# Proyecto personal análisis de datos

Eduardo Sánchez

2022-05-10

## Análisis Exploratorio de Datos de Metaloproteasas

### Importación de datos

A partir del set de datos *MetalloproteasesDB*, se crea el objeto denominado *Datos_Metaloproteasas*, especificando que para los valores faltantes se emplea la nomenclatura N/A, y se observan los datos contenidos en este.

```
Datos_Metaloproteasas <- read_excel("MetalloproteasesDB.xlsx", na = "N/A")

summary(Datos_Metaloproteasas)
```

```
##       ID               Acc_No             Family
##  Length:36          Length:36          Length:36
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Keratinolytic_activity Activity_type          Length      Active_site
##  Length:36              Length:36          Min.   :244.0   Length:36
##  Class :character       Class :character   1st Qu.:373.0   Class :character
##  Mode  :character       Mode  :character   Median :482.0   Mode  :character
##                                            Mean   :503.6
##                                            3rd Qu.:633.0
##                                            Max.   :993.0
##  Metal_binding_site   Mol_Weight           pI         Negative_AA
##  Length:36          Min.   : 25681   Min.   :4.600   Min.   : 16.00
##  Class :character   1st Qu.: 40533   1st Qu.:5.615   1st Qu.: 40.00
##  Mode  :character   Median : 52012   Median :6.005   Median : 49.50
##                     Mean   : 55456   Mean   :6.316   Mean   : 58.75
##                     3rd Qu.: 69377   3rd Qu.:6.845   3rd Qu.: 67.25
##                     Max.   :115429   Max.   :8.750   Max.   :126.00
##   Positive_AA     Extinction_coefficient Instability_index Aliphatic_index
##  Min.   : 19.00   Min.   : 11585         Min.   :17.68     Min.   : 60.11
##  1st Qu.: 34.75   1st Qu.: 38463         1st Qu.:26.44     1st Qu.: 69.42
##  Median : 44.50   Median : 70320         Median :29.07     Median : 75.84
##  Mean   : 51.31   Mean   : 71210         Mean   :30.68     Mean   : 78.70
##  3rd Qu.: 63.25   3rd Qu.:101136         3rd Qu.:34.48     3rd Qu.: 84.89
##  Max.   :119.00   Max.   :185765         Max.   :47.69     Max.   :105.66
##      GRAVY
```

```
##  Min.   :-0.7570
##  1st Qu.:-0.4710
##  Median :-0.3200
##  Mean   :-0.3010
##  3rd Qu.:-0.1705
##  Max.   : 0.1670
```

```
str(Datos_Metaloproteasas)
```

```
## tibble [36 x 16] (S3: tbl_df/tbl/data.frame)
##  $ ID                    : chr [1:36] "M0301KL" "M0302NK" "M0401KL" "M0402KL" ...
##  $ Acc_No                : chr [1:36] "AJD23200.1" "XP_533954.3" "ADP00718.1" "AJD77429.1" ...
##  $ Family                : chr [1:36] "M3" "M3" "M4" "M4" ...
##  $ Keratinolytic_activity: chr [1:36] "Keratinolytic" "Non_keratinolytic" "Keratinolytic" "Keratinoly
##  $ Activity_type         : chr [1:36] "oligo" "null" "endo" "endo" ...
##  $ Length                : num [1:36] 783 687 475 546 566 528 801 814 422 426 ...
##  $ Active_site           : chr [1:36] "EHYDGYY" "EHYDQYY" "E" "E" ...
##  $ Metal_binding_site    : chr [1:36] "HHE" "HHE" "HHE" "HHE" ...
##  $ Mol_Weight            : num [1:36] 88709 78301 51295 59667 60898 ...
##  $ pI                    : num [1:36] 6.3 5.72 6.15 5.36 5.92 4.6 4.79 5.78 7.7 5.66 ...
##  $ Negative_AA           : num [1:36] 110 97 49 59 64 82 126 113 40 47 ...
##  $ Positive_AA           : num [1:36] 101 83 44 45 58 44 83 100 41 37 ...
##  $ Extinction_coefficient: num [1:36] 92835 79605 65000 102110 75640 ...
##  $ Instability_index     : num [1:36] 41.5 47.7 26.6 30.6 27.8 ...
##  $ Aliphatic_index       : num [1:36] 79.6 84.6 71.1 74.7 70.7 ...
##  $ GRAVY                 : num [1:36] -0.492 -0.41 -0.346 -0.361 -0.479 -0.58 -0.757 -0.594 -0.331 -
```

### Transformación de datos

Del objeto *Datos_Metaloproteasas*, se transforman las variables *ID*, *Acc_No*, *Family*, *Keratinolytic_activity*, *Activity_type*, *Active_site* y *Metal_binding_site* a factores. Posteriormente, se explora nuevamente el set de datos para verificar.

```
Datos_Metaloproteasas$ID <- as.factor(Datos_Metaloproteasas$ID)

Datos_Metaloproteasas$Acc_No <- as.factor(Datos_Metaloproteasas$Acc_No)

Datos_Metaloproteasas$Family <- as.factor(Datos_Metaloproteasas$Family)

Datos_Metaloproteasas$Keratinolytic_activity <- as.factor(Datos_Metaloproteasas$Keratinolytic_activity)

Datos_Metaloproteasas$Activity_type <- as.factor(Datos_Metaloproteasas$Activity_type)

Datos_Metaloproteasas$Active_site <- as.factor(Datos_Metaloproteasas$Active_site)

Datos_Metaloproteasas$Metal_binding_site <- as.factor(Datos_Metaloproteasas$Metal_binding_site)

summary(Datos_Metaloproteasas)
```

```
##        ID            Acc_No       Family         Keratinolytic_activity
##  M0301KL: 1   AAQ21097.1: 1   M28    :10   Keratinolytic    :18
##  M0302NK: 1   AAS76669.1: 1   M36    : 6   Non_keratinolytic:18
```

```
##  M0401KL: 1   AAS76670.1: 1   M4     : 4
##  M0402KL: 1   ABG67896.1: 1   M14    : 2
##  M0403NK: 1   ABK17661.1: 1   M15    : 2
##  M0404NK: 1   ADP00718.1: 1   M16    : 2
##  (Other):30   (Other)   :30   (Other):10
##  Activity_type     Length       Active_site Metal_binding_site   Mol_Weight
##  endo  : 7     Min.   :244.0   E      :10   HHE    :15         Min.   : 25681
##  exo   : 9     1st Qu.:373.0   MCYCAIL: 3   HDEEDH : 6         1st Qu.: 40533
##  null  :18     Median :482.0   EH     : 2   HDEDH  : 4         Median : 52012
##  oligo : 1     Mean   :503.6   EHFY   : 2   HDH    : 2         Mean   : 55456
##  unspec: 1     3rd Qu.:633.0   LCYCAFS: 2   HEH    : 2         3rd Qu.: 69377
##                Max.   :993.0   (Other):16   (Other): 6         Max.   :115429
##                                NA's   : 1   NA's   : 1
##        pI         Negative_AA      Positive_AA     Extinction_coefficient
##  Min.   :4.600   Min.   : 16.00   Min.   : 19.00   Min.   : 11585
##  1st Qu.:5.615   1st Qu.: 40.00   1st Qu.: 34.75   1st Qu.: 38463
##  Median :6.005   Median : 49.50   Median : 44.50   Median : 70320
##  Mean   :6.316   Mean   : 58.75   Mean   : 51.31   Mean   : 71210
##  3rd Qu.:6.845   3rd Qu.: 67.25   3rd Qu.: 63.25   3rd Qu.:101136
##  Max.   :8.750   Max.   :126.00   Max.   :119.00   Max.   :185765
##
##  Instability_index Aliphatic_index      GRAVY
##  Min.   :17.68     Min.   : 60.11   Min.   :-0.7570
##  1st Qu.:26.44     1st Qu.: 69.42   1st Qu.:-0.4710
##  Median :29.07     Median : 75.84   Median :-0.3200
##  Mean   :30.68     Mean   : 78.70   Mean   :-0.3010
##  3rd Qu.:34.48     3rd Qu.: 84.89   3rd Qu.:-0.1705
##  Max.   :47.69     Max.   :105.66   Max.   : 0.1670
##
```

