TP final: Salarios en STEM

Jerónimo Barragán, Guido Rossi, Florencia Fontana Walser

2022-11-16

Análisis exploratorio

Dataset original

```
require(tidyverse)
require(dplyr)
require(usdata)
```

```
## Warning: package 'usdata' was built under R version 4.2.2
```

```
require(usmap)
```

```
## Warning: package 'usmap' was built under R version 4.2.2
```

```
require(ggridges)
```

```
## Warning: package 'ggridges' was built under R version 4.2.2
```

```
require("Hmisc")
require(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.2.2
```

```
require(rje)
```

Leemos el dataset, el cual fue sacado de Kaggle. Los datos provienen de _Levels.fyi_, un sitio que recopila salarios de muchos rubros y empresas en todo el mundo. Este dataset contiene salarios de trabajadores del área de tecnología, ingeniería, ciencia y matemática de las más grandes companías.

```
datos <- read_csv("Levels-Fyi-Salary-Data.csv")
```

```
## Rows: 62642 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): timestamp, company, level, title, location, tag, gender, otherdeta...
## dbl (19): totalyearlycompensation, yearsofexperience, yearsatcompany, basesa...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(datos)
```

```
## # A tibble: 6 x 29
##   timestamp    company level title total~1 locat~2 years~3 years~4 tag   bases~5
##   <chr>        <chr>   <chr> <chr>   <dbl> <chr>     <dbl>   <dbl> <chr>   <dbl>
## 1 6/7/2017 11~ Oracle  L3    Prod~  127000 Redwoo~     1.5     1.5 <NA>   107000
## 2 6/10/2017 1~ eBay    SE 2  Soft~  100000 San Fr~     5       3   <NA>        0
## 3 6/11/2017 1~ Amazon  L7    Prod~  310000 Seattl~     8       0   <NA>   155000
## 4 6/17/2017 0~ Apple   M1    Soft~  372000 Sunnyv~     7       5   <NA>   157000
## 5 6/20/2017 1~ Micros~ 60    Soft~  157000 Mounta~     5       3   <NA>        0
## 6 6/21/2017 1~ Micros~ 63    Soft~  208000 Seattl~     8.5     8.5 <NA>        0
## # ... with 19 more variables: stockgrantvalue <dbl>, bonus <dbl>, gender <chr>,
## #   otherdetails <chr>, cityid <dbl>, dmaid <dbl>, rowNumber <dbl>,
## #   Masters_Degree <dbl>, Bachelors_Degree <dbl>, Doctorate_Degree <dbl>,
## #   Highschool <dbl>, Some_College <dbl>, Race_Asian <dbl>, Race_White <dbl>,
## #   Race_Two_Or_More <dbl>, Race_Black <dbl>, Race_Hispanic <dbl>, Race <chr>,
## #   Education <chr>, and abbreviated variable names 1: totalyearlycompensation,
## #   2: location, 3: yearsofexperience, 4: yearsatcompany, 5: basesalary
```

Vemos que el dataset contiene alrededor de 62 mil filas y 29 columnas. Las `timestamps` (fecha y hora en que se registro la fila) van del 2017 al 2021, las variables catégoricas están "dummificadas", es decir, en el dataset aparecian con valores 0 y 1 (lo cual es incómodo), y las ubicaciones de los trabajos son de varios países (sólo nos van a interesar las de los Estados Unidos). Además, variables como `gender`, `Race` y `Education` contienen NAs. Limpiaremos todo esto para trabajar con un dataset más prolijo.

```
# retenemos las columnas relevantes
datos2 <- datos %>% select(company, title,bonus,
                                location,yearsofexperience,
                                yearsatcompany,
                                basesalary,gender,
                                Race,Education) %>%
  filter(!is.na(gender),!is.na(Race),!is.na(Education))

# En este caso queremos quedarnos solamente con los salarios de trabajos en Estados Unidos. En la colum
# Para quedarnos con los de EE.UU. retenemos las locaciones que tengan el código del estado al final co

datos_usa <- subset(datos2, grepl("{2}[A-Z]$", location))
# Creo columna con estados y le cambio la sigla por el nombre
datos_usa$state <- sapply(strsplit(datos_usa$location, ","), "[", 2)
datos_usa$state <- gsub( " ", "", datos_usa$state)
datos_usa$state <-  abbr2state(datos_usa$state)

# Modificamos columna con ciudades
colnames(datos_usa)[4] <- "city"
datos_usa$city <- sapply(strsplit(datos_usa$city, ","), "[", 1)

# Hacemos de la educación una variable ordinal

datos_usa$Education <- factor(datos_usa$Education, levels = c("Highschool", "Some College", "Bachelor's
```

Dataset limpio

Ahora tenemos una columna para la ciudad y otra para el estado. Tampoco quedaron NA's en el dataset. Las columnas con las que trabajaremos son las siguientes:

- `company`: la compañía donde trabaja el empleado.
- `title`: título.
- `bonus`: bonificación. Está basada mayormente en la performance anual de la compañía.
- `city`: ciudad.
- `yearsofexperience`: años de experiencia.
- `yearsatcompany`: años que lleva trabajando en la compañía.
- `basesalary`: salario base.
- `gender`: género.
- `Race`: etnia.
- `Education`: nivel educativo.
- `state`: Estado donde vive.

```
head(datos_usa)
```

```
## # A tibble: 6 x 11
##   company   title bonus city  years~1 years~2 bases~3 gender Race  Educa~4 state
##   <chr>     <chr> <dbl> <chr>   <dbl>   <dbl>   <dbl> <chr>  <chr> <fct>   <chr>
## 1 Google    Soft~ 45000 Sunn~       5       5  210000 Male   Asian PhD     Cali~
## 2 Microsoft Soft~ 11000 Redm~       3       2  124000 Male   Two ~ Bachel~ Wash~
## 3 Google    Soft~ 36000 San ~       6       6  177000 Male   Asian Bachel~ Cali~
## 4 Microsoft Soft~ 20000 Seat~       4       4  164000 Male   Asian Master~ Wash~
## 5 Blend     Soft~     0 San ~       5       0  165000 Male   White Bachel~ Cali~
## 6 Amazon    Soft~     0 Seat~      15       3  160000 Male   Asian Bachel~ Wash~
## # ... with abbreviated variable names 1: yearsofexperience, 2: yearsatcompany,
## #   3: basesalary, 4: Education
```

Objetivos y preguntas

- ¿Afecta el nivel de educación en un salario? ¿Es realmente necesario un PhD para tener un salario alto?
- ¿Qué variables inciden en la remuneración de una persona del rubro?
- Predecir el salario de un trabajador del sector.

