Limpieza de la base de datos: "Ask a Manager"

Visualización y storytelling

Shirley Sánchez Sedano

Todo el proceso presentado a continuación se realiza en RStudio

Limpieza de la base de datos:

Se procede primero a cargar la base de datos

```
Informacion <- read_excel("Ask A Manager Salary Survey 2021 (Responses).xlsx")
```

1.Variables en base de datos original:

La base de datos cuenta con 27.663 registros y 18 columnas, estas variables son:

How old are you?: Es una variable de tipo texto y se refiere al rango de edad de los encuestados. Entre las opciones de respuesta se tiene:

- under 18
- 18-24
- 25-34
- 45-54
- 55-64
- 65 or over

No existen valores nulos en esta columna

Industry Es una variable de tipo texto y se refiere al tipo de industria en el que labora, entre las que se encuentran: Computing or Tech,Education (Higher Education), Government and Public Administration,Insurance, entre otras. Hay 69 valores nulos.

Job title: Es una variable de tipo texto y se refiere al tipo de profesión de los encuestados, entre las que se encuentran:Software Engineer,Project Manager,Teacher,Data Analyst, entre otras.No hay valores nulos

Additional context on job title: Es una variable de tipo texto y se refiere a información adicional sobre la profesión. Hay 20.514 datos nulos

Annual salary: Es una variable de tipo númerico y se refiere al salario anual que reciben los profesionales. No existen valores nulos

Other monetary comp: Es una variable de tipo númerico y se refiere a otros ingresos adicionales que reciben los profesionales. Existen 7.158 valores nulos

Currency Es una variable de tipo texto y corresponde al tipo de moneda en la que se encuentra el salario anual. Estas divisas son:

- AUD/NZD

- CAD
- CHF
- EUR
- GBP
- HKD
- JPY
- SEK
- USD
- ZAR
- Other

No se encuentran valores nulos

Currency - other Es una variable de tipo texto y corresponde al tipo de moneda en la que se encuentra la variable correspondiente a otros ingresos. Se encuentran 27.471 valores nulos y además se presentan registros que no corresponden a nombres de divisas, por ejemplo: Additonal = Bonus plus stock, additional compensation is for overtime (i am paid hourly) so it varies. i have included an estimate, entre otros.

Additional context on income Es una variable de tipo texto y corresponde información adicional sobre los ingresos.

Country Es una variable de tipo texto y corresponde al pais de residencia del encuestado.No se encuentran valores nulos

State Es una variable de tipo texto y corresponde al estado en el que se encuentra ubicada la persona. Se encuentran 4.922 valores nulos

City Es una variable de tipo texto y corresponde a la ciudad en el que se encuentra ubicada la persona. Se encuentran 9 valores nulos

Overall years of professional experience: Es una variable de tipo texto y corresponde a los años de experiencia profesional del encuestado. Entre las opciones de respuesta se encuentran los siguientes rangos:

- 1 year or less
- 2 - 4 years
- 5-7 years
- 8 - 10 years
- 11 - 20 years
- 21 - 30 years
- 31 - 40 years
- 41 years or more

No hay valores nulos.

Years of experience in field: Es una variable de tipo texto y corresponde a los años de experiencia en el campo laboral actual. Entre las opciones de respuesta se encuentran los siguientes rangos:

- 1 year or less
- 2 - 4 years
- 5-7 years
- 8 - 10 years
- 11 - 20 years
- 21 - 30 years
- 31 - 40 years
- 41 years or more

No hay valores nulos.

Highest level of education completed: Es una variable de tipo texto y corresponde al nivel de estudio de la persona. Las opciones son las siguientes:

- High School
- College degree
- Some college
- Professional degree (MD, JD, etc.)
- Master's degree
- PhD

Se encuentran 207 valores nulos

Gender: Es una variable de tipo texto y corresponde al género de la persona. Las opciones de respuesta son las siguientes:

- Woman
- Man
- Non-binary Other or prefer not to answer
- Prefer not to answer

Se encuentran 164 valores nulos

Race: Es una variable de tipo texto y corresponde a la raza de la persona.

2. Variables luego de modeladas:

Para este punto se dejará tanto el tipo como el nombre de las variables igual a la base original, pero se crean 3 nuevas variables las cuales son:

Annual_salary_COP: Es una variable de tipo númerico que refleja el salario anual en pesos colombianos.

