Trabajo Final Business Performance Analysis

Roman Katz, Marité Ludoweig, Saulo Valdivia, Eva Suárez

12/3/2020

### Ejercicio 1:

Importar bases de datos en R y hacer las comprobaciones pertinentes. Importante describir y explicar cada paso en el Script.

## 'data.frame': 10000 obs. of 14 variables:  
## $ RowNumber : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ CustomerId : int 15634602 15647311 15619304 15701354 15737888 15574012 15592531 15656148 15792365 15592389 ...  
## $ Surname : chr "Hargrave" "Hill" "Onio" "Boni" ...  
## $ CreditScore : int 619 608 502 699 850 645 822 376 501 684 ...  
## $ Geography : chr "France" "Spain" "France" "France" ...  
## $ Gender : chr "Female" "Female" "Female" "Female" ...  
## $ Age : int 42 41 42 39 43 44 50 29 44 27 ...  
## $ Tenure : int 2 1 8 1 2 8 7 4 4 2 ...  
## $ Balance : num 0 83808 159661 0 125511 ...  
## $ NumOfProducts : int 1 1 3 2 1 2 2 4 2 1 ...  
## $ HasCrCard : int 1 0 1 0 1 1 1 1 0 1 ...  
## $ IsActiveMember : int 1 1 0 0 1 0 1 0 1 1 ...  
## $ EstimatedSalary: num 101349 112543 113932 93827 79084 ...  
## $ Exited : int 1 0 1 0 0 1 0 1 0 0 ...

## RowNumber CustomerId Surname CreditScore   
## Min. : 1 Min. :15565701 Length:10000 Min. :350.0   
## 1st Qu.: 2501 1st Qu.:15628528 Class :character 1st Qu.:584.0   
## Median : 5000 Median :15690738 Mode :character Median :652.0   
## Mean : 5000 Mean :15690941 Mean :650.5   
## 3rd Qu.: 7500 3rd Qu.:15753234 3rd Qu.:718.0   
## Max. :10000 Max. :15815690 Max. :850.0   
## Geography Gender Age Tenure   
## Length:10000 Length:10000 Min. :18.00 Min. : 0.000   
## Class :character Class :character 1st Qu.:32.00 1st Qu.: 3.000   
## Mode :character Mode :character Median :37.00 Median : 5.000   
## Mean :38.92 Mean : 5.013   
## 3rd Qu.:44.00 3rd Qu.: 7.000   
## Max. :92.00 Max. :10.000   
## Balance NumOfProducts HasCrCard IsActiveMember   
## Min. : 0 Min. :1.00 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 0 1st Qu.:1.00 1st Qu.:0.0000 1st Qu.:0.0000   
## Median : 97199 Median :1.00 Median :1.0000 Median :1.0000   
## Mean : 76486 Mean :1.53 Mean :0.7055 Mean :0.5151   
## 3rd Qu.:127644 3rd Qu.:2.00 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :250898 Max. :4.00 Max. :1.0000 Max. :1.0000   
## EstimatedSalary Exited   
## Min. : 11.58 Min. :0.0000   
## 1st Qu.: 51002.11 1st Qu.:0.0000   
## Median :100193.91 Median :0.0000   
## Mean :100090.24 Mean :0.2037   
## 3rd Qu.:149388.25 3rd Qu.:0.0000   
## Max. :199992.48 Max. :1.0000

## [1] 14

## [1] 10000 14

En el código anterior hemos cargado la base de datos y realizamos una primera aproximacion a los datos contenidos a través de diferentes comandos para ver la estructura y forma de los datos.

A partir de los comandos realizados pudimos observar 3 columnas tipo cáracter relacionadas con apellidos, país de procedencia y género; el resto de las columnas son de tipo cuantitativo.

Con la funcion summary, pudimos visualizar algunos parámetros principales, como son el rango de edades de nuestro dataset, el rango en el salario estimado y el credit score. También, de todas las variables cuantitativas pudimos observar los cuartiles, su media y su mediana.

Con las funciones Length y DIM, pudimos obtener las dimensiones de nuestros datos.

Y por último, eliminamos la columna ‘Row Number’ que contenía los índices y no proveía información relevante para el análisis.

### Ejercicio 2:

Describir de manera gráfica y estadística las variables de la base de datos. Importante explicar la eleccion de los graficos, así como lo que se observa en los descriptivos estadisticos. \*Puntos importantes: tendencia central, dispersion, forma de la distribucion (normalidad), valores perdidos, valores extremos

## BARRAS

A partir de la base de datos de 10000 observaciones, hemos realizado distintos gráficos para cada una de las variables, iniciando con los gráficos de barras para las variables categóricas. Luego se han realizado histogramas para aquellas variables cuantitativas.