Análisis exploratorio de nuestros datos

Ubicación

Veamos algunas relaciones entre las variables de nuestro dataset. Comencemos con la ubicación de cada puesto.

```
datos_states <- arrange(as.data.frame(table(datos_usa$state)), desc(Freq))
colnames(datos_states)[1] <- "states"
datos_states
```

```
##                  states Freq
## 1            California 6534
## 2            Washington 3552
## 3              New York 1671
## 4                 Texas 1143
## 5         Massachusetts  609
## 6              Virginia  388
## 7              Illinois  347
## 8                Oregon  260
## 9               Georgia  254
## 10 District of Columbia  220
## 11       North Carolina  215
## 12             Colorado  203
## 13         Pennsylvania  192
## 14              Arizona  153
## 15           New Jersey  144
## 16            Minnesota  127
## 17              Florida  118
## 18             Michigan  104
## 19             Missouri   90
## 20                 Utah   90
## 21                 Ohio   83
## 22              Indiana   49
## 23             Maryland   48
## 24            Wisconsin   47
## 25            Tennessee   40
## 26             Arkansas   38
## 27          Connecticut   38
## 28             Delaware   28
## 29               Kansas   24
## 30            Louisiana   16
## 31              Alabama   15
## 32                Idaho   14
## 33                 Iowa   12
## 34             Kentucky   12
## 35               Nevada   12
## 36             Nebraska   11
## 37       South Carolina   10
## 38        New Hampshire    9
## 39         Rhode Island    9
## 40             Oklahoma    8
## 41              Montana    5
## 42           New Mexico    5
## 43        West Virginia    4
## 44               Hawaii    2
## 45                Maine    2
## 46          Mississippi    2
## 47         North Dakota    2
## 48              Vermont    2
## 49              Wyoming    1
```

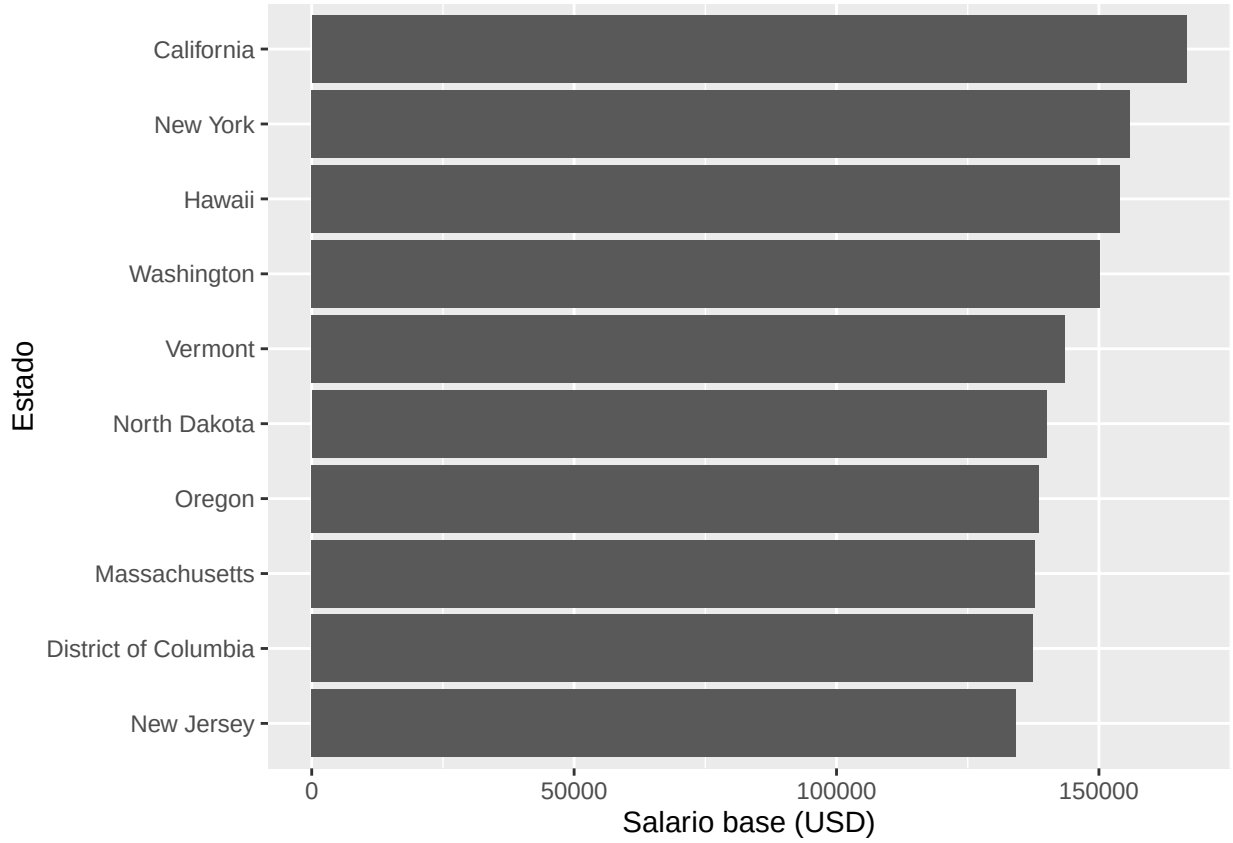
Observamos que la mayoría de puestos provienen de California, Washington y New York.

Veamos el salario promedio por Estado.

```r
prom_states <- datos_usa %>% select(state, basesalary) %>%
  group_by(state) %>%
  mutate(meansalary = mean(basesalary)) %>%
  arrange(desc(meansalary)) %>%
  select(state,meansalary) %>%
  distinct()

top_10 <- prom_states[1:10,]

ggplot(data=top_10, mapping=aes(x=reorder(state, meansalary), y=meansalary)) +
  stat_summary(fun.data=mean_sdl, geom="bar")+
  labs(x = "Estado", y = "Salario base (USD)")+
  coord_flip()
```



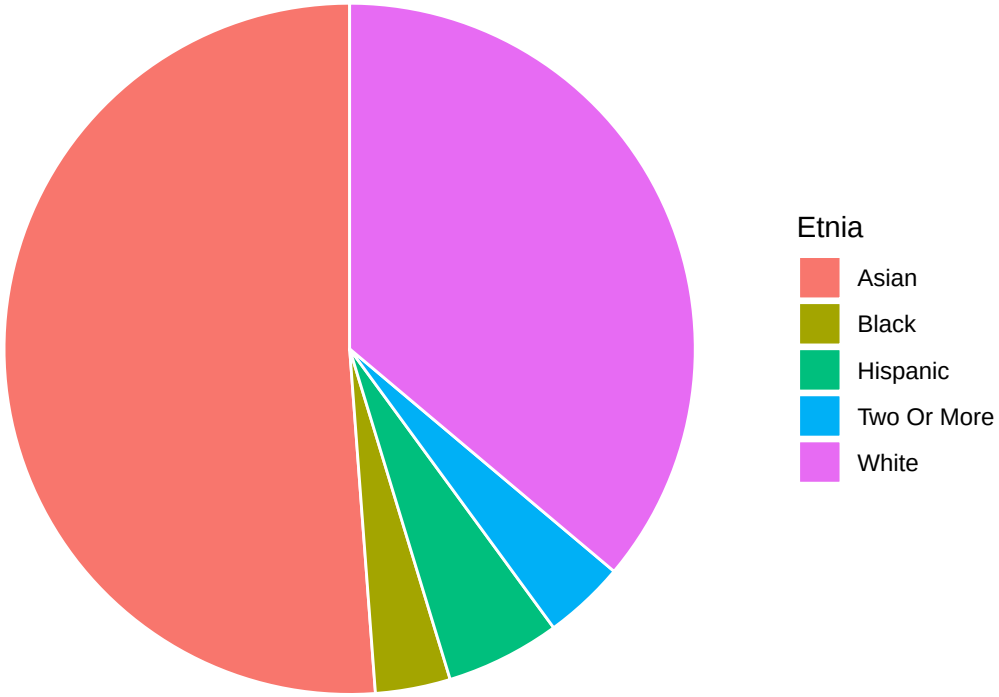
El salario promedio más alto es el de California.

