In [158]:
```
#Paqueterías a usar en el siguiente script
library(gridExtra) ; library(tidyr) ; library(dplyr) ; library(ggplot
2) ; library(lubridate) ; library(stringr) ;library(cowplot) ; librar
y(ggthemes) ; library(readxl); library("modeest"); library(corrplot)

options(repr.plot.width=10, repr.plot.height=8)

#Importación de los datos
Leña_Sur <- data.frame(read_excel("/home/carlos/Documentos/Tarea_chil
e/LeñaSur.xls", sheet = "Datos"))

display_png(file="/home/carlos/Documentos/Tarea_chile/leña.jpeg",widt
h = 10,height = 10)
```


Análisis de datos e Inferencia Estadística, Consumo de leña en chile (2013).

El siguiente conjunto de datos cuenta con 5 variables sobre el consumo de leña en chile.

- Sector: Sector de ubicación de la vivienda
- Aislacion: Variable categórica que registra si la vivienda posee o no aislación térmica
- Calefactor: Tiempo de uso (o antigüedad) del calefactor.
- Consumo: Cantidad media anual de leña, en metros cúbicos, que consume la vivienda
- Humedad: Nivel de humedad registrada en al leña al momento de visitar la vivienda

Objetivo: Conocer el consumo de leña, su relación con variables como el Sector de la vivienda, tiempo de uso del calefactor, asilación térmica, nivel de humedad.

Definiciones.

Atributo desconocido: Consumo de leña

Población: Chile

Unidades Muestrales: Ciertos Sectores en Chile

Estadistica Descriptiva: Promedio, medidas de tendencia central y dispersión, resumenes y visualizaciones.

Estadistica Inferencial: Funcion de distribución acumulada, pruebas de hipótesis, intervalos de confianza, probabilidad.

Comenzando con el Análisis de Datos (Estadística Descriptiva)

Encabezado de nuestros datos.

In [12]:
```
head(Leña_Sur)
```

A data.frame: 6 × 5

	Sector	Aislacion	Calefactor	Consumo	Humedad
	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1	Sector1	SI	10.0	11.337	18.77
2	Sector1	SI	4.1	9.129	18.46
3	Sector1	SI	3.8	8.873	15.46
4	Sector1	SI	2.3	9.193	16.96
5	Sector1	SI	5.6	10.336	18.74
6	Sector1	SI	12.3	13.307	22.02

Contamos con 670 registros, 2 variables categorias, 3 númericas y no contamos con valores nulos.

In [27]:
```
print(paste0("filas: ",dim(Leña_Sur)[1], "  columnas: ",dim(Leña_Sur)[2] ))
```

```
[1] "filas: 670  columnas: 5"
```

In [31]:
```
sum(is.na(Leña_Sur))
```

```
0
```

In [32]:
```
str(Leña_Sur)
```

```
'data.frame':   670 obs. of  5 variables:
 $ Sector    : chr  "Sector1" "Sector1" "Sector1" "Sector1" ...
 $ Aislacion : chr  "SI" "SI" "SI" "SI" ...
 $ Calefactor: num  10 4.1 3.8 2.3 5.6 12.3 3.7 6.4 8.8 11.2 ...
 $ Consumo   : num  11.34 9.13 8.87 9.19 10.34 ...
 $ Humedad   : num  18.8 18.5 15.5 17 18.7 ...
```

Resumenes generales

Antes de comenzar a mezclar variables, veamos algunos resumenes de cada variable catégorica y númerica.

Conteo por sectores, contamos con mayor muestra en el sector 1.

In [33]:
```
Leña_Sur %>%
  group_by(Sector) %>%
  summarise(Conteo = n())
```

A tibble: 3 × 2

Sector	Conteo
<chr>	<int>
Sector1	250
Sector2	220
Sector3	200

Conteo aislacion térmica.

In [34]:
```
Leña_Sur %>%
  group_by(Aislacion) %>%
  summarise(Conteo = n())
```