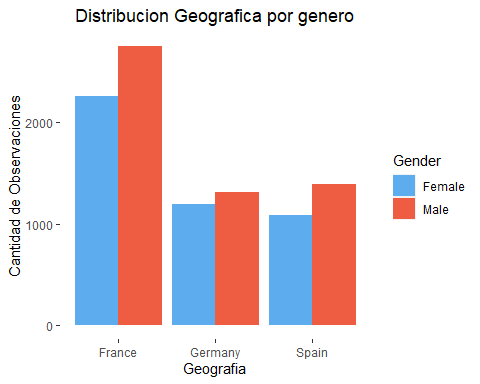
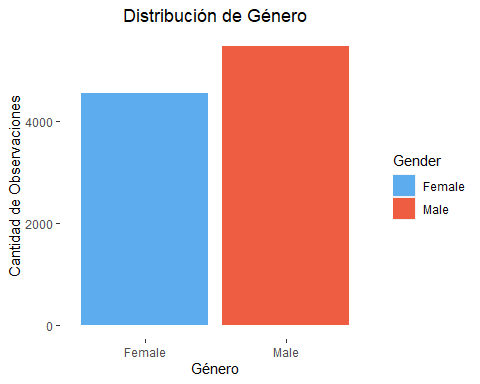
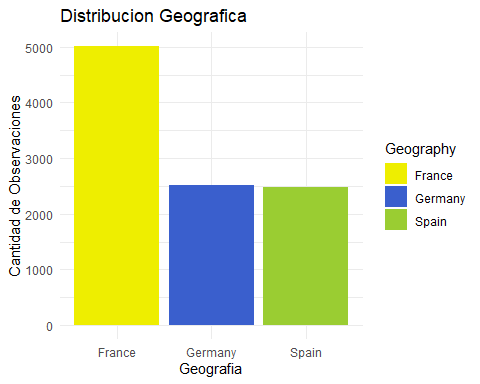
### Gráfico Barras – Geografía y Género

Iniciamos observando la variable geografía donde pudimos concluir que la data está concentrada en tres países siendo Francia con 5,014, el que lidera la cantidad de datos, luego Alemania con 2,509 y finalmente España con 2,477, siendo estos dos últimos muy similares. Tomando el gráfico anterior, quisimos observar la distribución por genero en cada país, para poder tener mayores detalles de éste, hemos realizado una relación entre la cantidad de hombre y mujeres donde vemos la tendencia en los tres países frente a la tendencia global, concluyendo lo siguiente: - Hay más hombres que mujeres, teniendo un ratio de 0.83 a nivel global. - El conteo de Francia es la muestra principal ya que tiene mayor cantidad de datos y por lo tanto su ratio toma mayor importancia en el promedio global. - Alemania, es el país que más se acerca a una relacion 1:1, ya que resulto tener un ratio de 0.906. - España es el país con mayor diferencia entre hombres y mujeres, siendo éste 0.78.

##   
## France Germany Spain   
## 5014 2509 2477

## data$Gender: Female  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 4543 1.54 0.6 1 1.49 0 1 4 3 0.85 0.89 0.01  
## ------------------------------------------------------------   
## data$Gender: Male  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 5457 1.52 0.56 1 1.49 0 1 4 3 0.63 0.17 0.01

## data$Geography: France  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 5014 1.53 0.57 1 1.5 0 1 4 3 0.68 0.54 0.01  
## ------------------------------------------------------------   
## data$Geography: Germany  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 2509 1.52 0.62 1 1.45 0 1 4 3 1.01 1.12 0.01  
## ------------------------------------------------------------   
## data$Geography: Spain  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 2477 1.54 0.56 2 1.51 1.48 1 4 3 0.52 -0.17 0.01



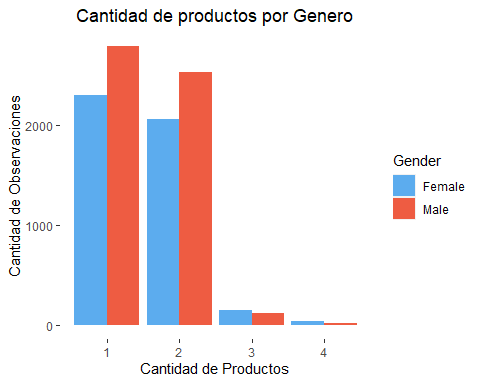
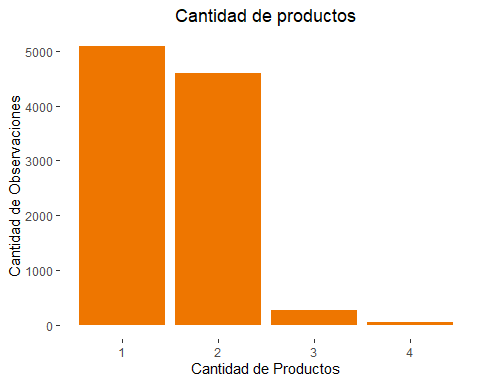
## Global France Spain Germany  
## 1 0.8325087 0.8212859 0.7845821 0.906535

### Gráfico de Barras – Number of Products y Género.

En este gráfico, podemos conocer la tendencia de la cantidad de productos que tiene cada cliente donde la media es 1.53 productos con una sd de 0.58; por lo tanto, se puede decir que la mayoría de la población cuenta con 1 y 2 productos.

