Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

Examen

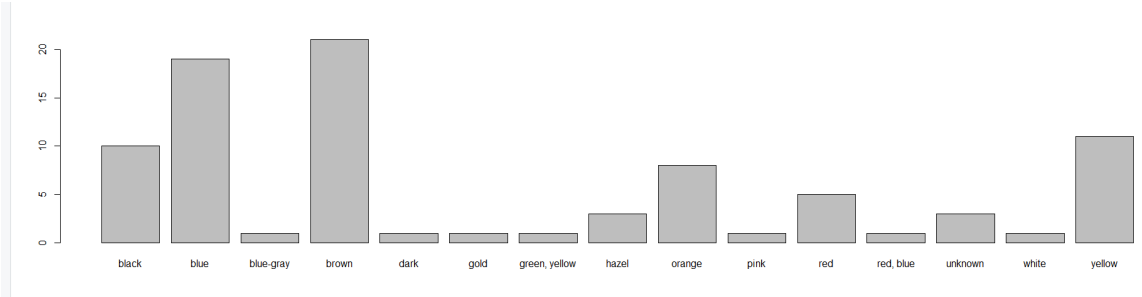
1.

```
mydata = read.csv("starwarsSmall.csv")
str(mydata)
typeof(mydata$height)
typeof(mydata$mass)
typeof(mydata$hair_color)
typeof(mydata$gender)
typeof(mydata$birth_year)
```

```
> typeof(mydata$height)
[1] "integer"

> typeof(mydata$mass)
[1] "double"

> typeof(mydata$hair_color)
[1] "character"

> typeof(mydata$gender)
[1] "character"

> typeof(mydata$birth_year)
[1] "double"
>
```

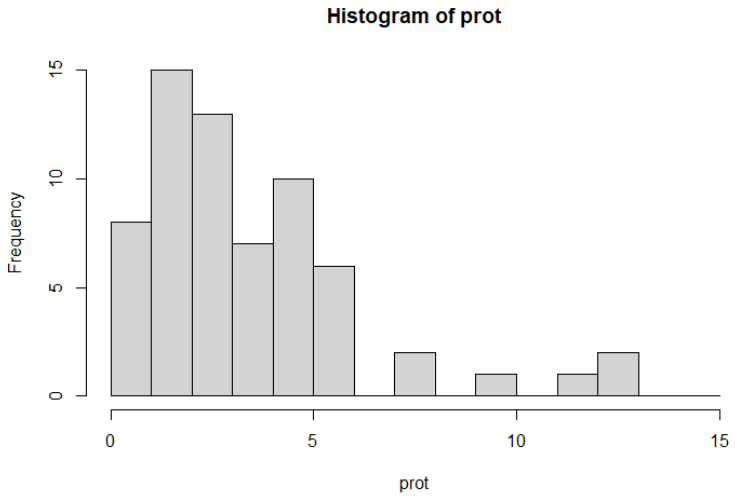
2.

```
x = mydata$eye_color
x_table = table(x)
x_table
barplot(x_table)
min(x_table)
which.min(x_table)
mins = names(x_table)[which(x_table==min(x_table))]
mins
```



3.

Proteins




```
> mode1
[1] "[2,3)"

> mean(prot)
[1] 3.683705

> median(prot)
[1] 3
```

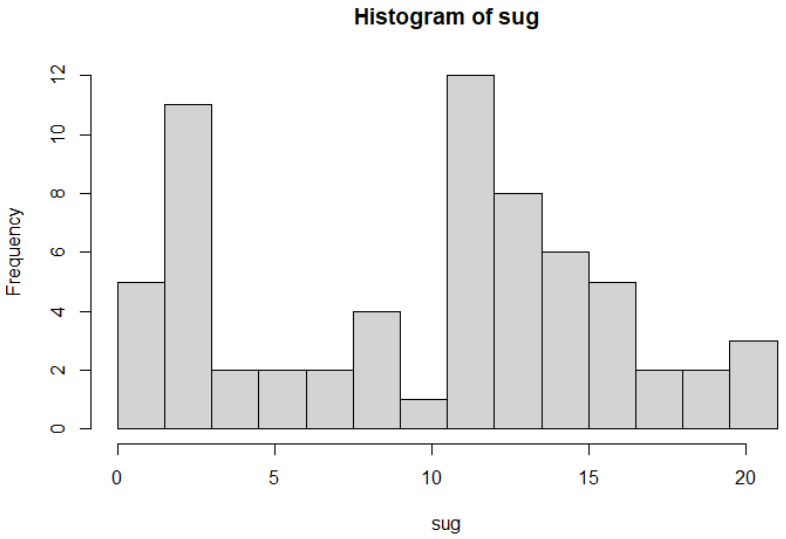
Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

```
pcut = cut(prot, breaks=seq(from=0, to=15, by=1), right=F) # same params as hist()
ptable = table(pcut)
barplot(ptable)
mode1 = names(ptable)[which(ptable==max(ptable))]
ptable
mode1
mean(prot)
median(prot)
```