Etnia

```r
etnias <- as.data.frame(table(datos_usa$Race))

raceplot_data <- datos_usa %>%
  count(Race) %>%
  mutate(percent = n/sum(n))
```

```
ggplot(raceplot_data, aes(x="", y=percent*100, fill = Race)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start = 0) +
  labs(fill = "Etnia")+
  theme_void()
```



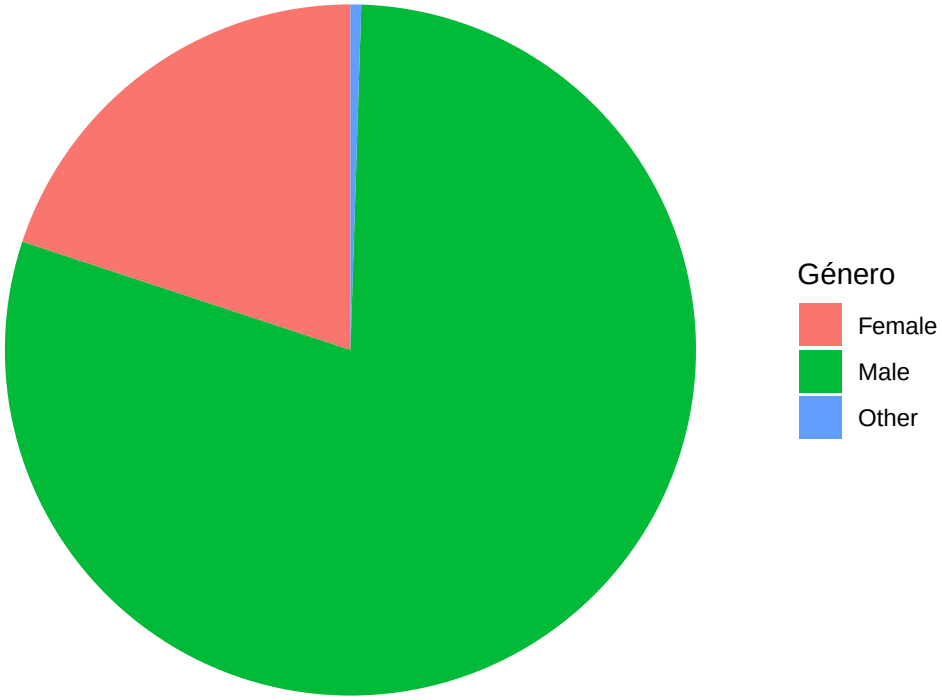
A simple vista, pareciera ser que los asiáticos predominan en el rubro, pues el piechart muestra que la cantidad de asiáticos supera al de cualquier otra etnia.

Género

Consideremos los porcentajes de cada género en nuestro dataset.

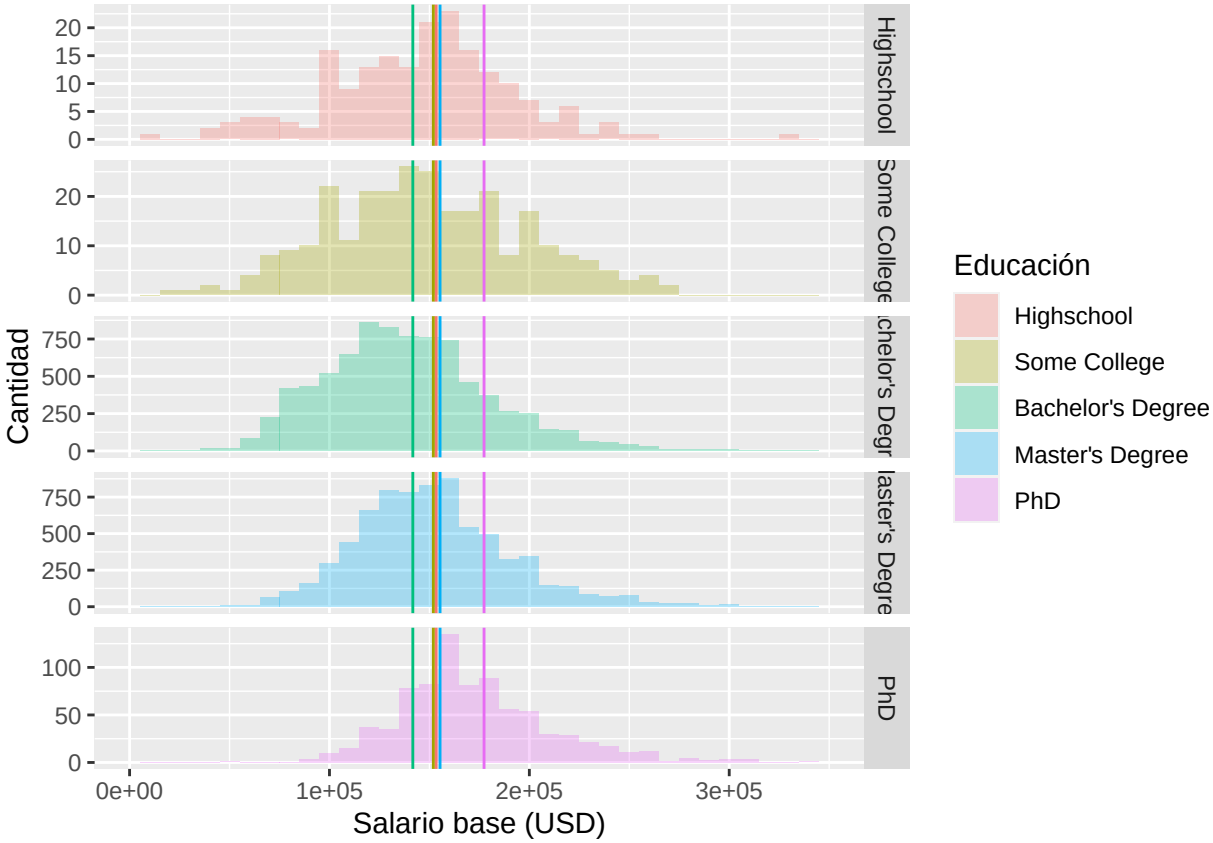
```
data <- as.data.frame(table(datos_usa$gender))

ggplot(data, aes(x="", y=Freq, fill=Var1)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  theme_void() +
  labs(fill = "Género")
```