A partir de esta conclusión, observamos que el número de hombres que poseen 1 o 2 productos es mayor que el de mujeres; sin embargo, conforme aumenta el número de productos (a partir de tres) la tendencia se invierte ligeramente (considerando únicamente la muestra que estamos analizando).

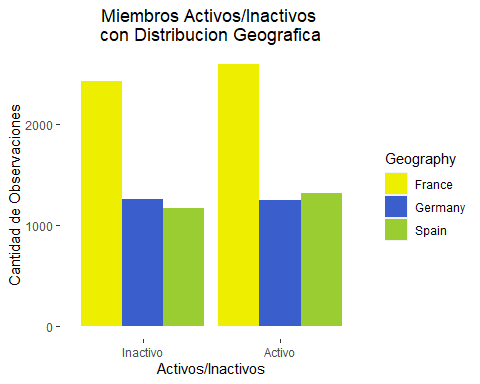
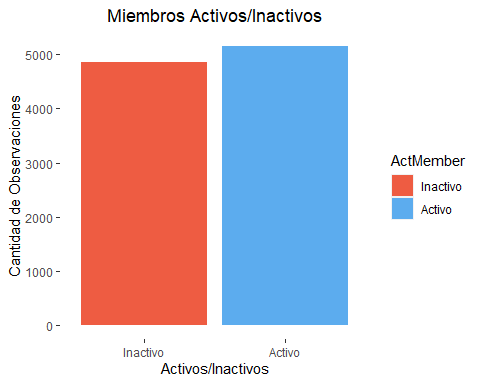
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 10000 1.53 0.58 1 1.49 0 1 4 3 0.75 0.58 0.01



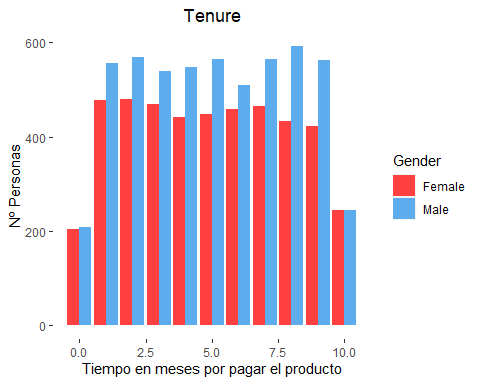
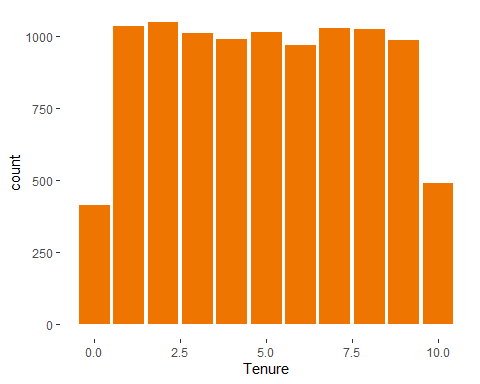
### Gráfico de Barras – Active Members

Del total de nuestra base de observables, podemos detectar que hay un gran numero de personas inactivas, analizando la variable, obtuvimos que la media es 0.52, siendo mayor el numero de personas con status activo.

Si lo vemos por geografía, podemos observar que están distribuidos similarmente en relación a la tendencia general de los 3 países.



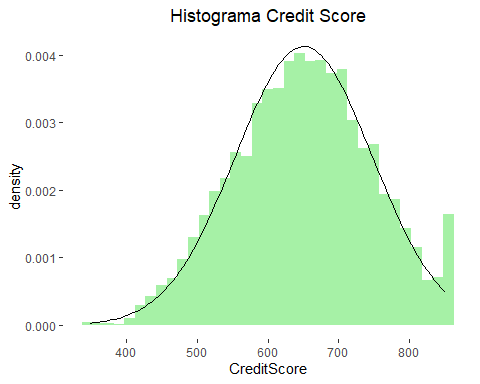
### Gráfico de Barras – Tenure



## Histogramas

### Gráfico de Histogramas – Credit Score

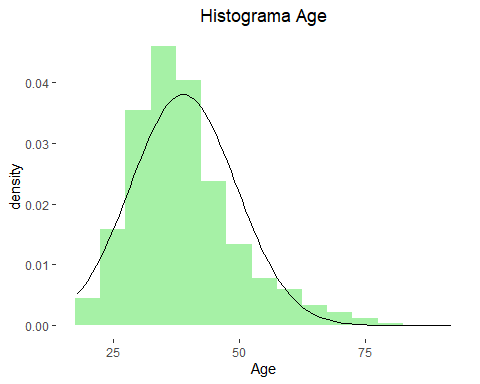
La media de Credit Score es 650 puntos, con un mínimo de 350 y máximo de 850; la media y desviación nos indica que la mayoría de la población está entre 550 y 750 puntos. Comparando el gráfico con una curva de distribución normal podemos observar que, en torno a la media de Credit Score, se distribuyen los datos como una normal.



## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 10000 650.53 96.65 652 651.01 99.33 350 850 500 -0.07 -0.43  
## se  
## X1 0.97

### Gráfico de Histogramas – Age

La edad promedio de la poblacion es de 39 años, con rango de edades comprendido entre 18 a 92 años. Esta distribución de edades no se ajusta a la de una normal, pues tiene un sesgo hacia la derecha con una cola de datos que se separan de la media más que la cola de la izquierda.



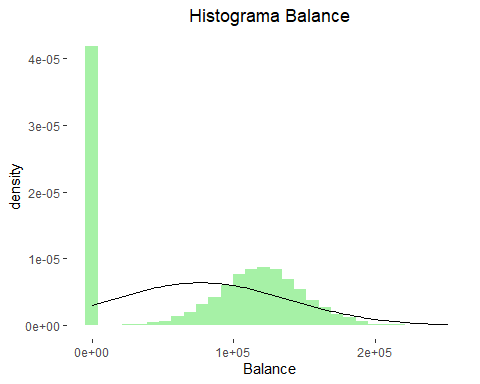
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 10000 38.92 10.49 37 37.91 8.9 18 92 74 1.01 1.39 0.1

### Gráfico de Histogramas – Balance

Con la variable Balance Score, hemos detectado que hay una gran cantidad de personas con un balance 0, lo que significa que no están debiendo dinero, lo cual podría distorsionar los datos de los deudores. Se observa que entre la media y la mediana existe una diferencia considerable, siendo éstas aprox 76,500 y 97,200 respectivamente.

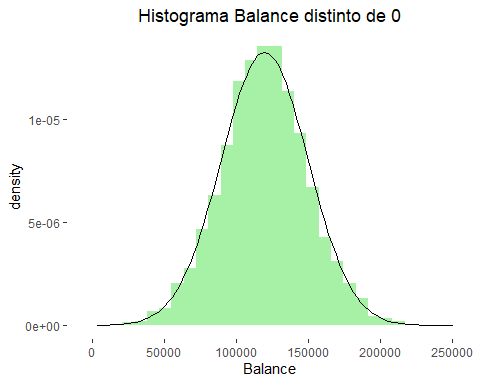
Generalmente no debemos excluir dichos datos, sin embargo, en este análisis, debemos enfocarnos en los deudores para poder formar una idea real de la deuda que esperamos recibir de vuelta. En este caso, los resultados indican que la deuda promedio de cada persona es de 119,800 lo cual sigue una distribución normal, basados en que la media y la mediana son iguales.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## vars n mean sd median trimmed mad min max range  
## X1 1 10000 76485.89 62397.41 97198.54 74827.8 69336.44 0 250898.1 250898.1  
## skew kurtosis se  
## X1 -0.14 -1.49 623.97

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

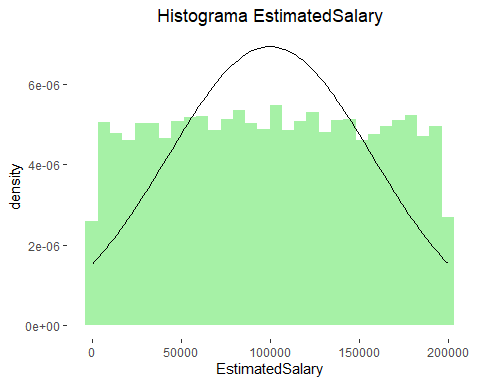


## vars n mean sd median trimmed mad min max  
## X1 1 6383 119827.5 30095.06 119839.7 119743.4 29163.01 3768.69 250898.1  
## range skew kurtosis se  
## X1 247129.4 0.03 0.2 376.69

### Gráfico de Histogramas – EstimatedSalary

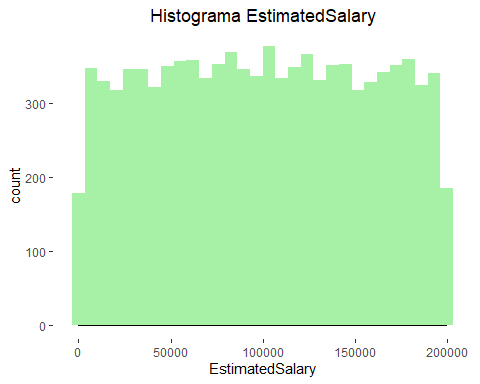
Podemos observar que los sueldos siguen una distribución uniforme entre 0 y 200,000 euros, siendo el sueldo promedio de 100,000 euros con una sd de 57,500.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## vars n mean sd median trimmed mad min max  
## X1 1 10000 100090.2 57510.49 100193.9 100114.9 72941.18 11.58 199992.5  
## range skew kurtosis se  
## X1 199980.9 0 -1.18 575.1

