1. **Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?**

   El dataset *juul* contiene datos estructurados en 1339 filas y 6 columnas con el objetivo de representar una muestra del factor de crecimiento insulínico IGF-I en niños de primaria.

   Se quiere analizar la relación entre este factor y el sexo, así como el estado de maduración sexual, que se recoge en las variables *menarche*, *tanner* y *testvol*. La primera hace referencia a si una chica ha tenido ya el periodo, la segunda al volumen testicular de los chicos y la tercera al estado puberal según una escala de 1 a 5.

   La seguiente tabla representa las principales características de las diferentes variables del dataset:

```
> dataset<-juul
> summary(dataset)
```

```
      age            menarche          sex             igf1       
 Min.   : 0.170   Min.   :1.000   Min.   :1.000   Min.   : 25.0  
 1st Qu.: 9.053   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:202.2  
 Median :12.560   Median :1.000   Median :2.000   Median :313.5  
 Mean   :15.095   Mean   :1.476   Mean   :1.534   Mean   :340.2  
 3rd Qu.:16.855   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:462.8  
 Max.   :83.000   Max.   :2.000   Max.   :2.000   Max.   :915.0  
 NA's   :5        NA's   :635     NA's   :5       NA's   :321    
     tanner        testvol      
 Min.   :1.00   Min.   : 1.000  
 1st Qu.:1.00   1st Qu.: 1.000  
 Median :2.00   Median : 3.000  
 Mean   :2.64   Mean   : 7.896  
 3rd Qu.:5.00   3rd Qu.:15.000  
 Max.   :5.00   Max.   :30.000  
 NA's   :240    NA's   :859     
```

2. **Integración y selección de los datos de interés a analizar.**

   En este caso la información de la edad no es de gran interés para el estudio que se quiere realizar. Por tanto, solo se analizarán el resto de datos.

```
> dataset<-dataset[,-1]
> summary(dataset)
```

```
    menarche          sex             igf1           tanner      
 Min.   :1.000   Min.   :1.000   Min.   : 25.0   Min.   :1.00  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:202.2   1st Qu.:1.00  
 Median :1.000   Median :2.000   Median :313.5   Median :2.00  
 Mean   :1.476   Mean   :1.534   Mean   :340.2   Mean   :2.64  
```

```
3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:462.8   3rd Qu.:5.00
Max.   :2.000   Max.   :2.000   Max.   :915.0   Max.   :5.00
NA's   :635     NA's   :5       NA's   :321     NA's   :240
   testvol
Min.   : 1.000
1st Qu.: 1.000
Median : 3.000
Mean   : 7.896
3rd Qu.:15.000
Max.   :30.000
NA's   :859
```

De hecho, si solo queremos acotar el problema al sexo femenino podemos eliminar la variable *testvol*, de la misma manera que si solo queremos analizar el sexo masculino se puede eliminar la variable *menarche*.

3. **Limpieza de los datos.**

   a) **¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?**

   Los datos contienen elementos vacíos representados con el valor NA. En la tabla anterior se puede ver la cantidad de NA's para cada variable. En este caso no podemos recuperar la información vacía ya que no se trata de una base de datos nuestra y, por tanto, es complicado intentar completar los campos vacíos volviendo a preguntar a las personas que se analizaron.

   Una posible forma de recuperar los valores perdidos de la variable *sex* habría sido posible si para estas muestras tuviéramos la información de *menarche* o *testvol*. Desgraciadamente, se comprueba que para estas muestras solo tenemos la información de *igf1*:

   ```
   > head(dataset,5)
     menarche sex igf1 tanner testvol
   1       NA  NA   90     NA       NA
   2       NA  NA   88     NA       NA
   3       NA  NA  164     NA       NA
   4       NA  NA  166     NA       NA
   5       NA  NA  131     NA       NA
   ```

   No obstante, solo se trata de un 0.37 % de los datos de sexo que no se pueden recuperar. Asímismo, solo el 23.97 % de los datos de *igf1* y el 17.92 % de *tanner* son NA's.

   Por otro lado, aunque no podamos recuperar los datos de *menarche* y *testvol*, podemos comprobar qué parte son datos perdidos y qué parte son NA solo por el hecho de que a una chica no se le puede medir el volumen testicular y un chico no puede tener el periodo.

