Análisis Exploratorio de datos utilizando R

11/5/2022

- A01245418 Andrés Sarellano Acevedo
- A00829837 Axel Amós Hernández Cárdenas
- A01281371 Izel María Ávila Rodríguez
- A01383422 Melissa Elvia Salazar Carrillo
- A00573134 Macías Romero Jorge Humberto

1. Cargando y mostrando el dataframe

```
setwd('C:/users/hplax/Desktop/ITESM/CuartoSemestre')
```

```
df2 <- wines
df2$Cultivar <- NULL
df2
```

```
## # A tibble: 178 x 13
##    Alcohol Malic.acid   Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoids
##      <dbl>      <dbl> <dbl>             <dbl>     <dbl>         <dbl>      <dbl>
##  1    14.2       1.71  2.43              15.6       127          2.8        3.06
##  2    13.2       1.78  2.14              11.2       100          2.65       2.76
##  3    13.2       2.36  2.67              18.6       101          2.8        3.24
##  4    14.4       1.95  2.5               16.8       113          3.85       3.49
##  5    13.2       2.59  2.87              21         118          2.8        2.69
##  6    14.2       1.76  2.45              15.2       112          3.27       3.39
##  7    14.4       1.87  2.45              14.6        96          2.5        2.52
##  8    14.1       2.15  2.61              17.6       121          2.6        2.51
##  9    14.8       1.64  2.17              14          97          2.8        2.98
## 10    13.9       1.35  2.27              16          98          2.98       3.15
## # ... with 168 more rows, and 6 more variables: Nonflavanoid.phenols <dbl>,
## #   Proanthocyanin <dbl>, Color.intensity <dbl>, Hue <dbl>,
## #   OD280.OD315.of.diluted.wines <dbl>, Proline <dbl>
```

2. Para todas las variables cuantitativas obtén las siguientes métricas: media, mediana, rango, y desviación estándar.

Media

A continuación se muestra la media de cada variable cuantitativa del dataframe.

```
apply(df2,2,mean)
```

```
##                     Alcohol                  Malic.acid
##                  13.0006180                   2.3363483
##                         Ash            Alcalinity.of.ash
##                   2.3665169                  19.4949438
##                   Magnesium                Total.phenols
##                  99.7415730                   2.2951124
##                  Flavanoids         Nonflavanoid.phenols
##                   2.0292697                   0.3618539
##               Proanthocyanin              Color.intensity
##                   1.5908989                   5.0580899
##                         Hue OD280.OD315.of.diluted.wines
##                   0.9574494                   2.6116854
##                     Proline
##                 746.8932584
```

```
#mean(df2$Alcohol)
```

Mediana

A continuación se muestra la mediana de cada variable cuantitativa del dataframe.

```
apply(df2,2,median)
```

```
##                     Alcohol                  Malic.acid
##                      13.050                       1.865
##                         Ash            Alcalinity.of.ash
##                       2.360                      19.500
##                   Magnesium                Total.phenols
##                      98.000                       2.355
##                  Flavanoids         Nonflavanoid.phenols
##                       2.135                       0.340
##               Proanthocyanin              Color.intensity
##                       1.555                       4.690
##                         Hue OD280.OD315.of.diluted.wines
##                       0.965                       2.780
##                     Proline
##                     673.500
```

```
#mean(df2$Alcohol)
```

Rango

A continuación se muestra el rango de cada variable cuantitativa del dataframe.

```
apply(df2,2,range)
```

```
##      Alcohol Malic.acid  Ash Alcalinity.of.ash Magnesium Total.phenols
## [1,]   11.03       0.74 1.36              10.6        70          0.98
## [2,]   14.83       5.80 3.23              30.0       162          3.88
##      Flavanoids Nonflavanoid.phenols Proanthocyanin Color.intensity  Hue
```

```
## [1,]       0.34                 0.13           0.41            1.28 0.48
## [2,]       5.08                 0.66           3.58           13.00 1.71
##      OD280.OD315.of.diluted.wines Proline
## [1,]                         1.27     278
## [2,]                         4.00    1680
```

```
#range(df2$Alcohol)
```

