Estadística Descriptiva – Data Analytic

Analytics AoZ

Como conocemos la estadística descriptiva es la ciencia que ***"recoge, organiza, presenta, analiza….datos"*** por lo cual debemos definir los ***métodos para organizar resumir y presentar los datos de manera informativa***, además se debe tomar en cuenta el ***conocimiento de los errores en los datos, detectar datos faltantes y prevenir errores en el análisis***. Dentro de esto tenemos la estadística descriptiva univariante (variables cualitativas y cuantitativas), dentro de las estas podemos usar medidas de resumen, tablas y/o gráficos que permiten conocer el comportamiento de las variables, considerando una a una, o la posible relación existente entre ellas.

- Frecuencias absolutas y relativas (resumir individualmente variables de tipo cualitativo)
- Diagrama de sectores (visualización de las frecuencias relativas de la variable cualitativa)
- Gráfico de barras (visualización en criterio de ojo humano adecuado para comparar longitudes, variable cualitativa)

Para las variables cuantitativas usamos metodologías más complejas en el análisis descriptivo univariante debido a que tenemos que considerar que existen más métodos y mayores cálculos, pero presentaremos a manera de resumen lo más importante:

- **Medidas de Centralización:** media aritmética, moda, media geométrica, mediana (*robusta*), media cortas (*robusta*).
- **Medidas de Localización:** mínimo, máximo, percentiles, cuartiles, deciles.
- **Medidas de Dispersión:** rango, rango intercuartílico (*robusta*), varianza, desviación estándar, desviación media absoluta (*robusta*), coeficiente de variación.

- **Medidas de Asimetría:** coeficiente de asimetría de Pearson, coeficiente de asimetría de Fisher, coeficiente (g) ajusta al tercer momento divido entre el cubo de la desviación estándar, asimetría coeficiente cuartil (Kenney and Keeping, 1954) (*robusta*), percentil 10 de medias cortas.
- **Valores Atípicos (*outlayers and faroutlayers*):** Estos son considerados los valores que son mucho mayores o menores al conjunto de observaciones tomadas, como consideración se usará el criterio básico de Tukey (1969), en el cual los atípicos son aquellos que son menores a el Q1-1.5RIQ o los mayores a Q3+1.5RIQ.

  Donde:

  Q1 : Es el primer cuartil; Q3: Es el tercer cuartil y RIQ: Es el rango intercuartílico.

  Estos valores pueden tener 3 causas principales:

  Una medida o grabado de error – Una observación de una población no similar a la mayoría de la data – Un evento raro desde una población simple que tiene alto sesgo.

Consideraciones:

- **Las medidas consideradas robustas no se ven afectadas generalmente por los valores extremos, por lo tanto en algunas situaciones pueden ser considerar como mejores instrumentos de medida al realizar el análisis estadístico.**
- **Los cálculos para obtener las medidas citadas deben ser estudiados y analizados aparte, las fuentes de información son proporcionadas por el curso y están subidas en el material de la clase.**
- **Existen más medidas y cálculos avanzados descriptivos (ej, como las transformaciones de las variables) pero esos no serán detallados en este curso.**

Medidas de Centralización (a veces conocidas como localización):

- **Media Aritmética ($\bar{x}$) (muestral):**

The arithmetic mean ($\bar{X}$), here referred to simply as the mean, is computed as the sum of all data values X_i, divided by the sample size n:

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \quad (1.1)$$

For data that are in one of k groups, equation 1.1 can be rewritten to show that the overall mean depends on the mean for each group, weighted by the number of observations (n_i) in each group:

$$\bar{X} = \sum_{i=1}^{k} \bar{X}_i \frac{n_i}{n} , \quad (1.2)$$

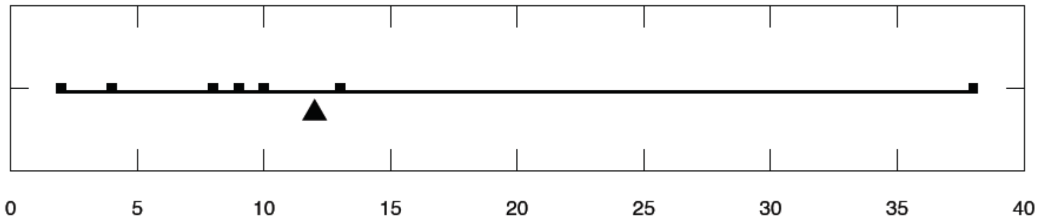


Figure 1.3. Graph showing the arithmetic mean (triangle) as the balance point of a dataset. The mean is 12.

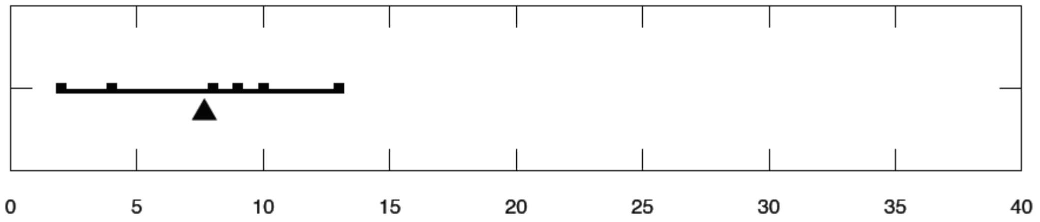


Figure 1.4. Graph showing the shift of the arithmetic mean (triangle) downward after removal of an outlier. The mean is 7.67.

Fuente: Statical Methods in Water Resource USGS 2020

Analytics AoZ

Características de la Media Aritmética:

- *Medida resumen de los datos que nos da una idea de aproximación o tendencia central.*
- *Los valores muy bajos o muy altos tienen una gran influencia en el cálculo de la media aritmética (generan desbalance).*
- *La media es una medida de tendencia central considerada no resistente, pero útil.*
- *Es influenciada por los valores no detectados además de los outliers.*

- **Mediana (Me) (Robusta): "The Middle"**

Analytics AoZ

The median, or 50th percentile ($P_{0.50}$), is the central value of the distribution when the data are sorted by magnitude. For an odd number of observations, the median is the data point that has an equal number of observations both above and below it. For an even number of observations, it is the arithmetic mean of the two central-most observations. To compute the median, first sort the observations from smallest to largest, so that $X(1)$ is the smallest observation and $X(n)$ is the largest observation. Then

$$median = P_{0.50} = \begin{bmatrix} X\left(\frac{n+1}{2}\right) & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(X\left(\frac{n}{2}\right)+X\left(\frac{n}{2}+1\right)\right) & \text{when } n \text{ is even} \end{bmatrix} . \tag{1.4}$$

The median is only minimally affected by the magnitude of any single observation. This resistance to the effect of a change in value or presence of outlying observations is often a desirable property. To demonstrate the resistance of the median, suppose the last value of the following dataset (a) of 7 observations was multiplied by 10 to obtain dataset (b):

Example 1.1–Resistance of the mean and median

Dataset (a) 2 4 8 9 11 11 12 $\bar{X} = 8.1$ $P_{50} = 9$

Dataset (b) 2 4 8 9 11 11 120 $\bar{X} = 23.6$ $P_{50} = 9$

Fuente: Statical Methods in Water Resource USGS 2020

Características de la Mediana:

- *Medida resumen de los datos que nos da una idea de aproximación o tendencia central.*
- *Los valores muy bajos o muy altos no tienen , en general, una gran influencia en el cálculo de la mediana (no generan desbalance).*
- *La mediana es una medida de tendencia central considerada resistente y muy útil.*
- *No está influenciada por valores no detectados.*
- *Puede ser usada con datos censurados.*

¿Calcular la mediana en la siguiente serie 1,2,98,99,100? ¿Qué se puede concluir?