   El porcentaje de chicas sin datos sobre el periodo y el de chicos sin datos sobre el volumen testicular es, respectivamente:

```
> sum(is.na(dataset$menarche[dataset$sex==2]))/length(dataset[dataset$sex==2,1])*10
[1] 1.949861
> sum(is.na(dataset$testvol[dataset$sex==1]))/length(dataset[dataset$sex==1,1])*100
[1] 23.32268
```

Una posible solución pasaría por eliminar las variables NA de *sex*, identificar los casos de *menarche* y *testvol* donde no proceda identificar un valor y modificarlos por un valor en común, como el cero, y aplicar un missForest al resultado.

Primero se eliminan las muestras sin información en *sex*.

```
> dataset<-dataset[!is.na(dataset$sex),]
```

Se identifican las muestras de chicas con un 0 en *testvol*.

```
> for (i in 1:nrow(dataset)){
+   if(dataset$sex[i]==2){
+       dataset$testvol[i]<-0
+   }
+ }
```

Se identifican las muestras de chicos con un 0 en *menarche*.

```
> for (i in 1:nrow(dataset)){
+   if(dataset$sex[i]==1){
+       dataset$menarche[i]<-0
+   }
+ }
```

Se imputan los datos restantes con missForest. Previamente, es necesario indicar las variables que son factores para que el algoritmo no proponga valores más allá de los posibles.

```
> dataset$menarche<-factor(dataset$menarche)
> dataset$sex<-factor(dataset$sex)
> dataset$tanner<-factor(dataset$tanner)
> dataset.imp<-missForest(dataset, variablewise = TRUE)
  missForest iteration 1 in progress...done!
  missForest iteration 2 in progress...done!
  missForest iteration 3 in progress...done!
  missForest iteration 4 in progress...done!
  missForest iteration 5 in progress...done!
  missForest iteration 6 in progress...done!
> dataset<-dataset.imp$ximp
> summary(dataset)
 menarche sex          igf1       tanner     testvol
 0:621    1:621   Min.   : 25.0   1:620   Min.   : 0.000
 1:378    2:713   1st Qu.:215.7   2:123   1st Qu.: 0.000
```

```
 2:335           Median :284.0   3: 81   Median : 0.000
                 Mean   :328.8   4: 81   Mean   : 3.349
                 3rd Qu.:442.0   5:429   3rd Qu.: 2.483
                 Max.   :915.0           Max.   :30.000
```

b) **Identificación y tratamiento de valores extremos.**

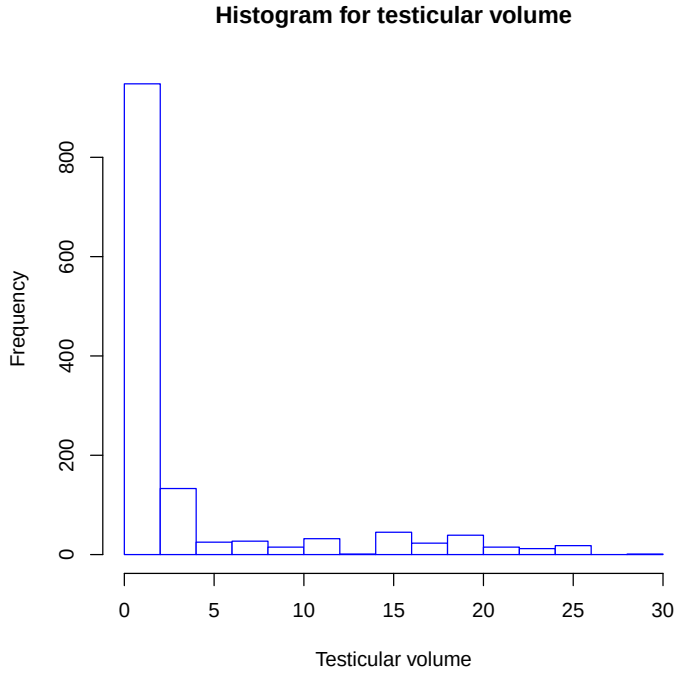
Respecto a los valores extremos, según algunas definiciones, estos se encuentran al menos a 3 desviaciones estándar alejados de la media. Así, para el volumen testicular, este umbral se encuentra en

```
> th_testvol<-mean(dataset$testvol,na.rm=T)+3*sd(dataset$testvol,na.rm=T)
> th_testvol
[1] 22.34905
```