Desviaciones Estandar

A continuación se muestran las desviaciones estándar de cada variable cuantitativa del dataframe.

```
apply(df2,2,sd)
```

```
##                     Alcohol                  Malic.acid
##                   0.8118265                   1.1171461
##                         Ash           Alcalinity.of.ash
##                   0.2743440                   3.3395638
##                   Magnesium               Total.phenols
##                  14.2824835                   0.6258510
##                  Flavanoids        Nonflavanoid.phenols
##                   0.9988587                   0.1244533
##              Proanthocyanin             Color.intensity
##                   0.5723589                   2.3182859
##                         Hue OD280.OD315.of.diluted.wines
##                   0.2285716                   0.7099904
##                     Proline
##                 314.9074743
```

3. Para variables cualitativas, genera gráficas de barras o de pastel para mostrar la distribución de frecuencias. ¿Cuál es el valor que más se repite? ¿Cuántas veces se repite? ¿Qué porcentaje de veces se repite?

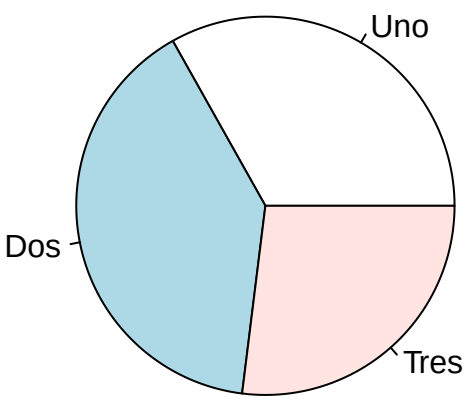
Table de frecuencias

```
table(wines$Cultivar)
```

```
##
##  1  2  3
## 59 71 48
```

Grafica de Pastel

```
pie(table(wines$Cultivar), labels = c("Uno","Dos","Tres"), main="Gráfica de Pastel de Frecuencias")
```



Porcentajes

```
percent1 <- 59/178 * 100
percent1
```

```
## [1] 33.14607
```

```
percent2 <- 71/178 * 100
percent2
```

```
## [1] 39.88764
```

```
percent3 <- 48/178 * 100
percent3
```

```
## [1] 26.96629
```

Se indica por la taba de frecuencias que el valor que más se repite en la columna del dato cualitativo es el 2. Esto mismo se puede apreciar en la gráfica de pastel creada a partir de la tabla de frecuencias.

Por otra parte, los porcentajes de cada frecuencia son:

- Frecuencia del 1: 33.14%
- Frecuencia del 2: 39.88%
- Frecuencia del 3: 26.96%

4. Para variables cuantitativas genera un diagrama de dispersión comparándolas (una en cada eje) y agrega una línea de regresión. ¿Cuál es la correlación entre las variables? ¿Es positiva o negativa? ¿Los datos se ajustan bien a la línea de regresión?