A partir de esta transformación y su posterior verificación, logra identificarse que hay un dato faltante tanto para *Active_site* como para *Metal_binding_site*.

### Normalización de datos

Con el fin de minimizar la influencia de la longitud de la secuencia aminoacídica (*Lenght*) en las variables *Negative_AA* y *Positive_AA*, se llevó a cabo la normalización de dichos parámetros mediante el cálculo de las variables derivadas *Negativity* y *Positivity* expresadas en proporciones, y se creo el nuevo objeto *Datos_Metaloproteasas_Norm*, que inclute dichas variables.

```
Datos_Metaloproteasas_Norm <- Datos_Metaloproteasas %>% mutate(Negativity = Negative_AA/Length)%>%
  mutate(Positivity = Positive_AA/Length)
```

### Balance de datos

Con el fin de identificar si los datos están balanceados o no entre tratamientos, se crearon tablas de frecuencia para estos, las cuales se muestran a continuación:

```
Tabla_queratinolisis <- table(Datos_Metaloproteasas$Keratinolytic_activity)
knitr::kable(Tabla_queratinolisis, col.names = c ("Queratinolisis", "Frecuencia"), caption = "Tabla 1. I
```

Table 1: Tabla 1. Frecuencias de enzimas segun el factor de actividad queratinolítica

| Queratinolisis | Frecuencia |
|---|---|
| Keratinolytic | 18 |
| Non_keratinolytic | 18 |

```
Tabla_Familias <- table(Datos_Metaloproteasas$Keratinolytic_activity, Datos_Metaloproteasas$Family)
knitr::kable(Tabla_Familias, caption = "Tabla 2. Frecuencias de enzimas por familia de proteasa de acue
```

Table 2: Tabla 2. Frecuencias de enzimas por familia de proteasa de acuerdo con clasificación MEROPS

| | M14 | M15 | M16 | M28 | M3 | M32 | M36 | M38 | M4 | M55 | M6 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Keratinolytic | 1 | 1 | 1 | 5 | 1 | 1 | 3 | 1 | 2 | 1 | 1 |
| Non_keratinolytic | 1 | 1 | 1 | 5 | 1 | 1 | 3 | 1 | 2 | 1 | 1 |

Se puede observar que los datos se encuentran balanceados tanto respecto a actividad queratinolítica, como a miembros con y sin actividad queratinolítica por cada familia. Cabe destacar que los datos no están balanceados por familia, debido a que para determinadas familias se han descrito una mayor cantidad de queratinasas que para otras.