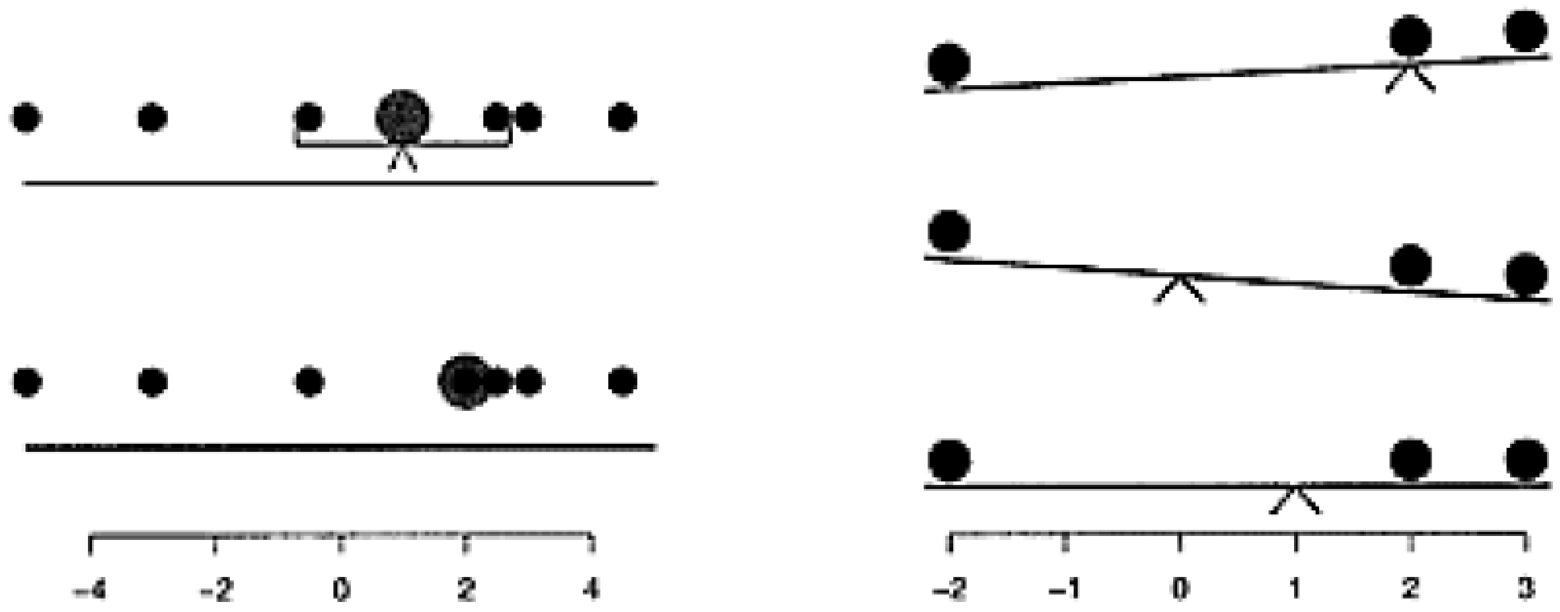


Figure 2.8 The median is the middle point, the mean the balance point

Fuente: Verzani, J. (2005). Using R for introductory statistics. CRC press.

- **Moda (Mo):** Definida como el dato más frecuente observado, es más aplicable con data discreta (mientras data solo posiblemente tiene valores enteros); en la data continua es más complicada su aplicación por lo cual se considera una pobre medida de localización en data continua. Es muy útil para data cualitativa.

Analytics AoZ

- **Media Geométrica (MG):**

The geometric mean (*GM*) is often reported for positively skewed datasets. It is only defined in cases where all data values are positive. By definition, it is the *n*th root of the product of the *n* values in the sample.

$$GM = \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n} \tag{1.5}$$

A simple way to calculate it is to take the mean of the logarithms of the data and then transform that value back to the original units.

$$GM = \exp\left(\overline{Y}\right) , \tag{1.6}$$

where

$$Y_i = \ln\left(X_i\right) \text{ ; and}$$

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} .$$

Fuente: Statical Methods in Water Resource USGS 2020

Características de la Media Geométrica:

- *Para data con asimetría positiva la media geométrica es aproximadamente igual a la mediana. Cuando los logaritmos de la data son simétricos, la media geométrica es una estimación no sesgada de la media.*
- *La media geométrica es siempre menor o igual que la media aritmética.*
- *Cuando se transforma a unidades originales la data (después de transformación log por ejemplo), la media geométrica continua teniendo la mitad de observaciones por debajo y la mitad por encima, y esto es porque la mediana esta debajo de la media aritmética.*

Analytics AoZ

- **Media Corta (α%-Robusta): "Trimmed Media":**

Una medida entre la mediana y media, en la cual se realiza un recorte (eliminación) de un α/2% de datos a la derecha y α/2% de datos la izquierda, por lo cual no es influenciada por valores extremos (quizás anómalos) de la muestra. La media corta es un estimador central resistente.

- **Media α-winsorizada muestral:**

Se sustituye un determinado porcentaje, α (20% generalmente) de valores extremos a cada lado de la muestra por el valor más próximo no sustituido.

- **Estimador de Hurber:**

Se encuentra dentro de los denominados M-estimadores, que generalizan al estimador de máxima verosimilitud con buenas propiedades de robustez y eficiencia. En este caso se descartan las observaciones que sean mayores (o menores) a una constante.

Medidas de Posición:

Analytics AoZ

- **Mínimo y Máximo:** El valor mínimo y máximo están representados para variables cualitativas como el menor y el más grande de ellos lo cual nos da una idea de la extensión de nuestra data.
- **Cuantiles:** son una posición 1+p(n-1) de la data ordenada. Cuando este no es un entero, se usa el promedio del peso. Este valor es esencial para dividir la data en 100p% su menor y 100(1-p)% su mayor. Aquí el p va de 0 a 1. La mediana es el cuantil 0.5.
- **Percentiles :** Son similares a los cuantiles, excepto por que la escala de 0 a 100 es usada, en vez de 0 a 1.
- **Cuartiles:** El termino cuartiles se refiere al 0.25, 0.75 y 100 percentiles, y el término quintiles se refiere a 0 , 20, 40, 60, 80 y 100 percentiles.

The exec.pay (UsingR) data set contains compensation to CEOs of 199 U.S. companies in the year 2000 in units of $10,000

```
> x = 0:5                      # 0,1,2,3,4,5
> length(x)
[1] 6
> sum(sort(x)[3:4])/2          # the median the hard way
[1] 2.5
> median(x)                    # easy way. Clearly the
middle
[1] 2.5
> quantile(x,.25)
25%
1.25
> quantile(x,c(0.25,0.5,0.75)) # more than 1 at a time
```

† There are other definitions used for the pth quantile implemented in the quantile() function. These alternatives are specified with the type= argument. The default is type 7. See ?quantile for the details.

```
25% 50% 75%
1.25 2.50 3.75
> quantile(x)               # default gives quartiles
0% 25% 50% 75% 100%
0.00 1.25 2.50 3.75 5.00
```

```
> sum(exec.pay > 100)/length(exec.pay) # proportion
more
[1] 0.09045                     # 9% make more than 1
million
> quantile(exec.pay,0.9)        # 914,000 dollars is 90
percentile
90%
91.4
> quantile(exec.pay,0.99)       # 9 million is top 1
percentile
997,
906.6
> sum(exec.pay <= 10)/length(exec.pay)
[1] 0.1457                      # 14 percent make
100,000 or less
> quantile(exec.pay,.10)        # the 10 percentile is
90,000
10%
9
```

Fuente: Using R for Introductory Statistics, John Verzani 2005

Medidas de Dispersión:

Se refiere a medidas de tendencia en la cual la información más importante es la extensión o dispersión de la data.

- **Rango (R):** Es una simple medida del dispersión el cual mide la distancia entre el más pequeño valor con el más grande. Cabe considerar que el rango no es muy informativo acerca de la dispersión de la data en caso de grandes bancos de datas o de presencia de datos anómalos.

Rango = Xmax-Xmin

- **Varianza (Var):** Mide la dispersión desde el data set . Un largo valor de variación indica que la data no está agrupada alrededor de la media, mientras pequeños valores indicar que está muy cercano a la media; la varianza se ve afectada notoriamente por los valores extremos y por números largos de data no detectada.

$$s^2 = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{(n-1)}$$

Fuente: Fuente: Statical Methods in Water Resource USGS 2020

- **Desviación Estándar (sd = s):** Es la raíz cuadrada del cálculo de la varianza.

Ojo: La desviación típica sólo es un buen estimador promedio de la desviación del conjunto de los datos con respecto al valor central, cuando la distribución es normal (gaussianna).

Analytics AoZ

- **Rango Intercuartilo (IQR-Robusta) : El rango intercuartil es la medida robusta más usada para medir variabilidad de la data. Esta medida contiene al 50% central de la data y no es influenciada por el 25% superior ni inferior. (Estudiar además el concepto de *plotting position*).**

Analytics AoZ

The IQR is defined as the 75th percentile minus the 25th percentile. The 75th, 50th (median), and 25th percentiles split the data into equal-sized quarters. The 75th percentile ($P_{0.75}$), also called the upper quartile, is a value that exceeds no more than 75 percent of the data and is therefore exceeded by no more than 25 percent of the data. The 25th percentile ($P_{0.25}$), or lower quartile, is a value that exceeds no more than 25 percent of the data and is therefore exceeded by no more than 75 percent. Consider a dataset ordered from smallest to largest: X_i, $i=1,2,...,n$. Percentiles (P_j) are computed using equation 1.8

$$P_j = X_{(n+1)\cdot j} \; , \tag{1.8}$$

where n is the sample size of X, and j is the fraction of data less than or equal to the percentile value (for the 25th, 50th, and 75th percentiles, $j=0.25$, 0.50, and 0.75, respectively).