Total fenoles y flavonoides

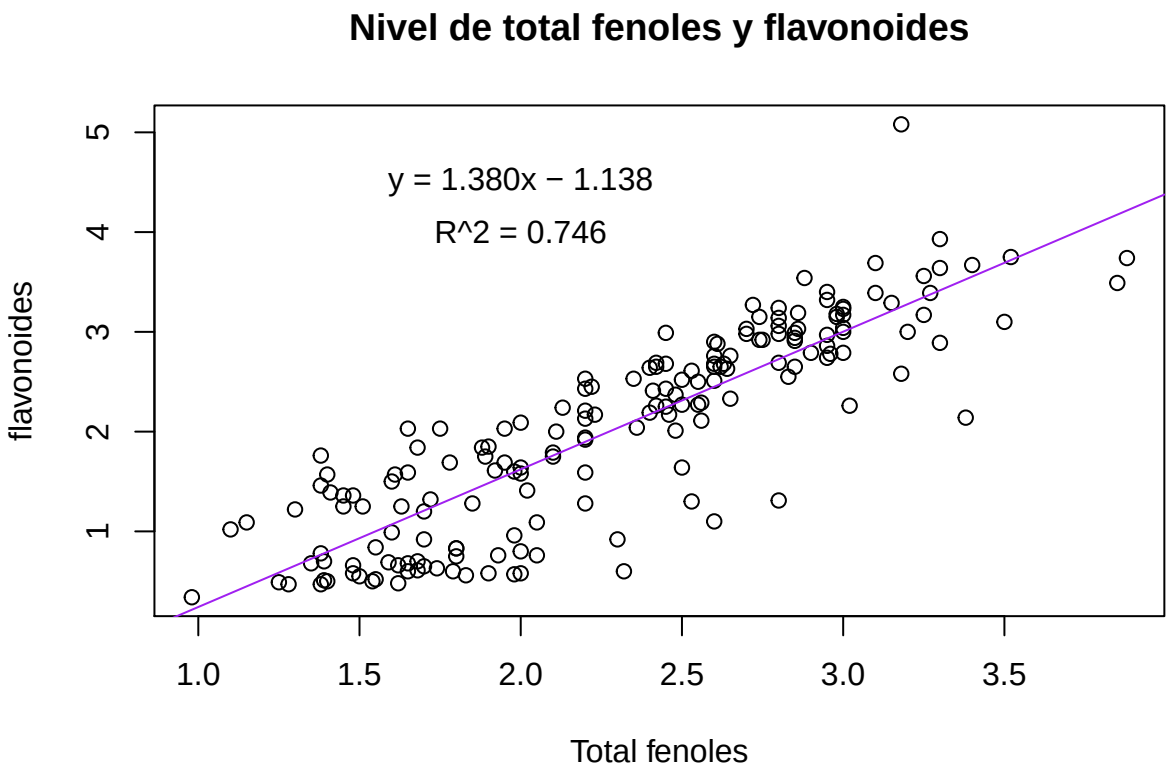
```
# Obteniendo los valores de X y B para la linea de regresión
summary(lm(Flavanoids~Total.phenols, wines))
```

```
##
## Call:
## lm(formula = Flavanoids ~ Total.phenols, data = wines)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46361 -0.28305  0.05922  0.37011  1.82972
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.13763    0.14379  -7.912 2.71e-13 ***
## Total.phenols  1.37984    0.06046  22.824  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5034 on 176 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.746
## F-statistic: 520.9 on 1 and 176 DF,  p-value: < 2.2e-16
```

```
plot(x=wines$Total.phenols , y=wines$Flavanoids, main="Nivel de total fenoles y flavonoides", ylab="fla

abline(a=-1.13763, b = 1.37984, col = "purple")

text(2, 4.5, "y = 1.380x - 1.138")
text(2,4, "R^2 = 0.746")
```




```
#r^2 = 0.746
```

Se puede observar que el par de "Total de Fenoles" y "Flavonoides" cuentan con un valor de su R^2 = 0.746, por lo que se podría concluir que para este par, la correlacion es buena y positiva, permitiendo que los datos se ajusten bien a la linea de regresión creada con abline.