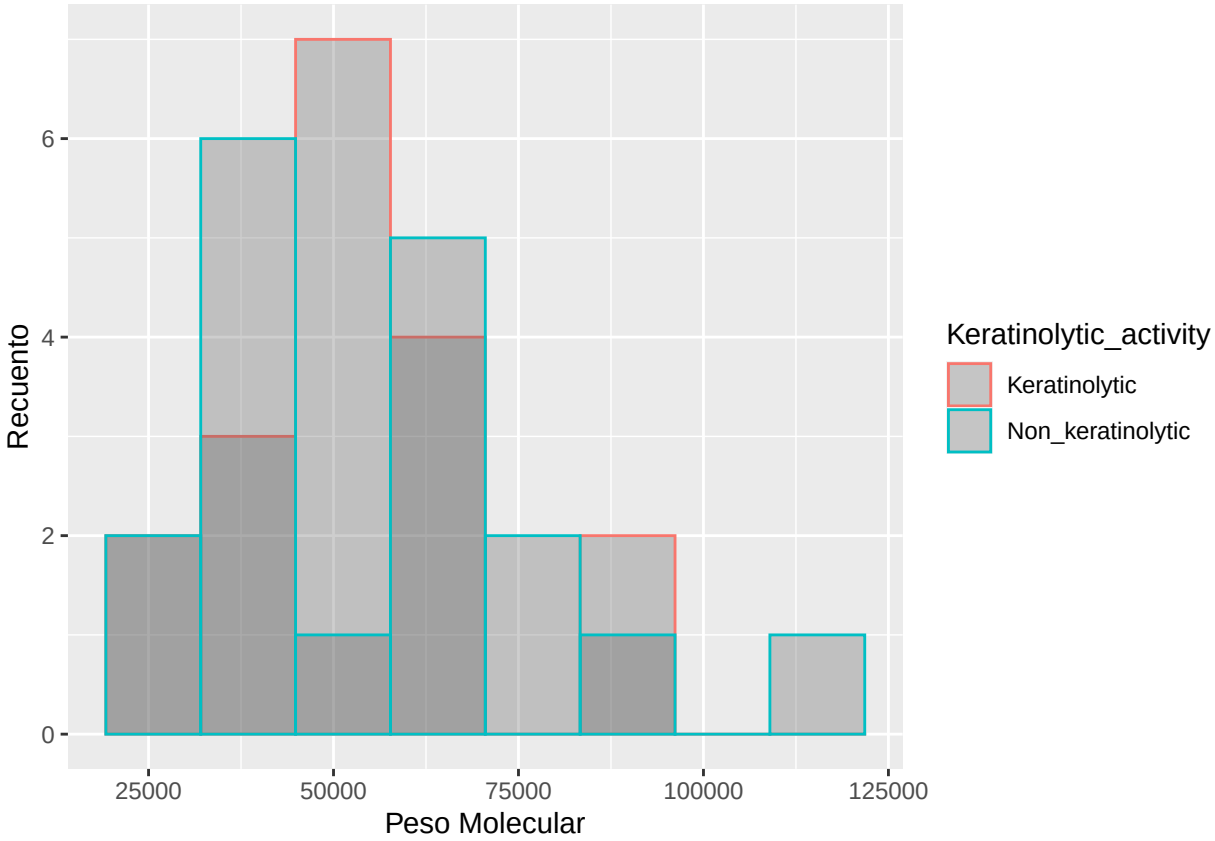
---

### Variación de las variables de estudio

A partir del set de datos se obtienen los histogramas correspondientes para analizar la variación de cada una de las variables de estudio.

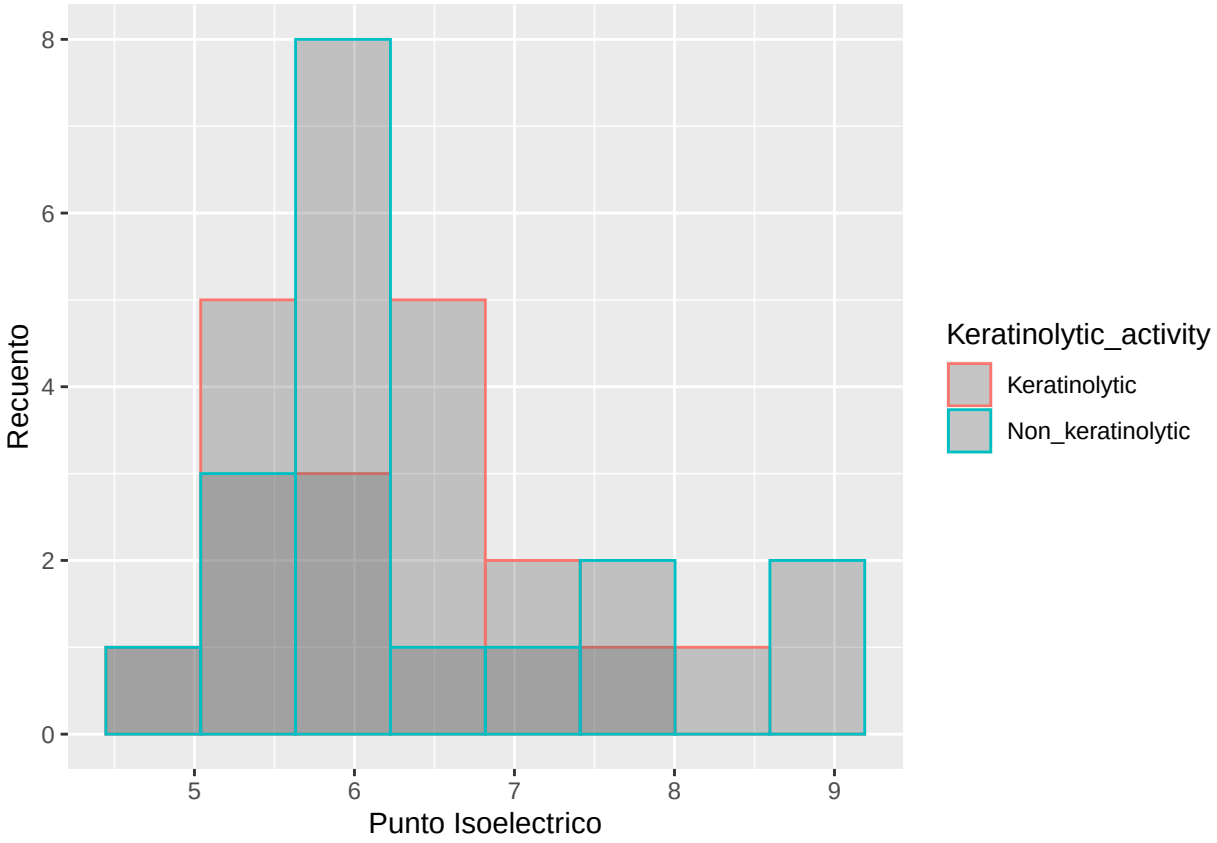
**Peso Molecular** El siguiente histograma de obtuvo a partir de los datos correspondientes al peso molecular de las metaloproteasas.

```
ggplot(Datos_Metaloproteasas_Norm, aes(x = Mol_Weight, color = Keratinolytic_activity)) +
  geom_histogram(position = "identity", bins = 8, alpha = 0.3) + labs(x= "Peso Molecular", y= "Recuento
```