Educación

```
ggplot(data = datos_usa,
  mapping = aes(
  x = basesalary, fill = Education)) +
  labs(x = "Salario base (USD)", y = "Cantidad", fill = "Educación") +
  geom_histogram(binwidth=10000, alpha = 0.3, position = "identity") +
  xlim(0,3.5e+05) +
  geom_vline(aes(xintercept = mean(datos_usa$basesalary[datos_usa$Education=="Highschool"])), col = "#F8
  geom_vline(aes(xintercept = mean(datos_usa$basesalary[datos_usa$Education=="Some College"])), col = "
  geom_vline(aes(xintercept = mean(datos_usa$basesalary[datos_usa$Education=="Bachelor's Degree"])), col
  geom_vline(aes(xintercept = mean(datos_usa$basesalary[datos_usa$Education=="Master's Degree"])), col =
  geom_vline(aes(xintercept = mean(datos_usa$basesalary[datos_usa$Education=="PhD"])), col = "#E76BF3",
  facet_grid(Education ~ ., scales = "free")
```




```
#Imprimimos la varianza de cada nivel educativo
for(nivel in levels(datos_usa$Education)) {
  print(nivel)
  print(var(datos_usa$basesalary[datos_usa$Education==nivel]))
}
```

```
## [1] "Highschool"
## [1] 4165604383
## [1] "Some College"
## [1] 3335617084
## [1] "Bachelor's Degree"
## [1] 2507037491
## [1] "Master's Degree"
## [1] 2107839283
## [1] "PhD"
## [1] 2542597094
```

Podemos ver que las distribuciones del salario base por nivel de educación se solapan considerablemente entre sí. Es decir, el nivel de educación no parece ser, a priori, un indicador determinante en el salario de una persona que trabaja en el sector. Sin embargo, podemos observar que, conforme el nivel de educación es más elevado, la varianza del salario base tiende a disminuir. Tal vez esto pueda deberse a que, a la hora de determinar las remuneraciones de sus empleados con niveles educativos mayores, las compañías tengan el camino "más claro".

Puesto de trabajo en la compañía

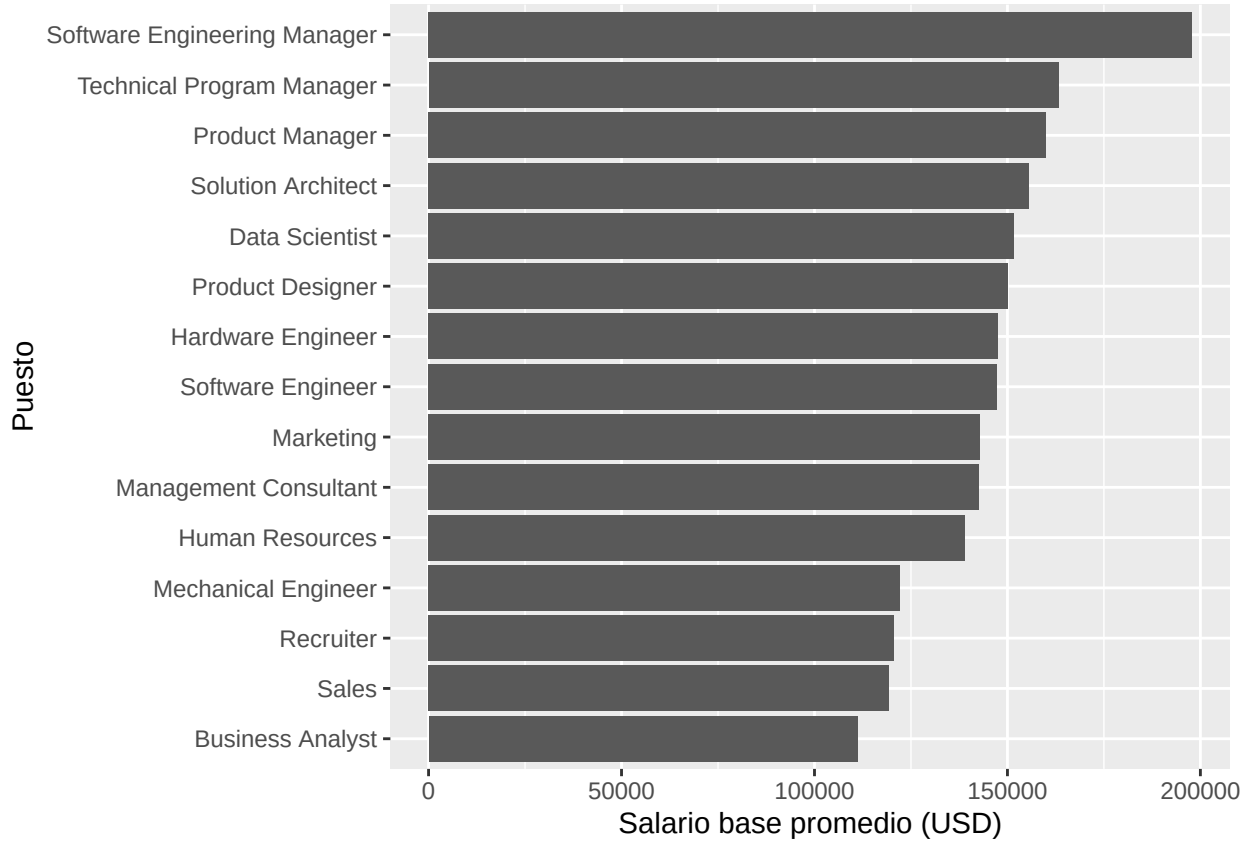
Estudiemos ahora si el salario base depende del puesto de trabajo en la compañía. ¿Qué puestos de trabajo tenemos registrados?

```
tabla <- arrange(as.data.frame(table(datos_usa$title)), desc(Freq))
colnames(tabla) <- c("Puesto", "Cantidad")
tabla
```

```
##                          Puesto Cantidad
## 1             Software Engineer    10373
## 2               Product Manager     1254
## 3  Software Engineering Manager      803
## 4                 Data Scientist      717
## 5              Hardware Engineer      656
## 6     Technical Program Manager      557
## 7               Product Designer      518
## 8         Management Consultant      368
## 9               Business Analyst      361
## 10                     Marketing      323
## 11           Solution Architect      302
## 12          Mechanical Engineer      234
## 13                     Recruiter      190
## 14                         Sales      155
## 15               Human Resources      151
```