Alcohol y Total fenoles

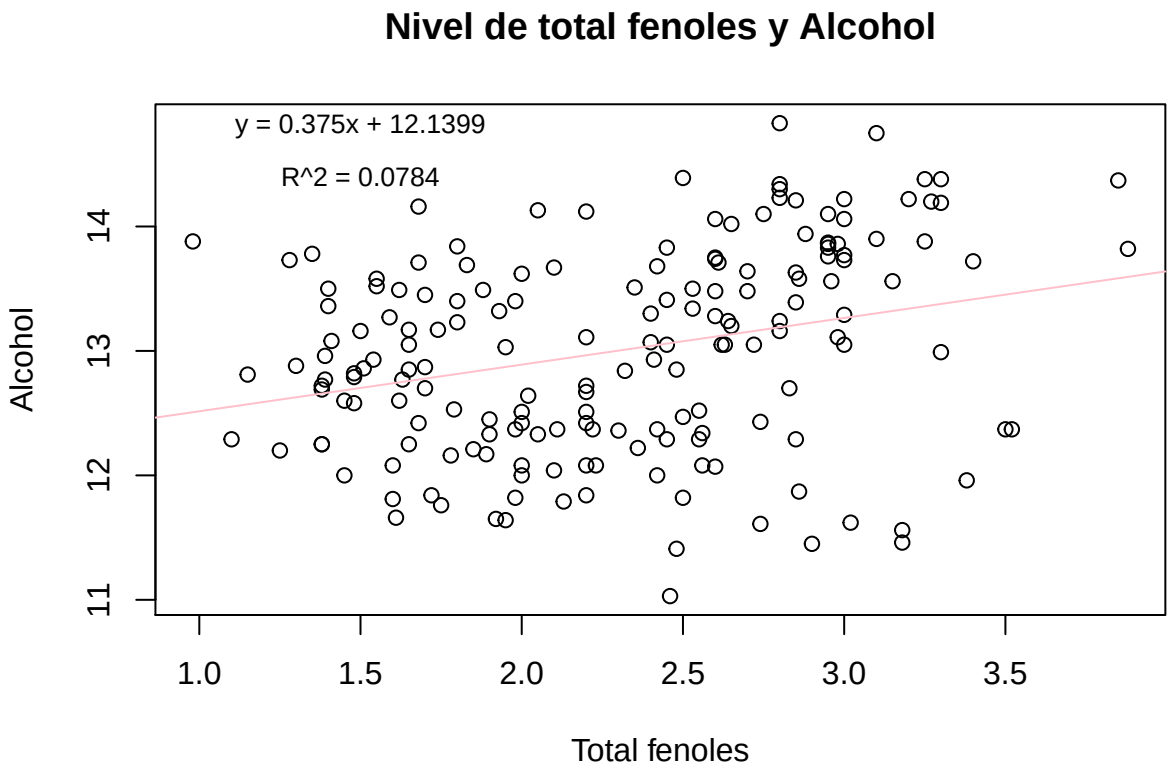
```
summary(lm(Alcohol~Total.phenols, df1))
```

```
##
## Call:
## lm(formula = Alcohol ~ Total.phenols, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03245 -0.57807  0.09381  0.60129  1.64004
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.1399     0.2226  54.530  < 2e-16 ***
## Total.phenols   0.3750     0.0936   4.006 9.08e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7794 on 176 degrees of freedom
## Multiple R-squared:  0.08358,    Adjusted R-squared:  0.07837
## F-statistic: 16.05 on 1 and 176 DF,  p-value: 9.085e-05
```

```
plot(x=wines$Total.phenols , y=df1$Alcohol, main="Nivel de total fenoles y Alcohol", ylab="Alcohol", xla
abline(a=12.1399, b = 0.3750 ,col = "pink")

text(1.5, 14.8, "y = 0.375x + 12.1399", cex = 0.8)
text(1.5,14.4, "R^2 = 0.0784", cex=0.8)
```




```
# r^2 = 0.07837
```

Se puede observar que el par de "Total de Fenoles" y "Alcohol" cuentan con un valor de su R^2 = 0.0784, por lo que se podría concluir que para este par, la correlacion es debil y positiva, por lo que los datos no se ajustarán completamente a la linea de regresion creada con abline.

