Bank Customer Churn

Gonzalo Cruz Gómez

Introducción

El conjunto de datos contiene datos sobre los clientes de un banco. Estos datos han sido extraídos del repositorio Bank Customer Churn Dataset de Kaggle. Objetivo Descripcion datos Tipo de problema

Business understanding

Planteamos preguntas sobre nuestros datos:

- ¿Influye el salario en si el cliente deja el banco?
- ¿Influye el tiempo que lleva el cliente en el banco en si este deja este banco?
- ¿Influye la edad?

Data understanding

Importamos las librerías

```
library(ggplot2)
library(readr)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Leemos los datos

```
Data <- read_csv("Bank Customer Churn Prediction.csv")
```

```
## Rows: 10000 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): country, gender
## dbl (10): customer_id, credit_score, age, tenure, balance, products_number, ...
##
## 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.
```

```
ntotal <- dim(Data)[1]
ptotal <- dim(Data)[2]
```

Podemos ver que tenemos n = 10000 observaciones y 12 variables en el dataset

Dividimos los datos en train-test-validate

```
set.seed(123)

# creamos los indices
indices <- 1:ntotal
ntrain <- ntotal *.6
ntest <- ntotal* .2
nval <- ntotal-(ntrain+ntest)
indices.train <- sample(indices, ntrain, replace= FALSE)
indices.test <- sample(indices[-indices.train], ntest, replace= FALSE)
indices.val <- indices[-c(indices.train, indices.test)]

# 60% para train, 20% para test y 20% para validate

train <- Data[indices.train,]
test <- Data[indices.test,]
validate <- Data[indices.val,]
```

Veamos las variables:

```
str(train)
```

```
## tibble [6,000 x 12] (S3: tbl_df/tbl/data.frame)
##  $ customer_id     : num [1:6000] 15704442 15607993 15635502 15631912 15788539 ...
##  $ credit_score    : num [1:6000] 672 625 443 840 501 755 698 850 552 558 ...
##  $ country         : chr [1:6000] "France" "France" "France" "France" ...
##  $ gender          : chr [1:6000] "Female" "Female" "Male" "Male" ...
##  $ age             : num [1:6000] 53 52 44 30 34 78 47 40 55 31 ...
##  $ tenure          : num [1:6000] 9 2 2 8 3 5 6 1 3 7 ...
##  $ balance         : num [1:6000] 169406 79469 0 136292 107748 ...
##  $ products_number : num [1:6000] 4 1 1 1 1 1 1 1 1 1 ...
##  $ credit_card     : num [1:6000] 1 1 1 1 1 1 1 1 1 1 ...
##  $ active_member   : num [1:6000] 1 1 0 0 0 1 0 0 1 0 ...
##  $ estimated_salary: num [1:6000] 147311 84606 159166 54113 9249 ...
##  $ churn           : num [1:6000] 1 0 0 0 0 0 1 0 0 0 ...
```

Podemos observar que la mayoria de nuestras variables son continuas, a excepción de aquellas que son char como country y gender, que son variables categóricas, y active_member, churn (variable objetivo) y credit card, que se trata de variables binarias.

Exploratory Data Analysis

```r
summary(train)
```

```
##   customer_id        credit_score     country             gender         
##  Min.   :15565706   Min.   :350.0   Length:6000        Length:6000       
##  1st Qu.:15630174   1st Qu.:583.0   Class :character   Class :character  
##  Median :15691008   Median :650.0   Mode  :character   Mode  :character  
##  Mean   :15691532   Mean   :649.3                                        
##  3rd Qu.:15754082   3rd Qu.:716.0                                        
##  Max.   :15815690   Max.   :850.0                                        
##       age         tenure          balance       products_number
##  Min.   :18   Min.   : 0.000   Min.   :     0   Min.   :1.000  
##  1st Qu.:32   1st Qu.: 2.000   1st Qu.:     0   1st Qu.:1.000  
##  Median :37   Median : 5.000   Median : 97441   Median :1.000  
##  Mean   :39   Mean   : 5.002   Mean   : 76748   Mean   :1.532  
##  3rd Qu.:44   3rd Qu.: 7.000   3rd Qu.:127928   3rd Qu.:2.000  
##  Max.   :92   Max.   :10.000   Max.   :238388   Max.   :4.000  
##   credit_card     active_member    estimated_salary        churn       
##  Min.   :0.0000   Min.   :0.0000   Min.   :    11.58   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 51552.11   1st Qu.:0.0000  
##  Median :1.0000   Median :1.0000   Median : 99644.16   Median :0.0000  
##  Mean   :0.7068   Mean   :0.5087   Mean   :100021.21   Mean   :0.2045  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:148733.11   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :199992.48   Max.   :1.0000  
```

Comprobamos si hay valores faltantes:

```r
sum(is.na(Data))
```

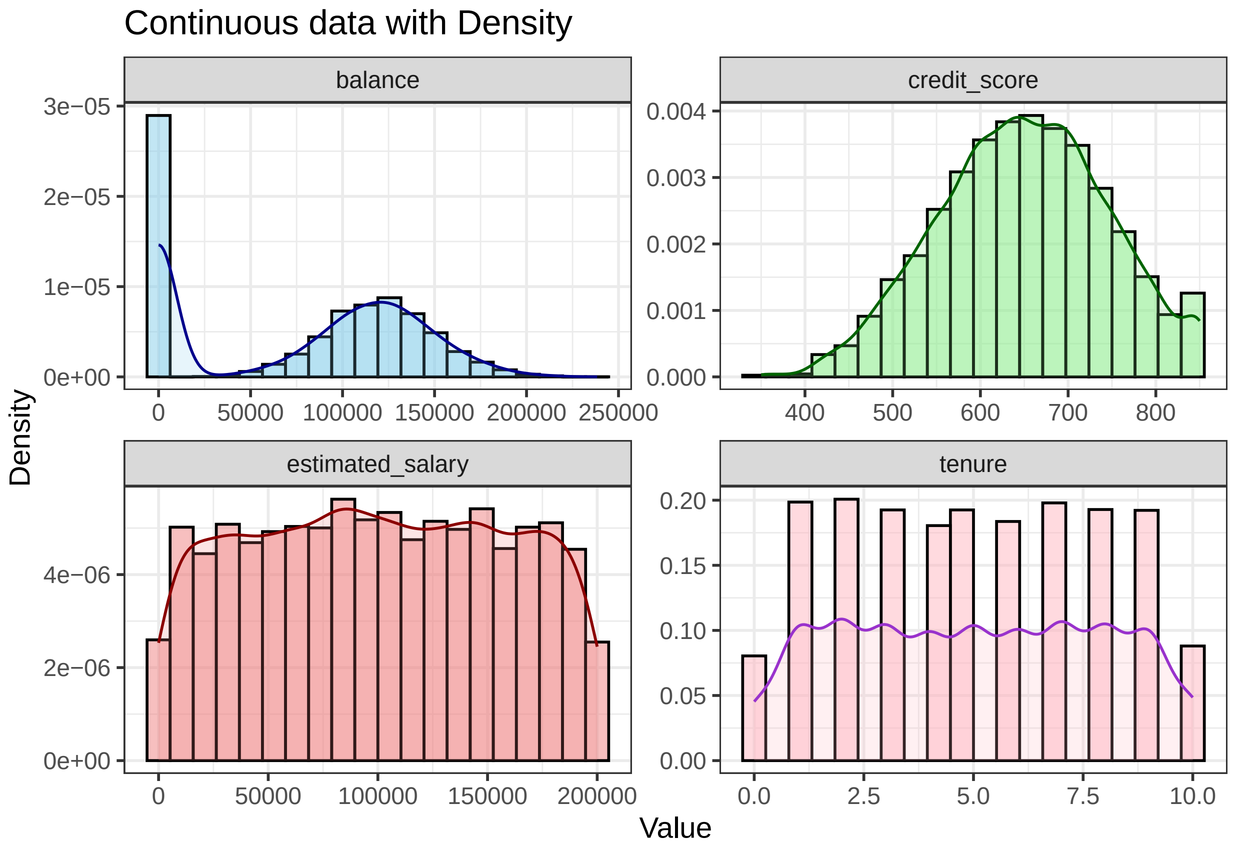
```
## [1] 0
```

Vemos que no tenemos valores faltantes.