Other_income_COP: Es una variable de tipo númerico que muestra otros ingresos en pesos colombianos.

Para realizar el cambio de moneda se usa el valor de las divisas para el 13 de febrero del 2022 según la información proporcionada por Google Finance. Además hay que aclarar que todos los cambios se realizarán con respecto a la variable "Currency" y cuando la divisa sea igual a "Other" se tomará como USD, esto debido a que no se cuenta con mucho tiempo para realizar todos los cambios.

Para crear las dos variables anteriores se ejecuta el siguiente código donde primero los valores NA de la variable "Other monetary comp" se convertiran a cero y posteriormente se realiza el cambio de divisa

```
## Cambiar los valores NA de la variable "Other monetary comp"

for (i in 1:nrow(Informacion)){
  if(is.na(Informacion$`Other monetary comp`[i])){
    Informacion$`Other monetary comp`[i]<-0
  }
}

## Cambio de divisa

for (i in 1:nrow(Informacion)){
  if(Informacion$Currency[i]=="CAD"){
    ## 1 CAD= 3.093,33 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*3093.33
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*3093.33
  }
  else if(Informacion$Currency[i]=="CHF"){
    ## 1 CHF= 4.259,49 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*4259.49
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*4259.49
  }
  else if(Informacion$Currency[i]=="EUR"){
    ## 1 EUR= 4.466,55 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*4466.55
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*4466.55
  }
  else if(Informacion$Currency[i]=="GBP"){
    ## 1 GBP= 5.341,87  COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*5341.87
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*5341.87
  }
  else if(Informacion$Currency[i]=="HKD"){
    ## 1 HKD= 505,04 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*505.04
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*505.04
  }
  else if(Informacion$Currency[i]=="JPY"){
    ## 1 JPY= 34,15 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*34.15
```

```
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*34.15
  }
  else if(Informacion$Currency[i]=="SEK"){
    ## 1 SEK= 422.24 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*422.24
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*422.24
  }
  else if(Informacion$Currency[i]=="USD"){
    ## 1 USD= 3939.71 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*3939.71
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*3939.71
  }
  else if(Informacion$Currency[i]=="ZAR"){
    ## 1 ZAR = 258,98 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*258.98
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*258.98
  }
  else if(Informacion$Currency[i]=="AUD/NZD"){
    ## 1 AUD = 2.808,72 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*2808.72
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*2808.72
  }
  else if(Informacion$Currency[i]=="Other"){
    ## 1 USD = 3939.71 COP
    Informacion$Annual_salary_COP[i]<-Informacion$`Annual salary`[i]*3939.71
    Informacion$Other_income_COP[i]<-Informacion$`Other monetary comp`[i]*3939.71
  }
}

##Redondear los valores

Informacion$Annual_salary_COP<-round(Informacion$Annual_salary_COP,0)
Informacion$Other_income_COP<-round(Informacion$Other_income_COP,0)
```

Total_COP Es una variable de tipo número y corresponde a la suma de las variables Annual_salary(COP) y Other_income(COP). Se ejecuta las siguientes lineas para crear la nueva variable

```
Informacion$Total_COP<-Informacion$Annual_salary_COP+Informacion$Other_income_COP
```

3.Limpieza de las columnas “Country” y "City:

Limpieza de la variable “Country”

Para realizar la limpieza de la variable Country se siguen los siguientes pasos:

1. El nombre de cada uno de los paises se pasará a letras minúsculas. Esto se realiza con la función “tolower”

```
Informacion$Country<-tolower(Informacion$Country)
```

2.Eliminar los espacios vacios al principio y al final del nombre del pais. Se usa la función trimws

```
Informacion$Country<-trimws(Informacion$Country)
```

3. Se eliminan los acentos con la función chartr

```
Informacion$Country<-chartr("áéíóú", "aeiou",(Informacion$Country))
```

4. Se extraen los valores únicos de la variable Country

```
paises<-unique(Informacion$Country)
```

Como se observa se cuentan con 250 datos. Con este último resultado es con el que se procede a realizar la limpieza