Sus 3 medidas de tendencia central estan bastante cercanas, aunque el salto entre barra también es pequeño, por lo que no podría decirse que las tres son exactamente iguales. Al ser la moda (2.5) la medida TC más chica, seguida por la mediana y la media, se puede concluir que tiene un skewness positivo, y esto se ve reflejado en el histograma: la mayor concentración de datos se da en la parte izquierda del eje x.

Sugars

```
> mode1
[1] "[12,13.5)"

> mean(sug)
[1] 10.05084

> median(sug)
[1] 12
```

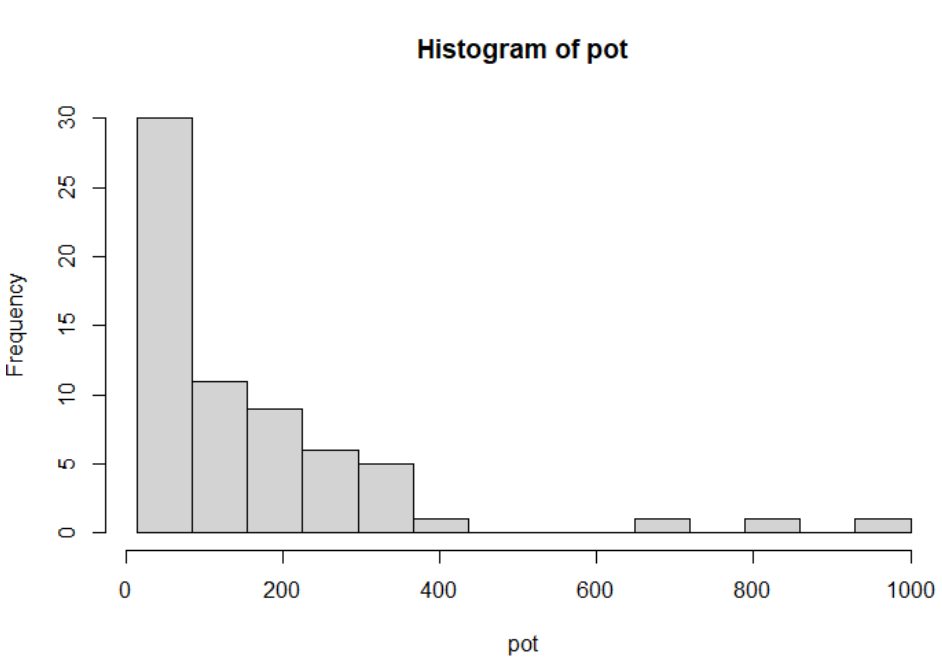



```
range(sug)
hist(sug, breaks=seq(from=0, to=21, length=15))
pcut = cut(sug, breaks=seq(from=0, to=21, length=15), right=F) # same params as hist()
ptable = table(pcut)
barplot(ptable)
mode1 = names(ptable)[which(ptable==max(ptable))]
ptable
mode1
mean(sug)
median(sug)
```

Aquí el tamaño que separa a las barras es de 1.5, por lo que también las medidas TC están cercanas. La medida más pequeña es el promedio (10.05), seguida por la mediana (12) y al final la moda (alrededor de 12.7), por lo que se podría decir que se tiene un skew negativo, y en el histograma se puede observar una ligera tendencia de los datos a la parte derecha. Probablemente lo que arrastra a la media es esa columna de 1.5-3 (segunda barra) que está alejada de los demás datos.

