# EjercicioTiposDatosMissing

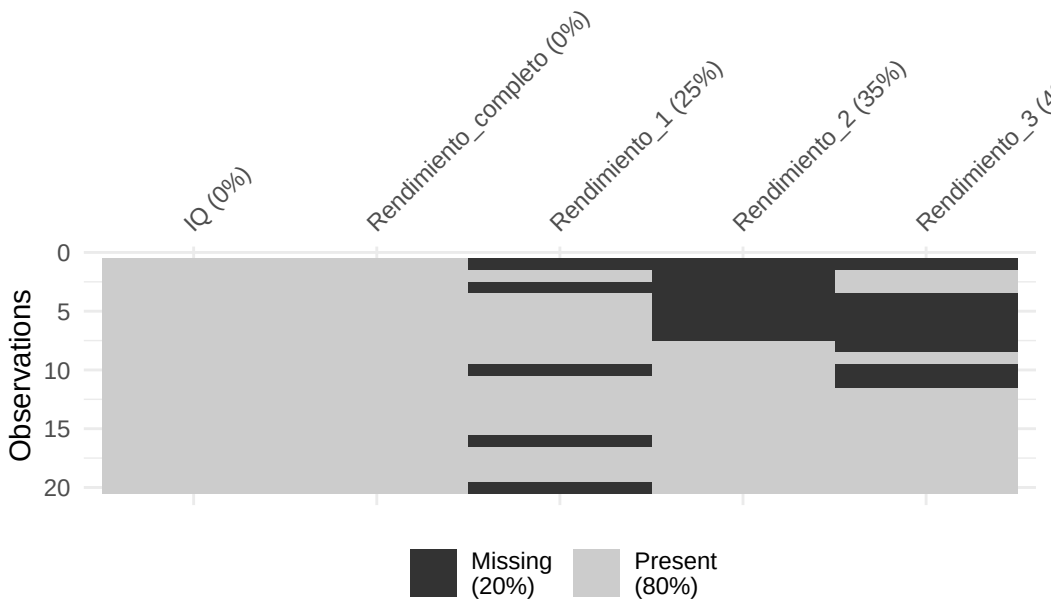
Silvia Pineda

## Lectura y carga de librerías

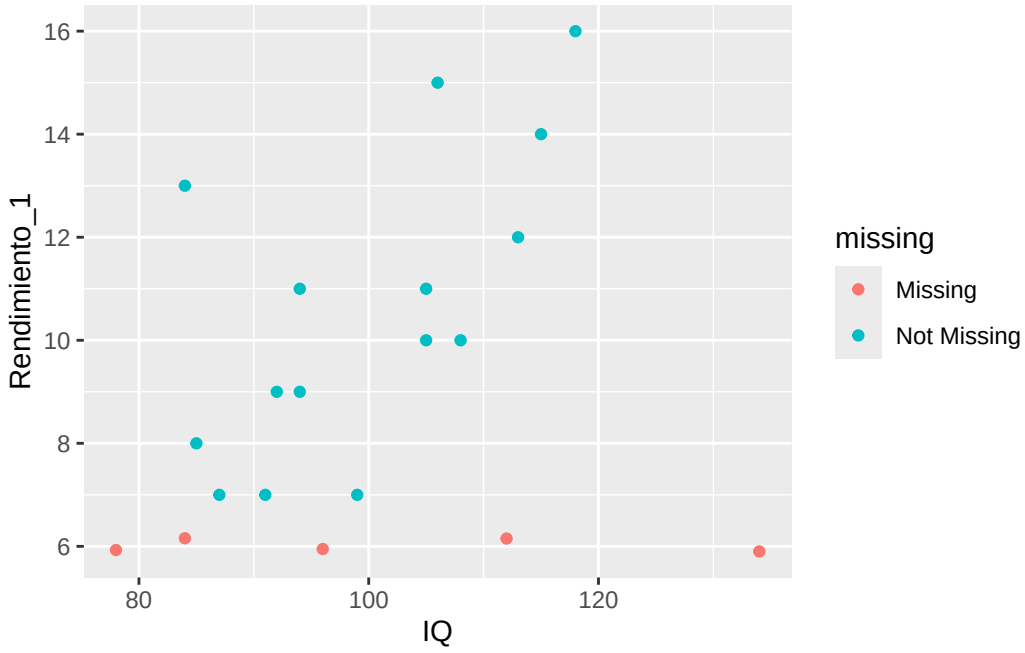
```
library(naniar)
library(ggplot2)
datos<-read.csv("EjemploTiposMissing.csv")
```

### 1. Averigua a qué columna corresponde los datos MCAR, MAR y MNAR y explica por qué. Después cambia el nombre de las columnas por `Rendimiento_MCAR`, `Rendimiento_MAR` y `Rendimiento_MNAR`.