A tibble: 2 × 2

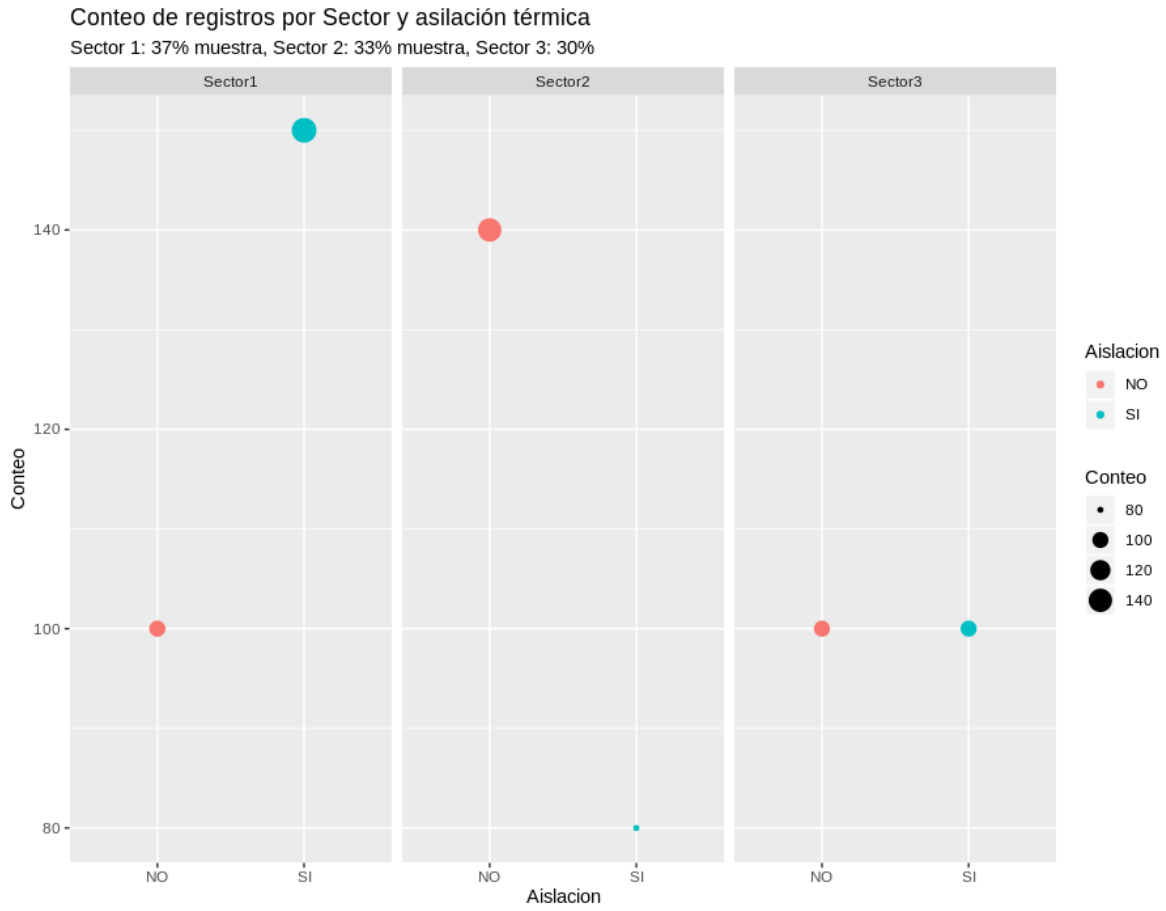
Aislacion	Conteo
<chr>	<int>
NO	340
SI	330

Conteo por sector y aislación térmica.

In [44]:
```
table(Leña_Sur$Sector, Leña_Sur$Aislacion) %>% prop.table() * 100
```

```
                NO       SI
Sector1 14.92537 22.38806
Sector2 20.89552 11.94030
Sector3 14.92537 14.92537
```

In [132]:
```
library("IRdisplay")
display_png(file="/home/carlos/Documentos/Tarea_chile/1.png")
```



Primeras obervaciones:

```
 1.El sector 1 tienen la mayor cantidad de muestra y en este predominan vivi
endas con asilación térmica.
 2.El sector 2 predominan las viviendas sin aislación térmica.
 3.El sector 3 se encuentra equilibrado en cuanto a viviendas con asilación
 térmica.
```

Más adelante veremos el consumo de leña por sector, con esto comienzan a surgir preguntas.

- ¿Que diferencias existen entre sectores?
- ¿De que depende que una vivienda cuente con aislación térmica?
- ¿Disminuye el consumo de leña por sector y repercute la aislación térmica?

Resumenes de las variables númericas, calefactor, humedad, consumo de leña.

Primeras estadísticas descriptivas de los datos.

Medidas de tendencia central y dispersión

- Tendencia central: Promedio, Mediana, Moda.
- Medidas de dispersión: Rango de variación, Desviación estándar, Coeficiente de variación.

In [92]:
```
summary(Leña_Sur)
```

```
    Sector           Aislacion           Calefactor       Consumo
 Length:670         Length:670         Min.   : 0.10   Min.   : 6.43
4
 Class :character   Class :character   1st Qu.: 7.30   1st Qu.:10.27
5
 Mode  :character   Mode  :character   Median :15.00   Median :19.37
1
                                       Mean   :13.46   Mean   :19.16
1
                                       3rd Qu.:18.38   3rd Qu.:25.26
1
                                       Max.   :29.90   Max.   :40.57
5
    Humedad
 Min.   : 7.62
 1st Qu.:13.21
 Median :15.97
 Mean   :15.62
 3rd Qu.:18.04
 Max.   :24.10
```

In [106]:
```
print(paste("Moda Consumo: ",mlv(round(Leña_Sur$Consumo,1), method =
"mfv")))

print(paste("Moda Calefactor: ",mlv(round(Leña_Sur$Calefactor,1), met
hod = "mfv")))

print(paste("Moda Humedad: ",mlv(round(Leña_Sur$Humedad,2), method =
"mfv")))
```

```
[1] "Moda Consumo:  9.9"
[1] "Moda Calefactor:  18.9"
[1] "Moda Humedad:  16.49"
```

In [113]:
```
print(paste("Rango Consumo: ",max(Leña_Sur$Consumo) - min(Leña_Sur$Co
nsumo)))

print(paste("Rango Calefactor: ",max(Leña_Sur$Calefactor) - min(Leña_
Sur$Calefactor)))

print(paste("Rango Humedad: ",max(Leña_Sur$Humedad) - min(Leña_Sur$Hu
medad)))
```

```
[1] "Rango Consumo:  34.141"
[1] "Rango Calefactor:  29.8"
[1] "Rango Humedad:  16.48"
```

In [127]:
```
print(paste0("Desviación estandar Consumo: ", round(sd(Leña_Sur$Consu
mo),1 )))

print(paste0("Desviación estandar Calefactor: ", round(sd(Leña_Sur$Ca
lefactor),1 )))

print(paste0("Desviación estandar Humedad: ", round(sd(Leña_Sur$Humed
ad),1 )))
```