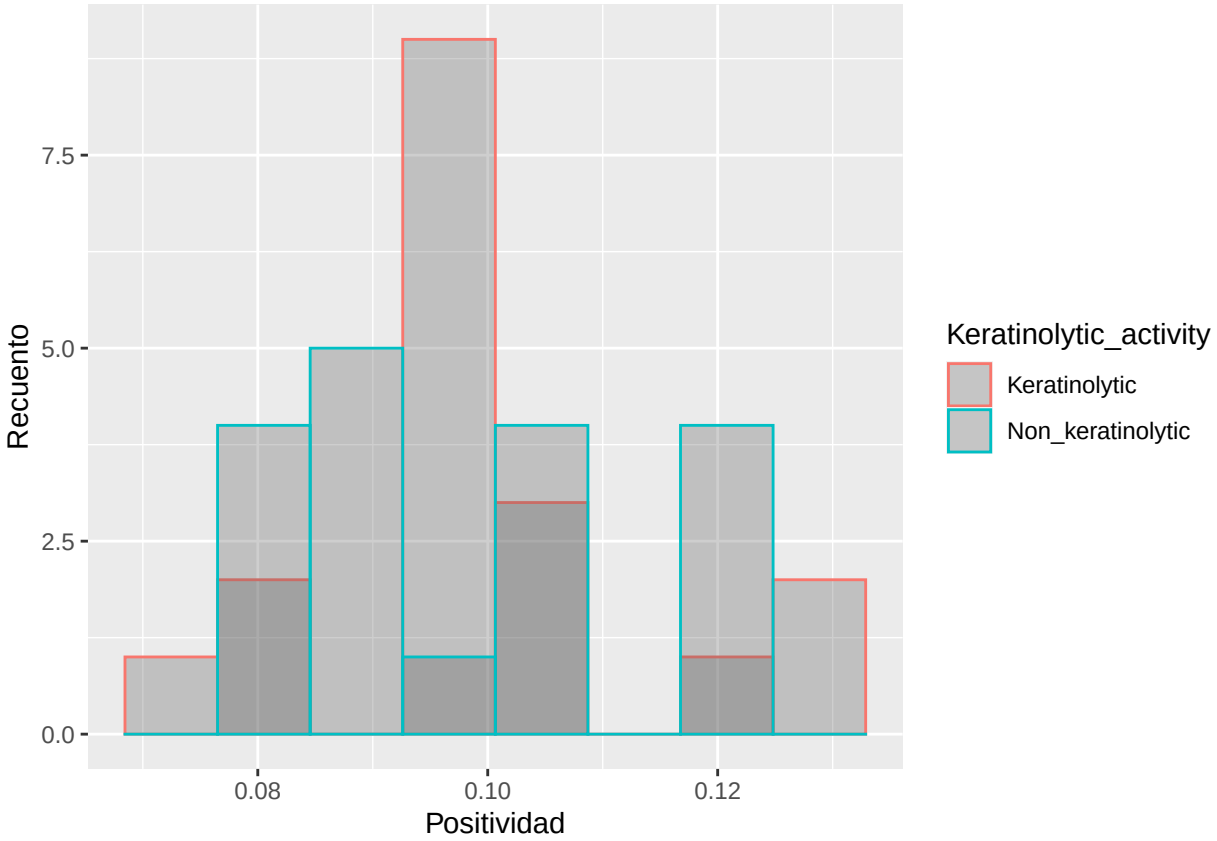

**Punto Isoeléctrico** El siguiente histograma de obtuvo a partir de los datos correspondientes al Punto Isoeléctrico de las metaloproteasas.

```
ggplot(Datos_Metaloproteasas_Norm, aes(x = pI, color = Keratinolytic_activity)) +
  geom_histogram(position = "identity", bins = 8, alpha = 0.3) + labs(x= "Punto Isoelectrico", y= "Recue
```



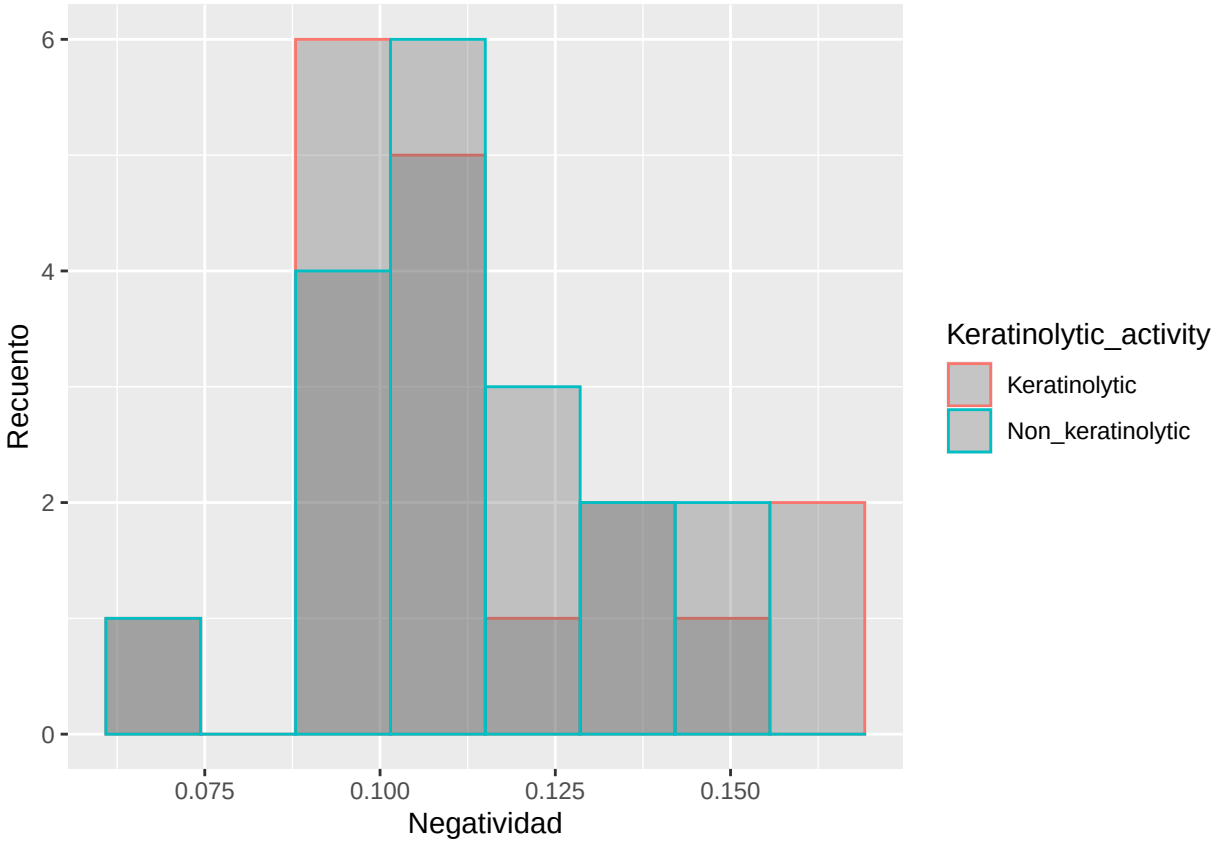

**Positividad** El siguiente histograma de obtuvo a partir de los datos correspondientes a la Positividad de las metaloproteasas.

```
ggplot(Datos_Metaloproteasas_Norm, aes(x = Positivity, color = Keratinolytic_activity)) +
  geom_histogram(position = "identity", bins = 8, alpha = 0.3) + labs(x= "Positividad", y= "Recuento")
```