Potassium (pot)

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce




```
> mode1
[1] "[14,84.4)"

> mean(pot)
[1] 159.1197

> median(pot)
[1] 96.59091
```

```
range(pot)
pcut = cut(pot, breaks=seq(from=14, to=1000, length=15), right=F)
ptable = table(pcut)
barplot(ptable)
mode1 = names(ptable)[which(ptable==max(ptable))]
ptable
mode1
mean(pot)
median(pot)
skewness(pot)
```

Ahora la moda es la medida más pequeña (14,84.4), seguida por la mediana y al último la media. Así, se concluye que se tiene un skewness positivo, y además en la gráfica se muestra que la mayor concentración de los datos está en el lado izquierdo del eje x.

4.

```
1  setwd("F:/DocumentsF/UP/1218_AD-2021/Estadistica/Exam1")
2  library(MASS)
3  prot = UScereal$protein
4  stem(prot)
5  sd(prot)
6  var(prot)
```

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

```
> stem(prot)

  The decimal point is at the |

   0 | 89000000333333
   2 | 00000000037777770000048
   4 | 000005555555558
   6 | 000000
   8 | 001
  10 |
  12 | 011


> sd(prot)
[1] 2.642618

> var(prot)
[1] 6.983432
```

```
 8  mydata = read.csv("Vitamin3.csv")
 9  eff = mydata$effort
10  stem(eff)
11  sd(eff)
12  var(eff)
```

```
> stem(eff)

  The decimal point is at the |

   3 | 5
   4 | 0266669
   5 | 011244577789
   6 | 11123444455556677777888888999
   7 | 000122334455555566667777788889999
   8 | 000111111111223333333344455555555556666666667777778888899999999
   9 | 00000011111122222333334444444444555555555666666667777778889999999
  10 | 000001111111122222333333334444444445555666666677777777788888888999999
  11 | 00000011111222222222233333334444445566666778888888888899999999999999
  12 | 000000000011112222222222333333444445555556666666667777788888888999999
  13 | 00000012222222222233334444445555555566666667777778888888888999999
  14 | 000000000011111111111222222222233333334444455555666666666667777788
  15 | 00000000111112222233333444555555666666677778888888899
  16 | 000000222223333344455556777777888899999
  17 | 000011122223444444555566668999
  18 | 011244566788999
  19 | 0122233344455566778
  20 | 1478
  21 | 03
  22 | 0
  23 | 7


> sd(eff)
[1] 3.555027

> var(eff)
[1] 12.63822
```

Protein

Su varianza es 6.98 y su desviación estándar es 2.64. La varianza eleva al cuadrado las diferencias con la media, por lo que es bastante fácil que sea un número elevado, por lo que podemos decir que la varianza de protein está siendo afectada con los atípicos de el

renglón 12, esto concuerda con el stem plot ya que se ve un eje ‘x’ pequeño (0-12) y los datos están agrupados, a excepción de los valores mayores a 12, que se separan un poco. Su desviación estándar es 2.64, y a comparación de la media (3.68) alcanza a ser un tanto elevado su coef de variación, por lo que existe cierta variación grande.

Effort

La varianza es (12.64) y su desviación estándar (3.55), y además su media es (12.19). Su desviación estándar es casi la tercera parte de la media (33%) por lo que es una varianza pequeña, lo cual concuerda con la forma del stem plot, ya que no presenta columnas alejadas como Protein, sino que están condensadas las barras.

5.
Protein (UScereal DB)

```
protM = mean(prot)
protSD = sd(prot)
protVariationCoeff = protSD/protM
mean(prot)
sd(prot)
protVariationCoeff
```

```
> mean(prot)
[1] 3.683705

> sd(prot)
[1] 2.642618

> protVariationCoeff
[1] 0.7173807
```

Se presenta un CV de 72%, lo que se considera como una gran dispersión en los datos, porque quiere decir que la mayoría de los datos puede variar hasta un 72% de la media, es decir, casi la media entera, lo cual es una variación grande.

Effort (Vitamin3 DB)

```
effM = mean(eff)
effSD = sd(eff)
effVariationCoeff = effSD/effM
mean(eff)
sd(eff)
effVariationCoeff
```

```
> mean(eff)
[1] 12.19806

> sd(eff)
[1] 3.555027

> effVariationCoeff
[1] 0.2914421
```

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

Se presenta ahora una variación de 29%, lo que consideramos como una variación un tanto pequeña o moderada, porque esto nos dice que desviación estándar es alrededor de 3 veces más pequeña que la media. Por lo tanto, la mayoría de los datos pueden variar hasta solamente un tercio de la media, lo cual es una variación pequeña en comparación a la anterior.