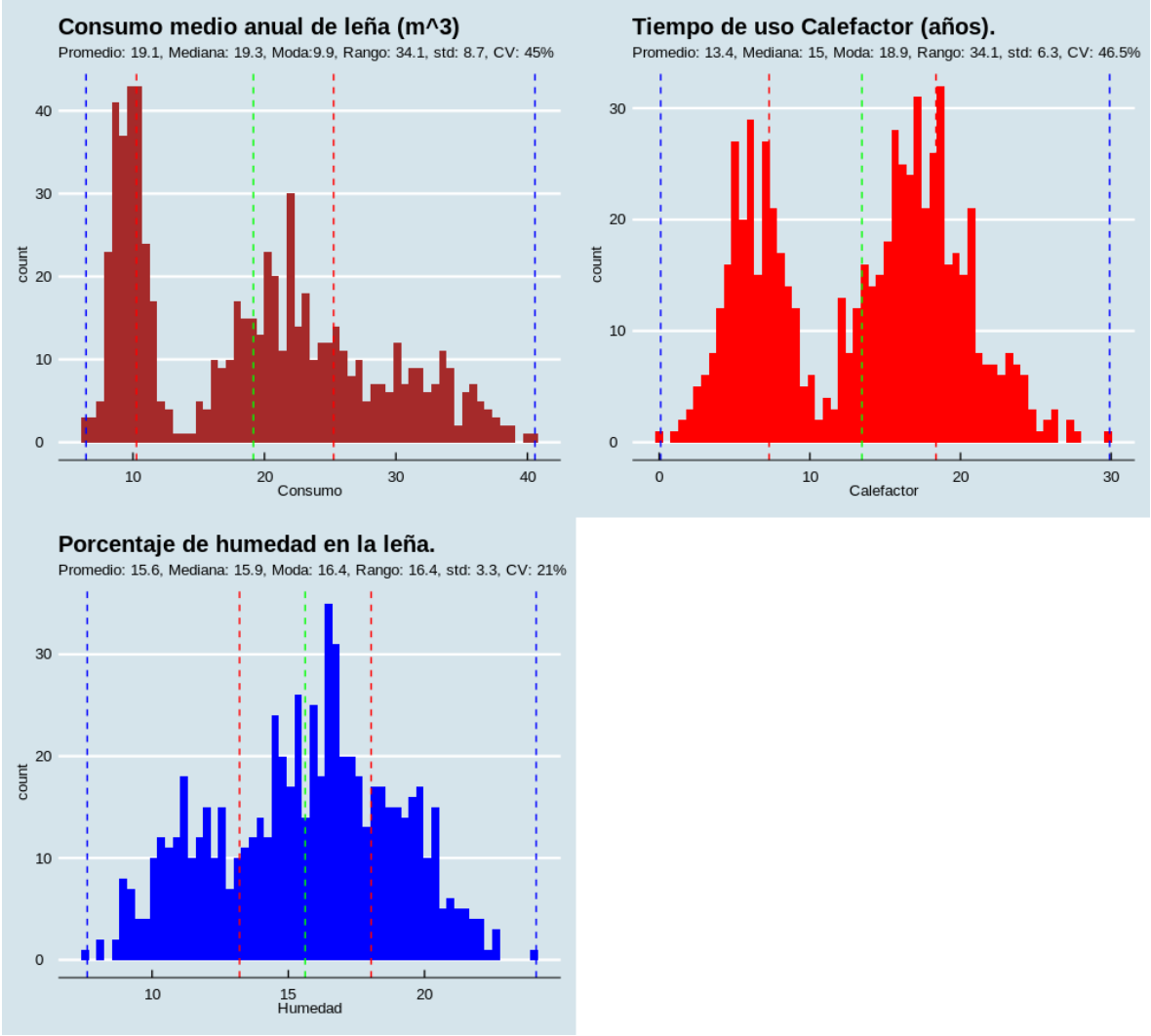
```
[1] "Desviación estandar Consumo: 8.7"
[1] "Desviación estandar Calefactor: 6.3"
[1] "Desviación estandar Humedad: 3.3"
```

In [128]:
```
print(paste0("Coefiente de Variación Consumo: ", round(sd(Leña_Sur$Co
nsumo) / mean(Leña_Sur$Consumo) *100,1),"%" ) )

print(paste0("Coefiente de Variación Calefactor: ", round(sd(Leña_Sur
$Calefactor) / mean(Leña_Sur$Calefactor) *100,1),"%" ) )

print(paste0("Coefiente de Variación Humedad: ", round(sd(Leña_Sur$Hu
medad) / mean(Leña_Sur$Humedad) *100),"%" ),1 )
```

```
[1] "Coefiente de Variación Consumo: 45.7%"
[1] "Coefiente de Variación Calefactor: 46.5%"
[1] "Coefiente de Variación Humedad: 21%"
```

Histogramas, medidas de tendencia central, disperción y quantiles variables númericas.

In [134]:
```
display_png(file="/home/carlos/Documentos/Tarea_chile/2.png")
```



Observando que en consumo de leña y uso del calefactor existe una gran dispersión en los datos de igual manera no se tienen una idea clara sobre algun tipo de distribución.

La humedad tienen menor disperción en los datos, con un buen tratamiento de los ouliers podríamos inferir alguna distribución en los datos.

Relación entre variables y correlación.

Ya que comprendimos las columnas de forma individual, relizare resumenes y estadisticas combinando las variables categoricas y númericas.