Alcohol y Acido Málico

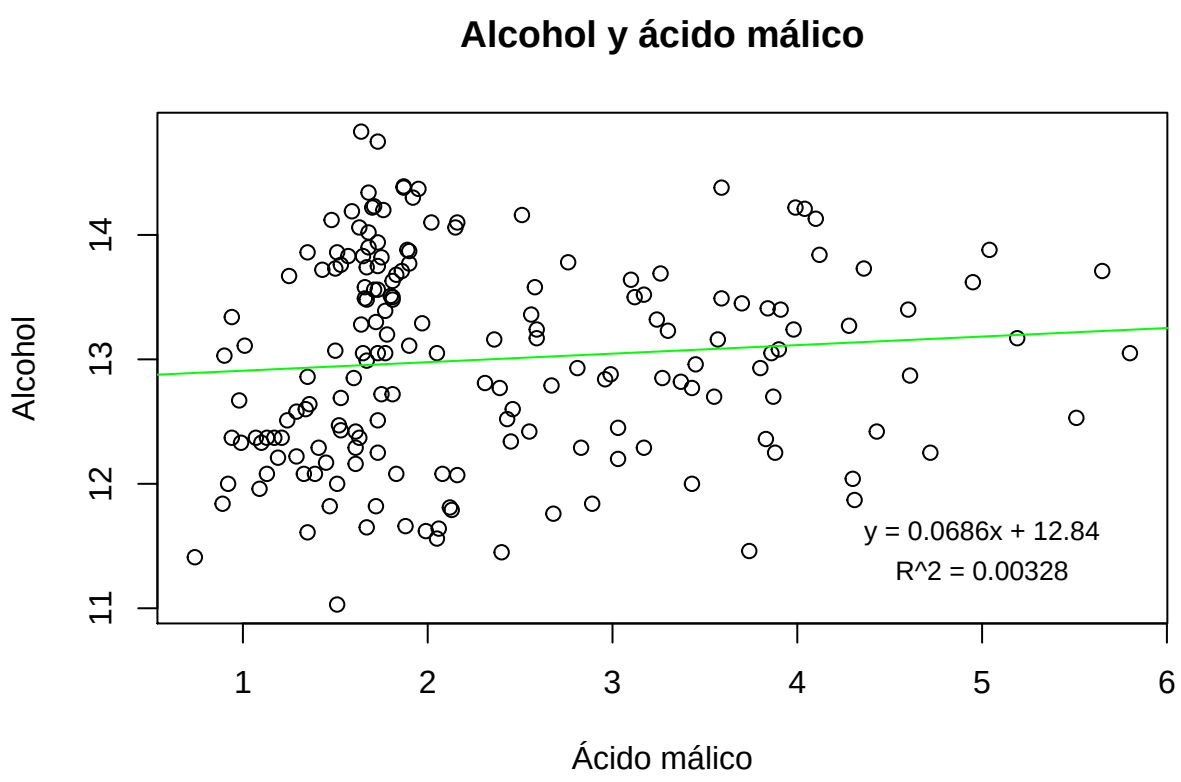
```
summary(lm(Alcohol~Malic.acid, df1))
```

```
##
```

```
## Call:
## lm(formula = Alcohol ~ Malic.acid, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91393 -0.63485  0.00436  0.62596  1.87715
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.84035    0.14115  90.971   <2e-16 ***
## Malic.acid   0.06860    0.05453   1.258     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8105 on 176 degrees of freedom
## Multiple R-squared:  0.008911,   Adjusted R-squared:  0.00328
## F-statistic: 1.582 on 1 and 176 DF,  p-value: 0.2101
```

```
plot(x=df1$Malic.acid , y=df1$Alcohol, main="Alcohol y ácido málico", ylab="Alcohol", xlab = "Ácido máli
abline(a=12.84035  ,b = 0.06860 ,col = "green")

text(5, 11.6, "y = 0.0686x + 12.84", cex = 0.8)
text(5,11.3, "R^2 = 0.00328", cex=0.8)
```



```
# r^2 = 0.00328
```

Se puede observar que el par de "Alcohol" y "Acido Málico" cuentan con un valor de su R^2 = 0.00328, por lo que se podría concluir que para este par, la correlacion es casi nula, pero positiva, por lo que los datos no se ajustan a la linea de regresion creada con abline.