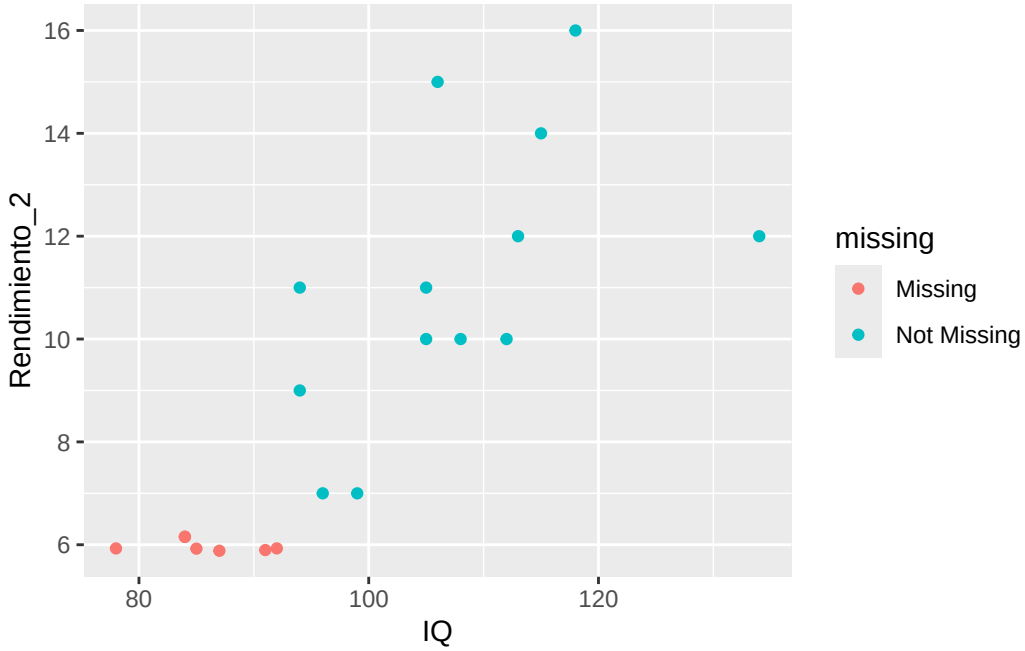
```
vis_miss(datos)
```



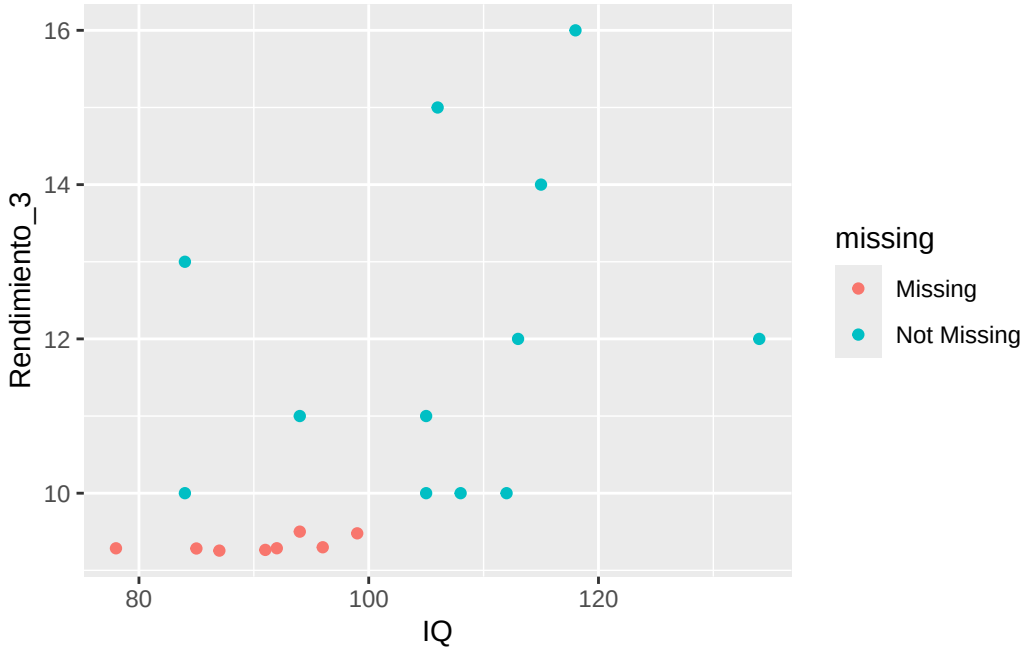

```
ggplot(data = datos, aes (x = IQ, y = Rendimiento_1)) + geom_miss_point()
```



```
ggplot(data = datos, aes (x = IQ, y = Rendimiento_2)) + geom_miss_point()
```




```
ggplot(data = datos, aes (x = IQ, y = Rendimiento_3)) + geom_miss_point()
```



```r
library(VIM)
```

```
Loading required package: colorspace
```

```
Loading required package: grid
```

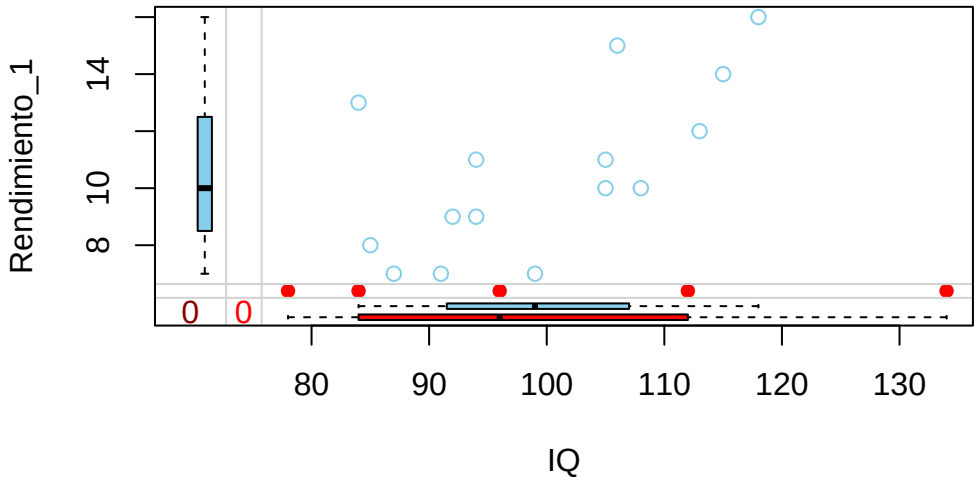
```
VIM is ready to use.
```

```
Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

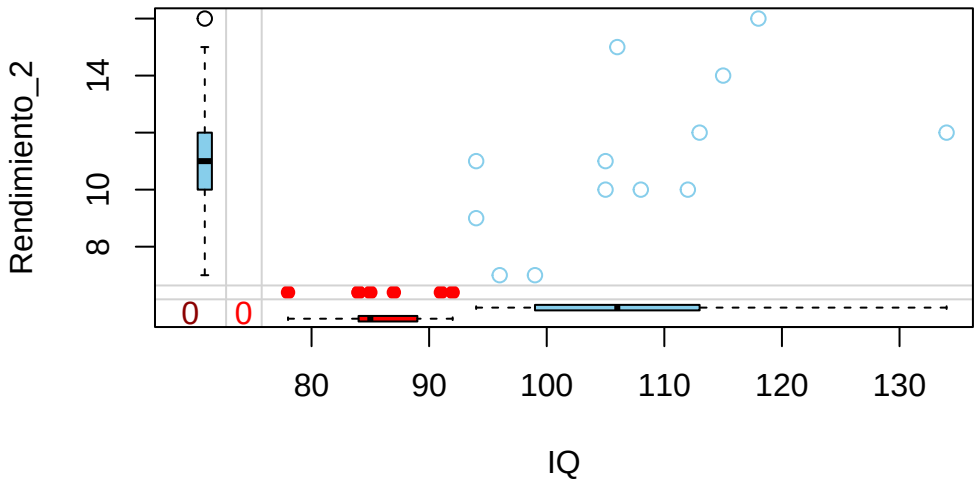
```
Attaching package: 'VIM'
```

```
The following object is masked from 'package:datasets':

    sleep
```

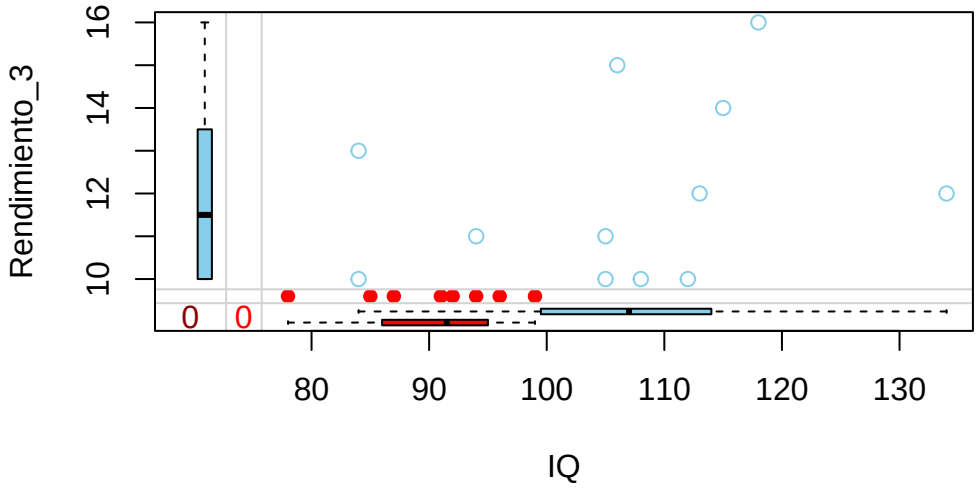
```r
marginplot(datos[c(1,3)])
```



```
marginplot(datos[c(1,4)])
```




```
marginplot(datos[c(1,5)])
```



**Rendimiento_1**: Este es de tipo MCAR. Los datos faltantes son completamente aleatorios. Están distribuidos a lo largo de la variable observada IQ y no corresponden a ningún patrón específico. Los boxplots de los missing vs. no missing no se diferencian.

**Rendimiento_2**: Este es de tipo MAR. Los datos faltantes corresponden a los candidatos con las puntuaciones IQ menores. Existe una clara asociación con la variable observada (IQ). Los boxplots de los datos missing vs. no missing se diferencian mucho y además no hay puntos que correspondan a los "not_missing" en la parte faltante.

**Rendimiento_3**: Este tipo es MNAR. Los datos faltantes corresponden a los candidatos con el rendimiento menor, pero CUIDADO esto no lo sabríamos si no tuvieramos la variable completa. Imagina por ejemplo que la compañía contrató a los 20 candidatos y a continuación despidió a un número de individuos por bajo rendimiento previo a las evaluaciones de los 6 meses. Los datos missing estarían asociados con una variable no observada (haber sido despedido). Sin tener la variable completa, esto lo podemos observar debido a que los datos missing del rendimiento_3 están asociados con IQ pero no solo porque también hay puntos "not_missing" para los de bajo IQ a diferencia del rendimiento_2.

```
colnames(datos)[3:5]<-c("Rendimiento_MCAR", "Rendimiento_MAR","Rendimiento_MNAR")
```

## 2. Compara la media entre las 3 variables de `Rendimiento_MCAR`, `Rendimiento_MAR` y `Rendimiento_MNAR` con `Rendimiento_completo`. ¿Qué observas?