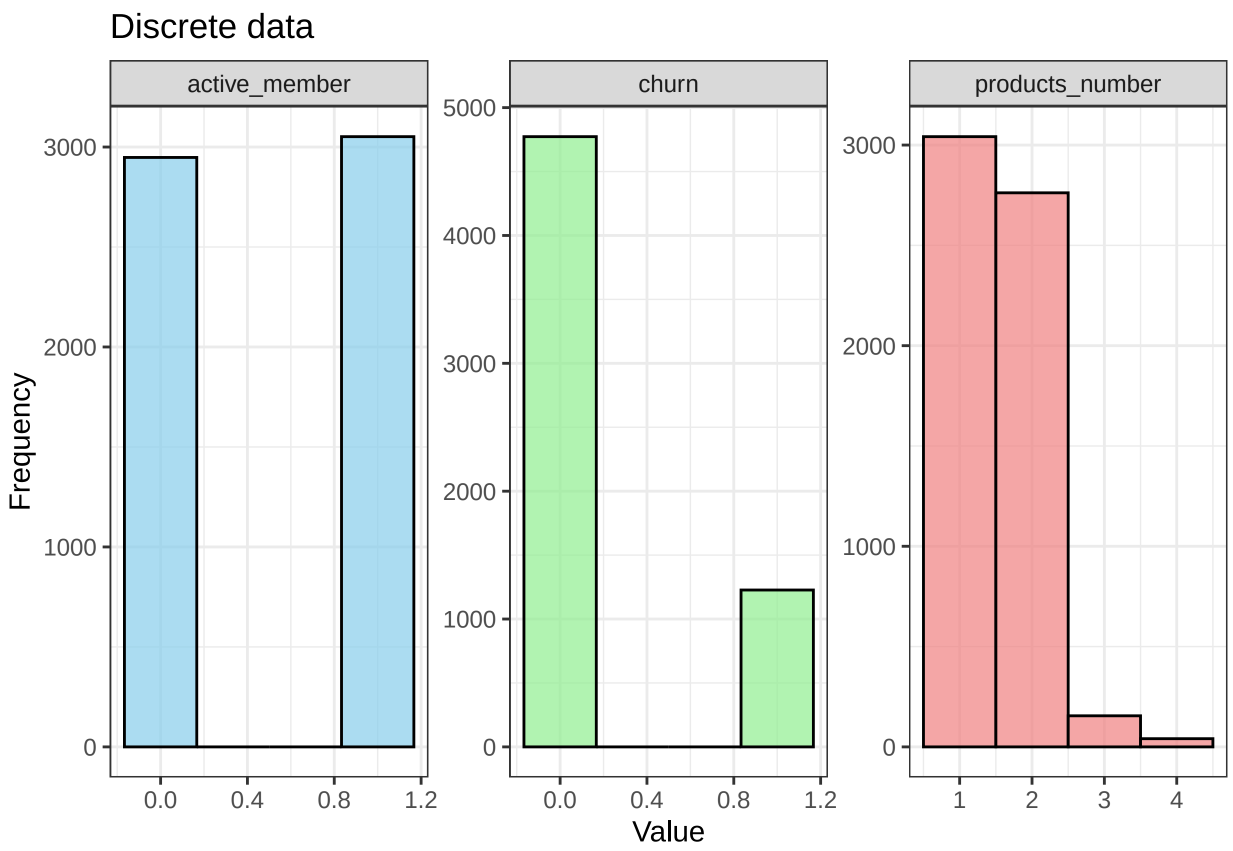
Visualizamos nuestros datos para ver como son y como están distribuidos