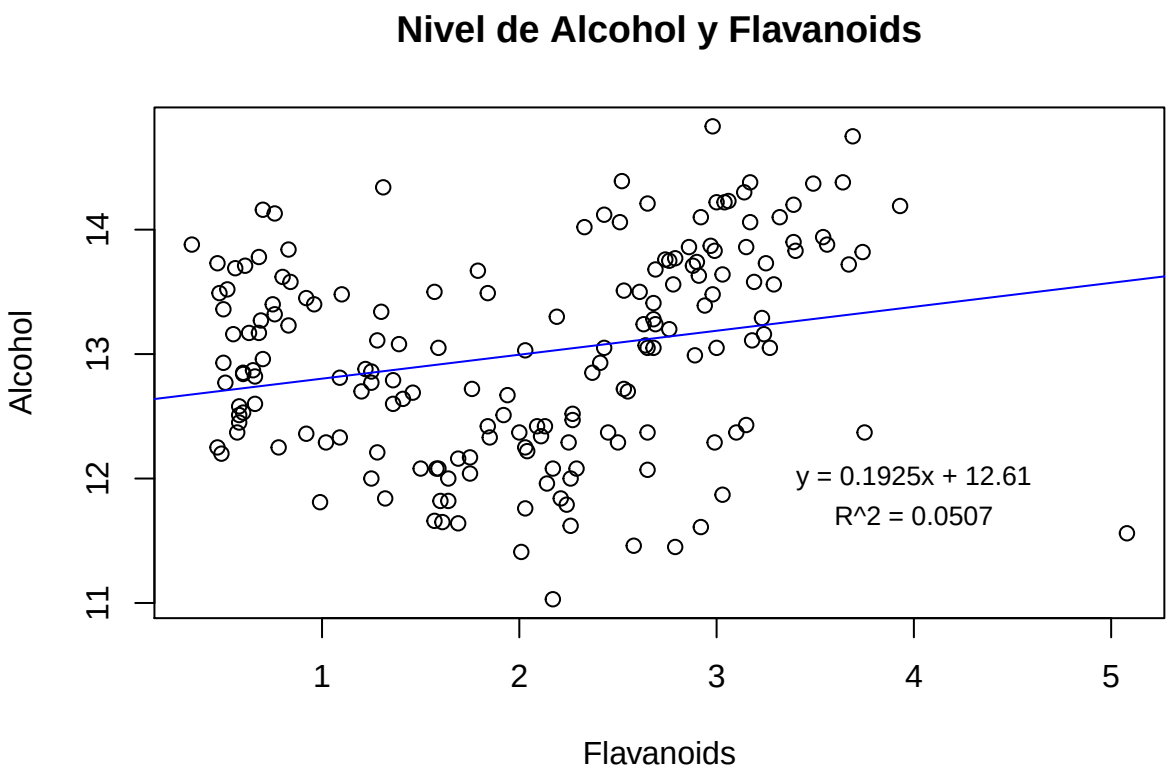
Alcohol y Flavanoids

```
summary(lm(Alcohol~Flavanoids, df1))
```

```
##
## Call:
## lm(formula = Alcohol ~ Flavanoids, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02780 -0.61874  0.05851  0.61912  1.64639
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.61004    0.13455  93.720  < 2e-16 ***
## Flavanoids   0.19247    0.05952   3.234  0.00146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.791 on 176 degrees of freedom
## Multiple R-squared:  0.05608,    Adjusted R-squared:  0.05072
## F-statistic: 10.46 on 1 and 176 DF,  p-value: 0.001459
```

```
plot(x=df1$Flavanoids , y=df1$Alcohol, main="Nivel de Alcohol y Flavanoids", ylab="Alcohol", xlab = "Fla
abline(a=12.61004  ,b = 0.19247 ,col = "blue")

text(4, 12, "y = 0.1925x + 12.61", cex = 0.8)
text(4,11.7, "R^2 = 0.0507", cex=0.8)
```




```
# r^2 = 0.05072
```

Se puede observar que el par de "Alcohol" y "Flavonoides" cuentan con un valor de su R^2 = 0.0507, por lo que se podría concluir que para este par, la correlacion es igual que la anterior: casi nula pero positiva, por lo que los datos no se ajustarán a la linea de regresion creada con abline.