Y podemos ver que esta variable presenta 31 valores extremos.

```
> sum(dataset$testvol>th_testvol,na.rm=T)
[1] 31
```

```
> hist(dataset$testvol,main="Histogram for testicular volume",
+      xlab="Testicular volume",
+      border="blue",
+      col="white")
```

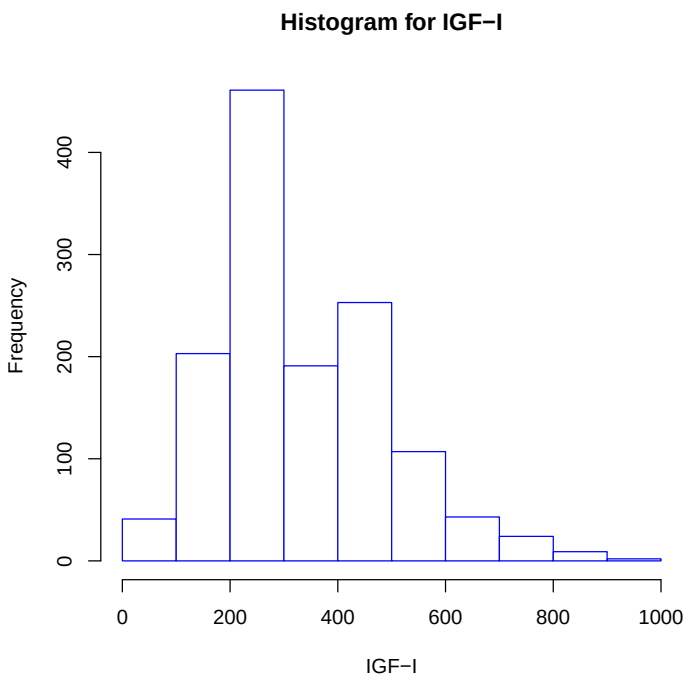


Por otro lado, el umbral de valores extremos para *igf1* es

```
> th_igf1<-mean(dataset$igf1,na.rm=T)+3*sd(dataset$igf1,na.rm=T)
> th_igf1
[1] 803.9074
```

Por tanto, podemos ver que 10 puntos se encuentran por encima de este umbral.

```
> sum(dataset$igf1>th_igf1,na.rm=T)
[1] 10
> hist(dataset$igf1,main="Histogram for IGF-I",
+      xlab="IGF-I",
+      border="blue",
+      col="white")
```



En función del origen de estos outliers tomaremos diferentes decisiones. Así, si suponemos que no ha habido errores en la adquisición de los datos y que, por lo tanto, se trata de puntos reales que provienen de la zona menos probable de la distribución de los datos, los incluiremos en el análisis. No obstante, si tenemos dudas sobre si los datos se recogieron correctamente, los eliminaremos del estudio.

Otra estrategia puede ser analizar los datos con y sin estos valores extremos y decidir qué resultado damos por correcto en función de las diferencias y las tendencias observadas en los resultados.

En este caso supondremos que los datos son correctos y los analizaremos ya que la muestra es suficientemente grande como para incluir valores extremos.

4. **Análisis de los datos.**

   a) **Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).**

   Se comparará el valor de *igf1* entre chicos/chicas (*sex*), entre chicas con/sin el periodo (*menarche*), entre chicos con diferente volumen testicular (*testvol*) y entre las diferentes etapas de maduración (*tanner*).