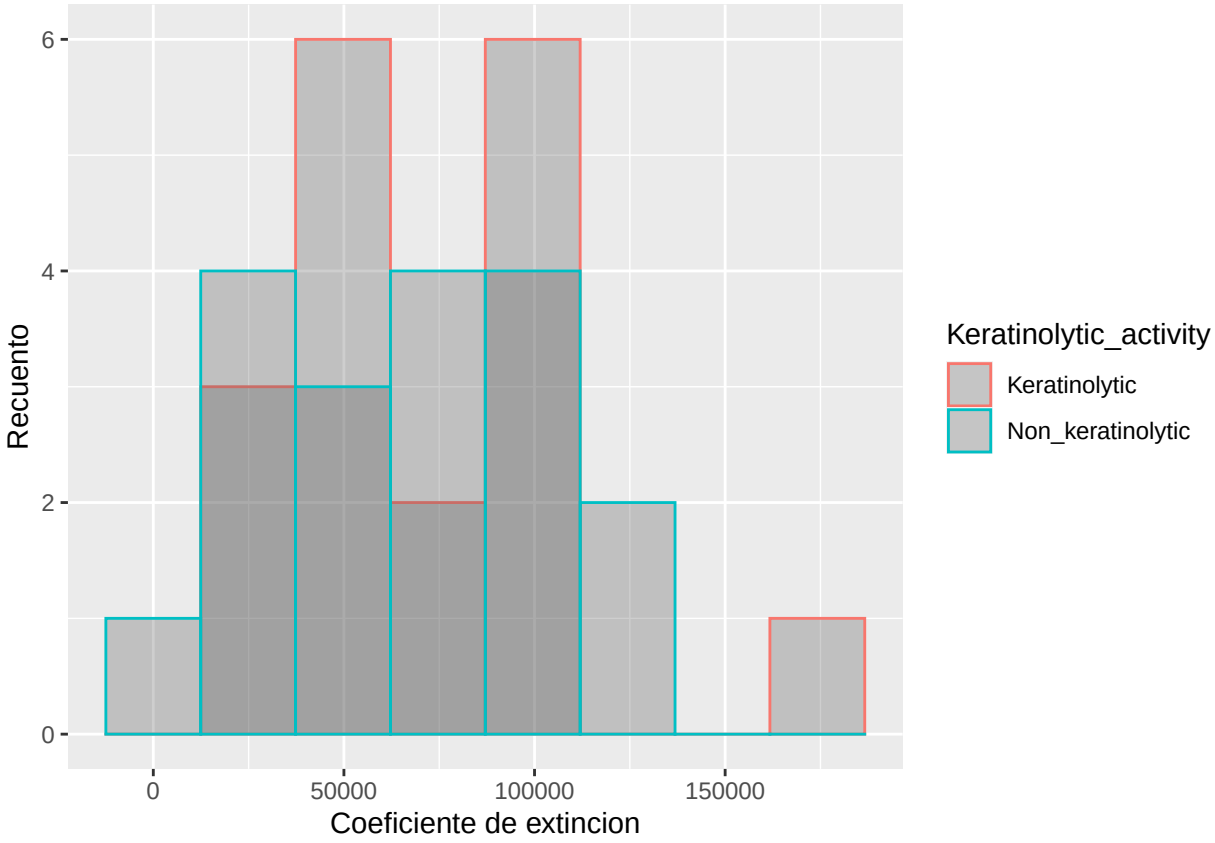

**Negatividad** El siguiente histograma de obtuvo a partir de los datos correspondientes a la Negatividad de las metaloproteasas.

```
ggplot(Datos_Metaloproteasas_Norm, aes(x = Negativity, color = Keratinolytic_activity)) +
  geom_histogram(position = "identity", bins = 8, alpha = 0.3) + labs(x= "Negatividad", y= "Recuento")
```




**Coeficiente de extinción** El siguiente histograma de obtuvo a partir de los datos correspondientes al Coeficiente de extinción de las metaloproteasas.

```
ggplot(Datos_Metaloproteasas_Norm, aes(x = Extinction_coefficient, color = Keratinolytic_activity)) + g
```