Centrandome en las siguientes preguntas:

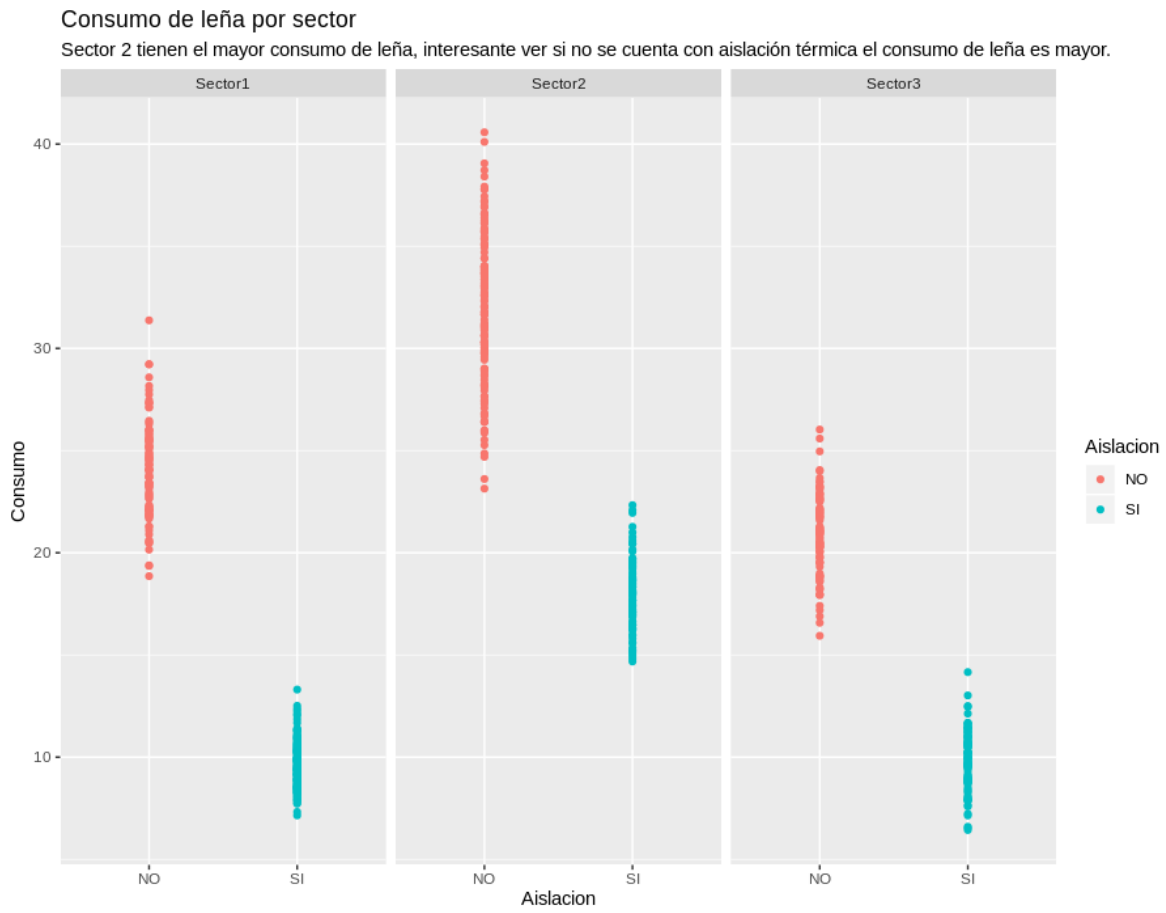
- ¿Consumo de leña, por sector y aislación térmica?
- ¿El tiempo de uso del calefactor afecta el consumo?
- ¿La humedad afecta el consumo?

Todo junto, el consumo de leña es mayor o menor dependiendo de la humedad y uso del calefactor, ayuda el asilamiento termico

- Coeficiente de correlación de Pearson.

¿Consumo de leña por sector y ailación térmica?

In [139]:
```
display_png(file="/home/carlos/Documentos/Tarea_chile/3.png")
```



Matriz de correlación.

In [153]:
```
cor(select(Leña_Sur,-Sector,-Aislacion))
```

A matrix: 3 × 3 of type dbl

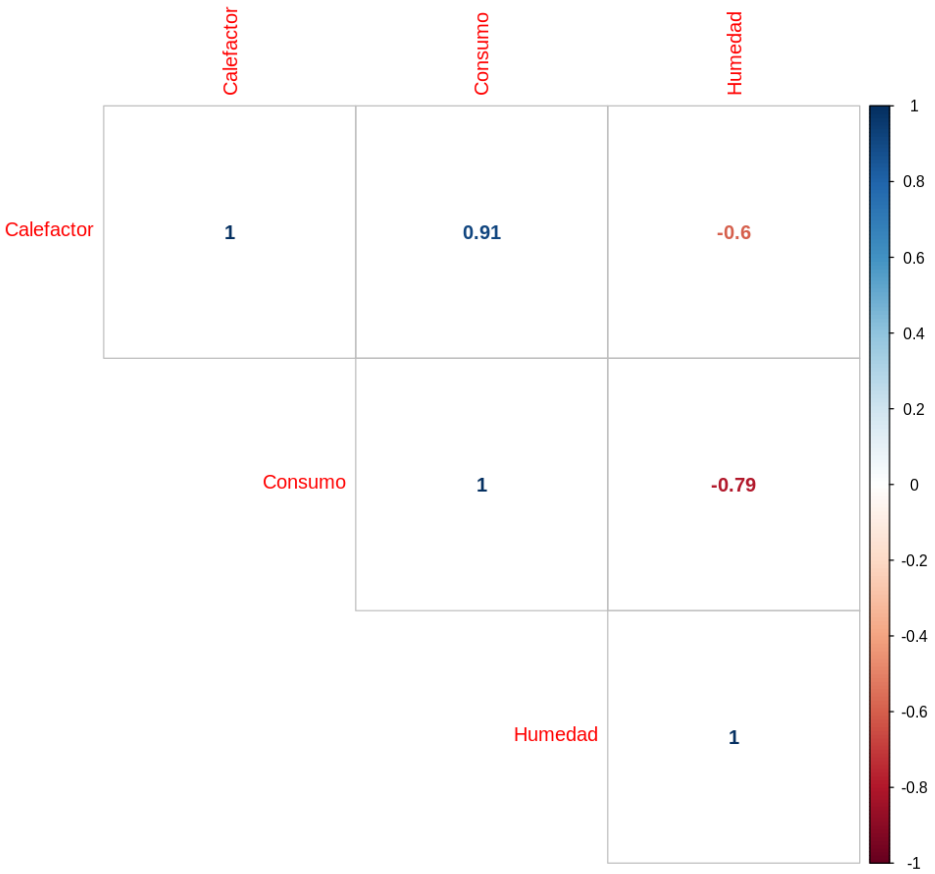
	Calefactor	Consumo	Humedad
Calefactor	1.0000000	0.9140235	-0.5998764
Consumo	0.9140235	1.0000000	-0.7851462
Humedad	-0.5998764	-0.7851462	1.0000000