6.

```
range(eff)
breaks = seq(3, 24, by=1.5)
effort_cut = cut(eff, breaks, right = FALSE)
effortTable = table(effort_cut)
effortCumulativeTable = cumsum(effortTable)
freqRelEffort = effortTable/sum(effortTable)*100
cumulativeTable = cumsum(freqRelEffort)
finalEffortTable = cbind(effortTable, effortCumulativeTable, freqRelEffort, cumulativeTable)
colnames(finalEffortTable) = c("Frequency", "Cumulative freq", "Relative freq", "Cumulative RelFreq")
finalEffortTable
```

```
> finalEffortTable
         Frequency Cumulative freq Relative freq Cumulative RelFreq
[3,4.5)          3               3     0.4166667          0.4166667
[4.5,6)         17              20     2.3611111          2.7777778
[6,7.5)         39              59     5.4166667          8.1944444
[7.5,9)         87             146    12.0833333         20.2777778
[9,10.5)       100             246    13.8888889         34.1666667
[10.5,12)      104             350    14.4444444         48.6111111
[12,13.5)      100             450    13.8888889         62.5000000
[13.5,15)      108             558    15.0000000         77.5000000
[15,16.5)       71             629     9.8611111         87.3611111
[16.5,18)       49             678     6.8055556         94.1666667
[18,19.5)       26             704     3.6111111         97.7777778
[19.5,21)       12             716     1.6666667         99.4444444
[21,22.5)        3             719     0.4166667         99.8611111
[22.5,24)        1             720     0.1388889        100.0000000
```

7.

Quantil 33:

```
[9,10.5)       100             246    13.8888889         34.1666667
```

Q33 = 9 +((0.33*720-146)/100)*1.5 = 10.374

Quantil 67:

```
[13.5,15)      108             558    15.0000000         77.5000000
```

Q67 = 13.5 +((0.67*720-450)/108)*1.5 = 13.95

Calculados en R:

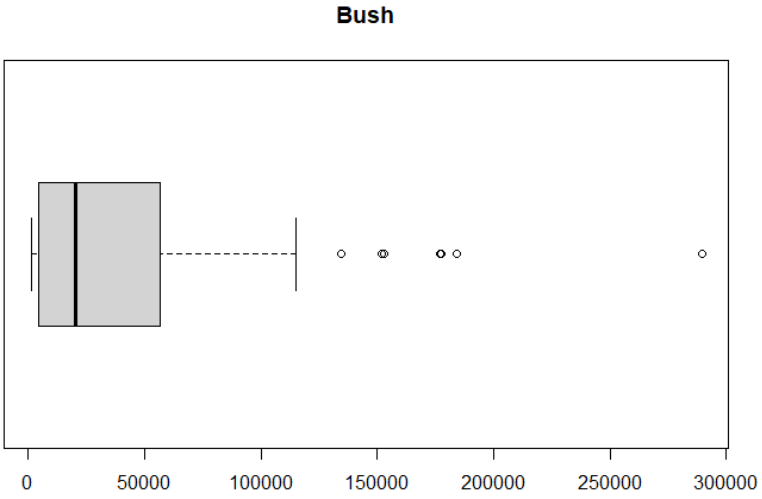
```
> quantile(eff, probs = c(0.33, 0.67))
 33%  67%
10.4 13.9
```

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

8.
Bush

```
mydata = Florida
str(Florida)
bush = mydata$BUSH
gore = mydata$GORE
moore = mydata$MOOREHEAD

IQR(bush)
quantile(bush)
boxplot(bush, horizontal=T, main="Bush")
```

```
> IQR(bush)
[1] 51795

> quantile(bush)
      0%      25%      50%      75%     100%
  1316.0   4746.5  20196.0  56541.5 289456.0

> boxplot(bush, horizontal=T, main="Bush")
```



Explicación:

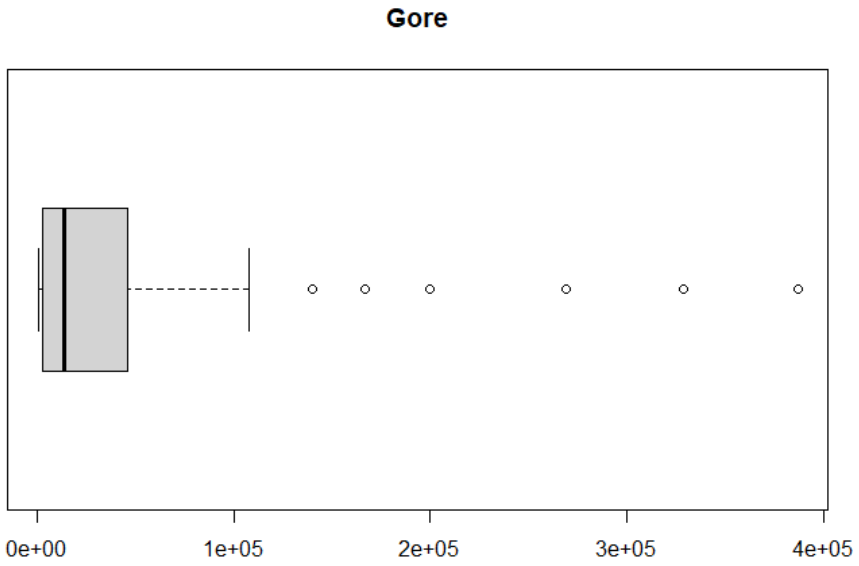
Q1 tiene una diferencia con el whisker inferior de 4000 unidades, Q2 con Q1 tiene una diferencia de 15000 unidades, Q3 con Q2 se diferencian con unos 36000 unidades, por lo que el tamaño de los cuantiles va incrementando conforme se avanza en el eje x, lo cual se ve en el tamaño de la caja y los bigotes, más cargado a la izquierda. La cantidad de outliers es grande, por lo que hay bastantes puntitos después de lo que ya era un largo whisker derecho, lo cual concuerda con la enorme diferencia de Q3 con el 100%, que es más o menos de 230 000 unidades. La mediana o Q2 se encuentra cargada al lado izquierdo de la gráfica por lo que condiciona a todo el boxplot a cargarse a dicho lado.

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

Gore

```
IQR(gore)
quantile(gore)
boxplot(gore, horizontal=T, main="Gore")
```

```
> IQR(gore)
[1] 42919

> quantile(gore)
    0%    25%    50%    75%   100%
   788   3055  14152  45974 386518

> boxplot(gore, horizontal=T, main="Gore")
```



Explicación

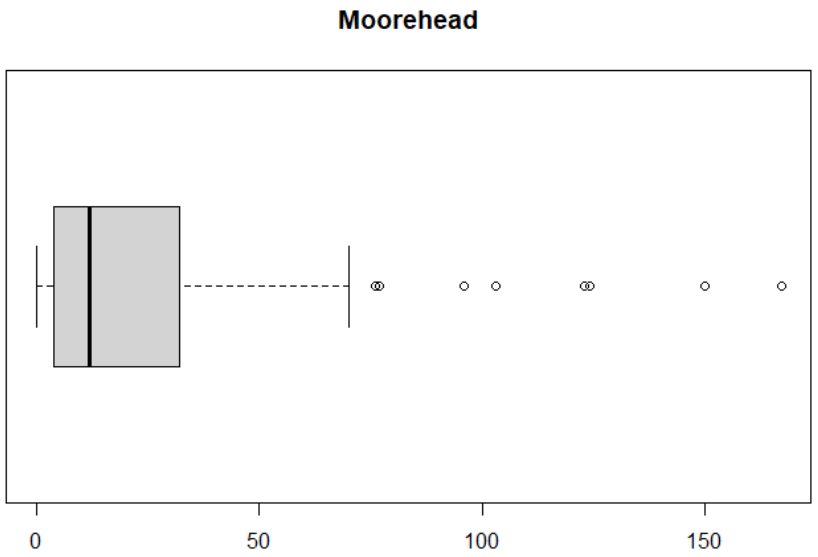
De igual manera, Q1, Q2 y Q3 se van separando incrementalmente en el cálculo de los cuantiles, lo cual se ve también en el tamaño incremental desde el primer whisker, la caja y hasta el whisker derecho. La diferencia entre Q3 y el 100% es de alrededor de 340 000 unidades, por lo que se ve que el whisker derecho y en sí los atípicos (6 atípicos) hacen un boxplot un poco más cargado a la izquierda que el anterior. La mediana está más pegada a q1 que a q3.