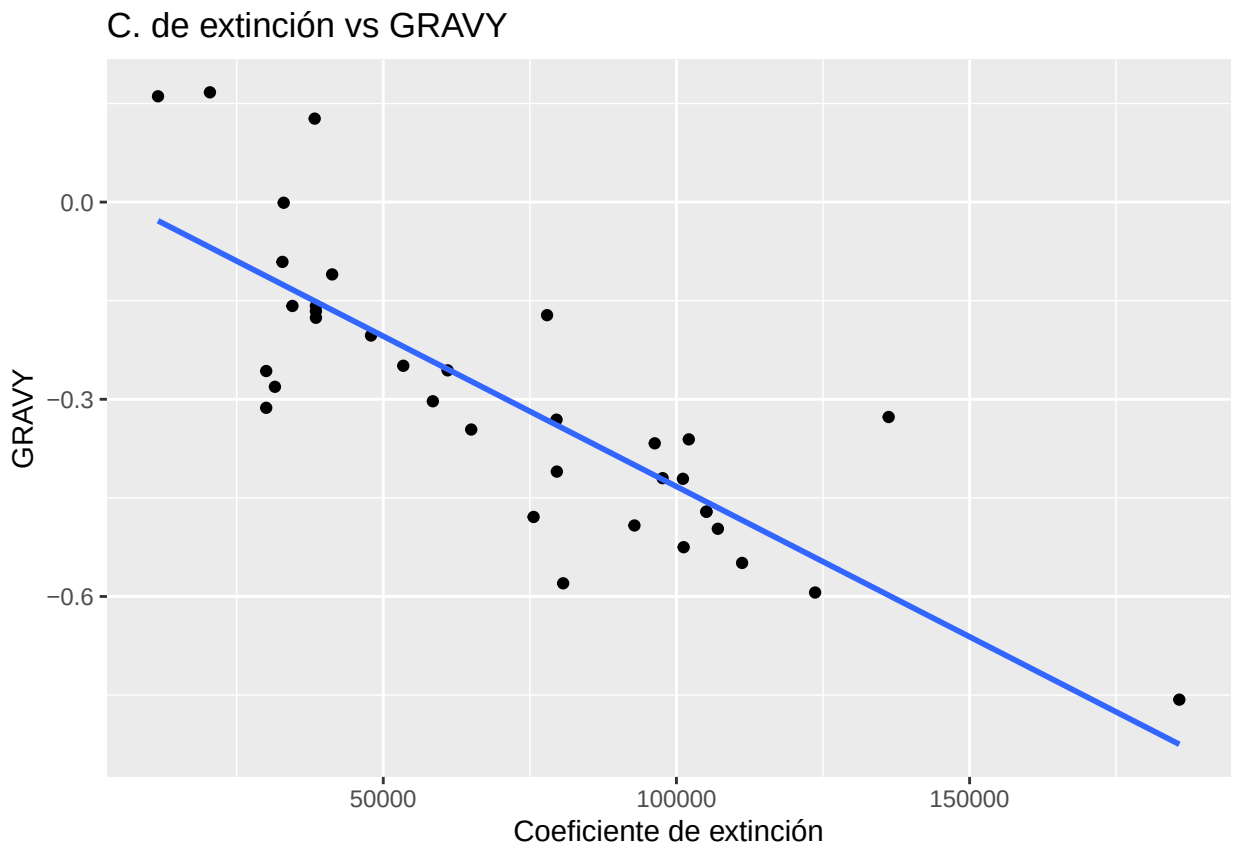



**Análisis de relación entre variables cuantitativas y factores**

**Correlaciones** Los siguientes gráficos muestran la correlación entre las diversas variables.

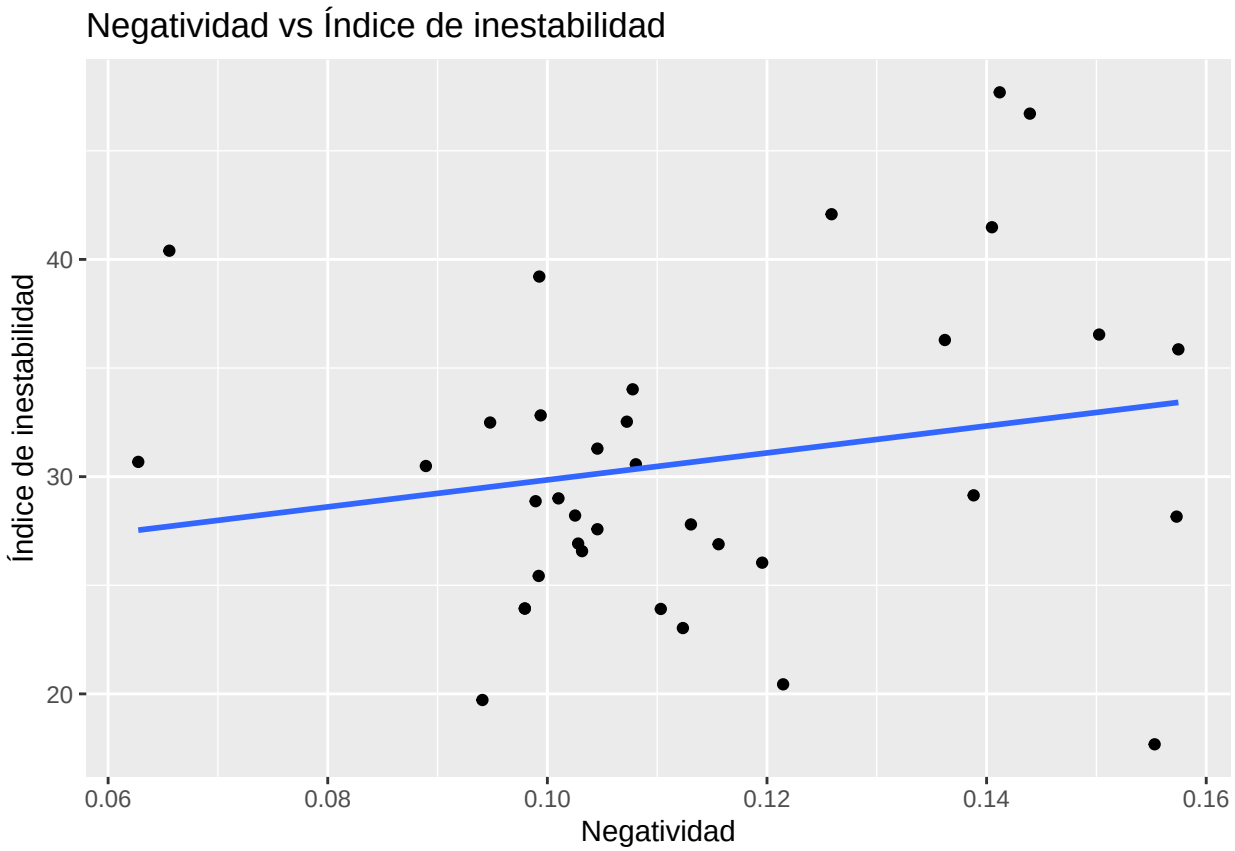
```
ggplot(Datos_Metaloproteasas_Norm, aes(x = Extinction_coefficient, y = GRAVY)) + geom_point() + geom_sm
```

```
## `geom_smooth()` using formula 'y ~ x'
```



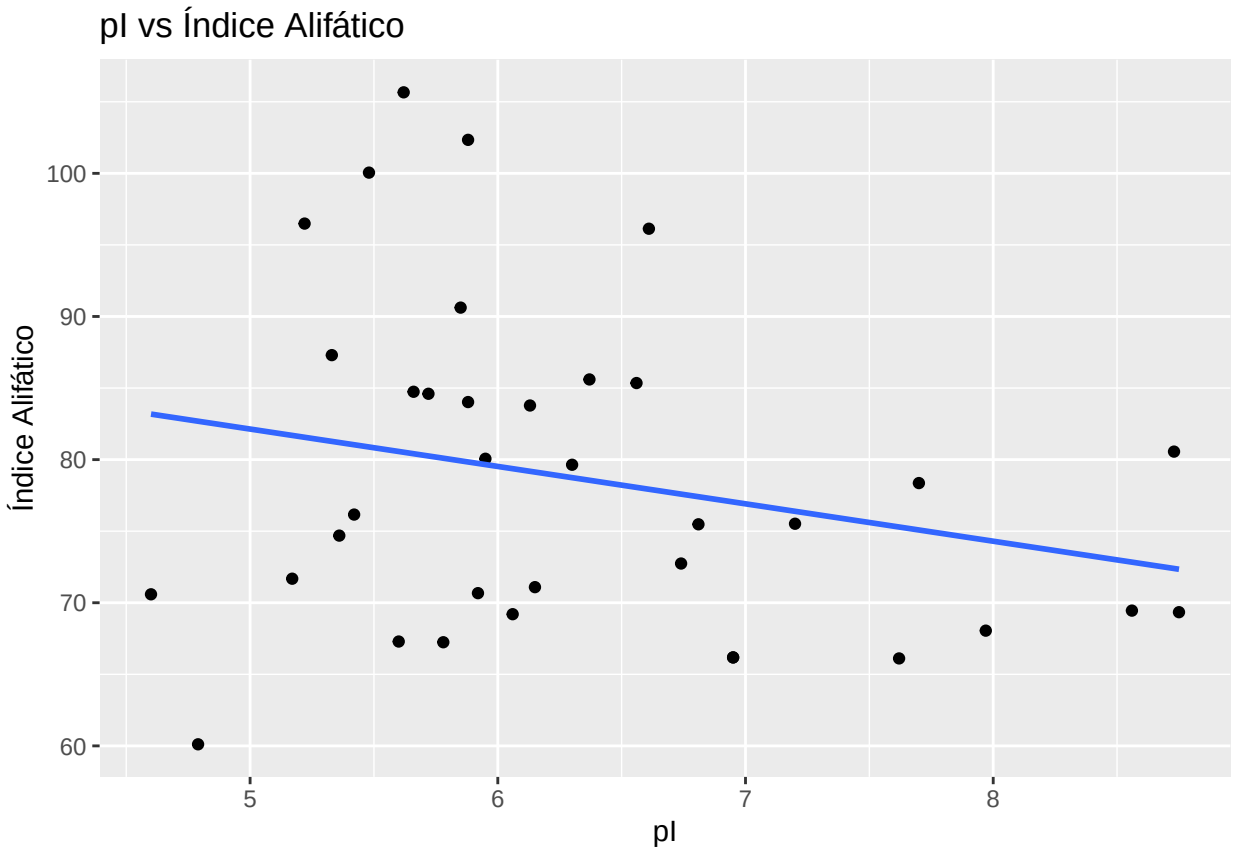

```
ggplot(Datos_Metaloproteasas_Norm, aes(x = Negativity, y = Instability_index)) + geom_point() + geom_sm
```

```
## `geom_smooth()` using formula 'y ~ x'
```