De momento podemos observar:

1. El consumo y calefactor tienen un coeficiente de correlación = .91, implica correlación positiva.
2. El consumo y la humedad tienen un coeficiente de correlación = -.78 implica correlación negativa.

Veamos los datos más adetalles para poder concluir mejor sobre correlación, segmentaremos por sector y aislación térmica.

In [161]:
```
correlacion<-round(cor(select(Leña_Sur,-Sector,-Aislacion)), 2)

corrplot(correlacion, method="number", type="upper")
```

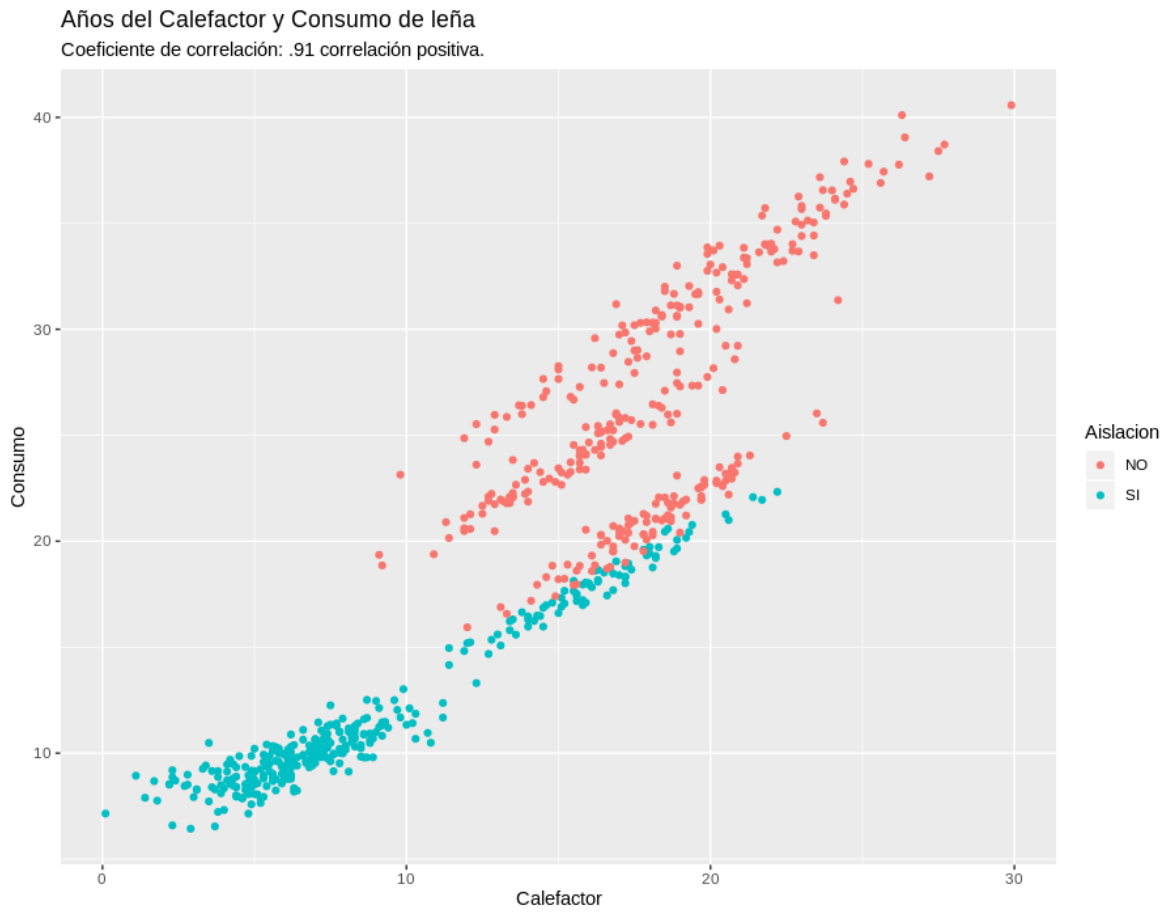


¿El tiempo de uso del calefactor afecta el consumo?