```
summary(datos)
```

```
       IQ        Rendimiento_completo Rendimiento_MCAR Rendimiento_MAR
 Min.   : 78.0   Min.   : 7.00        Min.   : 7.0     Min.   : 7.00
 1st Qu.: 90.0   1st Qu.: 8.75        1st Qu.: 8.5     1st Qu.:10.00
 Median : 97.5   Median :10.00        Median :10.0     Median :11.00
 Mean   :100.0   Mean   :10.35        Mean   :10.6     Mean   :11.08
 3rd Qu.:109.0   3rd Qu.:12.00        3rd Qu.:12.5     3rd Qu.:12.00
 Max.   :134.0   Max.   :16.00        Max.   :16.0     Max.   :16.00
                                      NA's   :5        NA's   :7
 Rendimiento_MNAR
 Min.   :10.00
 1st Qu.:10.00
 Median :11.50
 Mean   :12.00
 3rd Qu.:13.25
 Max.   :16.00
 NA's   :8
```

La media de rendimiento completo es 10.35. Si no hacemos nada con los datos y solo borramos los missing, vemos como la media de los MAR y MNAR son mucho más altas debido a que los datos faltantes corresponden a valores bajos de rendimiento.

### 3. Si calculamos los coeficientes de correlación y una regresión simple de cada una de las 3 variables `Rendimiento_MCAR`, `Rendimiento_MAR` y `Rendimiento_MNAR` para predecir el `IQ`. ¿Qué diferencias observas en comparación con `Rendimiento_completo`?

```
cor(datos$IQ,datos$Rendimiento_completo)
```

```
[1] 0.5419817
```

```
cor(datos$IQ,datos$Rendimiento_MCAR,use="complete.obs")
```

```
[1] 0.6375139
```

```
cor(datos$IQ,datos$Rendimiento_MAR,use="complete.obs")
```

```
[1] 0.5555719
```

```
cor(datos$IQ,datos$Rendimiento_MNAR,use="complete.obs")
```

```
[1] 0.2671781
```

```
summary(lm(IQ~Rendimiento_completo,data=datos))
```

```
Call:
lm(formula = IQ ~ Rendimiento_completo, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-23.569  -7.425   1.216   6.572  29.287

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)            70.439     11.143   6.322 5.87e-06 ***
Rendimiento_completo    2.856      1.044   2.736   0.0136 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.2 on 18 degrees of freedom
Multiple R-squared:  0.2937,    Adjusted R-squared:  0.2545
F-statistic: 7.487 on 1 and 18 DF,  p-value: 0.01357
```

```r
summary(lm(IQ~Rendimiento_MCAR,data=datos))
```

```
Call:
lm(formula = IQ ~ Rendimiento_MCAR, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-21.5608  -4.2046   0.0078   6.8673   9.8673

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       73.9955     8.9276   8.288 1.51e-06 ***
Rendimiento_MCAR   2.4281     0.8138   2.983   0.0106 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.9 on 13 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.4064,    Adjusted R-squared:  0.3608
F-statistic: 8.901 on 1 and 13 DF,  p-value: 0.01057
```

```r
summary(lm(IQ~Rendimiento_MAR,data=datos))
```

```
Call:
lm(formula = IQ ~ Rendimiento_MAR, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4416  -2.4416  -0.1827   2.8173  24.2995
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.594     11.608   7.115 1.95e-05 ***
Rendimiento_MAR    2.259      1.019   2.216   0.0487 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.719 on 11 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.3087,    Adjusted R-squared:  0.2458
F-statistic: 4.911 on 1 and 11 DF,  p-value: 0.0487
```

```
summary(lm(IQ~Rendimiento_MNAR,data=datos))
```

```
Call:
lm(formula = IQ ~ Rendimiento_MNAR, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-24.312  -7.125   3.188   5.469  27.500

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        84.750     25.149   3.370  0.00712 **
Rendimiento_MNAR    1.812      2.067   0.877  0.40119
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.32 on 10 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.07138,   Adjusted R-squared:  -0.02148
F-statistic: 0.7687 on 1 and 10 DF,  p-value: 0.4012
```

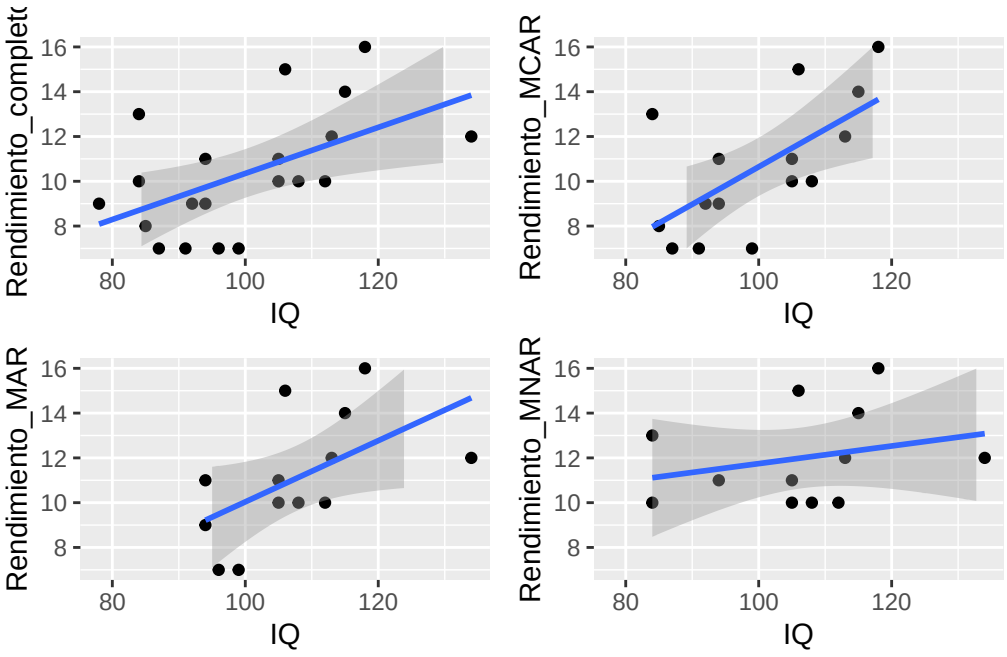
Tanto en la correlación como en el modelo de regresión vemos como varía la asociación según el tipo de datos faltantes siendo los de tipo MNAR los que provocan la pérdida total de correlación y asociación.

**4. Haz un gráfico para ver como se ajusta la recta de regresión en cada uno de los casos.**

```
library(ggplot2)
library(patchwork)


y_min <- min(datos$Rendimiento_completo)
y_max <- max(datos$Rendimiento_completo)


ggp1 <- ggplot(datos,aes(IQ, Rendimiento_completo)) + geom_point() +
 stat_smooth(method = "lm",se=TRUE, formula = y ~ x, geom = "smooth")  + ylim(y_min, y_max)


ggp2 <- ggplot(datos,aes(IQ, Rendimiento_MCAR)) + geom_point() +
 stat_smooth(method = "lm",se=TRUE, formula = y ~ x, geom = "smooth") + ylim(y_min, y_max)

ggp3 <- ggplot(datos,aes(IQ, Rendimiento_MAR)) + geom_point() +
stat_smooth(method = "lm",se=TRUE,formula = y ~ x, geom = "smooth") + ylim(y_min, y_max)


ggp4 <- ggplot(datos,aes(IQ, Rendimiento_MNAR)) + geom_point() +
stat_smooth(method = "lm",se=TRUE, formula = y ~ x, geom = "smooth")+ ylim(y_min, y_max)

ggp1+ggp2+ggp3+ggp4
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 5 rows containing missing values or values outside the scale range
(`geom_point()`).

Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 7 rows containing missing values or values outside the scale range
(`geom_point()`).

Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 8 rows containing missing values or values outside the scale range
(`geom_point()`).
```




## 5. Imputa por la media y por el modelo de regresión simple las tres variables rendimiento y compara los resultados.

### MCAR

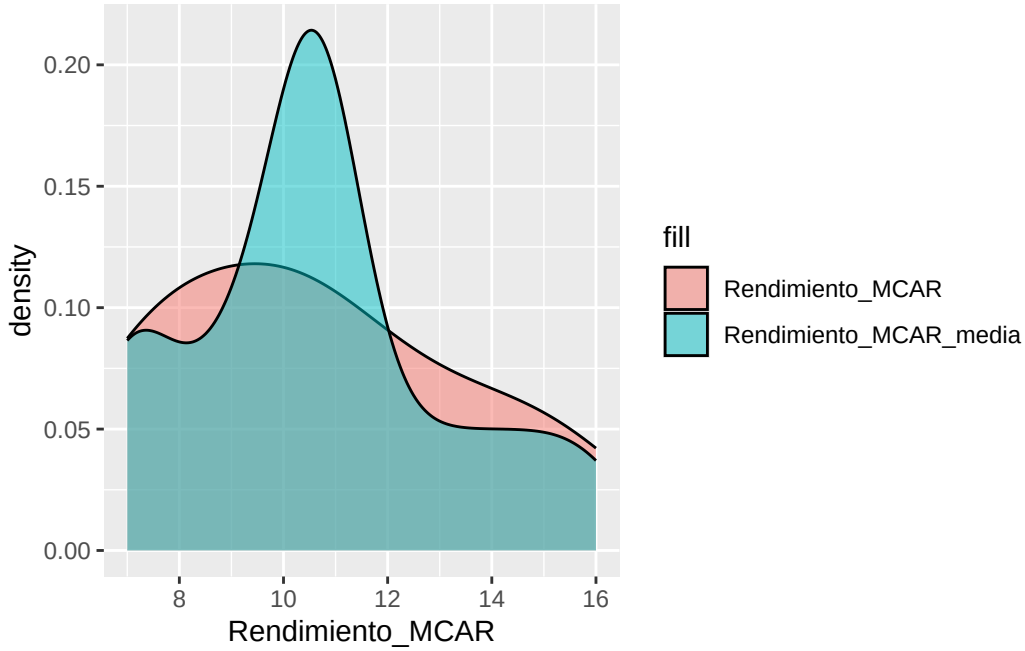
#### Imputación por la media

```
datos$Rendimiento_MCAR_media<-datos$Rendimiento_MCAR
mean(datos$Rendimiento_MCAR, na.rm = TRUE)
```

```
[1] 10.6
```

```
datos$Rendimiento_MCAR_media[is.na(datos$Rendimiento_MCAR_media)] <- mean(datos$Rendimiento_M

##Gráfico para representar las diferencias
ggplot(datos, aes(x = Rendimiento_MCAR, fill = "Rendimiento_MCAR")) +
geom_density(alpha = 0.5) +
geom_density(aes(x = Rendimiento_MCAR_media, fill = "Rendimiento_MCAR_media"), alpha = 0.5)
```

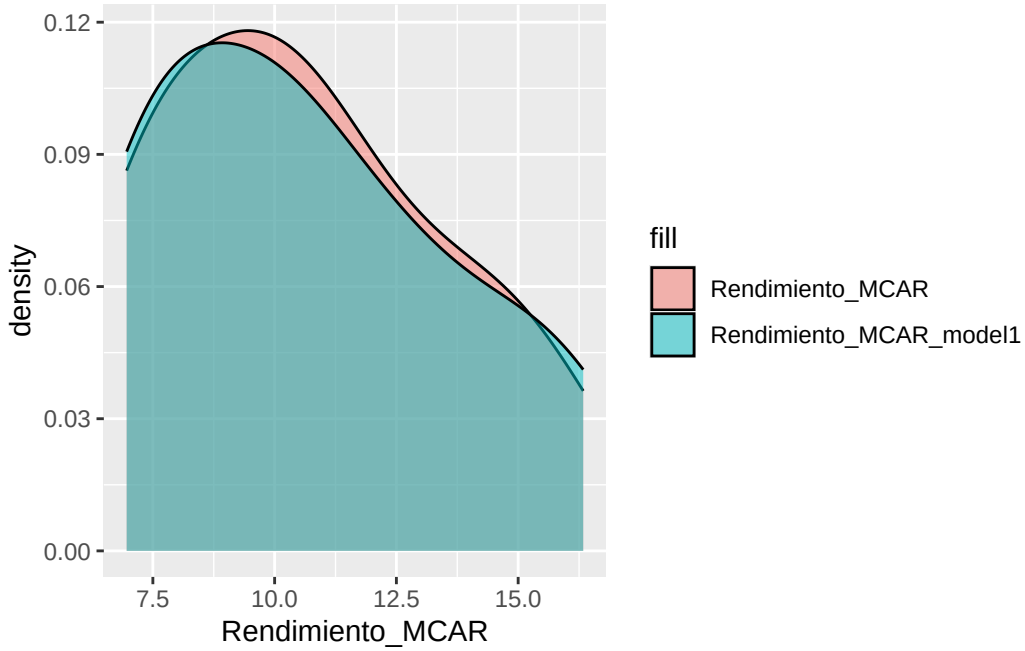
```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_density()`).
```




**Imputación por modelo de regresión lineal**

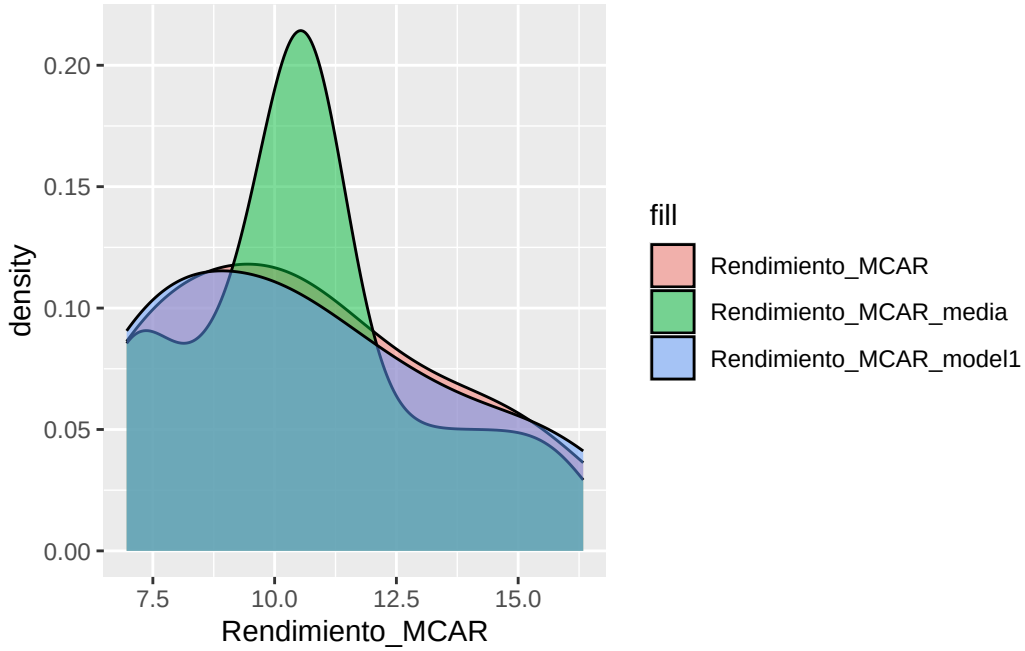
```
model1 <- lm(Rendimiento_MCAR ~ IQ, data = datos)
predictions <- predict(model1,newdata = datos [is.na(datos$Rendimiento_MCAR),])
datos$Rendimiento_MCAR_model1 <- datos$Rendimiento_MCAR
datos$Rendimiento_MCAR_model1[is.na(datos$Rendimiento_MCAR)]<- predictions