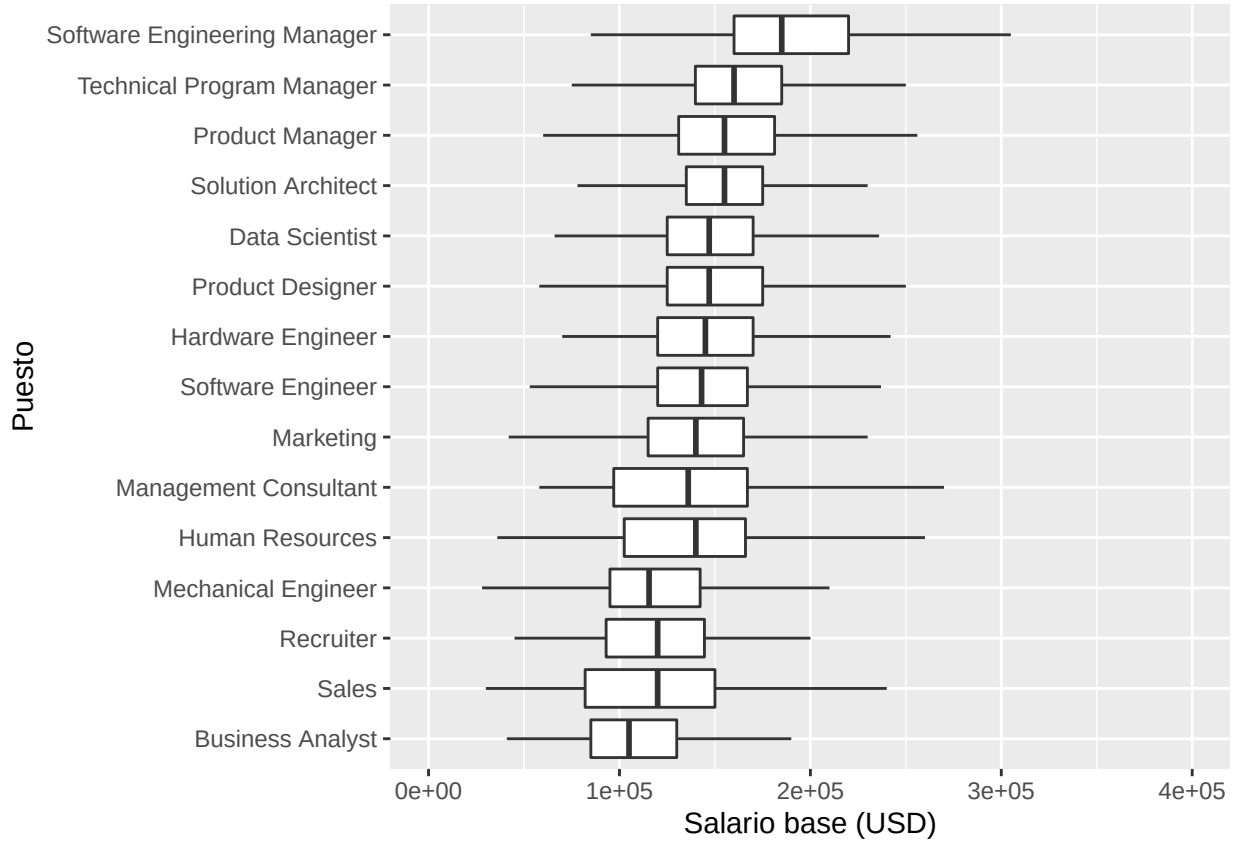
Podemos ver que hay 15 títulos registrados bien definidos en el dataset. Consideremos el salario BASE de las personas que ocupan cada puesto.

```
ggplot(data=datos_usa, mapping=aes(x=reorder(title,basesalary), y=basesalary)) +
  stat_summary(fun.data=mean_sdl, geom="bar")+
  labs(x = "Puesto", y = "Salario base promedio (USD)")+
  coord_flip()
```




```
ggplot(data=datos_usa, mapping=aes(x=reorder(title,basesalary), y=basesalary)) +
  geom_boxplot(outlier.shape = NA) +
  labs(x = "Puesto", y = "Salario base (USD)")+
  ylim(0,400000) +
  coord_flip()
```

```
## Warning: Removed 62 rows containing non-finite values (stat_boxplot).
```



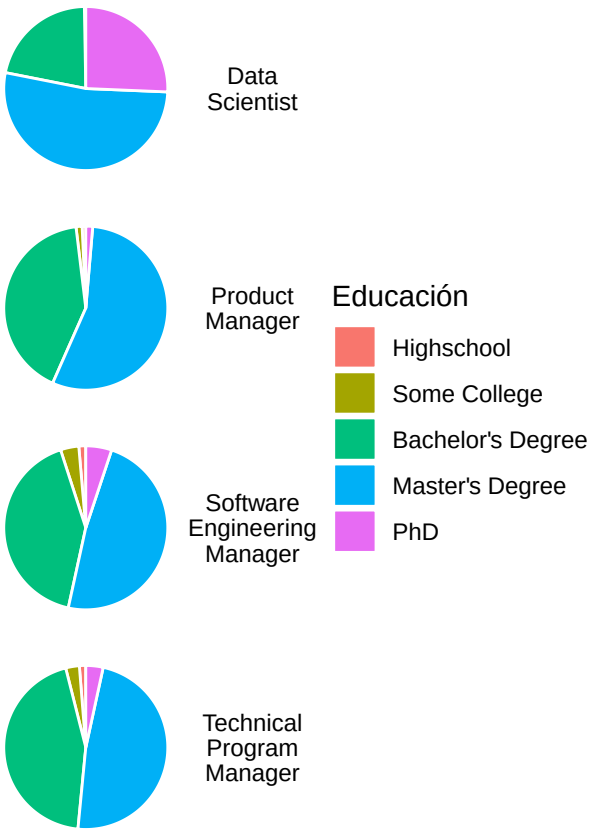
Mediante el boxplot, podemos ver que los cuartiles tienden a correrse hacia la derecha a medida que crece la jerarquía del puesto. Por otra parte, el barplot nos dice que el salario promedio más alto es el de Software Engineering Manager, que es casi de 200.000 dólares anuales. Parecería ser, como es esperado, que los cargos superiores (Managers, por ejemplo) conllevan un salario base superior. Por lo tanto, el puesto de trabajo debe ser un factor incidente en el salario base.