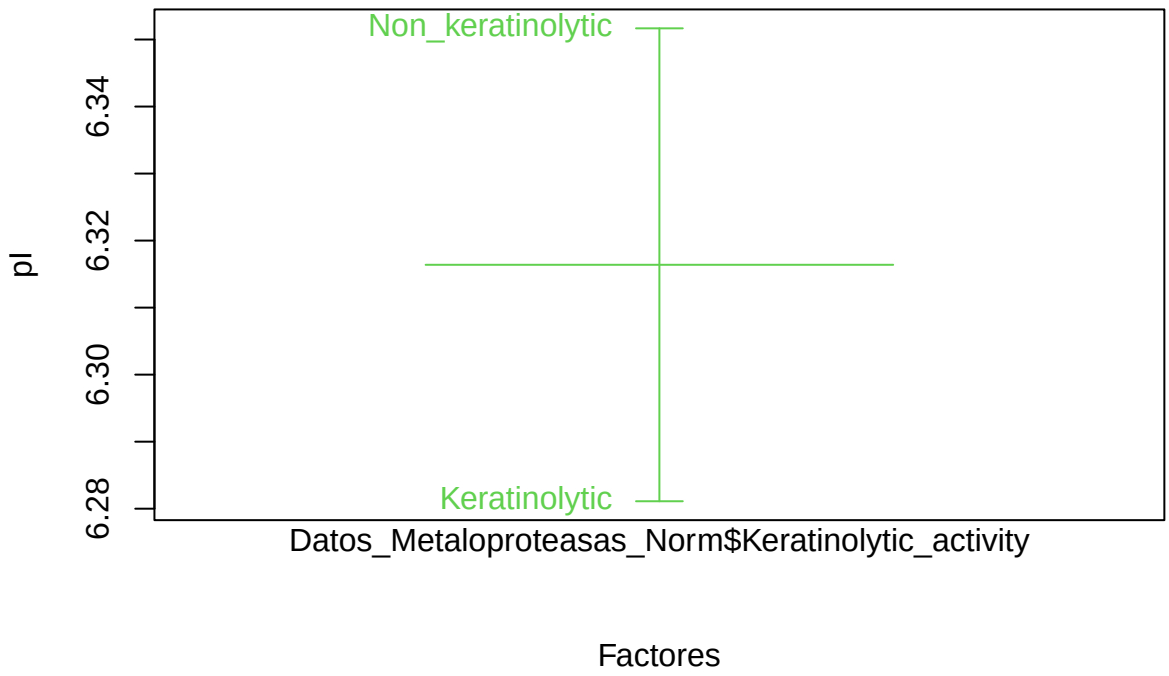

```
ggplot(Datos_Metaloproteasas_Norm, aes(x = pI, y = Aliphatic_index)) + geom_point() + geom_smooth(metho
```

```
## `geom_smooth()` using formula 'y ~ x'
```

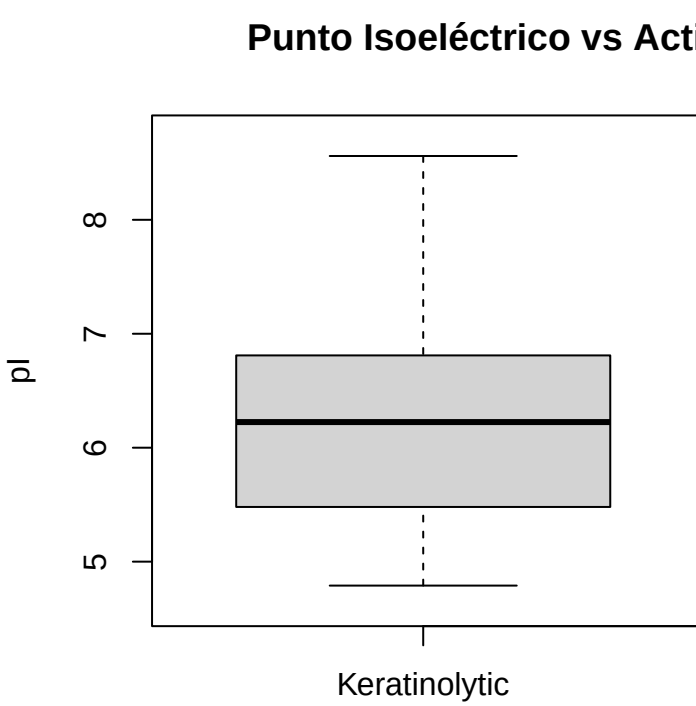



**Tamaño de los efectos** Los siguientes gráficos muestran el tamaño de los efectos para diversas variables.

```
plot.design(Datos_Metaloproteasas_Norm$pI ~ Datos_Metaloproteasas_Norm$Keratinolytic_activity, Datos_Me
```