   Por tanto, hay 3 comparaciones con variables categóricas que etiquetaremos de la siguiente manera:

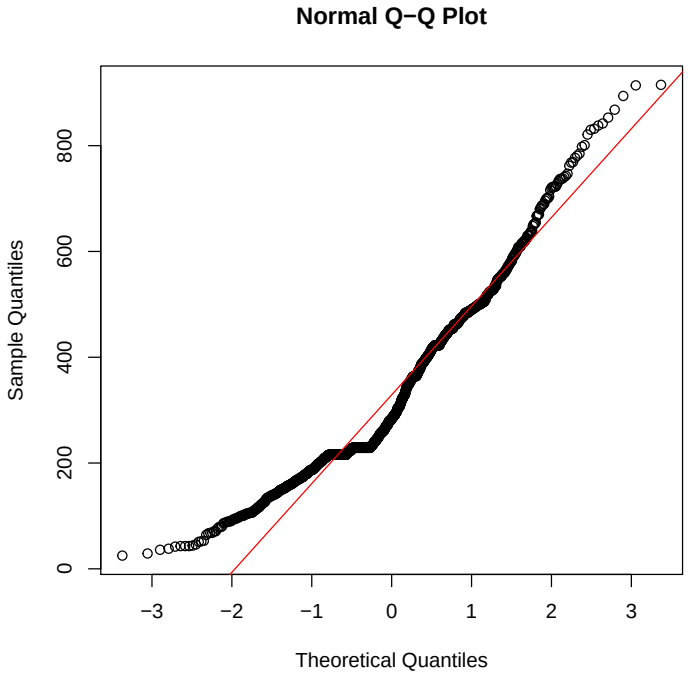
```
> dataset$sex<-factor(dataset$sex, labels=c("M","F"))
> dataset$menarche<-factor(dataset$menarche, labels=c("NP","No","Yes"))
> dataset$tanner<-factor(dataset$tanner, labels=c("I","II","III","IV","V"))
> summary(dataset)
```

```
 menarche  sex          igf1       tanner       testvol
 NP :621   M:621   Min.   : 25.0   I  :620   Min.   : 0.000
 No :378   F:713   1st Qu.:215.7   II :123   1st Qu.: 0.000
 Yes:335           Median :284.0   III: 81   Median : 0.000
                   Mean   :328.8   IV : 81   Mean   : 3.349
                   3rd Qu.:442.0   V  :429   3rd Qu.: 2.483
                   Max.   :915.0             Max.   :30.000
```

   b) **Comprobación de la normalidad y homogeneidad de la varianza.**

   Para comprobar la normalidad de los datos se pueden utilizar los QQplots ya que permiten observar la similitud entre las distribuciones de dos conjuntos de datos, la analizada y una distribución normal ideal. Así, para la variable *igf1*

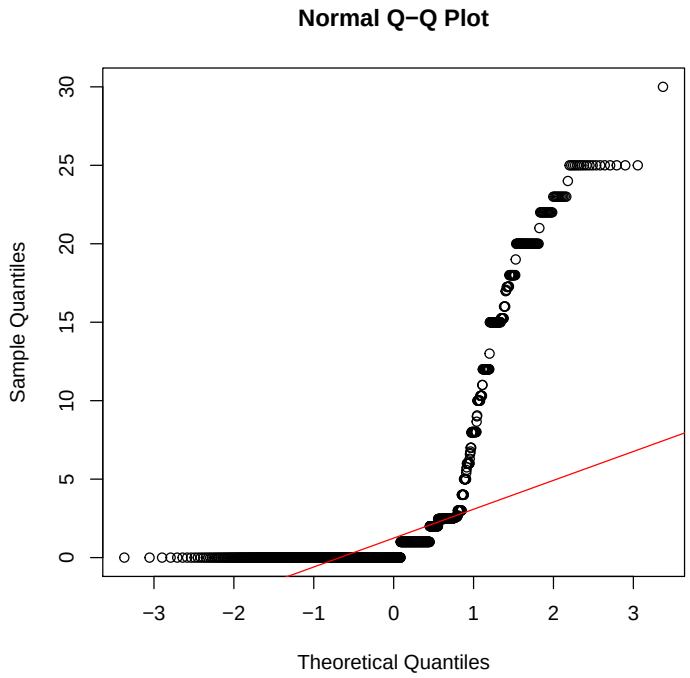
```
> qqnorm(dataset$igf1)
> qqline(dataset$igf1,col=2)
```



podemos observar que, como ya se podía intuir de la forma del histograma, la distribución no se aleja mucho de la normal. No obstante, habrá que comprobarlo con una prueba adicional ya que encontramos diversas muestras fuera de la recta de regresión.