¿Cómo se relaciona la educación con el puesto de trabajo?

```
subsetTitle <- datos_usa[datos_usa$title %in% c( "Data Scientist","Software Engineering Manager", "Techn

titleplot_data <- subsetTitle %>%
  count(Education, title) %>%
  group_by(title) %>%
  mutate(percent = n/sum(n))

ggplot(titleplot_data, aes(x="", y=percent*100, fill = Education)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start = 0) +
  theme_void() +
  labs(fill ="Educación") +
  facet_grid(title ~ .,  labeller = label_wrap_gen(width = 2, multi_line = TRUE))
```



Podemos ver que los puestos de trabajo por lo general mejor remunerados, están ocupados en su mayoría por personas que han alcanzado un Bachelor o un Master. En menor medida, siguen los que han alcanzado un PhD. Por lo tanto, podemos ver que los puestos más altos requieren un nivel educativo por lo menos universitario. En cuanto a los Data Scientists, la proporción de ellos que ha alcanzado un PhD es mayor que para los tres primeros puestos.

Correlaciones entre variables

Veamos las correlaciones entre las variables continuas para tratar de entender lo que nos muestra el dataset. Para eso utilizamos una matriz de correlaciones.

```
datos_usa_cont <- datos_usa %>% select(yearsofexperience,yearsatcompany,basesalary, bonus)
M <- cor(datos_usa_cont)
head(M)
```

```
##                   yearsofexperience yearsatcompany basesalary     bonus
## yearsofexperience         1.0000000      0.5485190  0.4886095 0.2887951
## yearsatcompany            0.5485190      1.0000000  0.1920869 0.1237051
## basesalary                0.4886095      0.1920869  1.0000000 0.3643855
## bonus                     0.2887951      0.1237051  0.3643855 1.0000000
```

```
corrplot(M, method="color")
```

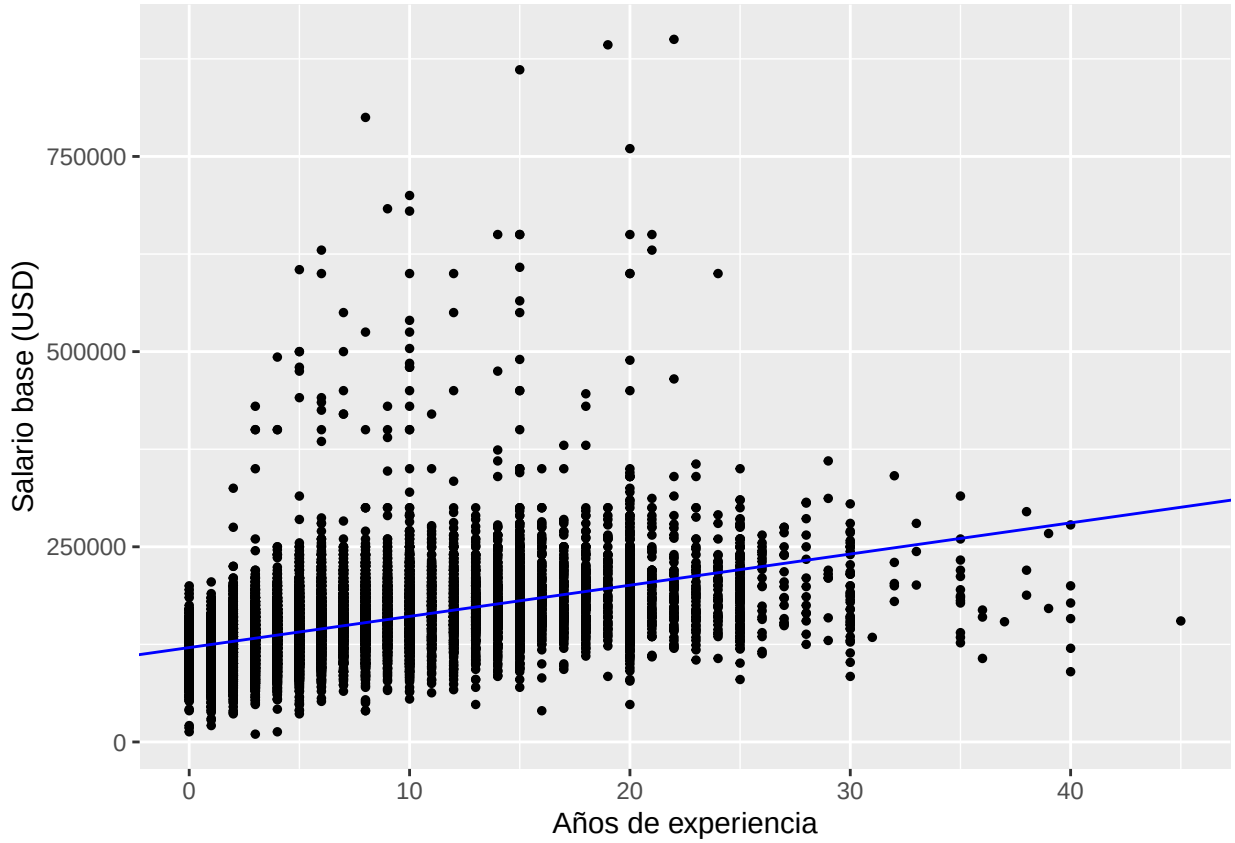
En lo que concierne a nuestro análisis, vemos que el salario base se encuentra positivamente correlacionado con los años de experiencia más que cualquier otra variable, con el bonus en segundo lugar, y un poco menos con los años en la compañía.

Años de experiencia y en la compañía

```
#datos_usa_ds <- datos_usa %>% filter(title == 'Data Scientist')

ajusM1 <- lm(basesalary ~ yearsofexperience, data = datos_usa)

ggplot(data = datos_usa, mapping = aes(x = yearsofexperience, y = basesalary)) +
  geom_point(size=1) +
  geom_abline(color="blue", slope = coef(ajusM1)[2], intercept = coef(ajusM1)[1]) +
  labs(y = "Salario base (USD)", x = "Años de experiencia", color = "Género"  )
```