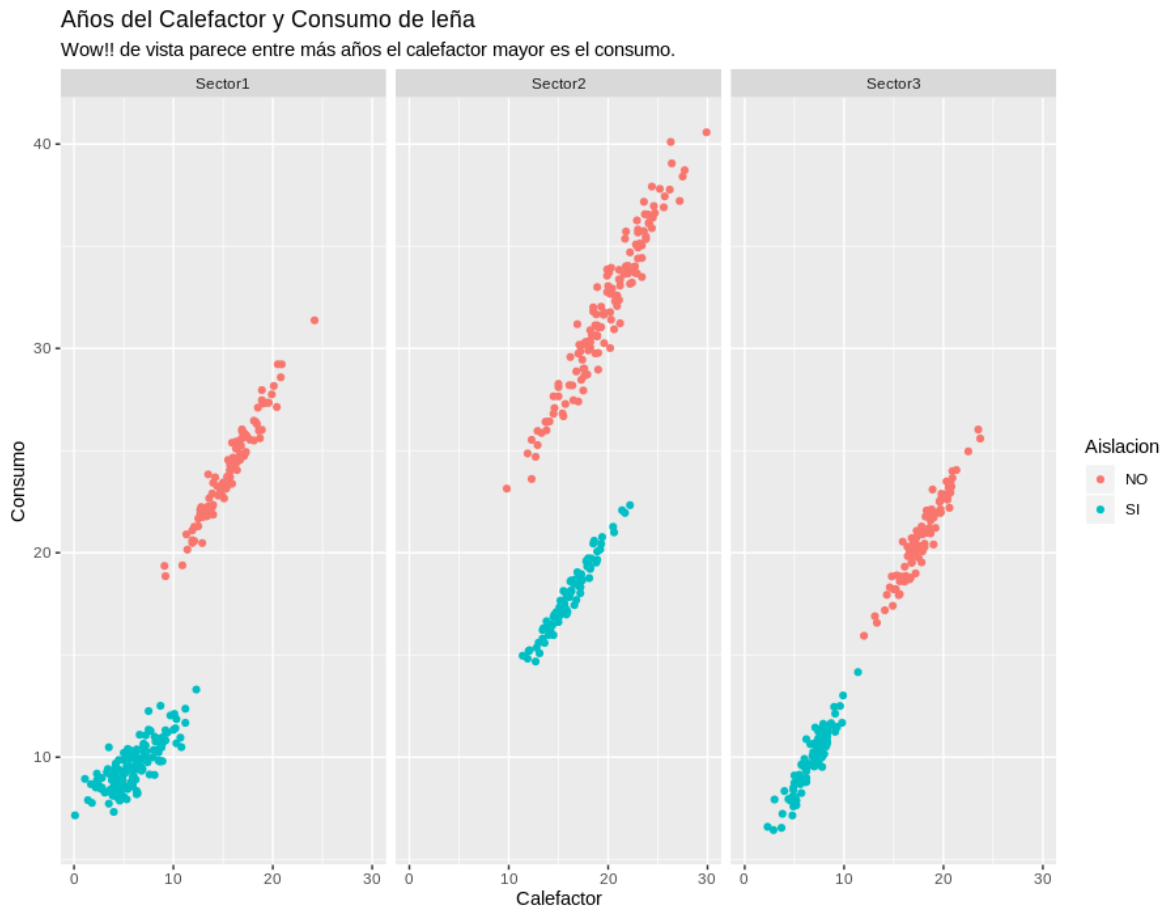
In [162]:
```
cor(Leña_Sur$Calefactor, Leña_Sur$Consumo, method = "pearson")
```

0.914023543668854

In [147]:
```
display_png(file="/home/carlos/Documentos/Tarea_chile/4_0.png")
```