For the datasets used in example 1.1, n=7, and therefore the 25th percentile is $X_{(7+1)\cdot 0.25}$ or $X_2=4$, the second lowest observation. The 75th percentile is X_6, the sixth lowest observation, or 11. The IQR is therefore $11-4=7$. When values of $(n+1)\cdot j$ are not integers, then some type of interpolation method is needed to compute the percentiles. There are several different ways to do this computation. The preference for this book is to use the R function `quantile`. This function allows the user to specify the type of interpolation. There are nine possible types of interpolation available in R for this function. With large sample sizes (greater than about 100 observations), the choice of type is of very little consequence. The choice preferred here, `type = 6`, is commonly known as the Weibull plotting position, which has a long history in hydrology (Weibull, 1939). In hydrology the term "plotting position" comes from the rules for constructing an empirical cumulative distribution function, which plots the individual observations in the sample (in ranked order) against an estimated probability of nonexceedance. That estimated probability is called the plotting position. Hydrologists have historically also used the Hazen plotting position (which corresponds to `type = 5`) (Hazen, 1914) and the Blom plotting position (Blom, 1958), (`type = 9`), which are both used in flood frequency analysis and in distribution fitting methods (as discussed in chap. 4). For an extensive discussion of these algorithms, see Hyndman and Fan (1996), which explains each `type` using the same system for the numbering for the choices of `type` that is used in R. Using our preferred choice, `type = 6`, the R commands for computing the IQR of a dataset (with the data stored as a vector called `x`) are the following:

```
> quant <- as.numeric(quantile(x, type = 6))
> IQR <- quant[4] – quant[2]
```

Note that the default for the `quantile` command is that it returns a set of five values representing, in this order, the minimum, lower quartile, median, upper quartile, and maximum. One of the arguments to the `quantile` function is `probs`, which can be either a scalar or vector for the probabilities we wish to estimate. When the argument `probs` is set equal to some other sequence of values (which are in the range of 0 to 1), then the function returns a set of values for each specified probability value. An alternative way to compute the IQR would be in a single line:

```
> IQR <- quantile(x, probs = 0.75, type = 6) – quantile(x, probs =
  0.25, type = 6)
```

There is a standard R function for the IQR (which is called `IQR`). When the default values of the function are used, the function will return a different value than what is defined above because the default is `type = 7`, known as the Gumbel plotting position. See section 2.1.2. for a discussion of these choices. However, it can return exactly the value of the IQR as defined here by calling it in this manner:

```
> IQR(x, type = 6)
```

In most cases, the difference between the results will be quite small if the sample size is larger than about 100 observations.

Fuente: Fuente: Statical Methods in Water Resource USGS 2020

- **Desviación Media Absoluta (MAD – Robusta):**

Es una medida de dispersión robusta la cual es calculada con la mediana del data set y luego se calcula todas las diferencias absolutas entre cada observación (Xi) y la mediana. La MAD es la mediana de todas esas diferencias absolutas.

$$MAD(X) = median\left|d_i\right| \ ,$$

where

$$d_i = X_i - median(X) \ .$$

Fuente: Fuente: Statical Methods in Water Resource USGS 2020

- **Coeficiente de Variación (CV):**

Es una medida adimensional que permite la comparación de dispersión alrededor de muchos set de data, se calcula dividiendo la desviación estándar de la muestra entre la media de la muestra. El CV es usado en aplicaciones medioambientales porque la variabilidad (expresada como desviación estándar) es usualmente proporcional a la media y una medida adimensional.

$$CV = \frac{\bar{x}}{sd(x)} * 100\%$$

Ojo: La desviación absoluta mediana estandarizada es una medida robusta para el cálculo de las medidas de dispersión.

Medidas de Asimetría:

El cálculo de los coeficientes de asimetría nos lleva a cuantificar la forma en que se prodúcela curva de la distribución de nuestros datos respecto a una referencia normal (a la derecha como positiva y a la izquierda como negativa), es decir ayuda a medir el sesgo de la información para poder entender el sesgo de la data y aplicar posibles transformaciones (ej. Transformación avanzada Box-Cox o Yeo-Johnson).

- **Coeficiente de Asimetría de Pearson (Asp):** El coeficiente de asimetría de Pearson mide esta asimetría comparando la moda (en caso no se pueda calcular la mediana) con la desviación típica de la data.
- **Coeficiente de Asimetría de Fisher (Asf):** Este realiza el cálculo respecto a la diferencia entre cada variables y la media elevado al cubo y generando una sumatoria que se divide entre la desviación estándar cúbica.

Asimetría de Pearson:

$$A_{p1} = \frac{\bar{x} - Mo}{s}$$

$$A_{p2} = \frac{3(\bar{x} - Me)}{s}$$

Asimetría de Fisher

$$A_F = \frac{\mu_3}{s^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Fuente: Fuente: Arnaldo Mangeaud (2014). Estadística aplicada a las Ciencias Geológicas. Universidad Nacional de Córdova.

- **Medida de Curtosis (Apuntalamiento):** Otorgan un valor numérico a la puntiagudez. Se dice que es mesocúrtica una distribución con valor cero, mientras que es platicúrtica una distribución “aplastada” y leptocúrtica una distribución puntiaguda.

$$K_F = \frac{\mu_4}{s^4} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$$

Fuente: Fuente: Arnaldo Mangeaud (2014). Estadística aplicada a las Ciencias Geológicas. Universidad Nacional de Córdova.

- **Momentos de Orden r:** La fórmulas de media, varianza, simetría (Fisher) y Curtosis se está ante fórmulas muy similares que cambian soo a qué valor están elevadas, se define momento de orden r a:

$$\overline{Mr} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n-1}$$

Fuente: Fuente: Arnaldo Mangeaud (2014). Estadística aplicada a las Ciencias Geológicas. Universidad Nacional de Córdova.

Los momentos de orden 1 corresponde a la Media, el de orden 2 a la varianza, el de orden 3 a la Asimetría y el de orden 4 a la Curtosis.

Una medida de Simetría Resistente —La asimetría cuartil

A resistant measure of symmetry is the quartile skew *qs* (Kenney and Keeping, 1954):

$$qs = \frac{\left(P_{0.75} - P_{0.50}\right) - \left(P_{0.50} - P_{0.25}\right)}{P_{0.75} - P_{0.25}} \qquad (1.11)$$

defined as the difference in distances of the upper and lower quartiles from the median, divided by the IQR. A right-skewed distribution again has a positive *qs*; a left-skewed distribution has a negative *qs*. Similar to the 25-percent trimmed mean and IQR, *qs* uses the central 50 percent of the data. For the example dataset, $qs = (11-9)-(9-4)/(11-4) = -0.43$ for both datasets (a) and (b). Note that this resistance may be a liability if sensitivity to a few observations is important.

Fuente: Fuente: Statical Methods in Water Resource USGS 2020

Cabe considerar que otros percentiles puedan ser usados para producir una serie de medidas de resistencia de localización, dispersión y asimetría. El caso ejemplo que se muestra en la figura de la izquierda nos indica usando el percentil 10 con media recortada puede ser usado para medir la dispersión en un rango de 10 a 90 como una medida de dispersión y una medida de asimetría.

$$qs_{0.10} = \frac{\left(P_{0.90} - P_{0.50}\right) - \left(P_{0.50} - P_{0.10}\right)}{P_{0.90} - P_{0.10}} \qquad (1.12)$$

to produce a consistent series of resistant statistics. Geologists have used the 16th and 84th percentiles for many years to compute a similar series of robust measures of the distributions of sediment particles (Inman, 1952). Measures based on quartiles have generally become standard, and other measures should be clearly defined prior to their use. The median, IQR, and quartile skew can be easily summarized graphically using a boxplot (see chap. 2) and are familiar to most data analysts.

Fuente: Fuente: Statical Methods in Water Resource USGS 2020

Outliers:

Los outliers son considerados datas que usualmente son bastante diferentes de los valores del dataset. Es Bueno tratar de responderse a estas preguntas al ver outliers:

1. ¿El resultado es product de un error (por ejemplo, instrument en mal funcionamientoo error de entrada de data?
2. ¿Ellos representan una precision razonable de una observación en una situación inusual?