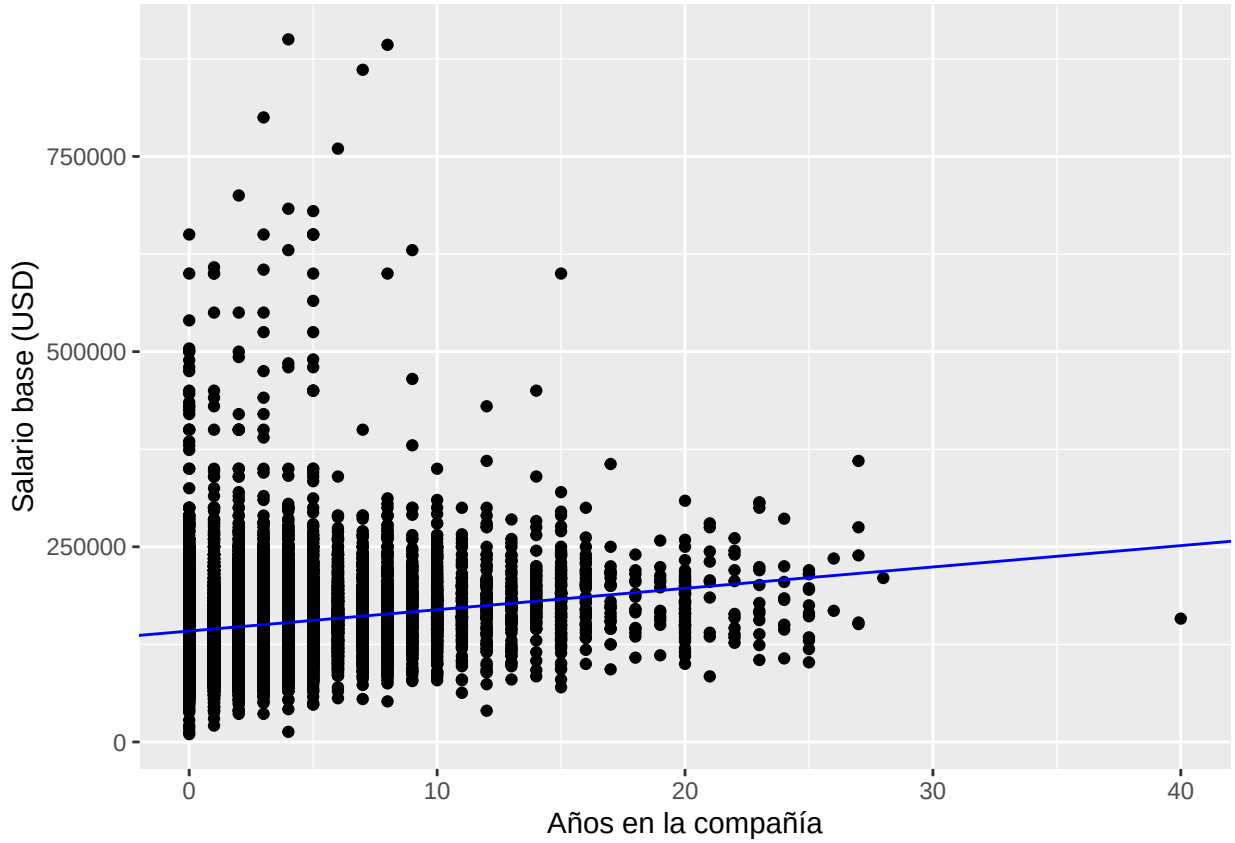


Vemos que, como era previsible, cuanto más años de experiencia, más alto parece ser el salario base.

Ahora veamos como se relaciona el salario base con los años que lleva en la empresa.

```
ajusM2 <- lm(basesalary ~ yearsatcompany, data = datos_usa)

ggplot(data = datos_usa, mapping = aes(x = yearsatcompany, y = basesalary)) +
  geom_point() +
  geom_abline(color="blue", slope = coef(ajusM2)[2], intercept = coef(ajusM2)[1]) +
  labs(y = "Salario base (USD)", x = "Años en la compañía", color = "Género") +
  scale_x_continuous(limits = c(0,40), breaks=seq(0,40,by=10))
```



Vemos una relación similar entre el salario base y los años en la compañía. Sin embargo, se pueden observar personas con 0 años en la empresa que tienen un salario mayor a personas que ya llevan 20, por ejemplo.

Modelado

Luego de explorar los datos y las relaciones entre las variables, observamos que: - La educación debe ser un factor incidente en el salario aunque no decisivo, dado que vimos que no podemos decidir el salario de una persona conociendo únicamente su nivel educativo, aunque un nivel educativo alto permite acceder a puestos mejor remunerados. - El comportamiento del salario base difiere marcadamente con cada puesto de trabajo desempeñado, por ende esta debe ser una variable incidente en el mismo. - Las correlaciones entre el salario base y los años de experiencia en la compañía resultan positivas, y al graficar, captamos cierta relación lineal entre ambas. - Si bien el bonus está correlacionado con el salario base, es una variable muy dependiente del contexto particular de la compañía (otra variable muy diversa) al momento en que se registró cada observación.

Por lo tanto, para construir un modelo del salario base e intentar predecirlo, elegimos construir modelos lineales utilizando las variables `yearsofexperience`, `yearsatcompany`, `title` y `Education`.

```
PMAE <- function(errores, datos, target) {
  return(mean(errores)/mean(datos[[target]]))
}

crossval <- function(datos, modelo, porcentajeEval, fun_error=PMAE, n_muestras=10){
  target <- strsplit(modelo, "[~]")[[1]][1]
  errores <- c()
```

```
  for(i in 1:n_muestras) {
    indicesEval <- sample(1:nrow(datos), size = (porcentajeEval/100)*nrow(datos))
    datosModelo <- datos[-indicesEval,]
    modeloLin <- lm(formula(modelo), data = datosModelo)
    abserrs <- c()
    for(o in indicesEval) {
      predicho <- predict(modeloLin, newdata = datos[o,])
      abserrs <- c(abserrs, abs(datos[[target]][o]-predicho))
    }
    errores <- c(errores, fun_error(abserrs, datos, target))
  }
  return(list(errores, mean(errores), var(errores), modelo, lm(formula(modelo), data = datos)))
}

set.seed(123)
```

```
#Creamos un vector con las variables predictoras
predictoras <- c("yearsofexperience", "yearsatcompany", "title", "Education")

partes <- powerSet(predictoras)

casos <- length(partes)-1

datos_modelo <- data.frame(
            'Modelo' = rep(NA, casos),
            'PMAE' = rep(NA, casos)
)

l <- 1
for(i in 1:4) {
  modelos <- partes[lapply(partes, length) == i]
  for(m in modelos) {
    formula <- paste("basesalary~",paste(m, collapse = '+'), sep = "")
    ajuste <- crossval(datos_usa, formula, 10, PMAE, 10)
    PMAEcv <- ajuste[2]
    datos_modelo[l,] <- c(formula, PMAEcv)
    l <- l+1
  }
}

# Ordenamos y mostramos los PMAEs obtenidos de cada modelo
datos_modelo <- datos_modelo %>% arrange(desc(PMAE))
datos_modelo
```