En cambio, como también se observaba en el histograma y como se puede comprobar del QQplot, *testvol* no tiene un comportamiento normal.

```
> qqnorm(dataset$testvol)
> qqline(dataset$testvol,col=2)
```



A pesar de que para *testvol* no sería ya necesario, lo mejor es aplicar a todas las variables un test de normalidad como el de Shapiro:

```
> shapiro.test(dataset$igf1)

        Shapiro-Wilk normality test

data:  dataset$igf1
W = 0.94715, p-value < 2.2e-16

> shapiro.test(dataset$testvol)

        Shapiro-Wilk normality test

data:  dataset$testvol
W = 0.58818, p-value < 2.2e-16
```

De los resultados significativos del test observamos que ninguna variable sigue una distribución normal.

Como los datos de *testvol* se alejan mucho de la distribución normal, los analizaremos a partir de pruebas que no presuponen estas características en los datos. No obstante, para *igf1* intentaremos transformar los datos para que sean normales.

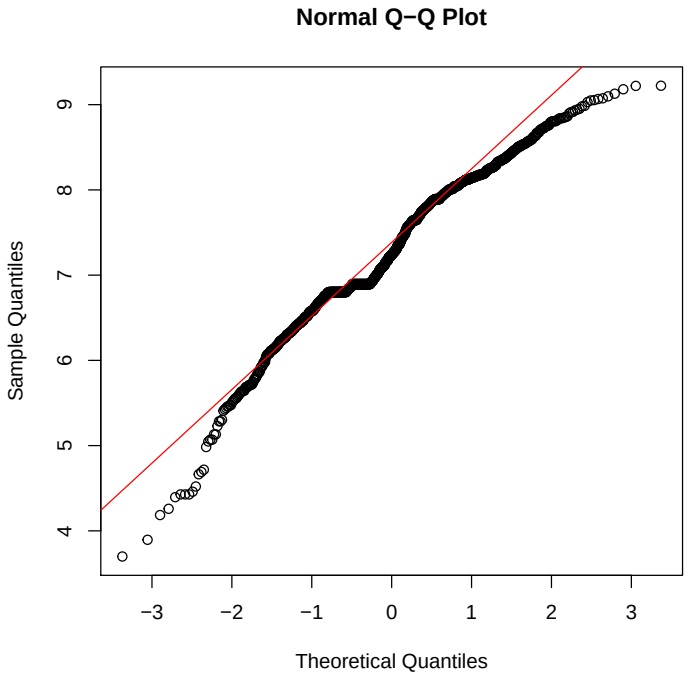
Lo haremos a partir de la transformación de BoxCox (*DescTools*).

```
> igf1.norm <- BoxCox(dataset$igf1, lambda = BoxCoxLambda(dataset$igf1))
> shapiro.test(igf1.norm)
```

```
        Shapiro-Wilk normality test

data:  igf1.norm
W = 0.97376, p-value = 7.442e-15
```

A pesar de que según el p-valor la transformación aplicada no es suficiente para tratar los datos como normales ($p<0.05$), observamos como el nuevo QQplot se aleja menos de la recta de regresión y, por tanto, se parece más a una distribución normal.

```
> qqnorm(igf1.norm)
> qqline(igf1.norm, col=2)
```



Se pueden probar tantas transformaciones como se desee, pero en este caso supondremos que los datos no se pueden normalizar y, por tanto, los analizaremos con pruebas que no presuponen estas características (pruebas no paramétricas).

En referencia a la homogeneidad de la varianza, dado que los datos no son normales, utilizaremos el método de Fligner-Killeen, para comparar la varianza de *igf1* entre los grupos de *sex* y *tanner*.

```
> fligner.test(igf1 ~ sex, data = dataset)

        Fligner-Killeen test of homogeneity of variances
```

```
data:  igf1 by sex
Fligner-Killeen:med chi-squared = 1.8822, df = 1, p-value = 0.1701
```

```
> fligner.test(igf1 ~ tanner, data = dataset)

        Fligner-Killeen test of homogeneity of variances

data:  igf1 by tanner
Fligner-Killeen:med chi-squared = 185.69, df = 4, p-value < 2.2e-16
```

De los resultados se puede concluir que la varianza de *igf1* es similar entre los chicos y chicas, pero diferente entre los grupos del estado de pubertad (*tanner*).

c) **Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.**

Dado que los datos a analizar no presentan una distribución normal, las pruebas estadísticas para analizar y comparar los diferentes grupos deberán ser no paramétricas.