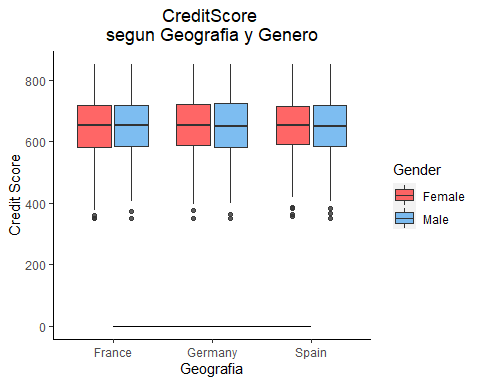
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## BoxPlots

### Boxplot: Credit Score - Geografia y Genero

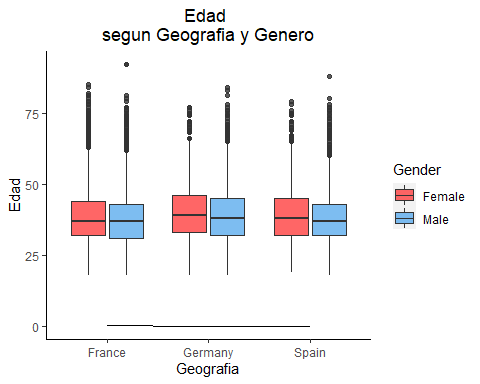
Se observa que tanto las medianas como el rango intercuatilico para todos los paises y ambos generos es similar, y se pueden ver algunos Outliers en la zona inferior.



## hombres$Geography: France  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 350.0 583.0 653.0 650.1 717.0 850.0   
## ------------------------------------------------------------   
## hombres$Geography: Germany  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 350.0 581.0 650.5 650.0 723.2 850.0   
## ------------------------------------------------------------   
## hombres$Geography: Spain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 350 584 650 651 716 850

## mujeres$Geography: France  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 350.0 580.0 652.0 649.2 716.0 850.0   
## ------------------------------------------------------------   
## mujeres$Geography: Germany  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 351.0 587.0 651.0 653.1 722.0 850.0   
## ------------------------------------------------------------   
## mujeres$Geography: Spain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 358.0 590.0 653.0 651.8 713.0 850.0

### Boxplot: Edad - Geografia y Genero

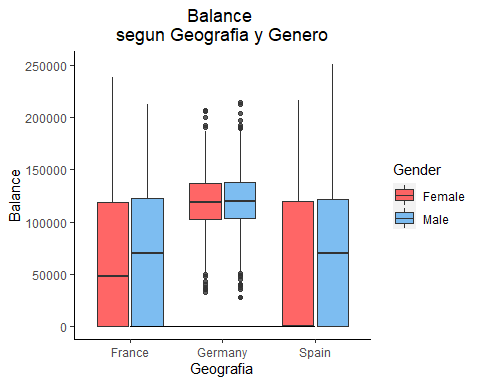
En este caso se observan las medianas tambien son similares entre si para ambos generos, pero que existen una cantidad importante de outliers en la zona superior del grafico, lo que muestra que aunque los datos en su mayoria se acumulan entre los 25 y 50 anos (aprox), existe una dispersion observable. 

## hombres$Geography: France  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.0 31.0 37.0 38.3 43.0 92.0   
## ------------------------------------------------------------   
## hombres$Geography: Germany  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 32.00 38.00 39.42 45.00 84.00   
## ------------------------------------------------------------   
## hombres$Geography: Spain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 32.00 37.00 38.65 43.00 88.00

## mujeres$Geography: France  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 32.00 37.00 38.77 44.00 85.00   
## ------------------------------------------------------------   
## mujeres$Geography: Germany  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 33.00 39.00 40.15 46.00 77.00   
## ------------------------------------------------------------   
## mujeres$Geography: Spain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.0 32.0 38.0 39.2 45.0 79.0

### Boxplot: Balance - Geografia y Genero

Podemos observar que para los casos de Espana y Francia la cantidad de balances en cero es tan importante que desplaza los boxplots hacia el cero, aunque el rango intercuartilico muestra una gran amplitud lo que significa que hay una gran dispersion de los datos. Por el contrario, en el caso de Alemania, se puede ver que los datos se concentran alrededor de la mediana con un rango intercuartilico mas acotado.