Estos outliers generalmente son removidos luego de describer la data o antes de aplicar algún test de hipótesis. Algo muy útil para no eliminarlos es calculary las medidas de resistencias en orden a realizar un análisis significativo y no eliminar directamente estos. Outliers son el punto más importante en el dataset, y deberán ser investigados a fondo. Cuando los outliers son eliminados crea el riesgo de que la data calculi solo lo que se quiere ver. Outliers tienen las siguientes causas principals:

- Una medida o grabado erróneo.
- Una observación desde una población diferente que la mayoría de la data.
- Un evento raro de una población simple.

Outliers no deben ser descartados sin un análisis riguroso previo, para ver que manera temenos de identificarlos.

Analytics AoZ

		Resumen de estadísticos de los datos obtenidos de elementos químicos en el Río Luna														
Porcentaje con datos mayor al 50% de L.D	**Temporada**	**Variable**	**Mínimo**	**Límite Inferior**	**Primer quartil (Q1)**	**Mediana**	**Media**	**Media Cortada**	**Tercer quartil (Q3)**	**Límite Superior**	**Máximo**	**Rango intercuartílico (RIC)**	**Desviación media absoluta (MAD)**	**Desviación estándar (SD)**	**Asimetría (As)**	**Curtosis (K)**
100%	Estiaje	pH	4.44000	5.62875	7.35000	7.97000	7.78500	7.78500	8.49750	10.21875	8.88000	1.14750	0.88956	0.96855	-0.57302	4.38441	12.44124
		Ce	45.04000	32.60000	101.31500	121.50000	156.26550	156.26550	147.12500	215.84000	490.60000	45.81000	30.64534	116.90859	0.89212	2.24728	74.81407
		SO4	1.40000	-16.82500	5.375zz00	10.25000	30.79500	30.79500	20.17500	42.37500	196.20000	14.80000	10.22994	55.26070	1.11535	3.63554	179.44698
		Ba	0.00300	-0.00946	0.00850	0.01440	0.01487	0.01487	0.02048	0.03844	0.03230	0.01198	0.00882	0.00821	0.16989	-0.86660	55.23704
		Mn	0.00040	-0.01434	0.00250	0.00620	0.04237	0.04237	0.01373	0.03056	0.43600	0.01123	0.00578	0.10565	1.02692	7.11848	249.38321
		S	0.50000	-6.12500	1.82500	3.85000	10.90000	10.90000	7.12500	15.07500	71.10000	5.30000	3.78063	19.47195	1.08618	3.82172	178.64174
		SiO2	14.60000	12.40000	17.95000	18.80000	20.92500	20.92500	21.65000	27.20000	34.00000	3.70000	2.59455	5.61313	1.13573	-0.21984	26.82500
>50% pero <100% (modelado)		Al	0.00048	-0.01563	0.00800	0.01200	0.55172	0.55172	0.02375	0.04738	10.44000	0.01575	0.00667	2.32804	0.69550	13.27649	421.95962
		As	0.00086	0.00009	0.00124	0.00200	0.00186	0.00186	0.00200	0.00314	0.00300	0.00076	0.00105	0.00072	-0.60423	-1.15299	38.61440
		Cu	0.00008	-0.00154	0.00033	0.00100	0.00244	0.00244	0.00158	0.00344	0.02860	0.00125	0.00102	0.00621	0.69414	12.75743	254.83834
		Fe	0.00346	-0.08500	0.02000	0.03500	0.13695	0.13695	0.09000	0.19500	1.60000	0.07000	0.02965	0.34977	0.87442	12.31450	255.40143
		Zn	0.00038	-0.00613	0.00175	0.00500	0.00679	0.00679	0.00700	0.01488	0.05300	0.00525	0.00445	0.01134	0.47262	10.83298	167.06381

BIBLIOGRAFÍA:

Practical Methods for Data Analysis (US EPA QA/G-9, 2000)

Helsel, D. R., & Hirsch, R. M. (2002). *Statistical methods in water resources* (Vol. 323). Reston, VA: US Geological Survey.

Salvador Figueras, M y Gargallo, P. (2003): "Análisis Exploratorio de Datos, 5campus.com, Estadística <http://www.5campus.com/leccion/aed>

Ramalle-Gómara, E., & De Llano, J. A. (2003). Utilización de métodos robustos en la estadística inferencial. *Atención Primaria*, *32*(3), 177-182.

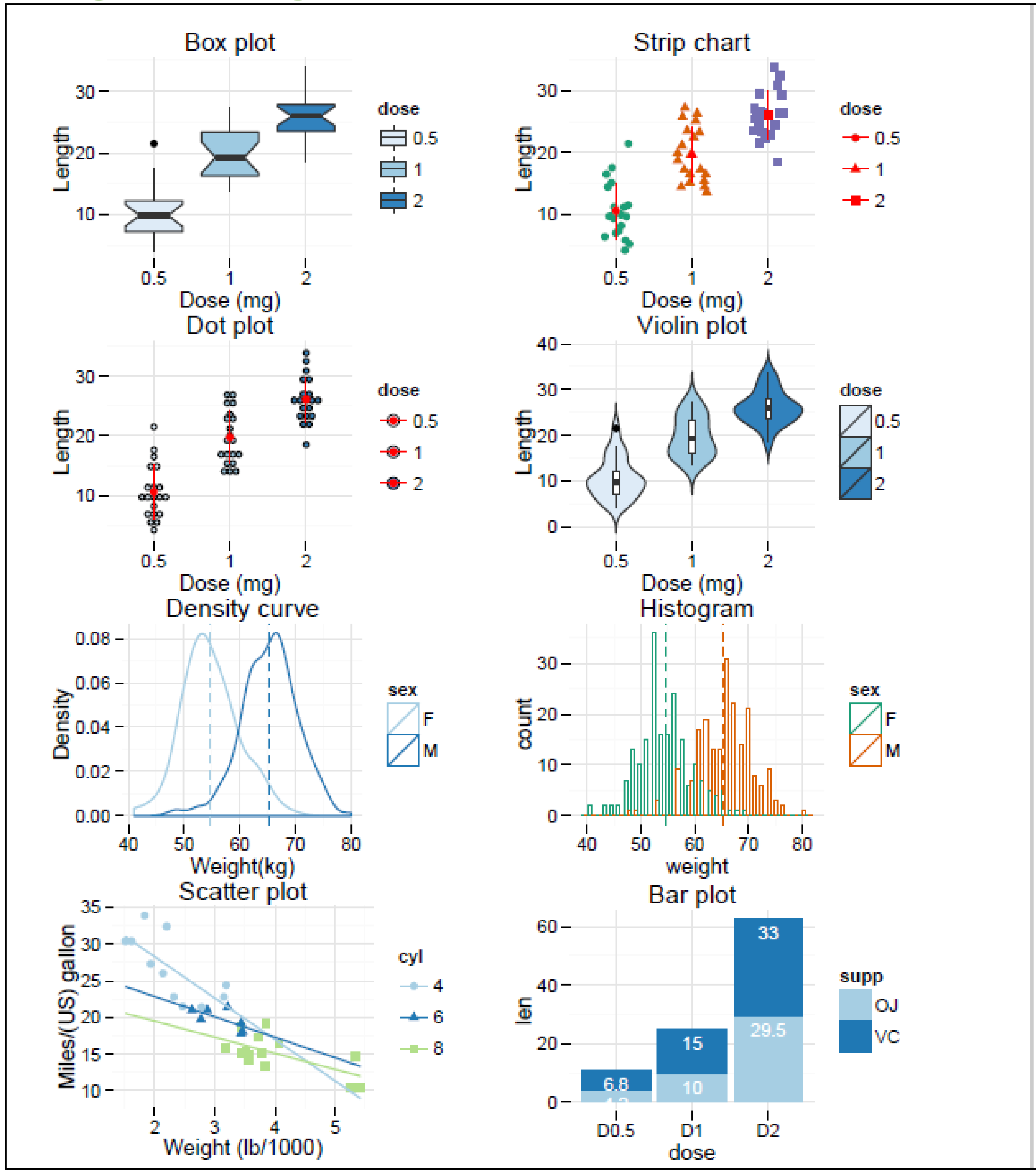
Verzani, J. (2005). *Using R for introductory statistics*. CRC press.