5. La limpieza de datos consiste en para cada uno de los paises con nombres similares se unirán en un solo grupo. Por ejemplo “australi”,“australia” y “australian” se unirán bajo el nombre de “australia”. Lo anterior se realizá a través de la funciones propiamente creadas

```
## Reemplazar por United States

valoresusa<- c("america","the us","u.s","u.s.","u.s","u.s.","u.s.a","u.s.a.","u.s>",
               "u.sa","uniited states","unite states","united  states","united sates",
               "united sates of america","united stares","united state",
               "united state of america","united statea","united stated",
               "united stateds","united statees","united states",
               "united states (i work from home and my clients are all
               over the us/canada/pr","united states is america","united states
               of america","united states of american","united states of americas",
               "united states- puerto rico","united statesp",
               "united statew","united statss","united stattes","united statues",
               "united status","united statws","united sttes","united y",
               "unitedstates","uniteed states","unitef stated","uniter statez",
               "unites states","unitied states","uniyed states",
               "uniyes states","unted states","untied states","us",
               "us govt employee overseas, country withheld","us of a",
               "usa","usa (company is based in a us territory, i work remote)",
               "usa tomorrow","usa-- virgin islands","usa, but for foreign gov't",
               "usaa","the united states","for the united states government,
               but posted overseas","uxz","uss","usd","usat","usab","usa","us","u. s.")


for (i in 1:length(valoresusa)){
  Informacion$Country[Informacion$Country==valoresusa[i]]<-"united states"
}

## Reemplazar por new zeland

valoresnz<- c("aotearoa new zealand","from new zealand but on projects across apac",
              "new zealand","new zealand aotearoa","nz")

for (i in 1:length(valoresnz)){
  Informacion$Country[Informacion$Country==valoresnz[i]]<-"new zeland"
```

```
}

## Reemplazar por argentina

valoresarg<- c("argentina","argentina but my org is in thailand",
               "i work for an us based company but i'm from argentina")

for (i in 1:length(valoresarg)){
  Informacion$Country[Informacion$Country==valoresarg[i]]<-"argentina"
}

## Reemplazar por australia

valoresaustralia<- c("australi","australia","australian")

for (i in 1:length(valoresaustralia)){
  Informacion$Country[Informacion$Country==valoresaustralia[i]]<-"australia"
}

## Reemplazar por austria

valoresaustria<- c("austria","austria, but i work remotely for a dutch/british company")

for (i in 1:length(valoresaustria)){
  Informacion$Country[Informacion$Country==valoresaustria[i]]<-"austria"
}

## Reemplazar por brasil

valoresbrasil<- c("brasil","brazil")

for (i in 1:length(valoresbrasil)){
  Informacion$Country[Informacion$Country==valoresbrasil[i]]<-"brazil"
}

## Reemplazar por canada

valorescanada<- c("can","canad","canada","canada and usa","canada,
                  ottawa, ontario","canadw","canda","csnada")

for (i in 1:length(valorescanada)){
  Informacion$Country[Informacion$Country==valorescanada[i]]<-"canada"
}

## Reemplazar por united kingdom

valoresukingdom<- c("england","england, gb","england, uk","england, uk.",
                    "england, united kingdom","england/uk","englang",
                    "northern ireland","northern ireland, united kingdom",
                    "scotland","scotland, uk","u.k","u.k.","u.k. (northern england)",
                    "uk","uk (england)","uk (northern ireland)","uk for u.s. company",
                    "uk, but for globally fully remote company","uk, remote",
                    "united kindom","united kingdom","united kingdom (england)",
```

```
                    "united kingdom.","united kingdomk","unites kingdom","wales"
                    ,"wales (uk)","wales (united kingdom)","wales, uk","britain",
                    "great britain")

for (i in 1:length(valoresukingdom)){
  Informacion$Country[Informacion$Country==valoresukingdom[i]]<-"united kingdom"
}

## Reemplazar por germany

valoresgermany<- c("company in germany. i work from pakistan.","germany")

for (i in 1:length(valoresgermany)){
  Informacion$Country[Informacion$Country==valoresgermany[i]]<-"germany"
}

## Reemplazar por costa de marfil

valorescm<- c("cote d'ivoire")

