Q     Max
-35.423  -9.161  -1.760   7.718  45.054

Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept) -7.81466057  8.04768489  -0.971   0.3337
Age          1.07474008  0.15887552   6.765 7.35e-10 ***
isMale       4.72709016  4.08041921   1.158   0.2492
EmployS     -0.00007064  0.00026810  -0.263   0.7927
AlgorithU    3.30420173  1.32085717   2.502   0.0139 *
LEmployT     0.63908289  0.37507237   1.704   0.0913 .
LEduc       -1.44398743  1.27957639  -1.128   0.2616
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.15 on 107 degrees of freedom
  (1 observation deleted due to missingness)
```

```
Multiple R-squared:  0.3744,  Adjusted R-squared:  0.3394
F-statistic: 10.67 on 6 and 107 DF,  p-value: 0.00000000284

> anova(modelo.SP)
Analysis of Variance Table

Response: CompensationAmount
           Df  Sum Sq Mean Sq F value    Pr(>F)
Age         1 12126.3 12126.3 52.8199 6.262e-11 ***
isMale      1   161.8   161.8  0.7048    0.4031
EmployS     1     4.6     4.6  0.0200    0.8878
AlgorithU   1  1533.7  1533.7  6.6803    0.0111 *
LEmployT    1   585.1   585.1  2.5486    0.1133
LEduc       1   292.4   292.4  1.2735    0.2616
Residuals 107 24564.9   229.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Donde apreciamos lo siguiente:

- Los residuos no aparecen centrados en 0 y algo escorados a la izquierda(no son normales centrados en 0).
- El punto de interceptación estimado se situa cercano a los -7000 euros.
- Los valores de los parametros estimados son positivos y negativos como era de esperar a partir de lo visto en la grafica, siendo algunos significativos ((Pr(>|t|) bajo) para Edad(Age) y conocimiento algoritmos(AlgorithU) ,
- El error es muy grande al (15) y los coeficientes de Regresion R bajos (deberían ser por encima de 0.6 para empezar a valorar un ajuste bueno).
- Viendose el tamaño del error medio claramente en el analisis ANOVA.

Si procedemos a una optimización del modelo con step/AIC:

```
modelo.resulSP<-step(modelo.SP, direction="both")
summary(modelo.resulSP)
anova(modelo.resulSP)
```

```
Start:  AIC=626.51
CompensationAmount ~ Age + isMale + EmployS + AlgorithU + LEmployT +
    LEduc

            Df Sum of Sq   RSS    AIC
- EmployS    1      15.9 24581 624.58
- LEduc      1     292.4 24857 625.86
- isMale     1     308.1 24873 625.93
<none>                   24565 626.51
- LEmployT   1     666.5 25231 627.56
- AlgorithU  1    1436.7 26002 630.99
- Age        1   10505.7 35071 665.10

Step:  AIC=624.58
CompensationAmount ~ Age + isMale + AlgorithU + LEmployT + LEduc

            Df Sum of Sq   RSS    AIC
- isMale     1     294.7 24876 623.94
- LEduc      1     297.0 24878 623.95
<none>                   24581 624.58
- LEmployT   1     651.1 25232 625.56
+ EmployS    1      15.9 24565 626.51
- AlgorithU  1    1422.4 26003 628.99
- Age        1   10539.7 35120 663.26

Step:  AIC=623.94
CompensationAmount ~ Age + AlgorithU + LEmployT + LEduc

            Df Sum of Sq   RSS    AIC
- LEduc      1     254.4 25130 623.10
<none>                   24876 623.94
- LEmployT   1     483.3 25359 624.13
+ isMale     1     294.7 24581 624.58
+ EmployS    1       2.5 24873 625.93
- AlgorithU  1    1526.6 26402 628.73
```

```
- Age        1   10763.3 35639 662.93

Step:  AIC=623.1
CompensationAmount ~ Age + AlgorithU + LEmployT

            Df Sum of Sq   RSS    AIC
- LEmployT   1     424.7 25555 623.01
<none>                   25130 623.10
+ LEduc      1     254.4 24876 623.94
+ isMale     1     252.1 24878 623.95
+ EmployS    1       5.1 25125 625.08
- AlgorithU  1    1418.6 26549 627.36
- Age        1   10604.0 35734 661.23

Step:  AIC=623.01
CompensationAmount ~ Age + AlgorithU

            Df Sum of Sq   RSS    AIC
<none>                   25555 623.01
+ LEmployT   1     424.7 25130 623.10
+ LEduc      1     195.8 25359 624.13
+ isMale     1     110.4 25444 624.52
+ EmployS    1       0.0 25555 625.01
- AlgorithU  1    1587.8 27142 627.88
- Age        1   10512.2 36067 660.29
> summary(modelo.resulSP)

Call:
lm(formula = CompensationAmount ~ Age + AlgorithU, data = datos.SP2)

Residuals:
    Min      1Q  Median      3Q     Max
-36.595  -8.840  -0.627   7.662  44.163

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.7915     6.3614  -1.068  0.28801
Age           1.0345     0.1531   6.757 6.79e-10 ***
AlgorithU     3.4220     1.3031   2.626  0.00985 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.17 on 111 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3492,  Adjusted R-squared:  0.3375
F-statistic: 29.78 on 2 and 111 DF,  p-value: 4.415e-11

> anova(modelo.resulSP)
Analysis of Variance Table

Response: CompensationAmount
           Df  Sum Sq Mean Sq F value    Pr(>F)
Age         1 12126.3 12126.3 52.6723 5.709e-11 ***
AlgorithU   1  1587.8  1587.8  6.8967  0.009854 **
Residuals 111 25554.6   230.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Deduce dejar solo dos paraetros Edad(Age) y conocimiento de algoritmos.

D.5. Conclusiones.Resolución del problema.

La primera conclusión evidente es que la bondad del modelo no es buena como tal, dado seguramente por:

- este falto de más parámetros,
- no se ha cumplido la homocedasticidad ,
- alguno de los parametros existentes no presentan comportamiento lineal
- requiere un analisis más detallado de covarianzas y correlaciones no lineales(logaritmicas...)

De todas formas el análisis de los datos si permite ver un comportamiento en la media de sueldos relevante:

- Se aprecia crecimiento por la mayoría de factores.
- Se aprecia un sesgo importante por genero.

- En el mercado español se aprecia claramente el efecto de la crisis.

E - REFERENCIAS

- Fuente de datos:
  - https://www.kaggle.com/kaggle/kaggle-survey-2017
- Proyecto GITHUB
  - https://github.com/jnavajasb/UOCTPCVD_PRAC2/
- Herramientas de R utilizadas:
  - Rcmdr
  - dplyr
  -
- Fuente datos adicionales:
  - Paises:http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators#
  -
- Documentos:
  - Seleccion de modelos:http://www.stat.umn.edu/geyer/5931/mle/sel.pdf
  - Seleccion de modelos:https://sites.ualberta.ca/~lkgray/uploads/7/3/6/2/7362679/slides_-_multiplelinearregressionaic.pdf
  -