##Gráfico para representar las diferencias
ggplot(datos, aes(x = Rendimiento_MCAR, fill = "Rendimiento_MCAR")) +
geom_density(alpha = 0.5) +
geom_density(aes(x = Rendimiento_MCAR_model1, fill = "Rendimiento_MCAR_model1"), alpha = 0.5)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_density()`).
```



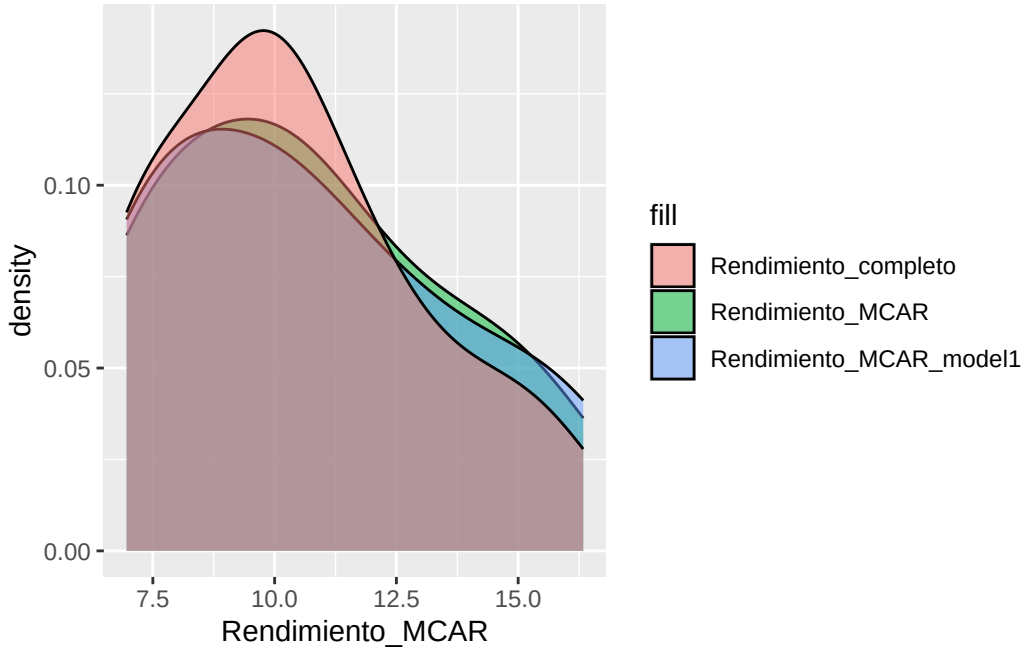

```
##Gráfico para representar las diferencias
ggplot(datos, aes(x = Rendimiento_MCAR, fill = "Rendimiento_MCAR")) +
geom_density(alpha = 0.5) +
  geom_density(aes(x = Rendimiento_MCAR_media, fill = "Rendimiento_MCAR_media"), alpha = 0.5)
geom_density(aes(x = Rendimiento_MCAR_model1, fill = "Rendimiento_MCAR_model1"), alpha = 0.5)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_density()`).
```




```
##Gráfico para representar las diferencias
ggplot(datos, aes(x = Rendimiento_MCAR, fill = "Rendimiento_MCAR")) +
geom_density(alpha = 0.5) +
   geom_density(aes(x = Rendimiento_MCAR_model1, fill = "Rendimiento_MCAR_model1"), alpha = 
geom_density(aes(x = Rendimiento_completo, fill = "Rendimiento_completo"), alpha = 0.5)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_density()`).
```




En este caso la imputación por regresión lineal simple parece correcta

## MAR

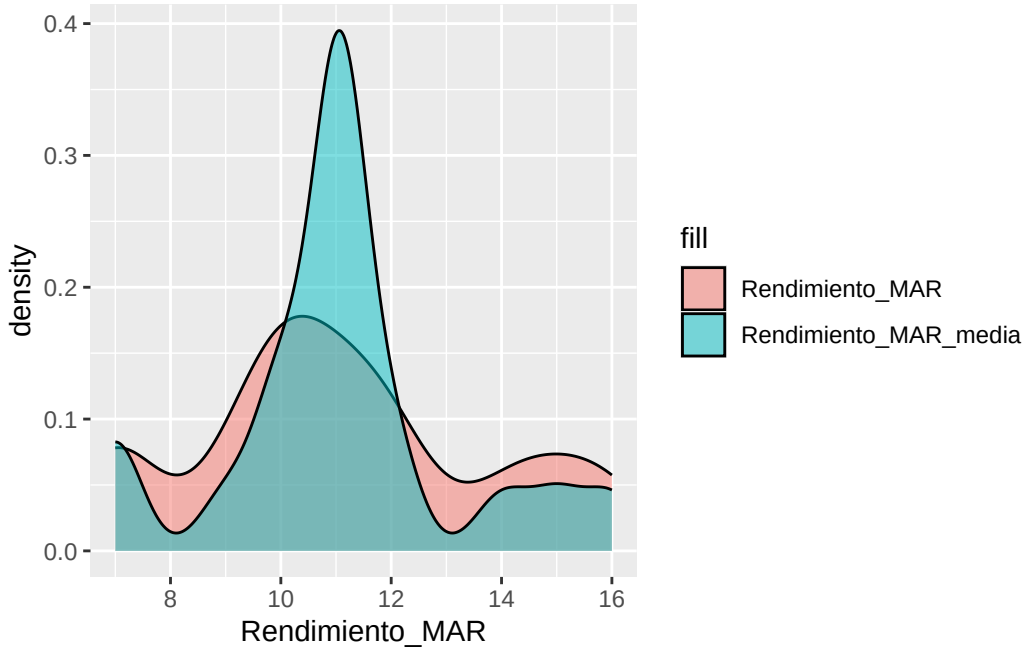
### Imputación por la media

```
datos$Rendimiento_MAR_media<-datos$Rendimiento_MAR
mean(datos$Rendimiento_MAR, na.rm = TRUE)
```

```
[1] 11.07692
```

```
datos$Rendimiento_MAR_media[is.na(datos$Rendimiento_MAR_media)] <- mean(datos$Rendimiento_MAI


##Gráfico para representar las diferencias
ggplot(datos, aes(x = Rendimiento_MAR, fill = "Rendimiento_MAR")) +
geom_density(alpha = 0.5) +
geom_density(aes(x = Rendimiento_MAR_media, fill = "Rendimiento_MAR_media"), alpha = 0.5)
```

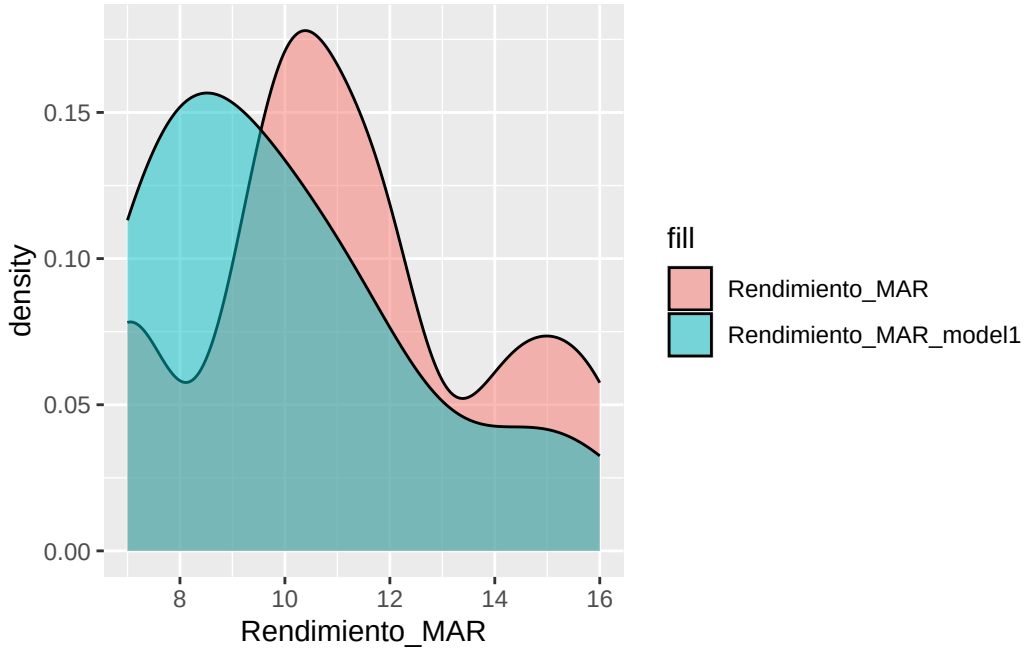
```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_density()`).
```




**Imputación por regresión lineal**

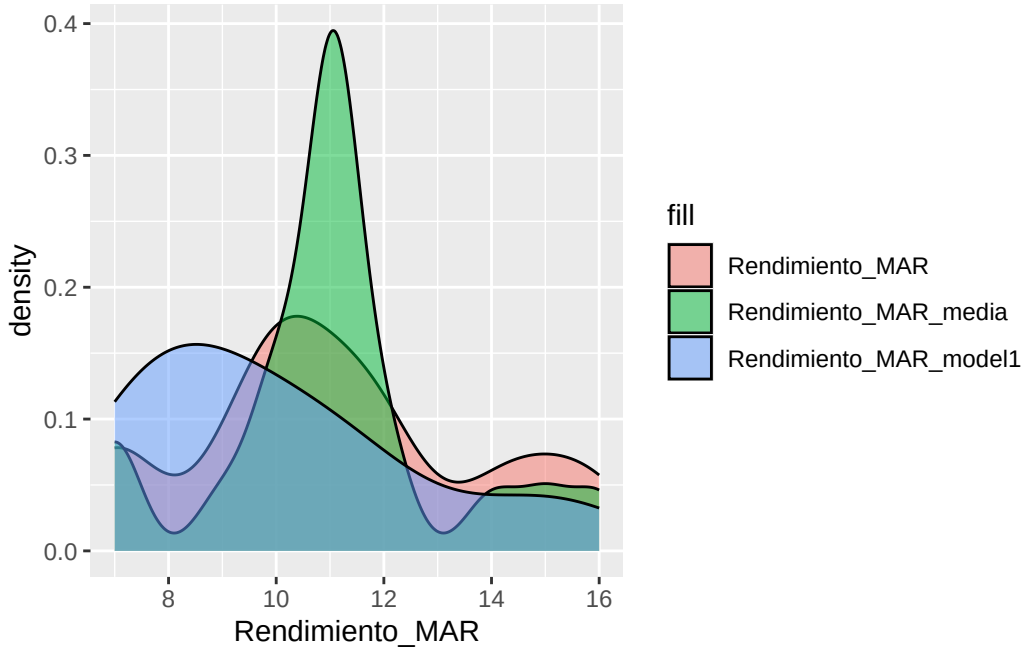
```
model1 <- lm(Rendimiento_MAR ~ IQ, data = datos)
predictions <- predict(model1,newdata = datos [is.na(datos$Rendimiento_MAR),])
datos$Rendimiento_MAR_model1 <- datos$Rendimiento_MAR