Alcohol y Alcalinity.of.ash

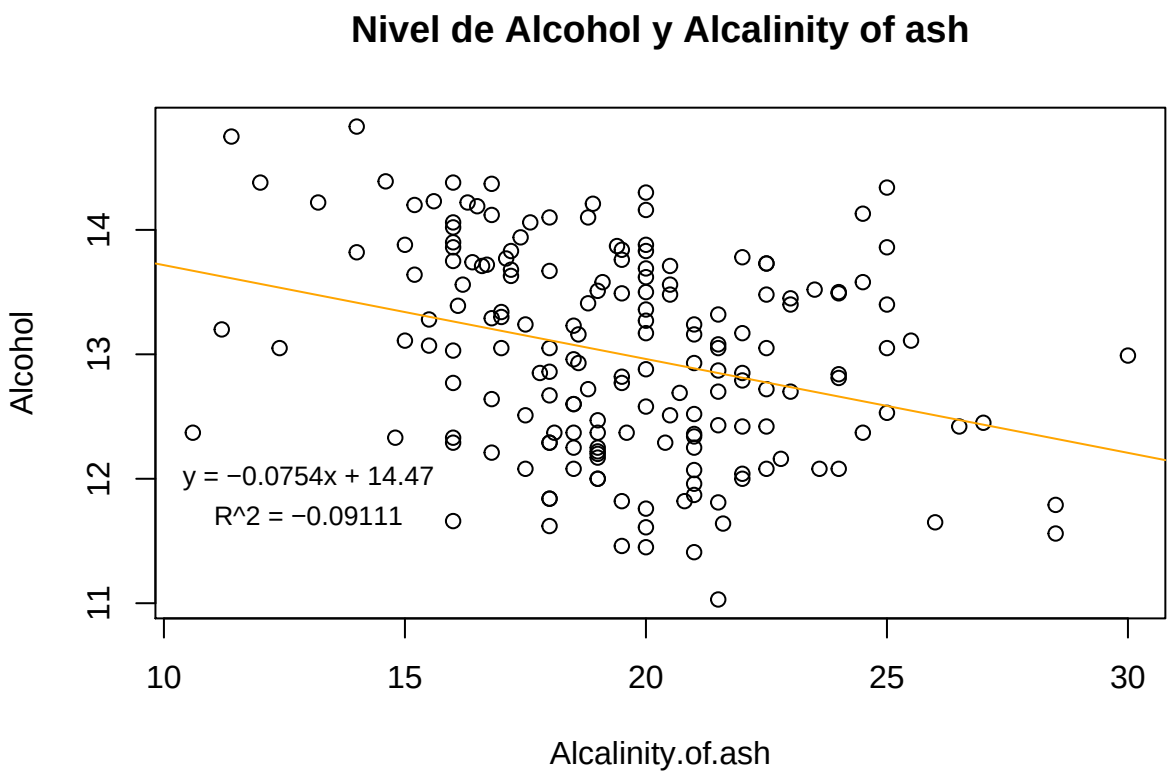
```
summary(lm(Alcohol~Alcalinity.of.ash, df1))
```

```
##
## Call:
## lm(formula = Alcohol ~ Alcalinity.of.ash, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81940 -0.61978  0.02945  0.65118  1.75455
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       14.47085    0.34452  42.003  < 2e-16 ***
## Alcalinity.of.ash -0.07542    0.01742  -4.329 2.51e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.774 on 176 degrees of freedom
## Multiple R-squared:  0.09625,    Adjusted R-squared:  0.09111
## F-statistic: 18.74 on 1 and 176 DF,  p-value: 2.505e-05
```

```
plot(x=df1$Alcalinity.of.ash , y=df1$Alcohol, main="Nivel de Alcohol y Alcalinity of ash", ylab="Alcohol
abline(a=14.47085  ,b = -0.07542 ,col = "orange")


text(13, 12, "y = -0.0754x + 14.47", cex = 0.8)
text(13,11.7, "R^2 = -0.09111", cex=0.8)
```




```
# r^2 = 0.09111
```

Se puede observar que el par de "Alcohol" y "Alcalinity of Ash" cuentan con un valor de su R^2 = -0.0911, por lo que se podría concluir que para este par, la correlacion casi nula pero negativa, por lo que los datos no se ajustarán a la linea de regresion creada con abline.