Moorehead

```
IQR(moore)
quantile(moore)
boxplot(moore, horizontal=T, main="Moorehead")
```

```
> IQR(moore)
[1] 28

> quantile(moore)
  0%  25%  50%  75% 100%
   0    4   12   32  167

> boxplot(moore, horizontal=T, main="Moorehead")
```

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce



Explicación

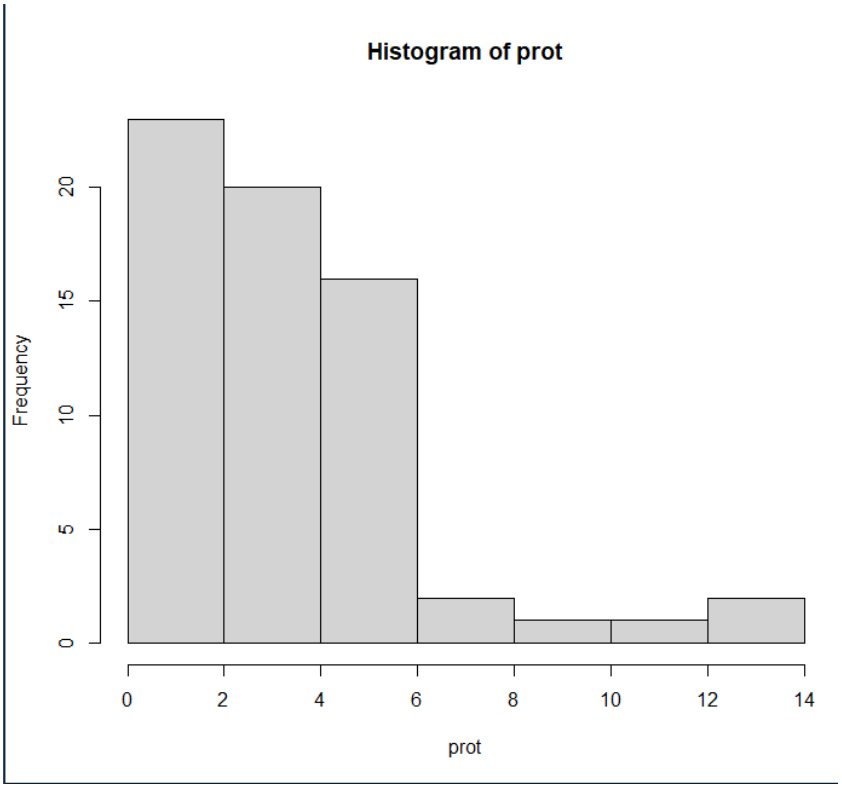
Ahora son alrededor de 8 atípicos (bolitas), y la mediana o q2 (12) está a 8 unidades de q1 y a 20 unidades de q3, por lo que está la caja más cargada a la izquierda que a la derecha. La mediana está en el lado derecho de la caja también. El whisker derecho está alrededor de 70, mientras que el 100% está en 167, de ahí que los atípicos se logran ver después de dicho whisker

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

9.
Protein (UScereal DB)

```
library(e1071)
protSk = skewness(prot)
protKurt = kurtosis(prot, type = 1) + 3
hist(prot)
protSk
protKurt
```

```
> protSk
[1] 1.5677

> protKurt
[1] 5.648528
```