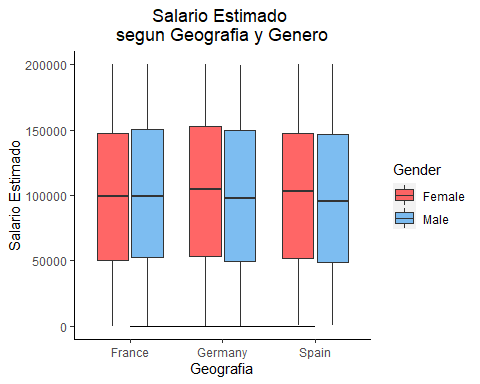


## hombres$Geography: France  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 69413 63546 122807 212693   
## ------------------------------------------------------------   
## hombres$Geography: Germany  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 27288 103398 120120 120260 137815 214347   
## ------------------------------------------------------------   
## hombres$Geography: Spain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 69857 63353 121348 250898

## mujeres$Geography: France  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 47536 60323 119014 238388   
## ------------------------------------------------------------   
## mujeres$Geography: Germany  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 32198 102287 118908 119146 137327 206869   
## ------------------------------------------------------------   
## mujeres$Geography: Spain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 59862 119799 216110

### Boxplot: Salario Estimado - Geografia y Genero

Nuevamente en este grafico se observa una distribucion pareja de los datos, es decir, en todos los casos los datos se distribuyen practicamente de forma homogenea alrededor de la mediana con un rango intercuartilico amplio y con bigotes entre los maximos y minimos de cada uno de los datos, y gracias a esta distribucion no se observan Outliers en el grafico.



## hombres$Geography: France  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 90.07 52525.15 99017.34 100174.25 150468.35 199929.17   
## ------------------------------------------------------------   
## hombres$Geography: Germany  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 11.58 49601.94 97892.70 99905.03 149590.84 199693.84   
## ------------------------------------------------------------   
## hombres$Geography: Spain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 417.4 49179.5 95672.4 98425.7 146773.1 199953.3

## mujeres$Geography: France  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 178.2 50457.2 99504.0 99564.2 147545.6 199862.8   
## ------------------------------------------------------------   
## mujeres$Geography: Germany  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 91.75 53217.30 104617.99 102446.42 153012.13 199970.74   
## ------------------------------------------------------------   
## mujeres$Geography: Spain  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 479.5 52138.5 102889.3 100734.1 147606.7 199992.5

1. Llevar a cabo un analisis estadistico en R y saber interpretar los resultados.

Para dicho analisis, el alumno/a podra elegir entre las dos siguientes opciones: a) Realizar un analisis de comparacion, siendo la hipotesis inicial: “el salario de los consumidores es igual en los diferentes paises donde se han recogido los datos” b) Realizar un analisis de relacion, siendo la hipotesis inicial: “la edad predice la solvencia (credit score) de los consumidores” opcional: en cualquiera de las elecciones se podra incluir covariables para realizar un analisis mas completo (justificar la eleccion e incluir en la interpretacion de los resultados) Pautas para el analisis:

* ¿Cual es la hipotesis nula y la hipotesis alternativa?
* ¿Que analisis es correcto para testar dicha hipotesis? ¿Por qué?
* Mostrar los resultados y representarlos graficamente
* Interpretar los resultados
* Sacar conclusiones en base a los resultados obtenidos y en relacion con la hipotesis inicial.

### Opcion B: Analisis de Hipotesis.

describe(data$Age)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 10000 38.92 10.49 37 37.91 8.9 18 92 74 1.01 1.39 0.1

summary(data$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 32.00 37.00 38.92 44.00 92.00

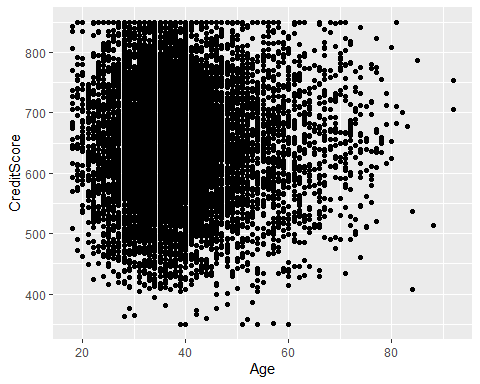
describe(data$CreditScore)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 10000 650.53 96.65 652 651.01 99.33 350 850 500 -0.07 -0.43  
## se  
## X1 0.97

summary(data$CreditScore)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 350.0 584.0 652.0 650.5 718.0 850.0

p+geom\_point(mapping = aes(Age,CreditScore))



cor.test(data$Age,data$CreditScore)