Así, para los datos dicotómicos como *sex* y *menarche*, se aplica el test de Wilcoxon.

```
> wilcox.test(dataset$igf1~dataset$sex)

        Wilcoxon rank sum test with continuity correction

data:  dataset$igf1 by dataset$sex
W = 173430, p-value = 8.187e-12
alternative hypothesis: true location shift is not equal to 0
```

```
> #eliminamos los datos donde el menarche no procede antes del analisis
> dataset_menarche<-dataset[-which(dataset$menarche=="NP"),]
> wilcox.test(dataset_menarche$igf1~dataset_menarche$menarche)

        Wilcoxon rank sum test with continuity correction

data:  dataset_menarche$igf1 by dataset_menarche$menarche
W = 28513, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Ambas variables muestran que las diferencias entre los grupos de datos son estadísticamente significativas ($p<0.05$). Es decir, que el factor IGF-I es estadísticamente diferente entre chicos y chicas y entre chicas con y sin el periodo.

El test equivalente cuando se tienen 3 o más grupos de datos, como es el caso de *tanner*, es el test de Kruskal-Wallis:

```
> kruskal.test(dataset$igf1~dataset$tanner)
```

```
        Kruskal-Wallis rank sum test

data:  dataset$igf1 by dataset$tanner
Kruskal-Wallis chi-squared = 774.1, df = 4, p-value < 2.2e-16
```

También en este caso los resultados muestran que los grupos son estadísticamente diferentes entre ellos.

Finalmente, un posible análisis que se puede hacer con los datos del volumen testicular es analizar la correlación, a partir del método de Spearman, entre este y el factor IGF-I.

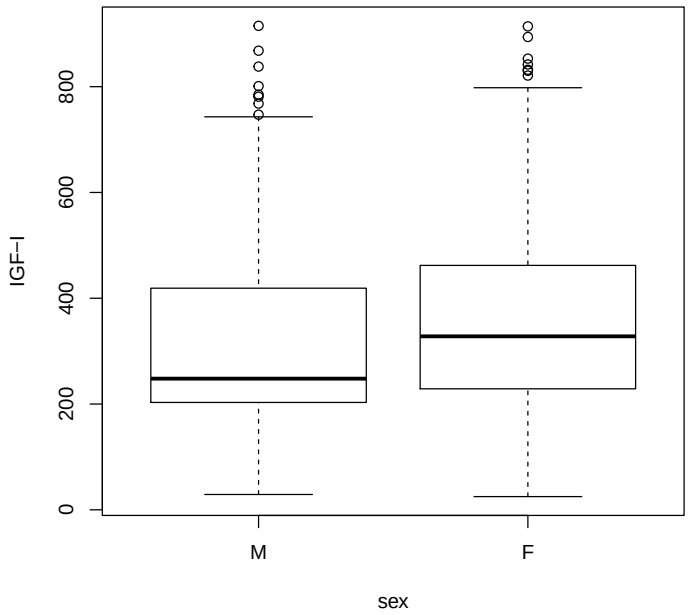
```
> #eliminamos los datos donde el testvol no procede antes del analisis
> dataset_testvol<-dataset[-which(dataset$testvol==0),]
> cor.test(dataset_testvol$igf1,dataset_testvol$testvol,method="spearman")