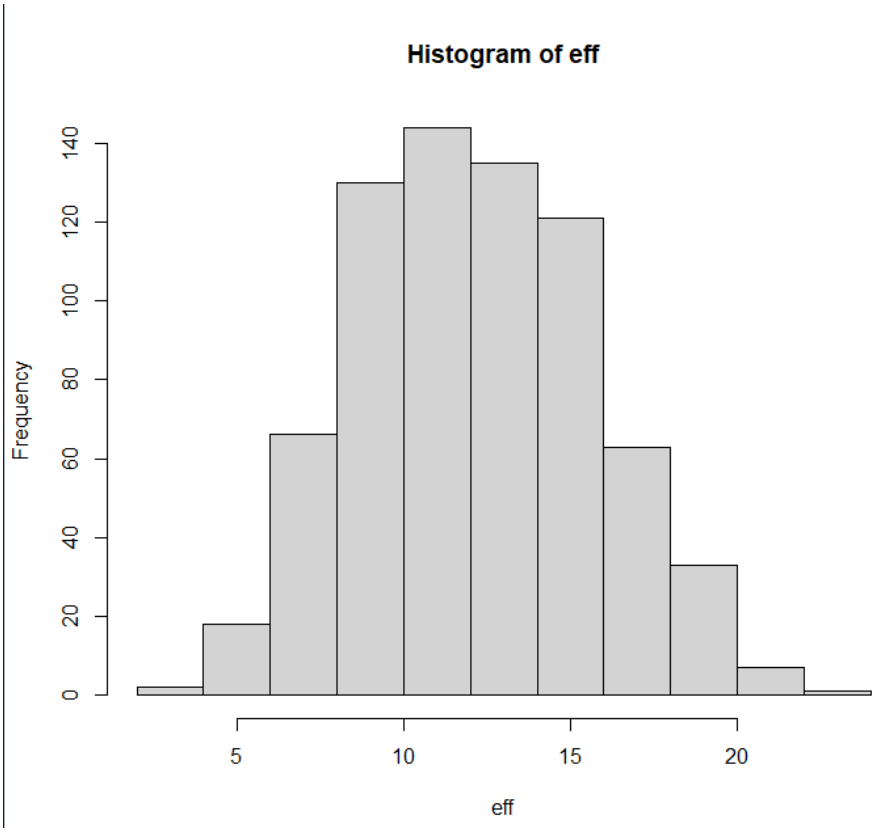
Explicación
El skewness da 1.56 positivo, lo que significa que se tiene Skewness Positivo, por lo que la mayor concentración a los datos está en la parte izquierda del histograma. Esto también anticipa que la kurtosis de 5.64, lo cual es mayor a 3. Al tener kurtosis mayor a 3, significa una forma leptokúrtica, o con tails largas. En este caso, por la skewness positiva tenemos

sólo la cola derecha, lo cual hace que la cola se contabilice como 'larga' en el cálculo numérico de la kurtosis.

Effort (Vitamin3 DB)

```
effortSk = skewness(eff)
effortKurt = kurtosis(eff, type = 1) + 3
hist(eff)
effortSk
effortKurt
```

```
> effortSk
[1] 0.1781856

> effortKurt
[1] 2.57875
```



Explicación

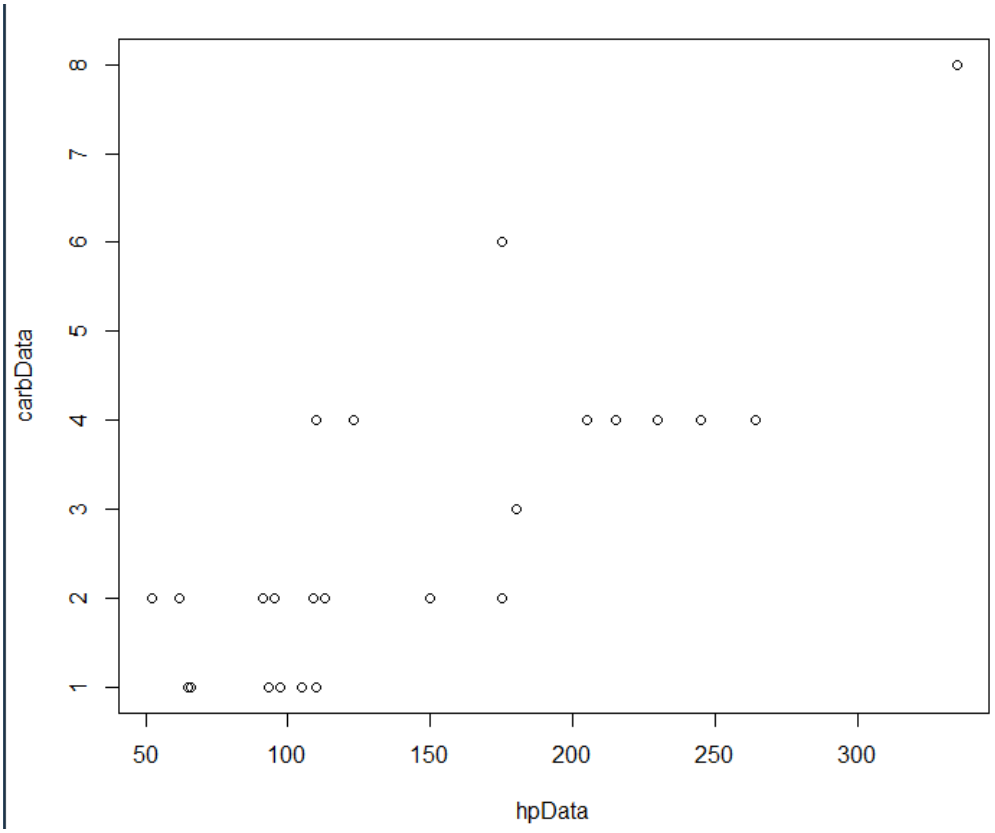
La skewness es casi 0, con un valor de 0.17, por lo que la distribución es simétrica y muy parecida a la normal, lo cual se ve perfectamente en el histograma. Su kurtosis es 2.57, lo cual es bastante cercano a 3, aunque si se es riguroso, es menor que 3, por lo que esta distribución es platicúrtica aunque casi mesokúrtica. Así, sus colas son cortas pero cercanas a la medida de las colas de distribución normal, como el histograma muestra.