for (i in 1:length(valorescm)){
  Informacion$Country[Informacion$Country==valorescm[i]]<-"ivory coast"
}

## Reemplazar por czech republic

valoresrc<- c("czech republic","czechia")

for (i in 1:length(valoresrc)){
  Informacion$Country[Informacion$Country==valoresrc[i]]<-"czech republic"
}

## Reemplazar por denmark

valoresdm<- c("denmark","danmark")

Informacion$Country[grepl(valoresdm, Informacion$Country)]<-"denmark"

for (i in 1:length(valoresdm)){
  Informacion$Country[Informacion$Country==valoresdm[i]]<-"denmark"
}

## Reemplazar por romania

valoresromania<- c("from romania, but for an us based company","romania")

for (i in 1:length(valoresromania)){
  Informacion$Country[Informacion$Country==valoresromania[i]]<-"romania"
}

## Reemplazar por hong kong

valoreshong<- c("hong kong","hong konh")
```

```r
for (i in 1:length(valoreshong)){
  Informacion$Country[Informacion$Country==valoreshong[i]]<-"hong kong"
}

## Reemplazar por iceland

valoresiceland<- c("i.s.","is","isa")

for (i in 1:length(valoresiceland)){
  Informacion$Country[Informacion$Country==valoresiceland[i]]<-"iceland"
}

## Reemplazar por italy

valoresitaly<- c("italy","italy (south)")

for (i in 1:length(valoresitaly)){
  Informacion$Country[Informacion$Country==valoresitaly[i]]<-"italy"
}


## Reemplazar por japan

valoresjapan<- c("japan","japan, us gov position")

for (i in 1:length(valoresjapan)){
  Informacion$Country[Informacion$Country==valoresjapan[i]]<-"japan"
}

## Reemplazar por jersey

valoresjersey<- c("jersey, channel islands")

for (i in 1:length(valoresjersey)){
  Informacion$Country[Informacion$Country==valoresjersey[i]]<-"jersey"
}

## Reemplazar por luxembourg

valoreslux<- c("luxembourg","luxemburg")

for (i in 1:length(valoreslux)){
  Informacion$Country[Informacion$Country==valoreslux[i]]<-"luxembourg"
}

## Reemplazar por china

valoreschina<- c("china","mainland china")

for (i in 1:length(valoreschina)){
  Informacion$Country[Informacion$Country==valoreschina[i]]<-"china"
}
```

```r
## Reemplazar por philippines

valoresphilippines<- c("philippines","remote (philippines)")

for (i in 1:length(valoresphilippines)){
  Informacion$Country[Informacion$Country==valoresphilippines[i]]<-"philippines"
}


## Reemplazar por ukraine

valoresukraine<- c("ukraine","u.a.","ua")

for (i in 1:length(valoresukraine)){
  Informacion$Country[Informacion$Country==valoresukraine[i]]<-"ukraine"
}

## Reemplazar por united arab emirates

valoresea<- c("united arab emirates","uae")

for (i in 1:length(valoresea)){
  Informacion$Country[Informacion$Country==valoresea[i]]<-"united arab emirates"
}

## Reemplazar por netherlands

valoresnt<- c("the netherlands","netherlands")

for (i in 1:length(valoresnt)){
  Informacion$Country[Informacion$Country==valoresnt[i]]<-"netherlands"
}

## Reemplazar por na

valoresna<- c("$2,175.84/year is deducted for benefits",
              "bonus based on meeting yearly goals set w/ my supervisor",
              "contracts","currently finance",
              "i am located in canada but i work for a company in the us",
              "i earn commission on sales. if i meet quota, i'm guaranteed
              another 16k min. last year i earned an additional 27k. it's
              not uncommon for people in my space to earn 100k+ after commission.",
              "i was brought in on this salary to help with the ehr and very
              quickly was promoted to current position but compensation was
              not altered.","i work for a uae-based organization, though i
              am personally in the us.","international","n/a (remote from
              wherever i want)","na","remote","we don't get raises, we get
              quarterly bonuses, but they periodically asses income in the
              area you work, so i got a raise because a 3rd party assessment
              showed i was paid too little for the area we were located",
              "worldwide (based in us but short term trips aroudn the world)",
              "y","global","nl","policy")
```

```
for (i in 1:length(valoresna)){
  Informacion$Country[Informacion$Country==valoresna[i]]<-"na coast"
}
```

Limpieza de la variable "City"

El mismo procedimiento planteado anteriomente con la variable "Country" se realiza para la homogenización de esta columna. Al ser un proceso extenso y por cuestión no tiempo no será presentado en este informe.

Posteriormente se extrae la base de datos para realizar en dashboard en google data studio