        Spearman's rank correlation rho

data:  dataset_testvol$igf1 and dataset_testvol$testvol
S = 13155000, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6704043
```

Los resultados demuestran que la correlación es significativa y en un factor de poco más del 65 %.

5. **Representación de los resultados a partir de tablas y gráficas.**

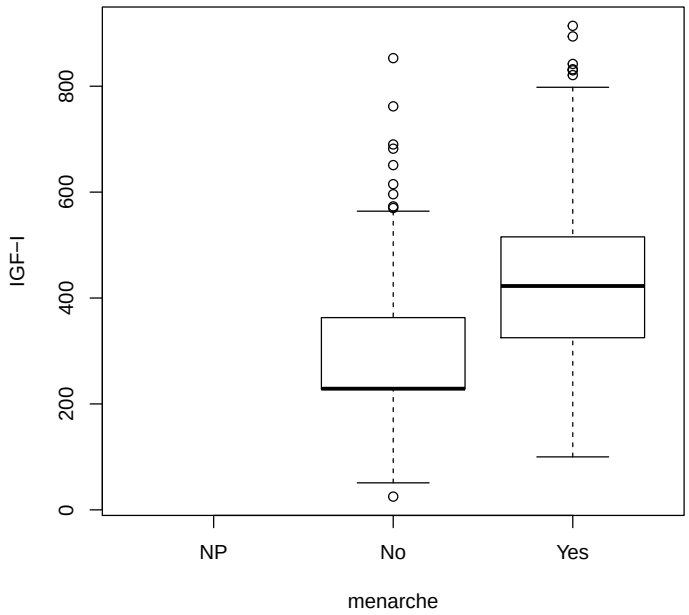
   Las diferencias encontradas entre los grupos se pueden analizar a través de boxplots, para ver más claramente la dirección de estas diferencias. Así, se puede observar que las chicas (F) presentan valores más elevados de IGF-I que los chicos (M).

```
> boxplot(dataset$igf1~dataset$sex, ylab="IGF-I",xlab="sex")
```



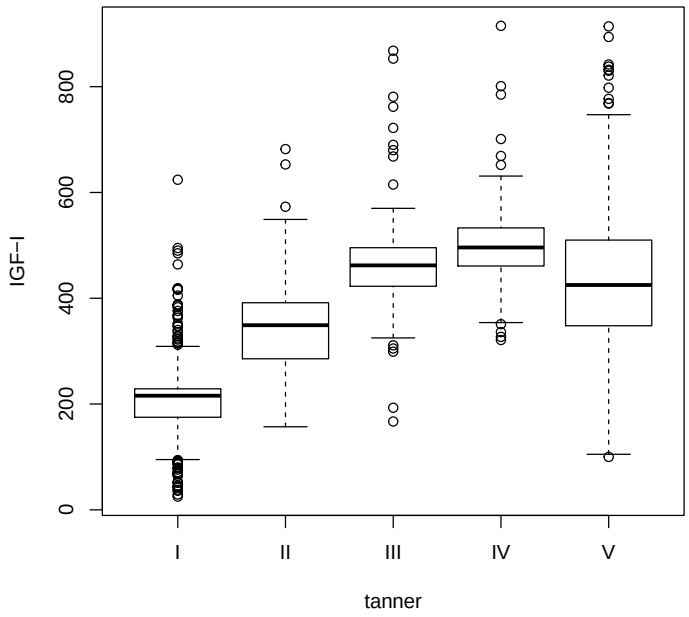
Que las chicas con el periodo presentan también valores más elevados.

```
> boxplot(dataset_menarche$igf1~dataset_menarche$menarche, ylab="IGF-I",xlab="menarche"
```



Y que el factor IGF-I aumenta a medida que aumenta el estado puberal entre I y IV, y disminuye hasta valores similares al estado III en la última fase de la pubertad.

```
> boxplot(dataset$igf1~dataset$tanner, ylab="IGF-I",xlab="tanner")
```



6. **Resolución del problema. A partir de los resultados obtenidos, ¿Cuáles son las conclusiones? ¿Los resultados permiten responder al problema?**

   De los resultados se puede observar que las niñas presentan valores de IGF-I más elevados que los ninos. Además, dentro de las ninas, aquellas que ya han tenido el periodo presentan valores más elevados del factor de crecimiento. Cuando tenemos en cuenta el estado de maduración según la escala de la variable *tanner*, se puede observar la misma tendencia en general, excepto que en el punto de máxima maduración (fase V) los niveles de factor IGF-I parecen bajar a valores similares a los de la fase III.

   Por tanto, el factor IGF-I representa una buena aproximación del estado de maduración, especialmente en las etapas más iniciales de la pubertad. Así, aunque hay que tener en cuenta las diferencias relacionadas con el sexo de la persona, puede ser utilizado para detectar enfermedades que dan lugar a desequilibrios en el proceso de desarrollo físico de una persona.