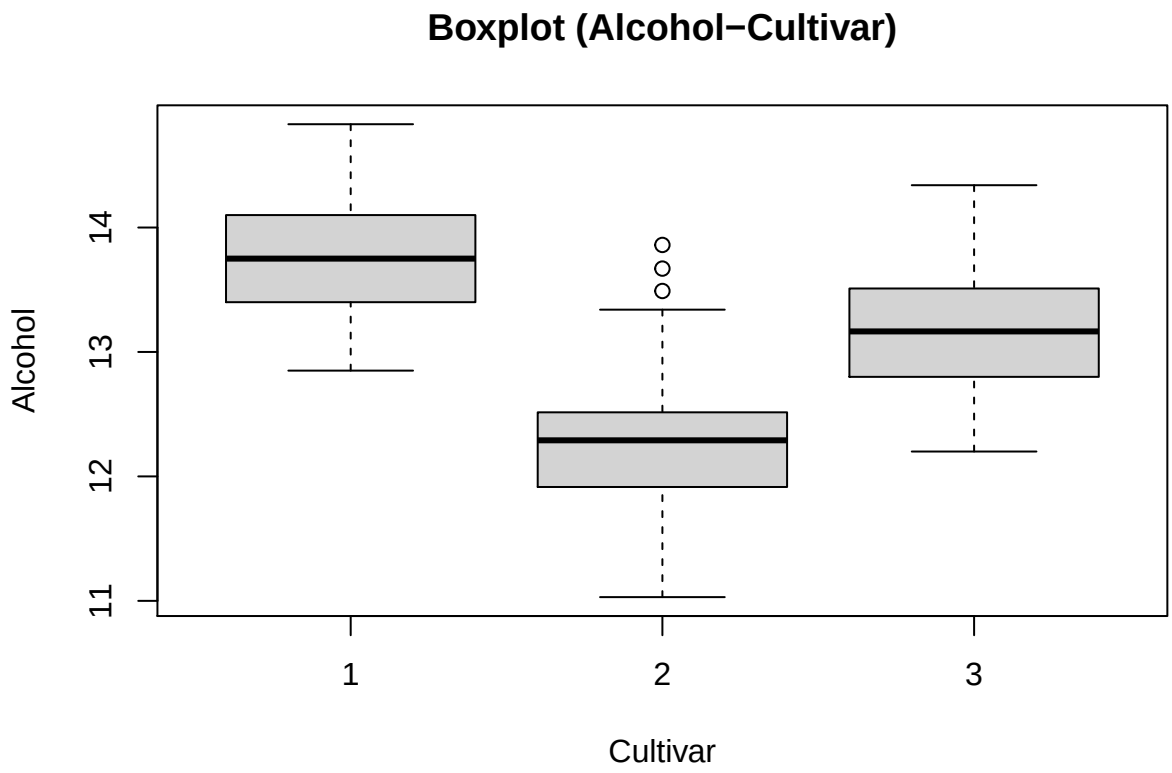
5. Para un par de variable cuantitativa-cualitativa, y elabora un boxplot. Explica qué significan las partes del boxplot (son 5 partes). ¿Qué información me dan los boxplots? ¿Las distribuciones de valores en el eje "y" son los mismos para todas las categorías?

El boxplot está compuesto por 5 partes:

- Mínimo: representa el valor mínimo de los datos.
- Primer Cuartil (Q1): el 25% de los valores son menores o igual a este valor.
- Mediana: divide en dos partes iguales la distribución; de forma que el 50% de los valores son menores o igual a este punto.
- Tercer Cuartil (Q3): el 75% de los valores son menores o igual a este valor.
- Máximo: representa el valor máximo de los datos.

Los boxplots son una representación visual que describe varias características importantes al mismo tiempo. Nos permite identificar valores atípicos y comparar distribuciones.(Montes, 2019) (Estadistica para Todos, 2018)

```
list.values1 <- unlist(wines[1:59,2])
list.values2 <- unlist(wines[60:130,2])
list.values3 <- unlist(wines[131:178,2])
boxplot(x=list(list.values1,list.values2,list.values3),names=c("1","2","3"),  main = "Boxplot (Alcohol-C
```



Para este boxplot, se puede observar que la distribución de valores en el eje "y" no son los mismos ya que cada cultivo tiene diferentes valores, los cual los hacen tener un máximo y un mínimo diferentes.

6. Escribe tus conclusiones. ¿Cuál es la información más importante que te dan los datos?

Consideramos que la información mas importante que muestran los datos para este set, sería todo lo relacionado con los modelos de regresion lineal y boxplot. Si es importante considerar aspectos de la media,

mediana, moda, frecuencia, etc, pero se cuentan con herramientas mas completas para el análisis de un conjunto de datos lo cual nos puede ayudar a generar predicciones y determinar posibles futuras tendencias. En conclusión, las herramientas y analisis de regresion lineal o boxplot, ofrecer un análisis poderoso sobre los datos y son muy útiles para cualquier trabajo que requiera análisis de datos.

Referencias

Estadistica Para Todos. (2008). Diagrama de Caja y Bigotes. Titapg. https://www.estadisticaparatodos.es/taller/graficas/cajas.html

Montes, D. (2019). Diagrama Box Plot. PCG. https://www.pgconocimiento.com/diagrama-boxplot/