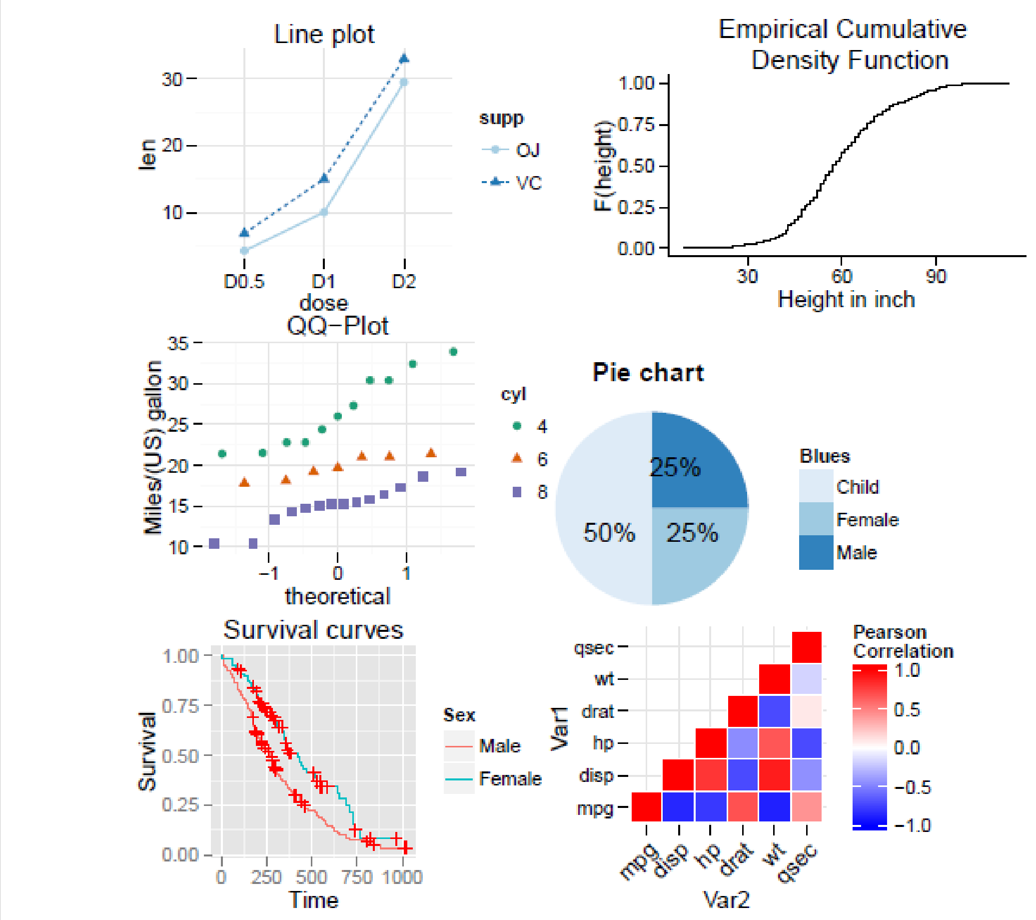
Cohen, Y., & Cohen, J. Y. (2008). *Statistics and Data with R: An applied approach through examples*. John Wiley & Sons.

Arnaldo Mangeaud (2014). *Estadística aplicada a las Ciencias Geológicas*. Universidad nacional de Córdova.

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, *Statistical methods in water resources: U.S. Geological Survey Techniques and Methods*, book 4, chapter A3, 458 p.