In [140]:
```
display_png(file="/home/carlos/Documentos/Tarea_chile/4.png")
```

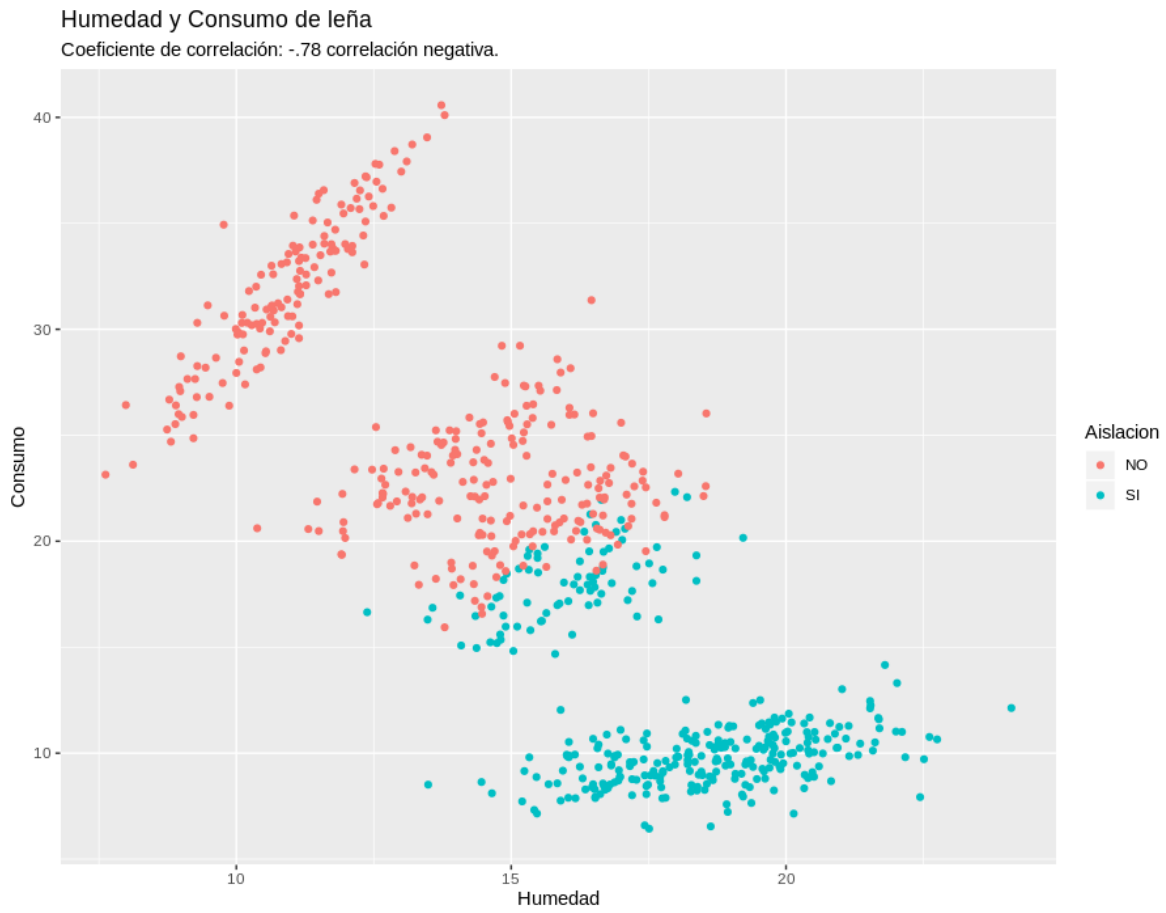


En promedio por sector el coeficiente de correlación se encuentra entre [.85-.93].

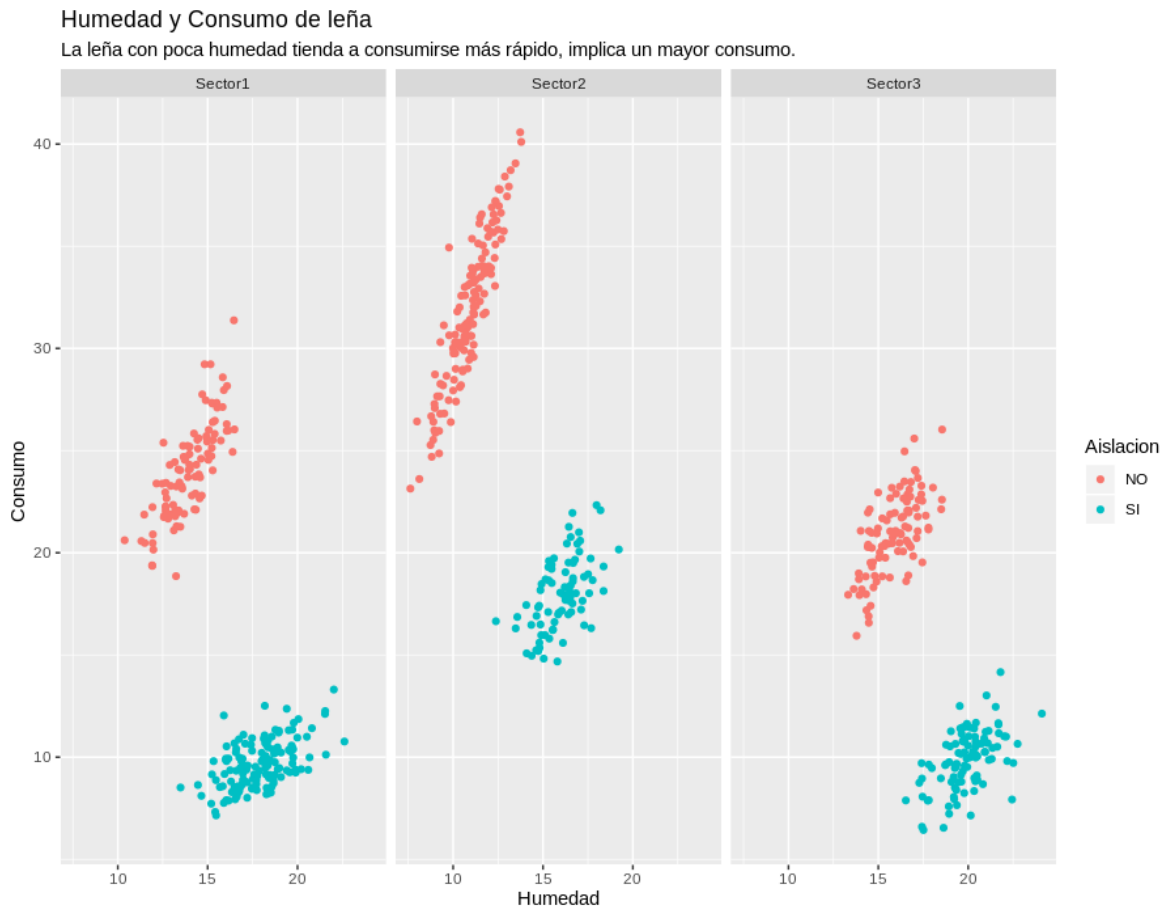
Wow!!, parece que entre menos año tenga nuestro calefactor menor consumo de leña. ¿Que pasa con la humedad?

¿La humedad afecta el consumo?

In [164]:
```
display_png(file="/home/carlos/Documentos/Tarea_chile/5_0.png")
```



In [141]:
```
display_png(file="/home/carlos/Documentos/Tarea_chile/5.png")
```



En promedio por sector el coeficiente de correlación se encuentra entre [-.65, -.82].

Wow!!, parece que la humedad es importante para disminuir el consumo de leña, investiguemos un poco la importancia sobre la humedad en la leña.

Qué sucede cuando la madera está húmeda

Si los troncos de madera que utilizamos para encender el fuego no está bien secos, notaremos que en la puerta de cristal se acumula más suciedad y la leña tarda más de lo normal en consumirse.

En realidad esta no es ninguna ventaja, porque la capacidad de producir calor será mucho más baja de lo normal.

Además, los residuos generados como el hollín o creosota o el humo serán mayores.

Hay que destacar que la leña siempre tiene un grado de humedad, así que no podemos esperar que haya troncos absolutamente secos.