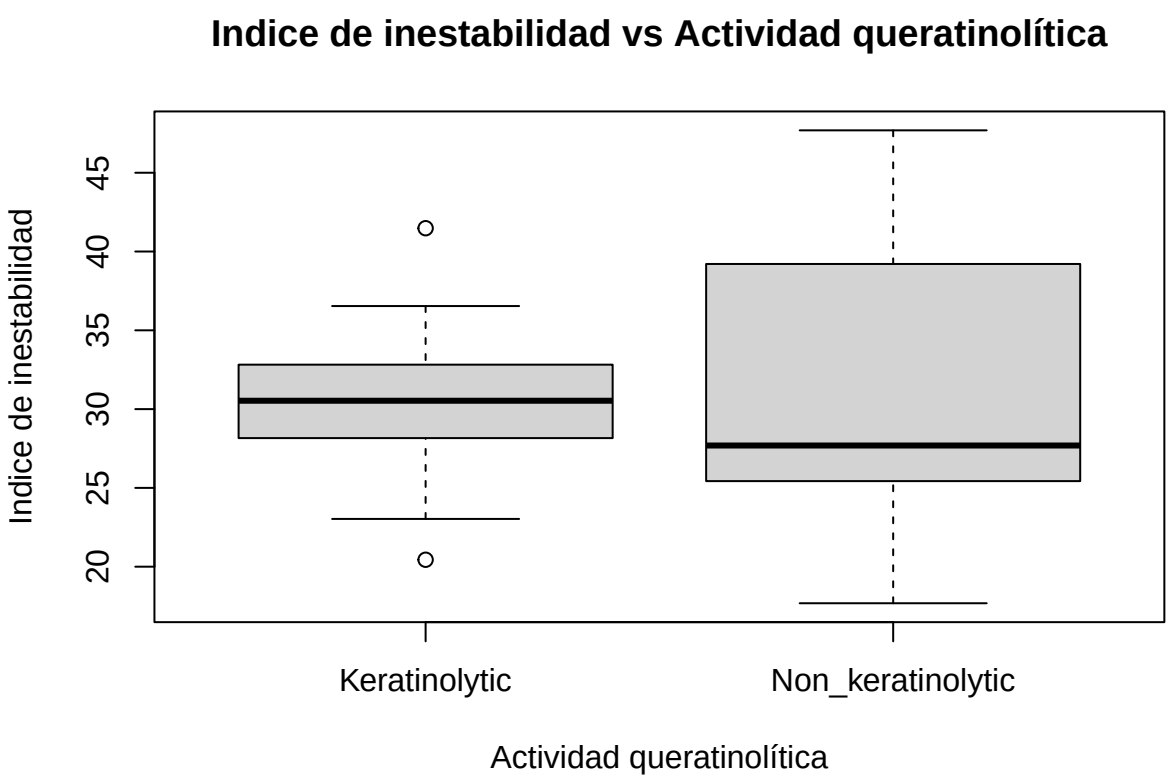



```
boxplot(Datos_Metaloproteasas_Norm$pI ~ Datos_Metaloproteasas_Norm$Keratinolytic_activity, main = "Punt
```

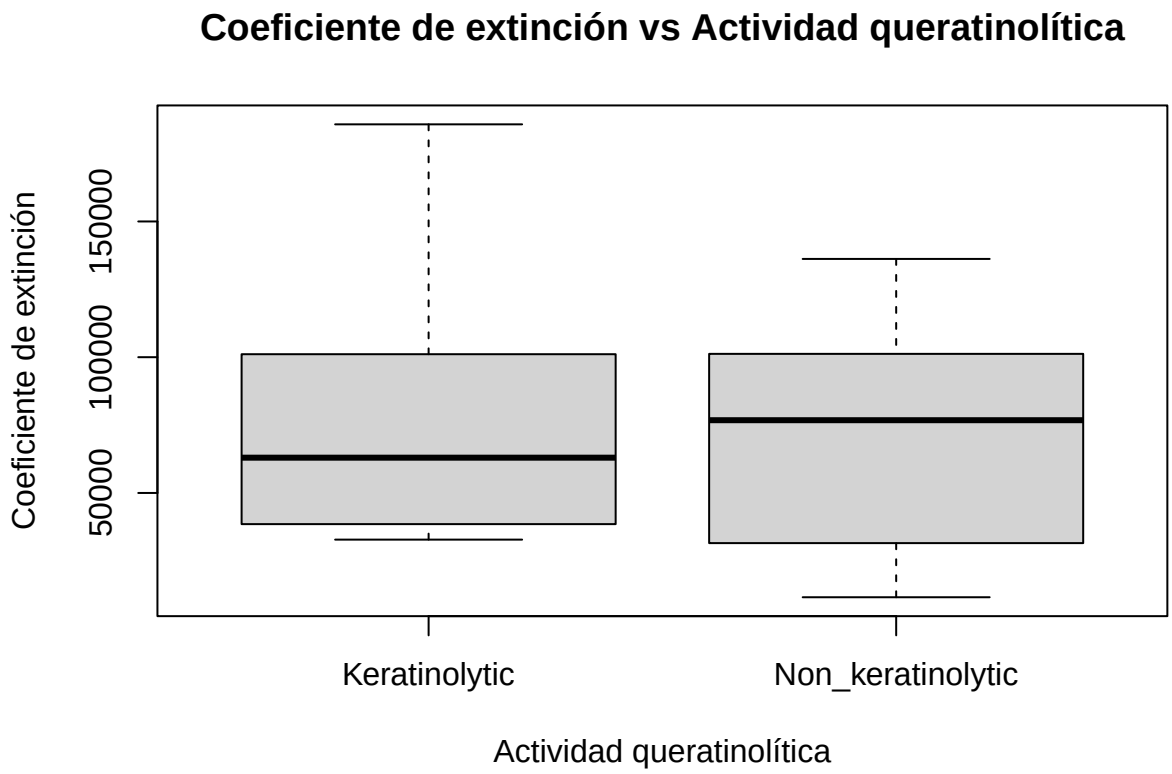



**Diferencias entre queratinolíticas y no-queratinolíticas**

```
boxplot(Datos_Metaloproteasas_Norm$Instability_index ~ Datos_Metaloproteasas_Norm$Keratinolytic_activity
```




```
boxplot(Datos_Metaloproteasas_Norm$Extinction_coefficient ~ Datos_Metaloproteasas_Norm$Keratinolytic_act
```




**Diferencias entre queratinolíticas y no-queratinolíticas**