Gráficos para Análisis

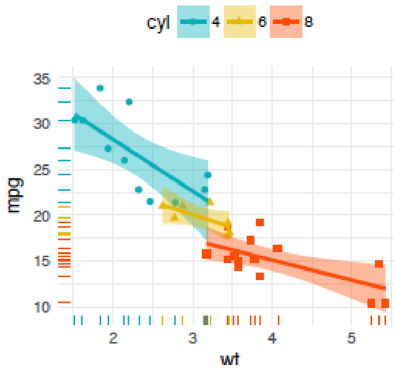


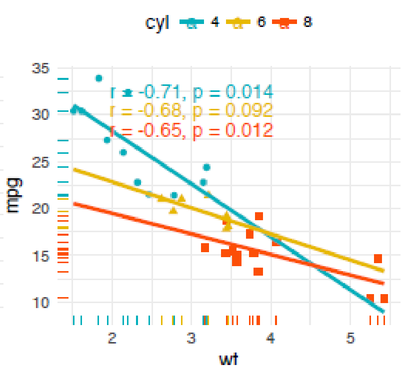


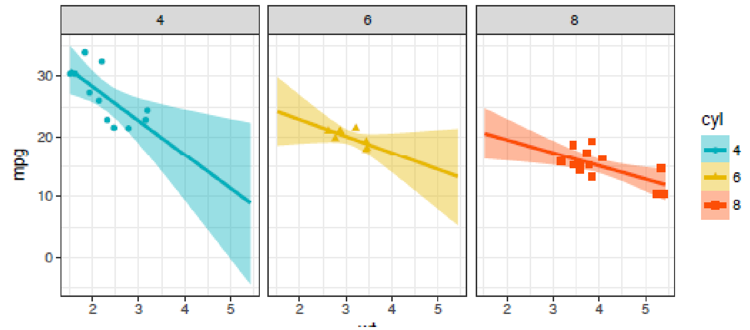
GRÁFICOS SIMPLES PARA UNA VARIABLE CUALITATIVA Y CUANTITATIVA

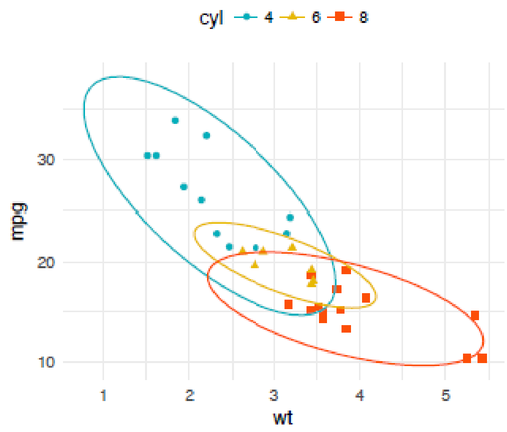
Analytics AoZ

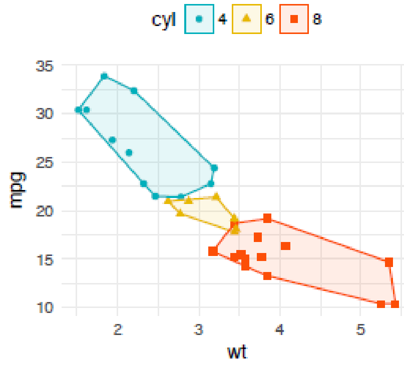
GRÁFICOS SIMPLES PARA DOS VARIABLES CONTINUAS

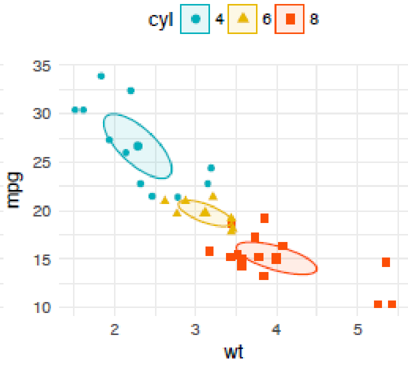


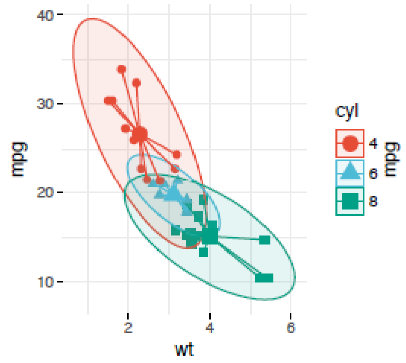


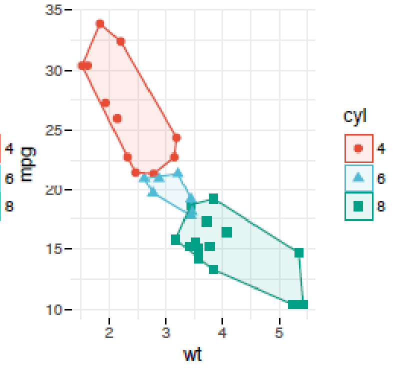


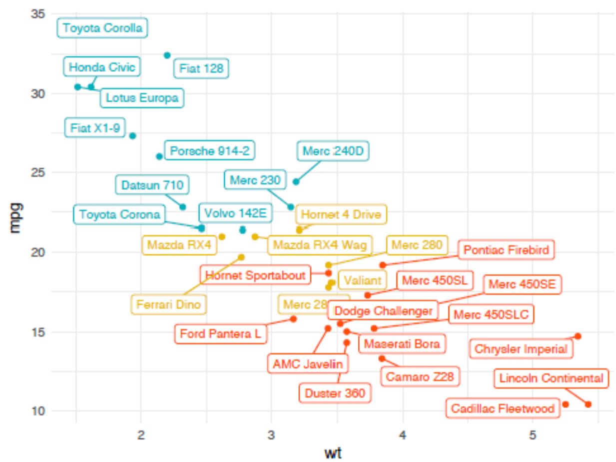


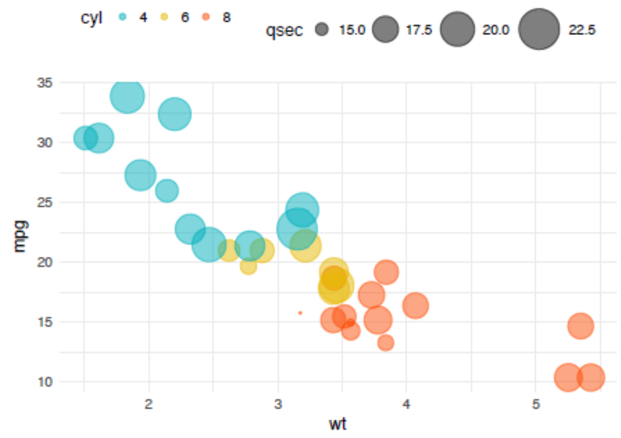


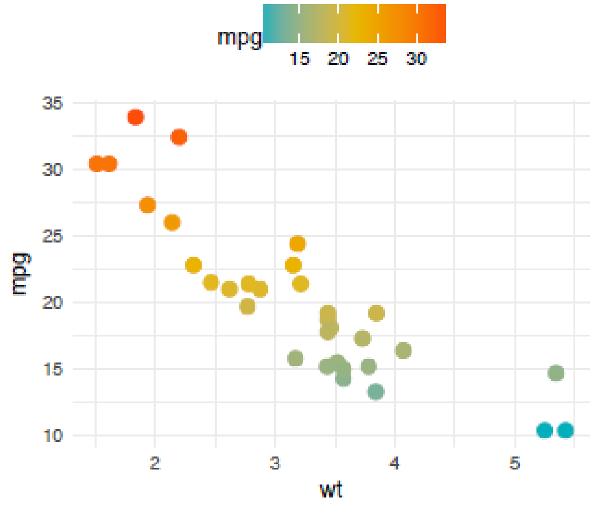