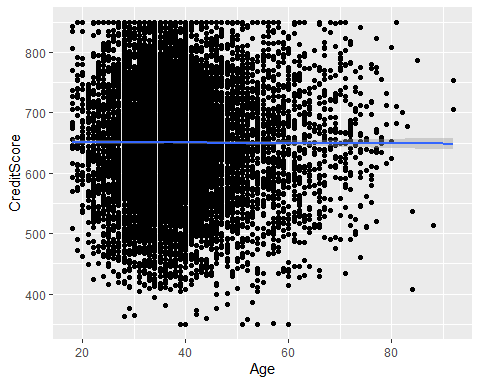
##   
## Pearson's product-moment correlation  
##   
## data: data$Age and data$CreditScore  
## t = -0.39645, df = 9998, p-value = 0.6918  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.02356314 0.01563638  
## sample estimates:  
## cor   
## -0.003964906

lineal <-lm(data$CreditScore~data$Age)  
summary(lm(data$CreditScore~data$Age))

##   
## Call:  
## lm(formula = data$CreditScore ~ data$Age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -300.526 -66.785 1.358 67.045 201.009   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 651.95099 3.71522 175.481 <2e-16 \*\*\*  
## data$Age -0.03654 0.09217 -0.396 0.692   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 96.66 on 9998 degrees of freedom  
## Multiple R-squared: 1.572e-05, Adjusted R-squared: -8.43e-05   
## F-statistic: 0.1572 on 1 and 9998 DF, p-value: 0.6918

p+geom\_point(mapping = aes(Age,CreditScore))+geom\_smooth(mapping=aes(Age,CreditScore),method = 'lm')

## `geom\_smooth()` using formula 'y ~ x'



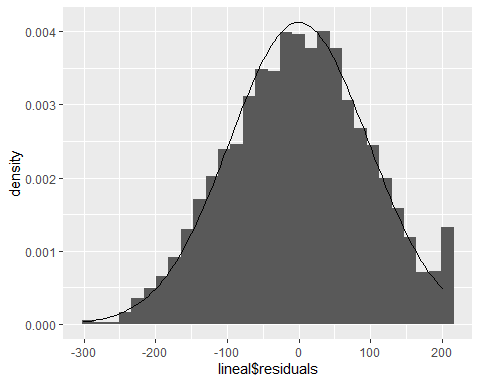
#Linealidad : Con el Cor.Test.  
#Normalidad  
library(DescTools)

##   
## Attaching package: 'DescTools'

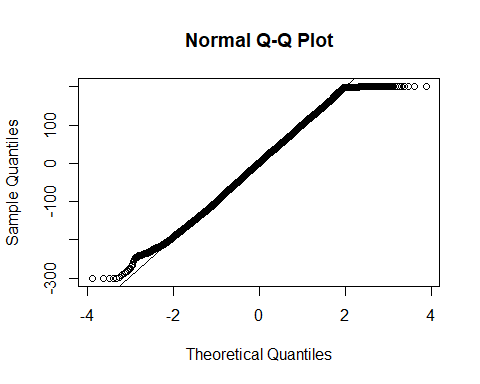
## The following objects are masked from 'package:psych':  
##   
## AUC, ICC, SD

g <- ggplot(lineal)  
g+geom\_histogram(mapping = aes(lineal$residuals, y=..density..))+stat\_function(fun = dnorm, args = list(mean=mean(lineal$residuals),sd = sd(lineal$residuals)))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



qqnorm(lineal$residuals);qqline(lineal$residuals)



ks.test(lineal$residuals,pnorm)

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: lineal$residuals  
## D = 0.49209, p-value < 2.2e-16  
## alternative hypothesis: two-sided

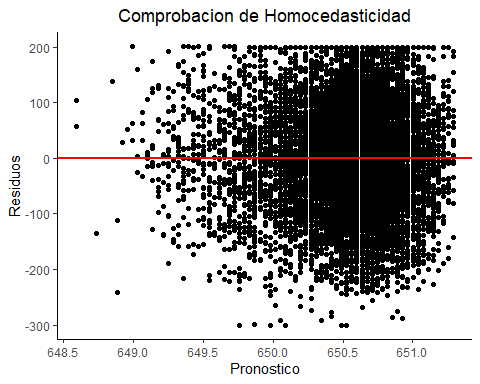
LillieTest(lineal$residuals)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: lineal$residuals  
## D = 0.018776, p-value = 1.983e-08

describe(lineal$residuals)

## vars n mean sd median trimmed mad min max range skew  
## X1 1 10000 0 96.65 1.36 0.48 99.32 -300.53 201.01 501.53 -0.07  
## kurtosis se  
## X1 -0.43 0.97

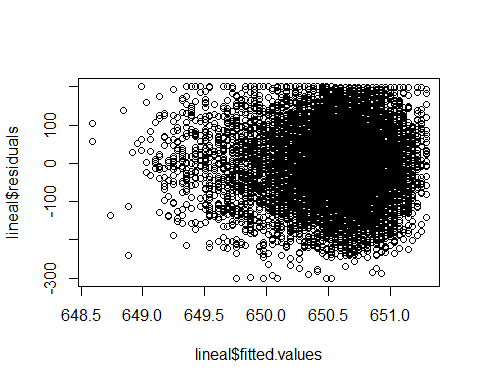
#Homocedasticidad  
g+geom\_point(mapping =aes(x=lineal$fitted.values,y=lineal$residuals))+geom\_hline(aes(yintercept = 0),color = 'Red',size=1)+theme(plot.title=element\_text(hjust = 0.5),panel.background = element\_rect(fill="white"),axis.line.y=element\_line(colour="black",size=0.2),  
 axis.line.x=element\_line(colour="black",size=0.2))+labs(title = 'Comprobacion de Homocedasticidad',x='Pronostico',y='Residuos')