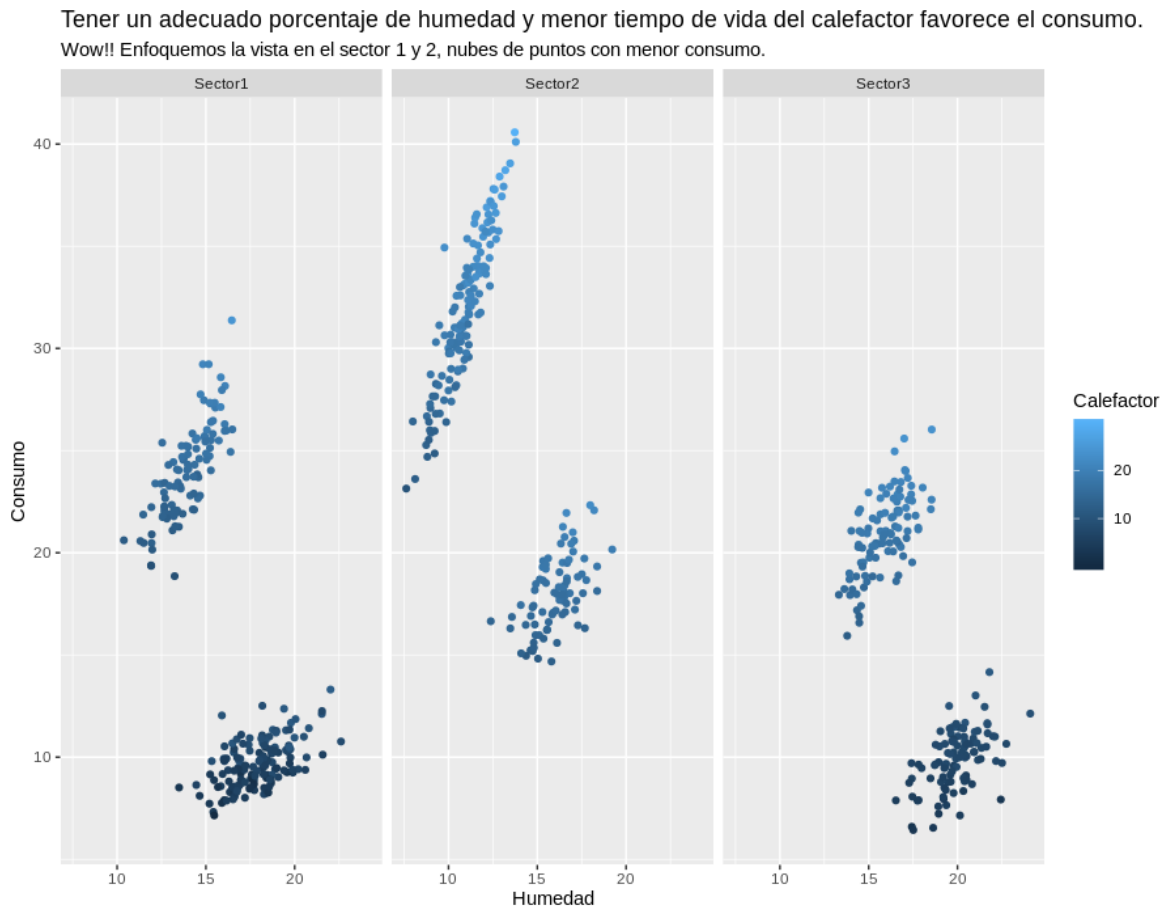
De hecho, si los hubiera, tardarían muy poco en consumirse y no nos servirían.

Cuando miremos qué leña debo usar para la chimenea buscaremos una con aproximadamente un 20 por ciento de humedad.

Ojo!!, el consumo de leña y la humedad no siempre tendrán una correlación negativa, no significa que a mayor humedad menor consumo siempre, ya que se tienen un ideal del porcentaje de humedad.

Todo junto,consumo de leña dependiendo de la humedad, años del calefactor y asilamiento térmico y Sector.

In [146]:
```
display_png(file="/home/carlos/Documentos/Tarea_chile/6.png")
```



El analis descriptivo nos ayudo a conocer más sobre el consumo de leña.

Ojo!! al ser un análisis inicial no podemos concluir de forma definitiva sobre el consumo de la leña apartir del sector, humedad, calefacción y asilamiento térmico.

Resumen del análisis.

1. El consumo de leña tienen una relación positiva con respecto a los años del calefactor.

2 El consumo de leña tienen una relación negativa con respecto a humedad.
  Aunque debemos tener cuidado con esta hipotesis ya que se tiene que buscar un equilibrio en cuanto al porcentaje ideal de humedad

3.Mantener en buen estado el calefactor y la leña ayudan a optimizar el consumo de leña.

4.Los consumos de leña por sector y viviendas con aislación térmica son diferentes.

Conclusiones del analisis:

No podemos concluir con certeza el consumo de leña, pero conocimos mejor nuestros datos muestrales, deberíamos definir adecuadamente la población y la muestra.

En este ejercicio la población fue toda la base con los 3 sectores y las muestras los sectores.

Generamos un mayor número de preguntas las cuales nos obligarían a tomar en cuenta un mayor número de variables como, temperatura, tipo de calefactor, cuidado que se le da a le leña (en almacen, días lluviosos, días soleados, momento en que las personas compran la leña, etc..).

Siguiente pasos, realizar Inferencia estadística.

La inferencia estadística es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra, cual es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad.

Los métodos más utilizados en inferencia estadística son los intervalos de confianza y las pruebas de significancia.

Para hallar la inferencia estadística de nuestro objeto de estudio tendremos que acudir al muestreo probabilístico, que consiste en la elección de una muestra de la población al azar.