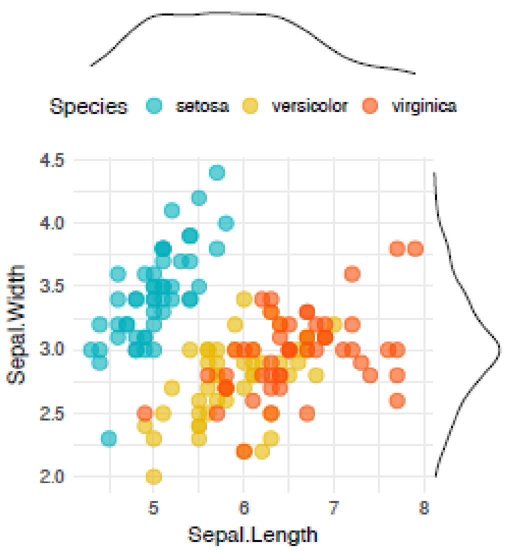


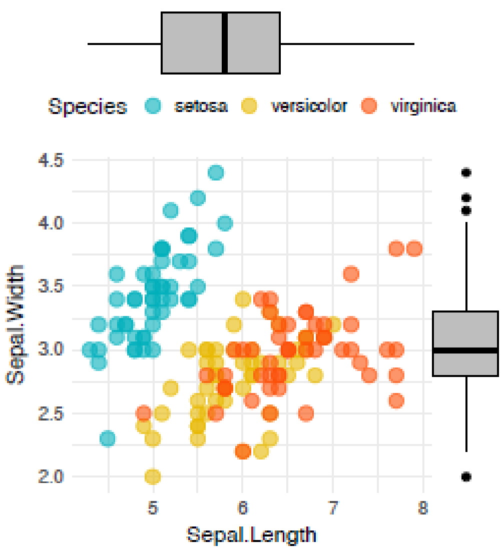


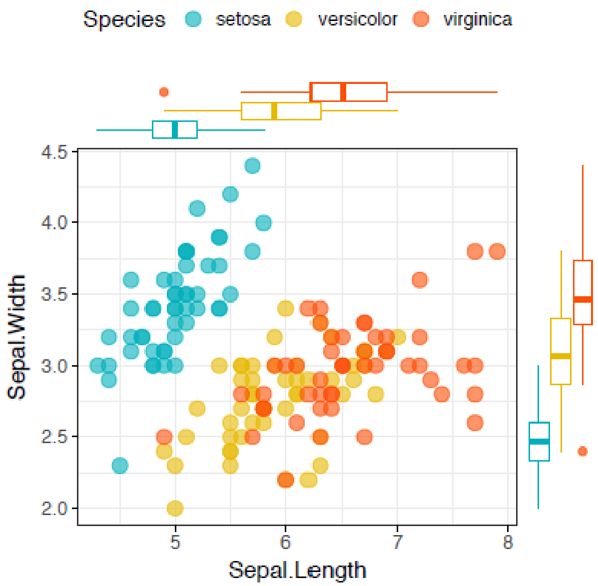


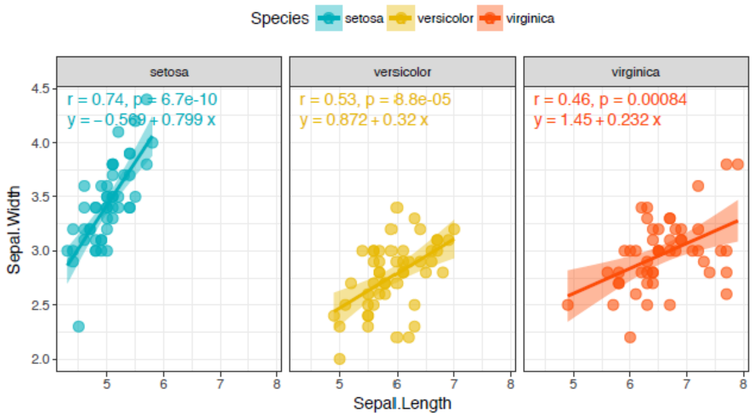


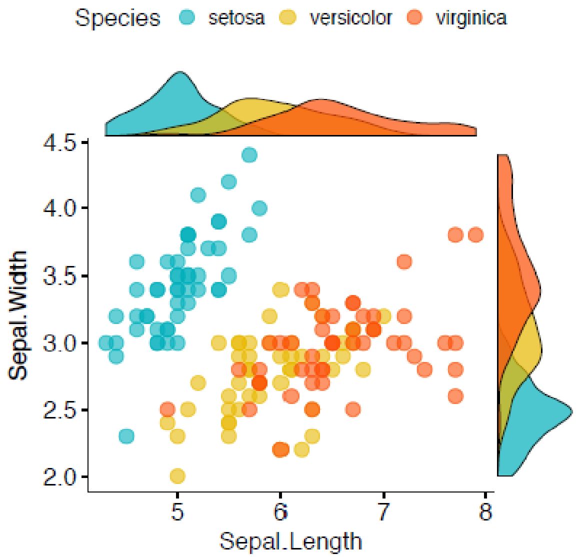
GRÁFICOS SIMPLES PARA 3 VARIABLES CONTINUAS

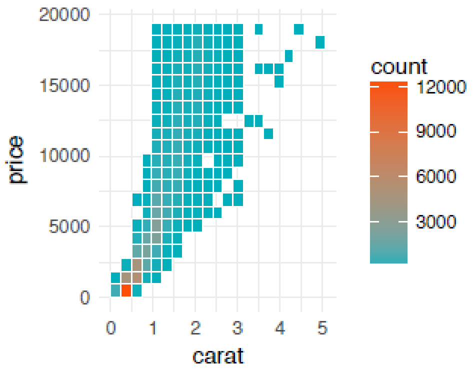


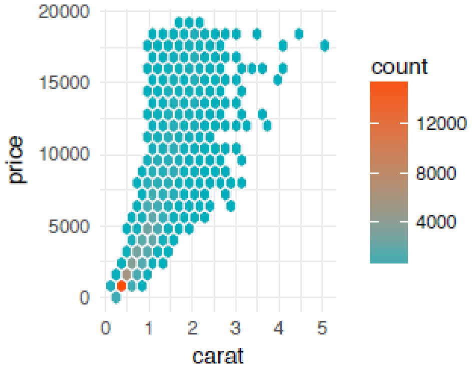


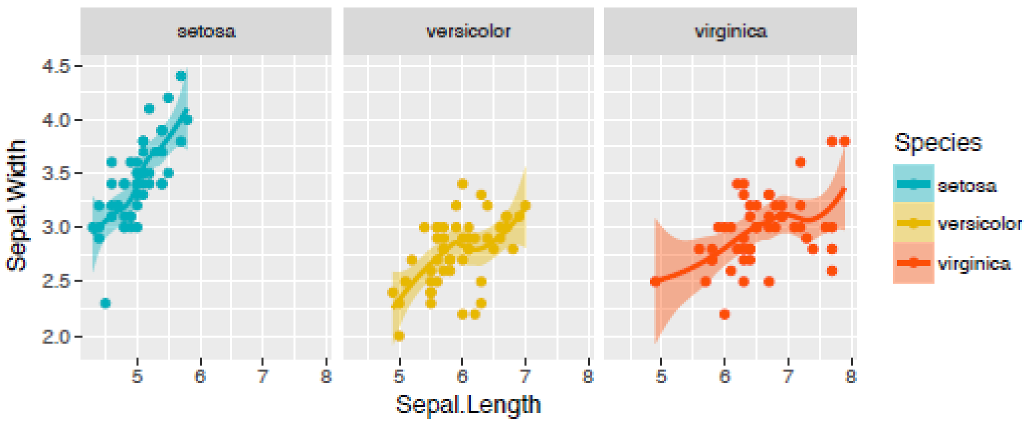


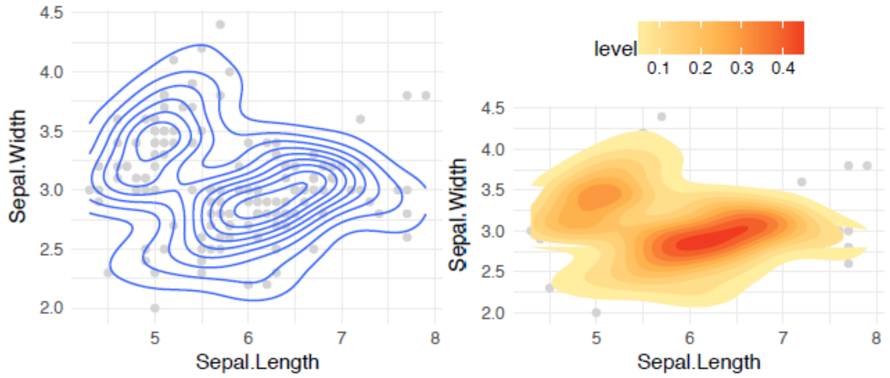




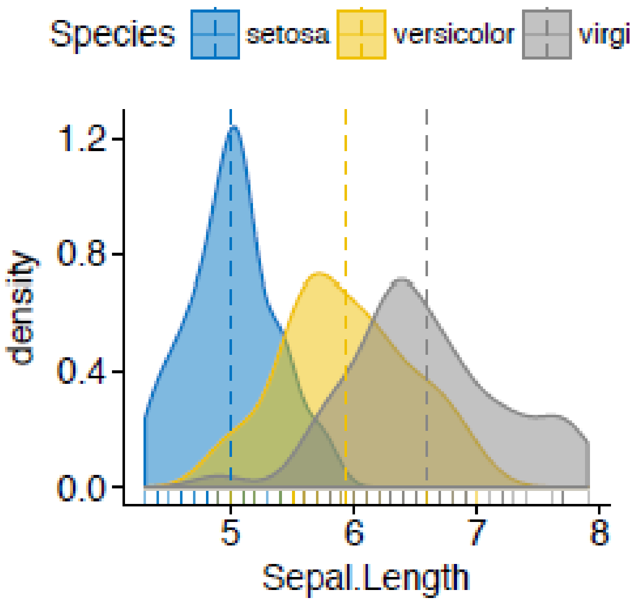


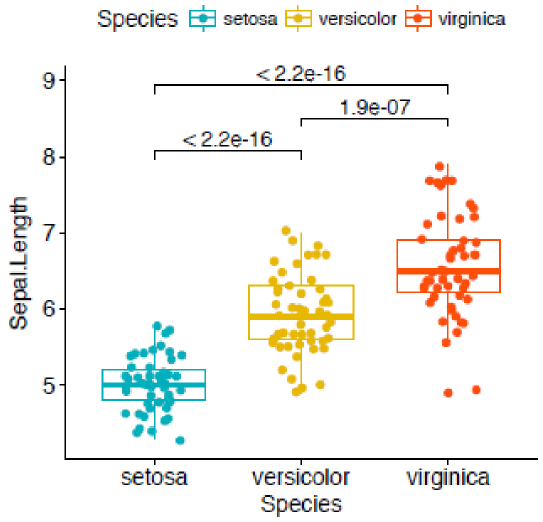


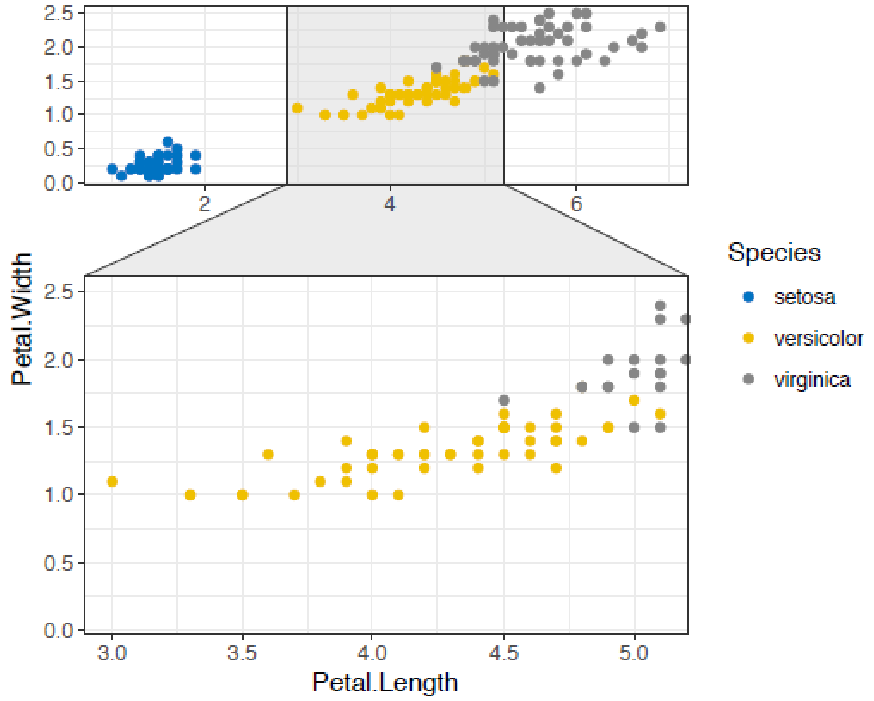