#Independencia  
DurbinWatsonTest(lineal)

##   
## Durbin-Watson test  
##   
## data: lineal  
## DW = 2.0139, p-value = 0.7569  
## alternative hypothesis: true autocorrelation is greater than 0

plot(lineal$fitted.values,lineal$residuals)



Como primera aproximación a las variables que vamos analizar, realizamos un descripción estadística básica de ellas para enter sus propiedades. Luego, realizamos un grafico de dispersion donde pudimos tener una aproximacion visual de como se comportaban los datos ente variables.

Para poder entender si existía una correlación lineal entre las variables a analizar, hicimos un cor.test (Test de Pearson) para rechazar o no su H0 de que la correlación lineal o B1 sean igual a 0.

Al realizar el test, obtuvimos un p-valor mayor a 0.05, lo que nos llevo a mantener la hipotesis nula, y por lo tanto concluir que no hay correlacion lineal estadisticamente significativa.

Sin embargo, decidimos proceder con la estimacion de la regresion lineal para obtener un resultado visual de la recta. Visualmente arribamos a la misma conclusion ya que se observa un coeficiente B1 muy cercano a 0 y una recta practicamente horizontal, y ademas, con el R cuadrado vemos quela variable independiente (Edad/Age) solamente puede explicar el 0.0015% de la variable dependiente (CreditScore).

Mas allá de que con los resultados anteriores podemos concluir que no hay casi relación entre las variables, decidimos testear todos los supuestos.

La linealidad la testeamos anteriormente con el cor.test y vimos que no cumplia con ella. Continuamos con la normalidad de los residuos y a traves de los graficos y descriptivos estadisticos pudimos concluir que cumple con este supuesto. Posteriormente al testear la Homocedasticidad con un grafico de dispersion entre las predicciones y los residuos, se observo una dispersion sin tendencia y alrededor del cero, con lo que concluimos que cumple dicho supuesto. Por ultimo, se probo la Independencia a traves del DurbinWatson Test obteniendo un valor de 2.013 con lo que se puede concluir que cumple con el supuesto de independencia de los residuos.

Respondiendo las preguntas:

La H0 seria que la edad NO predice la solvencia económica (Credit Score). Esto se traduce en que la correlación lineal o B1 es igual a 0. Por otra parte, la H1 sería que la edad si predice (hasta cierto punto) la solvencia economica (CreditScore). Para testear dichas hipótesis, realizamos un análisis estadístico de regresión lineal, basándonos en los supuestos de linealidad, normalidad, homocedasticidad e independencia que dicha regresion requiere para ser aplicada. Como conclusion, a partir de los resultados obtenidos, podemos afirmar que no hay una relacion lineal entre la edad y la solvencia económica, y por lo tanto no es posible predecir una en funcion de la otra.

data\_france\_Spain <- filter(data,data$Geography!='Germany')  
t.test(data\_france\_Spain$EstimatedSalary~data\_france\_Spain$Geography ,var.equal=T)

##   
## Two Sample t-test  
##   
## data: data\_france\_Spain$EstimatedSalary by data\_france\_Spain$Geography  
## t = 0.32612, df = 7489, p-value = 0.7443  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2298.034 3215.251  
## sample estimates:  
## mean in group France mean in group Spain   
## 99899.18 99440.57

data\_france\_Germany <- filter(data,data$Geography!='Spain')  
t.test(data\_france\_Germany$EstimatedSalary~data\_france\_Germany$Geography ,var.equal=T)

##   
## Two Sample t-test  
##   
## data: data\_france\_Germany$EstimatedSalary by data\_france\_Germany$Geography  
## t = -0.86135, df = 7521, p-value = 0.3891  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3977.678 1549.170  
## sample estimates:  
## mean in group France mean in group Germany   
## 99899.18 101113.44

data\_germany\_Spain <- filter(data,data$Geography!='France')  
t.test(data\_germany\_Spain$EstimatedSalary~data\_germany\_Spain$Geography ,var.equal=T)

##   
## Two Sample t-test  
##   
## data: data\_germany\_Spain$EstimatedSalary by data\_germany\_Spain$Geography  
## t = 1.0238, df = 4984, p-value = 0.306  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1530.578 4876.304  
## sample estimates:  
## mean in group Germany mean in group Spain   
## 101113.44 99440.57