```r
train_long <- train %>%
  dplyr::select(tenure, estimated_salary, balance, credit_score) %>%
  tidyr::gather(key = "Variable", value = "Value")

ggplot(train_long, aes(x = Value, fill = Variable)) +
  geom_histogram(bins = 20, color = "black", alpha = 0.5, aes(y =..density..)) +
  geom_density(aes(y =..density.., color = Variable), linewidth = 0.5, alpha = 0.2) +
  facet_wrap(~ Variable, scales = "free") +
  labs(title = "Continuous data with Density", x = "Value", y = "Density") +
  scale_fill_manual(values = c("skyblue", "lightgreen", "lightcoral", "lightpink")) +
  scale_color_manual(values = c("darkblue", "darkgreen", "darkred", "darkorchid")) +
  theme_bw() +
  theme(legend.position = "none")
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
train_long <- train %>%
  dplyr::select(products_number, active_member, churn) %>%
  tidyr::gather(key = "Variable", value = "Value")

ggplot(train_long, aes(x = Value, fill = Variable)) +
  geom_histogram(bins = 4, color = "black", alpha = 0.7) +
  facet_wrap(~ Variable, scales = "free") +
  labs(title = "Discrete data", x = "Value", y = "Frequency") +
  scale_fill_manual(values = c("skyblue", "lightgreen", "lightcoral")) +
  theme_bw() +
  theme(legend.position = "none")
```

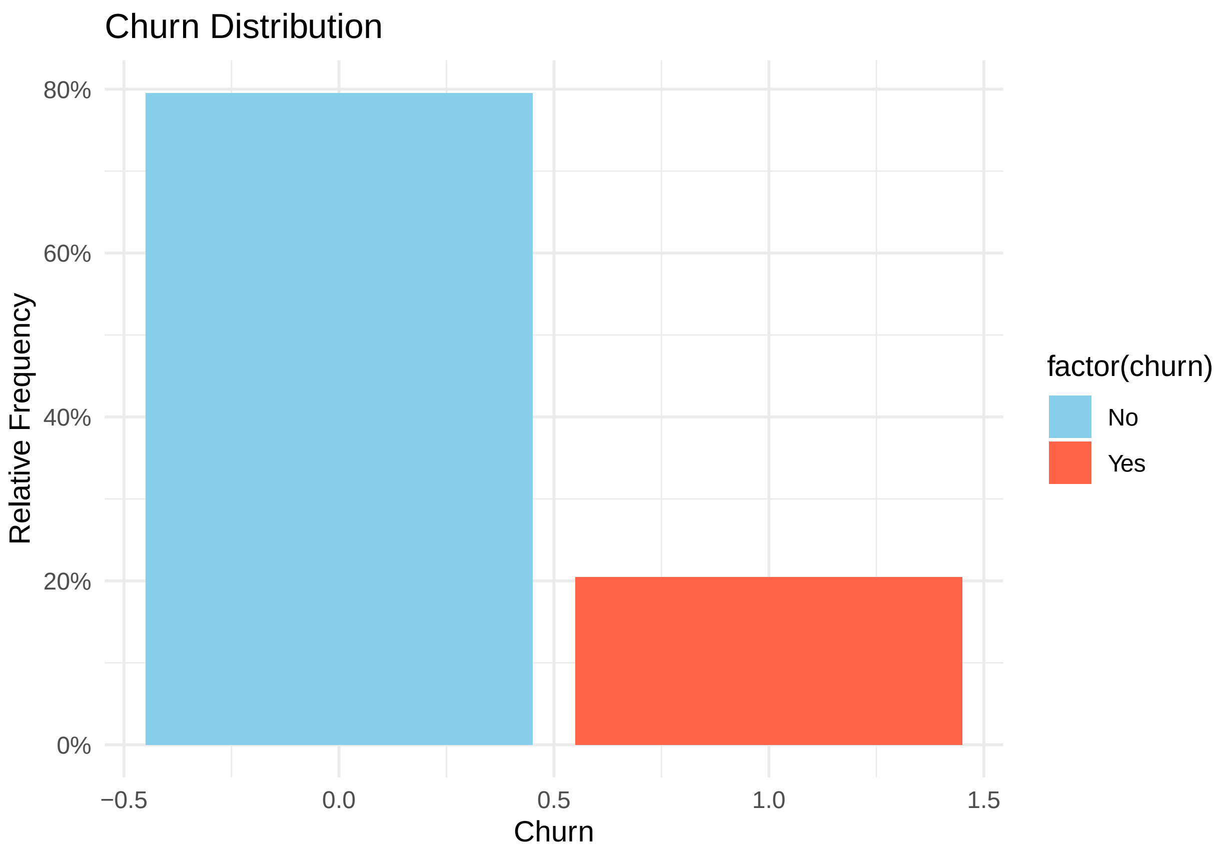


Vamos a ver qué nos dice nuestra variable objetivo