datos$Rendimiento_MAR_model1[is.na(datos$Rendimiento_MAR)]<- predictions


ggplot(datos, aes(x = Rendimiento_MAR, fill = "Rendimiento_MAR")) +
geom_density(alpha = 0.5) +
  geom_density(aes(x = Rendimiento_MAR_model1, fill = "Rendimiento_MAR_model1"), alpha = 0.5)
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_density()`).
```



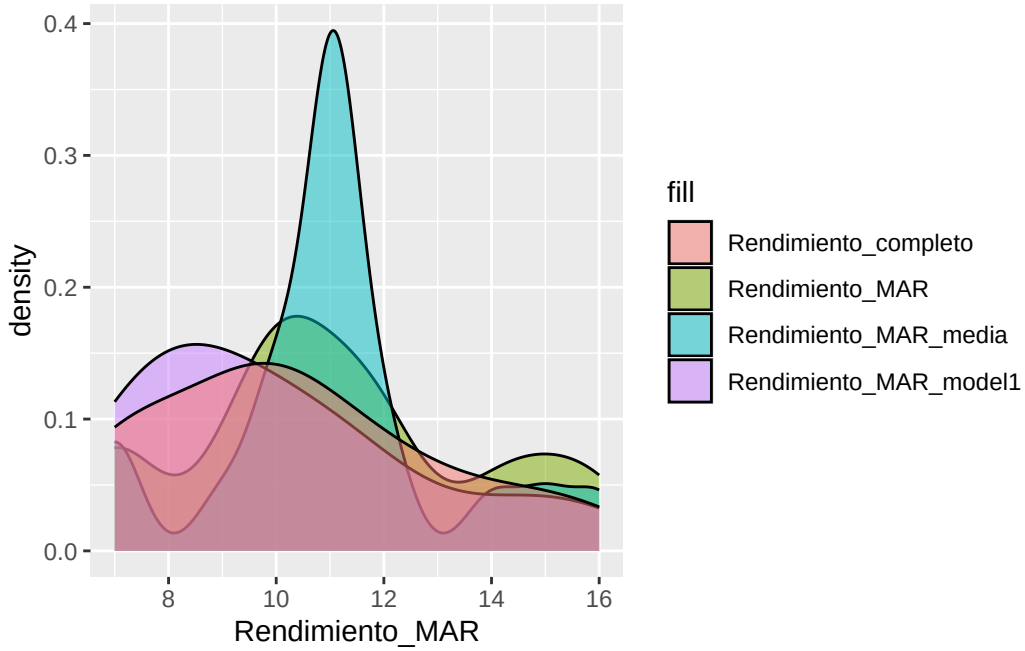

```
##Los datos faltantes corresponden a los que tienen IQs más bajos, por tanto imputa todos lo


ggplot(datos, aes(x = Rendimiento_MAR, fill = "Rendimiento_MAR")) +
geom_density(alpha = 0.5) +
  geom_density(aes(x = Rendimiento_MAR_media, fill = "Rendimiento_MAR_media"), alpha = 0.5) +
geom_density(aes(x = Rendimiento_MAR_model1, fill = "Rendimiento_MAR_model1"), alpha = 0.5)
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_density()`).
```




```
ggplot(datos, aes(x = Rendimiento_MAR, fill = "Rendimiento_MAR")) +
geom_density(alpha = 0.5) +
geom_density(aes(x = Rendimiento_MAR_media, fill = "Rendimiento_MAR_media"), alpha = 0.5) +
geom_density(aes(x = Rendimiento_MAR_model1, fill = "Rendimiento_MAR_model1"), alpha = 0.5) +
geom_density(aes(x = Rendimiento_completo, fill = "Rendimiento_completo"), alpha = 0.5)
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_density()`).
```




##Aquí lo que ocurre es que como todos los datos missing corresponden a los que tienen bajo IQ, ##se imputan todos como muy bajo rendimiento.

### Imputación con incentidumbre

```
model1 <- lm(Rendimiento_MAR ~ IQ, data = datos)

summary(model1)
```

```
Call:
lm(formula = Rendimiento_MAR ~ IQ, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8997 -1.6760 -0.2165  1.7835  4.1438

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.62794    6.66849  -0.544   0.5973
IQ           0.13664    0.06166   2.216   0.0487 *
---
```

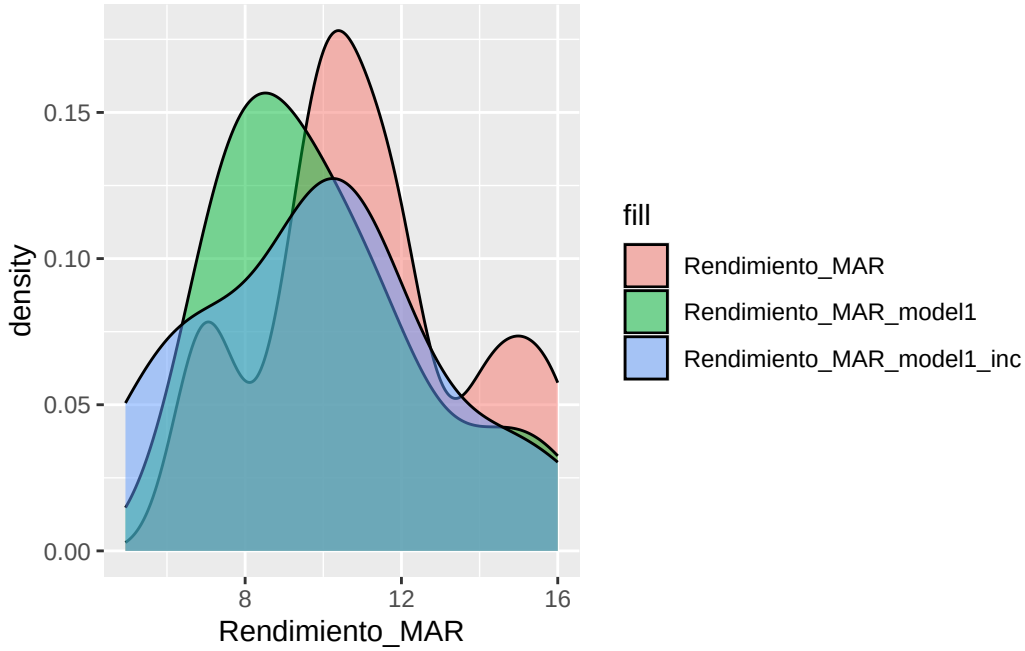
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.39 on 11 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.3087,    Adjusted R-squared:  0.2458
F-statistic: 4.911 on 1 and 11 DF,  p-value: 0.0487
```

```r
set.seed(3)
rnorm(sum(is.na(datos$Rendimiento_MAR)), 0, sd = 2.39)
```

```
[1] -2.29902086 -0.69913648  0.61850384 -2.75359521  0.46792095  0.07199623
[7]  0.20414838
```