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce

10.

HP vs. Carb

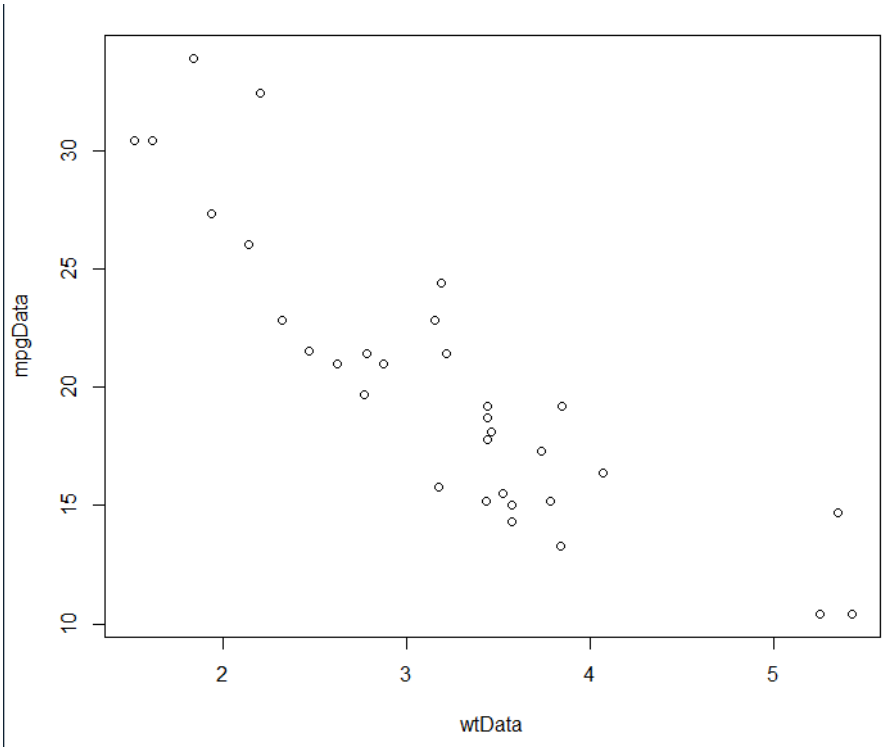



```
> hpCarbCov
[1] 83.03629

> hpCarbCor
[1] 0.7498125
```

Estas variables presentan una covarianza positiva bastante alta, por lo que cuando una de ellas incrementa su la otra se comporta de la misma manera: se ve una leve línea recta que va hacia arriba en el scatter.

Wt vs. Mpg

Daniel Heráclito Pérez Díaz
Mariana Ávalos Arce




```
> wtMpgCov
[1] -5.116685

> wtMpgCor
[1] -0.8676594
```

En el scatter plot se muestra una línea hacia abajo, y se confirma con el valor de correlación de -0.86, lo cual es una fuerte asociación negativa: cuando wtData aumenta, mgData disminuye, y por lo mismo está relación opuesta se visualiza como una línea recta hacia abajo en el scatter plot.