```
table(train$churn)
```

```
##
##    0    1
## 4773 1227
```

```
ggplot(data = train, aes(x = churn, fill = factor(churn))) +
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous(labels = scales::percent) +
  theme(legend.position = "none") +
  ylab("Relative Frequency") +
  xlab("Churn") +
  theme_minimal() +
  scale_fill_manual(values = c("skyblue", "tomato"),
                    labels = c("No", "Yes")) +
  labs(title = "Churn Distribution")
```

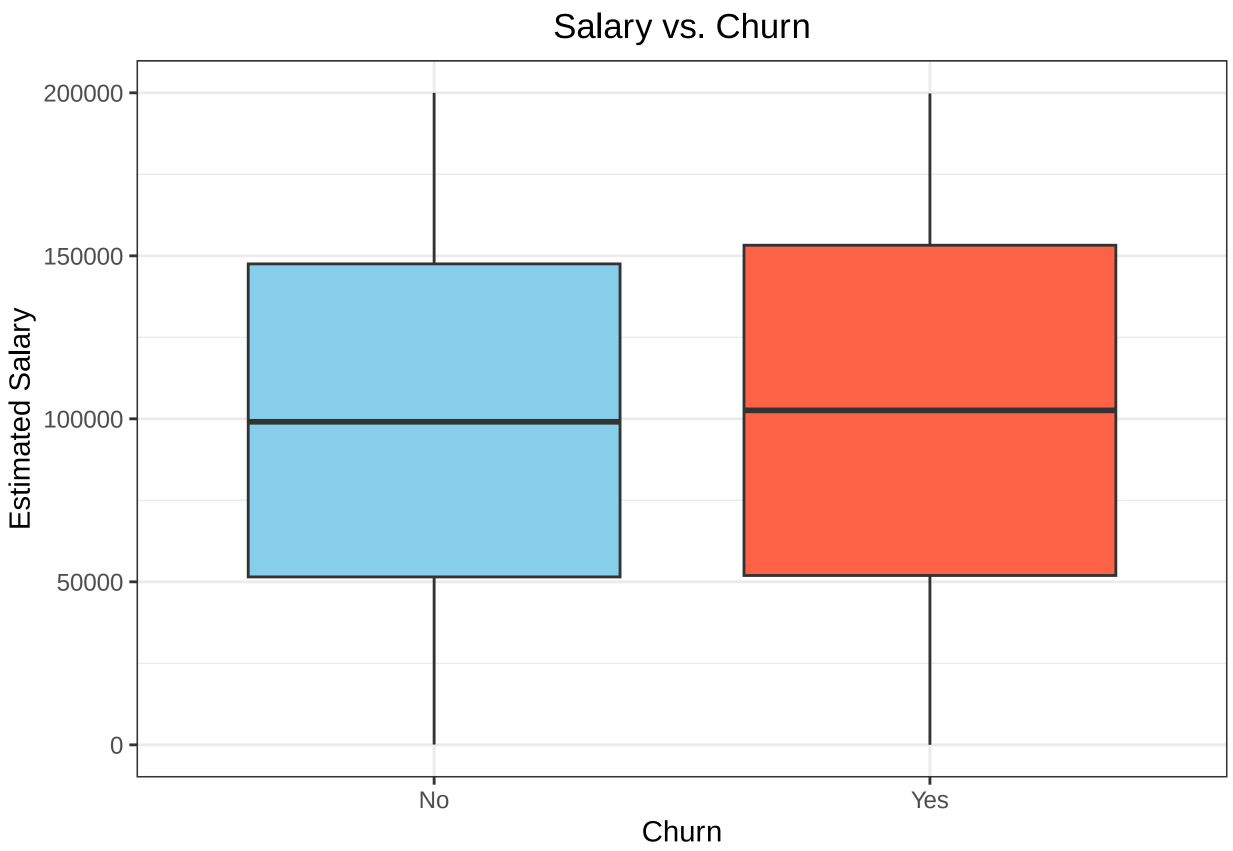


Podemos ver como la mayoría de los clientes no se van del banco pero hay un porcentaje significativo que si lo hace

Vamos a intentar contestar las preguntas que nos hicimos al inicio:

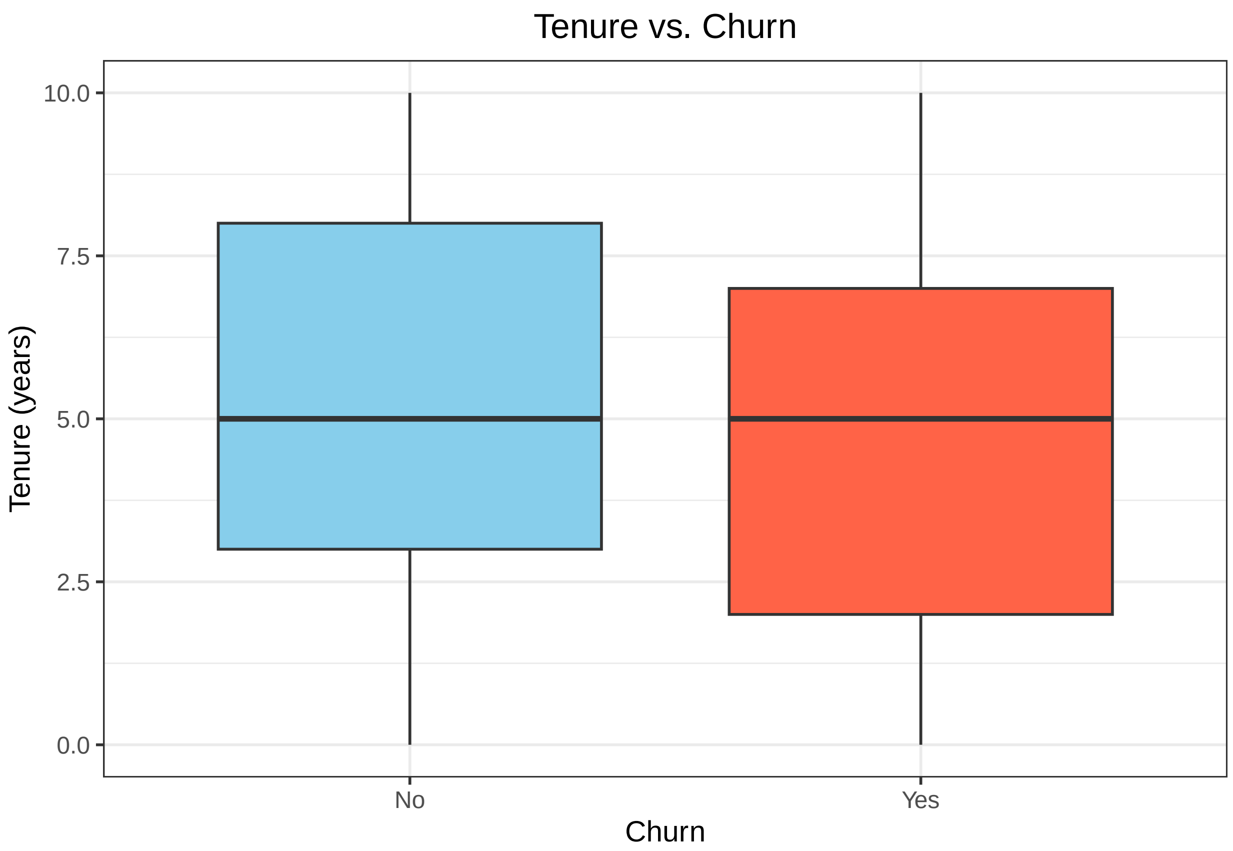
- ¿Influye el salario en si el cliente deja el banco?