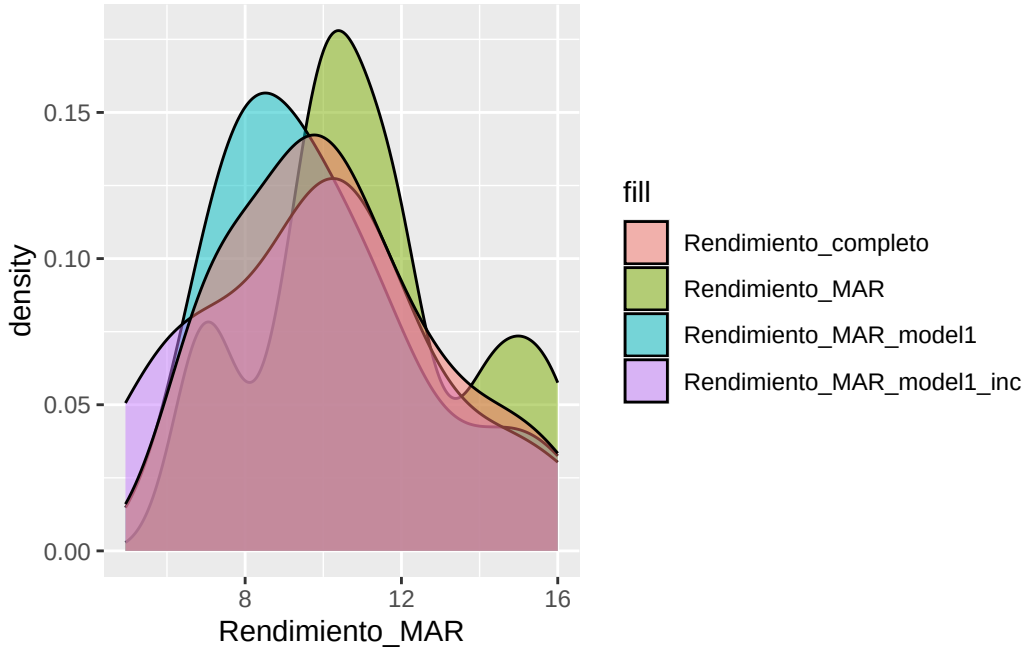
```r
datos$Rendimiento_MAR_model1_inc <- datos$Rendimiento_MAR
datos$Rendimiento_MAR_model1_inc[is.na(datos$Rendimiento_MAR)]<- predictions + rnorm(sum(is.n



ggplot(datos, aes(x = Rendimiento_MAR, fill = "Rendimiento_MAR")) +
geom_density(alpha = 0.5) +
  geom_density(aes(x = Rendimiento_MAR_model1, fill = "Rendimiento_MAR_model1"), alpha = 0.5)
geom_density(aes(x = Rendimiento_MAR_model1_inc, fill = "Rendimiento_MAR_model1_inc"), alpha
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_density()`).
```




```
ggplot(datos, aes(x = Rendimiento_MAR, fill = "Rendimiento_MAR")) +
geom_density(alpha = 0.5) +
  geom_density(aes(x = Rendimiento_MAR_model1, fill = "Rendimiento_MAR_model1"), alpha = 0.5
geom_density(aes(x = Rendimiento_MAR_model1_inc, fill = "Rendimiento_MAR_model1_inc"), alpha
geom_density(aes(x = Rendimiento_completo, fill = "Rendimiento_completo"), alpha = 0.5)
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_density()`).
```




La imputación con incertidumbre mejora bastante

### MNAR

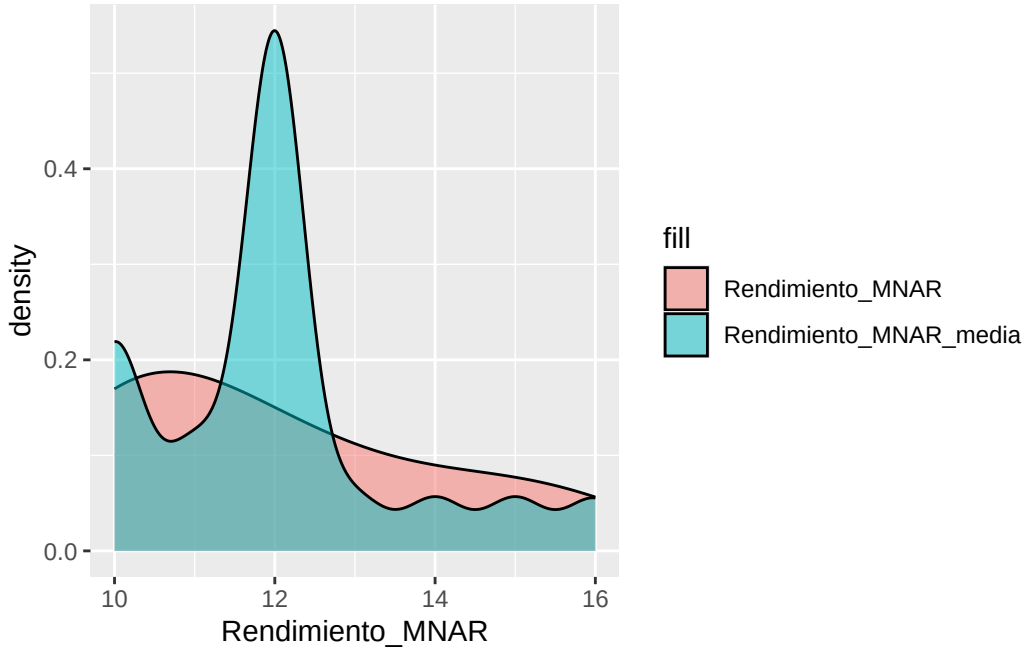
#### Imputación por la media

```
datos$Rendimiento_MNAR_media<-datos$Rendimiento_MNAR
mean(datos$Rendimiento_MNAR, na.rm = TRUE)
```

```
[1] 12
```

```
datos$Rendimiento_MNAR_media[is.na(datos$Rendimiento_MNAR_media)] <- mean(datos$Rendimiento_M


##Gráfico para representar las diferencias
ggplot(datos, aes(x = Rendimiento_MNAR, fill = "Rendimiento_MNAR")) +
geom_density(alpha = 0.5) +
  geom_density(aes(x = Rendimiento_MNAR_media, fill = "Rendimiento_MNAR_media"), alpha = 0.5)
```

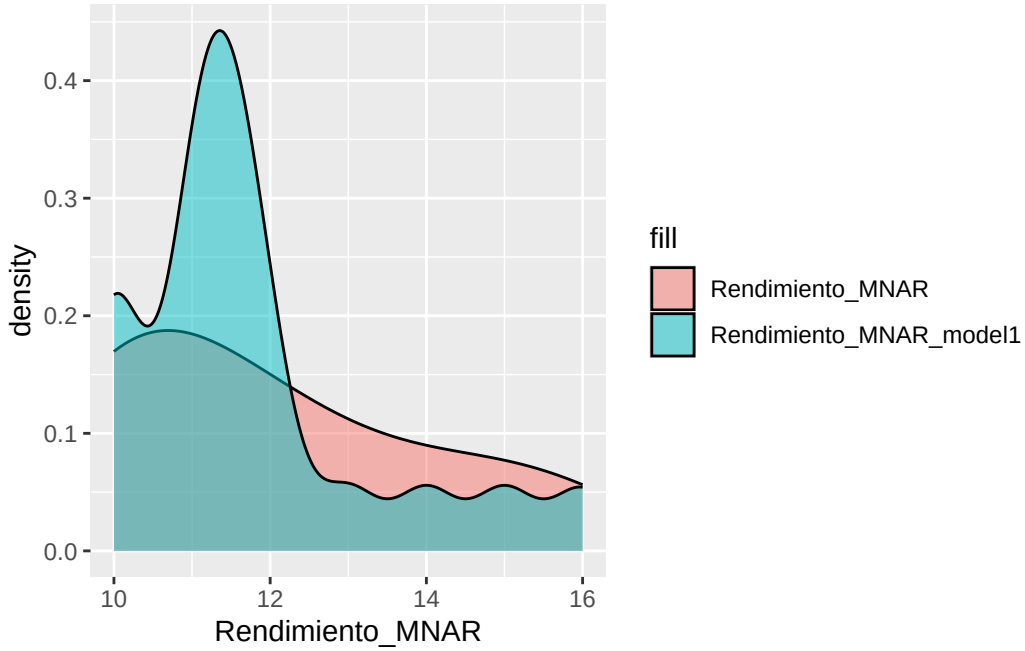
```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_density()`).
```