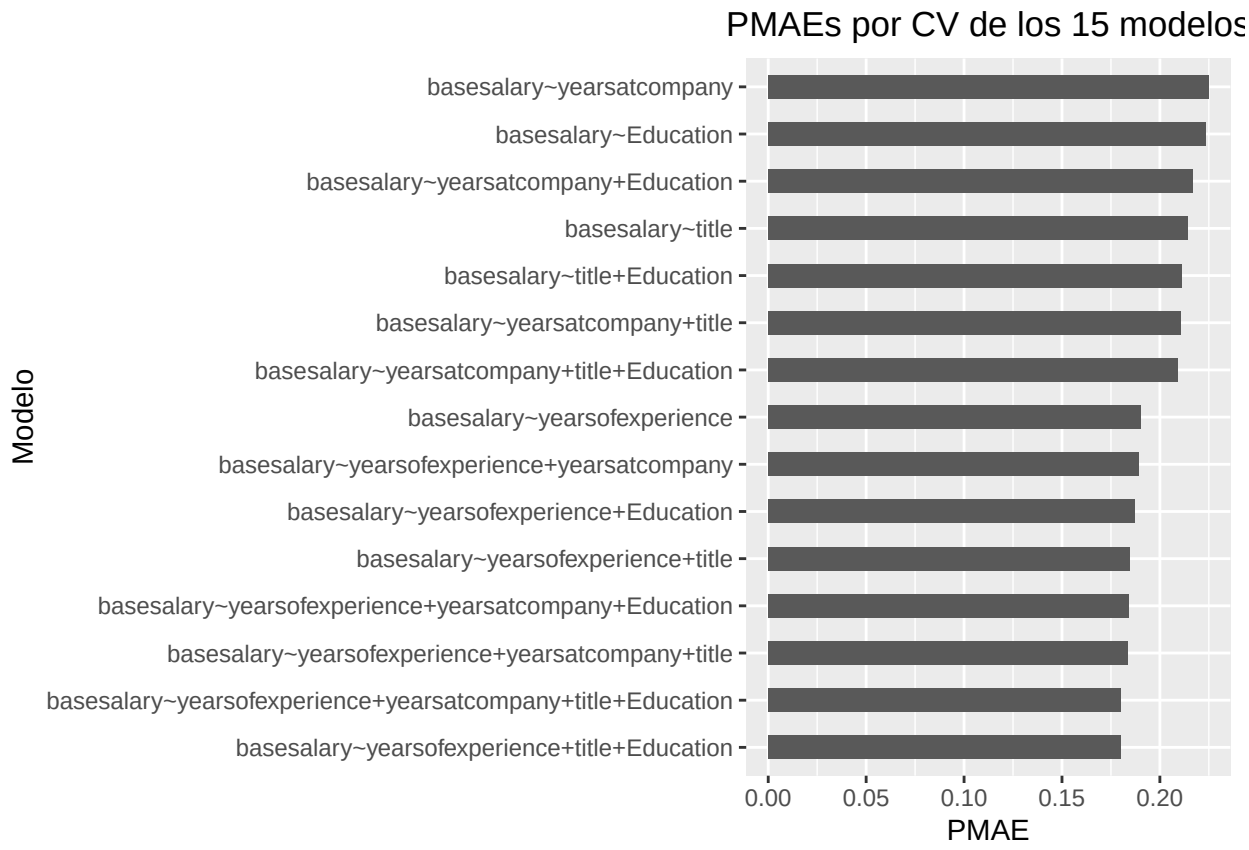
```
##                                                      Modelo      PMAE
## 1                                 basesalary~yearsatcompany 0.2247605
## 2                                     basesalary~Education 0.2232607
## 3                          basesalary~yearsatcompany+Education 0.2165936
## 4                                       basesalary~title 0.2141614
## 5                                  basesalary~title+Education 0.2109427
## 6                              basesalary~yearsatcompany+title 0.2104372
## 7                    basesalary~yearsatcompany+title+Education 0.2092252
## 8                                 basesalary~yearsofexperience 0.1903823
```

```
## 9                  basesalary~yearsofexperience+yearsatcompany 0.1892239
## 10                     basesalary~yearsofexperience+Education 0.1871789
## 11                         basesalary~yearsofexperience+title 0.1844087
## 12       basesalary~yearsofexperience+yearsatcompany+Education 0.1841140
## 13           basesalary~yearsofexperience+yearsatcompany+title 0.1836043
## 14 basesalary~yearsofexperience+yearsatcompany+title+Education 0.1799426
## 15                basesalary~yearsofexperience+title+Education 0.1799202
```

```
# Devolvemos la fila de menor MAE
print(datos_modelo[which.min(datos_modelo[,2]),])
```

```
##                                         Modelo      PMAE
## 15 basesalary~yearsofexperience+title+Education 0.1799202
```

```
datos_modelo %>%
  ggplot(aes(x=PMAE, y=reorder(Modelo, PMAE))) +
  geom_bar(stat = "identity", width=0.5) +
  theme(plot.title = element_text(hjust=0.5)) +
  ylab("Modelo") +
  ggtitle("PMAEs por CV de los 15 modelos")
```



Una observación importante es que el modelo que considera todas las variables arroja un PMAE muy similar al que considera todas menos `yearsatcompany`, es decir que tienen capacidades predictivas casi idénticas.

Por otro lado, podemos ver que los modelos que difieren tan solo en si consideran la variable `title` o `Education` o ambas, también tienen capacidades predictivas muy similares. De hecho, performan mejor a

la hora de predecir el salario, aquellos que consideran solo el puesto de trabajo (`title`) en lugar del nivel educativo (`Education`).

Finalmente mostramos el modelo que mejor predice el salario con sus respectivos coeficientes.

```
# Mostramos el modelo con los coeficientes

ajuste_mejor <- crossval(datos_usa, "basesalary~yearsofexperience+title+Education", 10, PMAE, 10)
#coeficientes_mejor <- coef(ajuste_mejor)
#coeficientes_mejor
ajuste_mejor[5]
```

```
## [[1]]
##
## Call:
## lm(formula = formula(modelo), data = datos)
##
## Coefficients:
##                          (Intercept)                   yearsofexperience
##                              88439.1                              3944.3
##                titleData Scientist              titleHardware Engineer
##                              34810.3                             21445.6
##               titleHuman Resources          titleManagement Consultant
##                               5407.1                             22949.6
##                    titleMarketing            titleMechanical Engineer
##                              17726.5                              9029.5
##              titleProduct Designer                titleProduct Manager
##                              31953.3                             35788.2
##                    titleRecruiter                            titleSales
##                               1821.5                              -557.2
##             titleSoftware Engineer  titleSoftware Engineering Manager
##                              34973.5                             50234.6
##           titleSolution Architect     titleTechnical Program Manager
##                              18364.6                             24782.0
##              EducationSome College         EducationBachelor's Degree
##                             -10664.8                             -4088.3
##          EducationMaster's Degree                          EducationPhD
##                               3885.4                             29184.7
```

Conclusiones

- El mejor modelo es el que predice el salario base según los años de experiencia, el título y la educación, con un error del 17% aproximadamente.
- El nivel educativo afecta al salario positivamente, aunque no es determinante, ya que vimos personas con título secundario tener el mismo salario que personas con doctorado. Observando los coeficientes del modelo, notamos que el coeficiente del PhD supera ampliamente al resto de los niveles de educación, con lo cual, afirmamos que un nivel educativo alto aumenta las posibilidades de tener un salario alto.
- A la hora de estimar el salario vimos que, los años de experiencia es la variable numérica más importante (los PMAEs de modelos con esta variable son menores), y más atrás quedan el puesto, la educación y por último los años en la compañía.