```r
ggplot(train, aes(x = factor(churn), y = estimated_salary, fill = factor(churn))) +
  geom_boxplot() +
  labs(title = "Salary vs. Churn", x = "Churn", y = "Estimated Salary") +
  scale_x_discrete(labels = c("No", "Yes")) +
  scale_fill_manual(values = c("skyblue", "tomato"), labels = c("No", "Yes")) +
  theme_bw() +
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5))
```



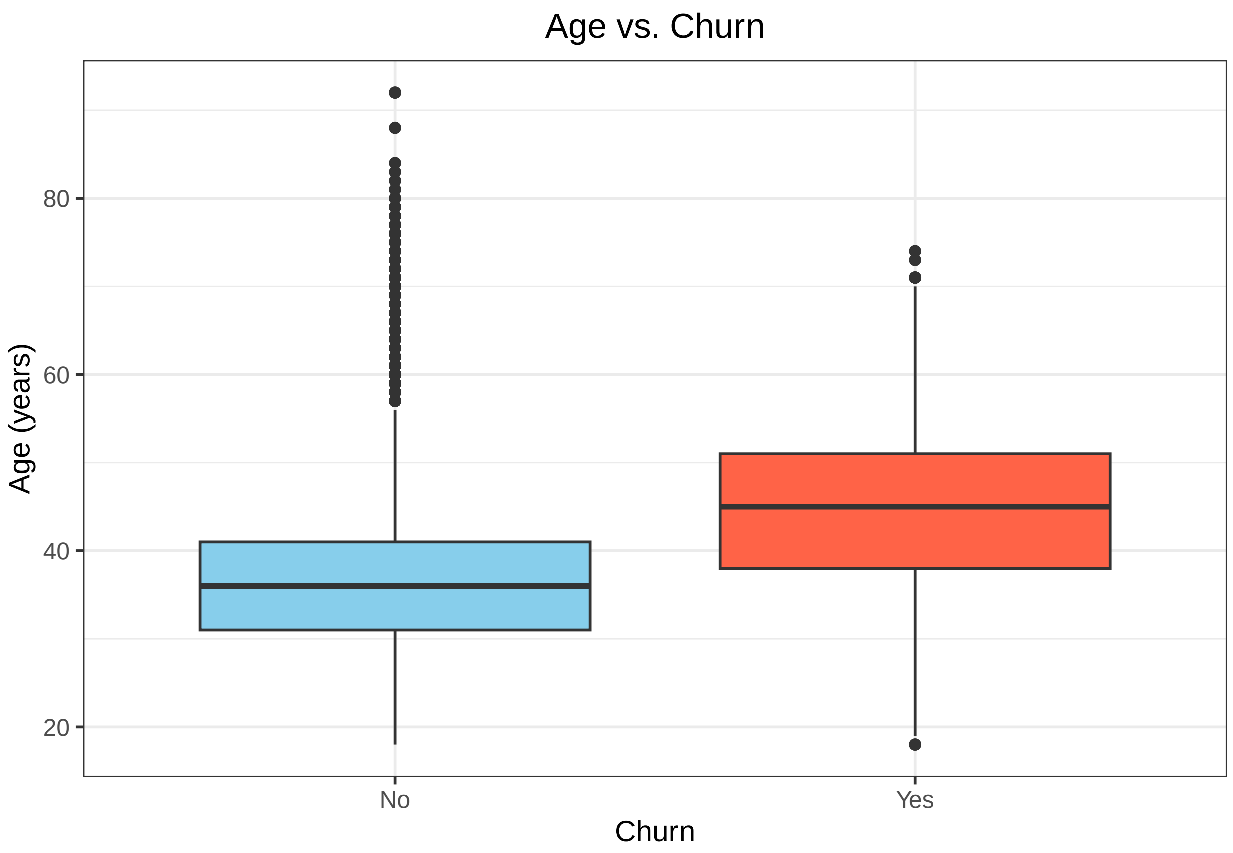
- ¿Influye el tiempo que lleva el cliente en el banco en si este deja el banco?

```
ggplot(train, aes(x = factor(churn), y = tenure, fill = factor(churn))) +
  geom_boxplot() +
  labs(title = "Tenure vs. Churn", x = "Churn", y = "Tenure (years)") +
  scale_x_discrete(labels = c("No", "Yes")) +
  scale_fill_manual(values = c("skyblue", "tomato"), labels = c("No", "Yes")) +
  theme_bw() +
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5))
```



- ¿Influye la edad?

```
ggplot(train, aes(x = factor(churn), y = age, fill = factor(churn))) +
  geom_boxplot() +
  labs(title = "Age vs. Churn", x = "Churn", y = "Age (years)") +
  scale_x_discrete(labels = c("No", "Yes")) +
  scale_fill_manual(values = c("skyblue", "tomato"), labels = c("No", "Yes")) +
  theme_bw() +
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5))
```


Age vs. Churn
Age (years)
80
60
40
20
No
Yes
Churn