**Imputación por regresión lineal**

```
model1 <- lm(Rendimiento_MNAR ~ IQ, data = datos)
predictions <- predict(model1,newdata = datos [is.na(datos$Rendimiento_MNAR),])
datos$Rendimiento_MNAR_model1 <- datos$Rendimiento_MNAR
datos$Rendimiento_MNAR_model1[is.na(datos$Rendimiento_MNAR)]<- predictions


ggplot(datos, aes(x = Rendimiento_MNAR, fill = "Rendimiento_MNAR")) +
geom_density(alpha = 0.5) +
geom_density(aes(x = Rendimiento_MNAR_model1, fill = "Rendimiento_MNAR_model1"), alpha = 0.5)
```

```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_density()`).
```




Pasa un poco lo mismo que en el caso de MAR, se imputan los que tienen bajo rendimiento con un modelo que asume bajo rendimiento para los que tiene IQ

### Imputación con incentidumbre

```
model1 <- lm(Rendimiento_MNAR ~ IQ, data = datos)
summary(model1)
```

```
Call:
lm(formula = Rendimiento_MNAR ~ IQ, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2166 -1.3206 -0.7243  1.7205  3.5471

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.80557    4.82266   1.619    0.137
IQ           0.03938    0.04492   0.877    0.401

Residual standard error: 2.111 on 10 degrees of freedom
```

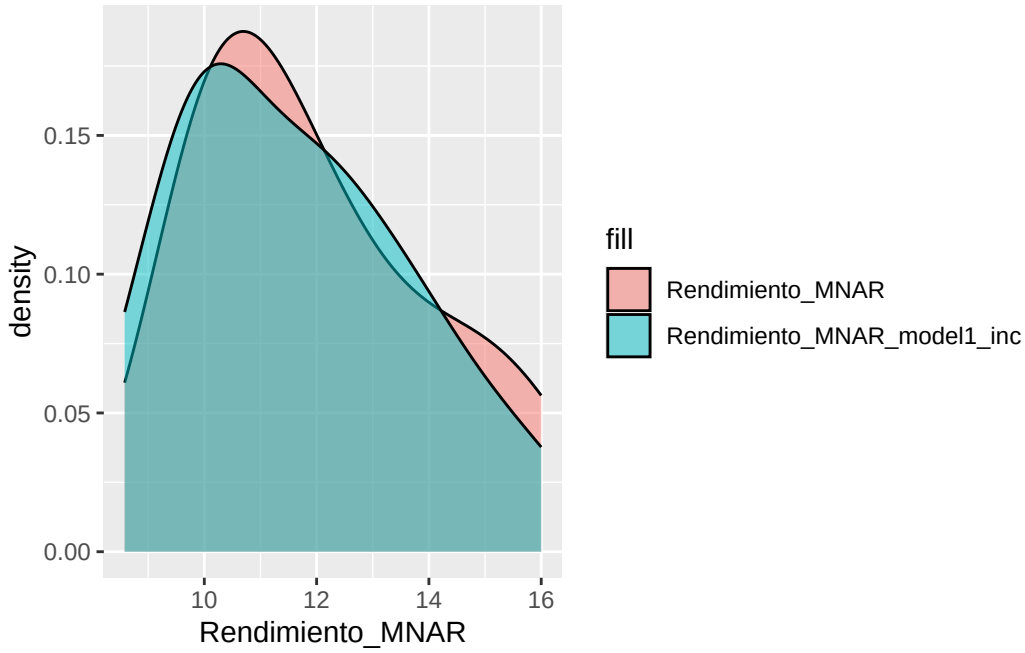
```
  (8 observations deleted due to missingness)
Multiple R-squared:  0.07138,   Adjusted R-squared:  -0.02148
F-statistic: 0.7687 on 1 and 10 DF,  p-value: 0.4012
```

```r
set.seed(3)
rnorm(sum(is.na(datos$Rendimiento_MAR)), 0, sd = 2.11)
```

```
[1] -2.02967951 -0.61722928  0.54604314 -2.43099828  0.41310176  0.06356152
[7]  0.18023141
```

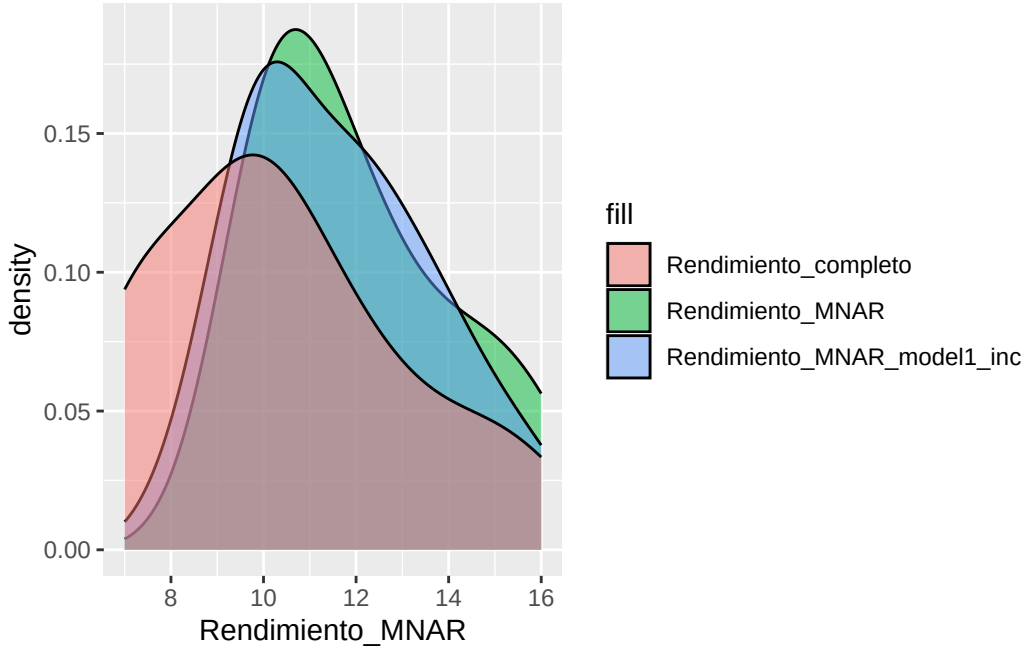
```r
datos$Rendimiento_MNAR_model1_inc <- datos$Rendimiento_MNAR
datos$Rendimiento_MNAR_model1_inc[is.na(datos$Rendimiento_MNAR)]<- predictions + rnorm(sum(is


ggplot(datos, aes(x = Rendimiento_MNAR, fill = "Rendimiento_MNAR")) +
geom_density(alpha = 0.5) +
geom_density(aes(x = Rendimiento_MNAR_model1_inc, fill = "Rendimiento_MNAR_model1_inc"), alpl
```

```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_density()`).
```



```r
ggplot(datos, aes(x = Rendimiento_MNAR, fill = "Rendimiento_MNAR")) +
geom_density(alpha = 0.5) +
geom_density(aes(x = Rendimiento_MNAR_model1_inc, fill = "Rendimiento_MNAR_model1_inc"), alp
geom_density(aes(x = Rendimiento_completo, fill = "Rendimiento_completo"), alpha = 0.5)
```

```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_density()`).
```




El modelo con incertidumbre se ajusta bien a los datos de MNAR pero no a la realidad porque imputar una relación no observada es muy difícil.