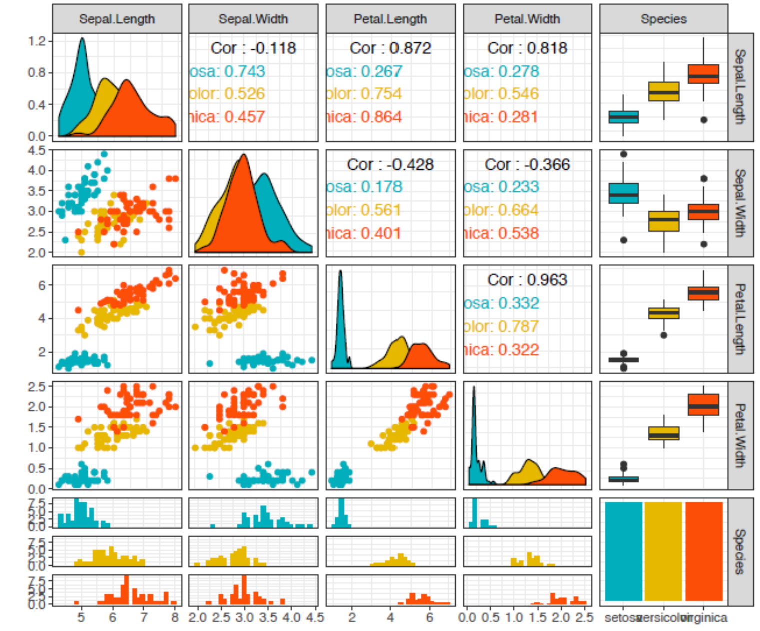


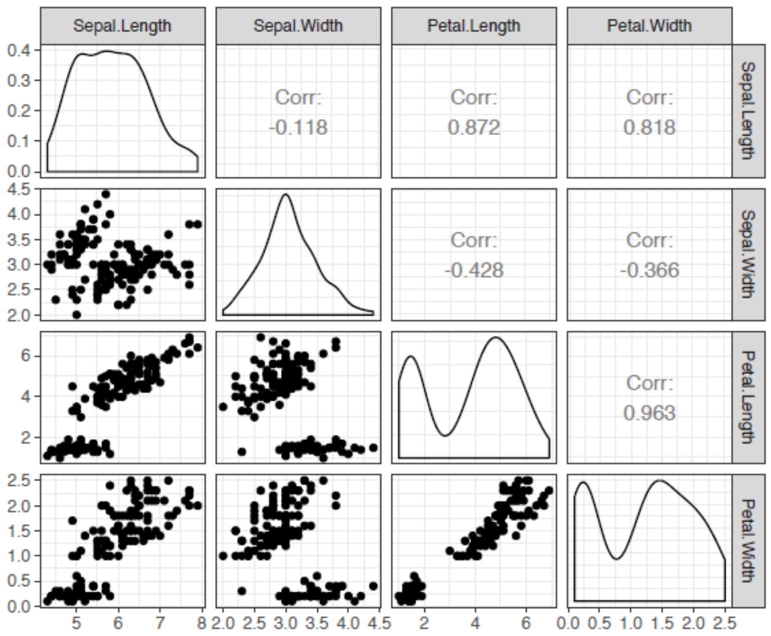
Analytics AoZ

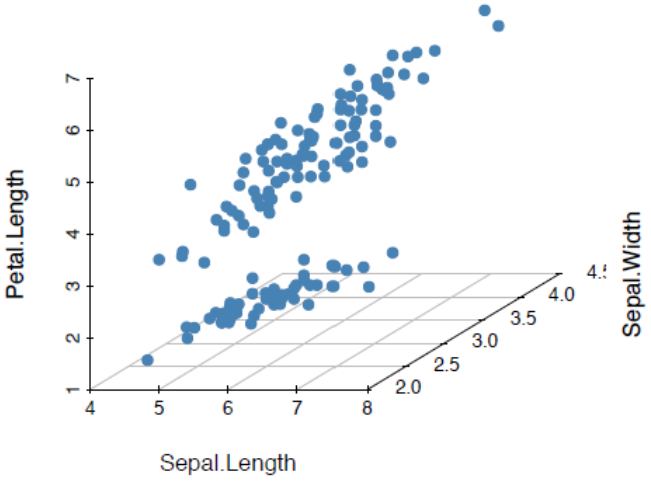


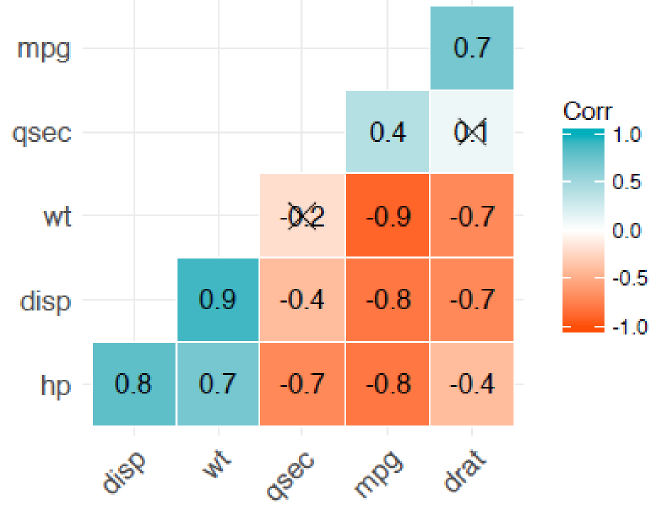


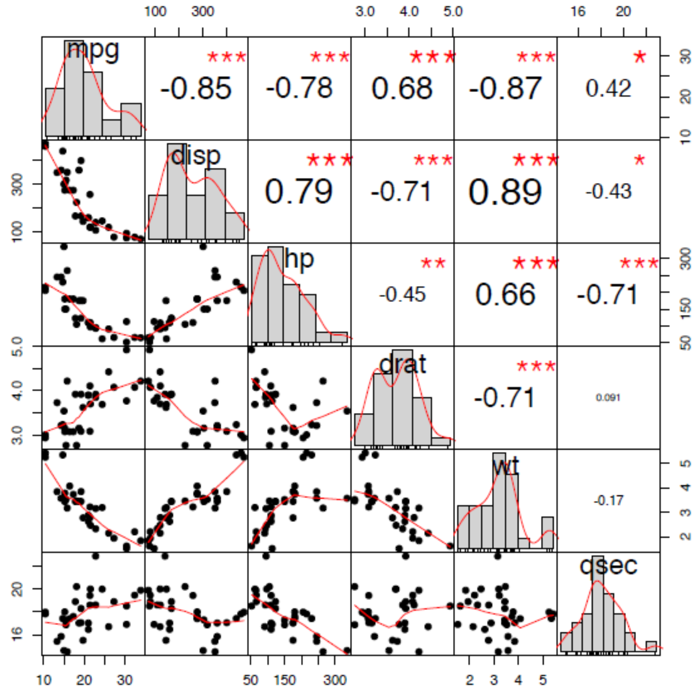



Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
Cor : -0.118
Cor : 0.872
Cor : 0.818
Cor : -0.428
Cor : -0.366
Cor : 0.963

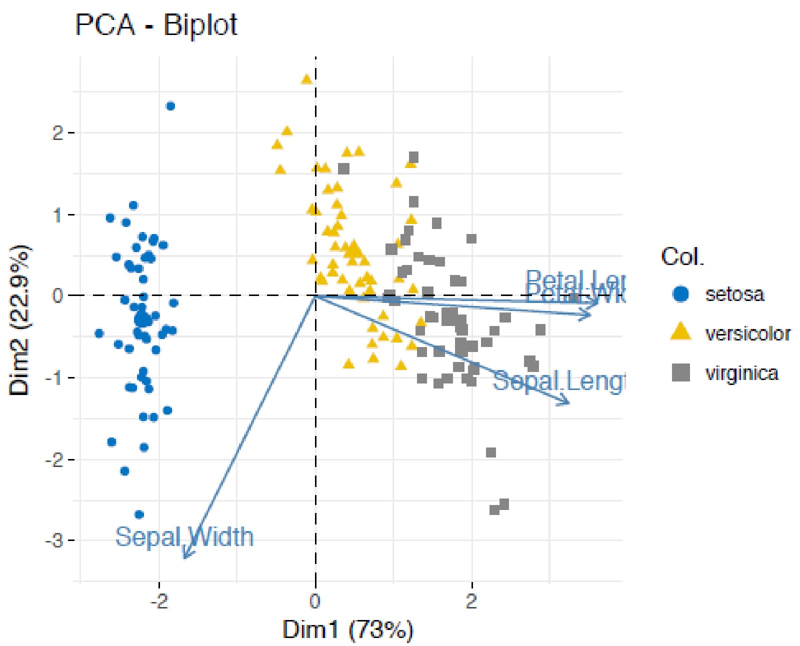

Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Corr: -0.118
Corr: 0.872
Corr: 0.818
Corr: -0.428
Corr: -0.366
Corr: 0.963

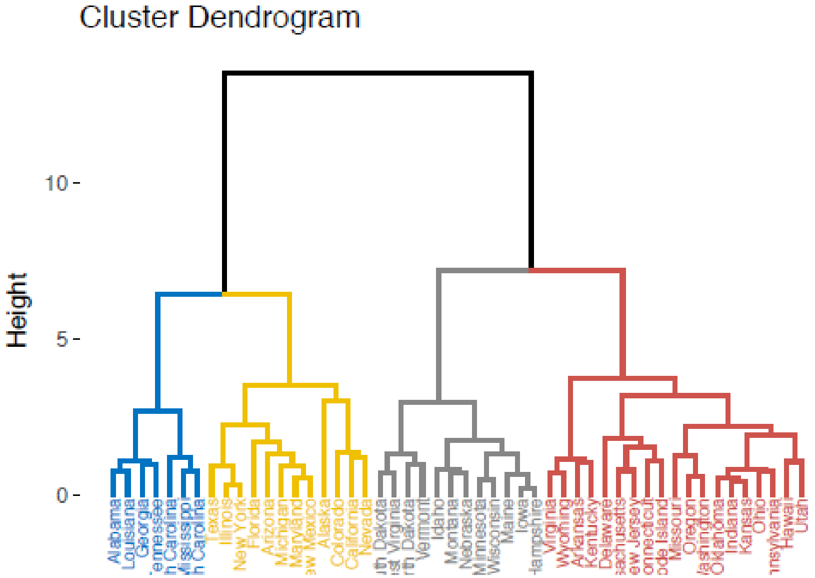

Petal.Length
Sepal.Length
Sepal.Width

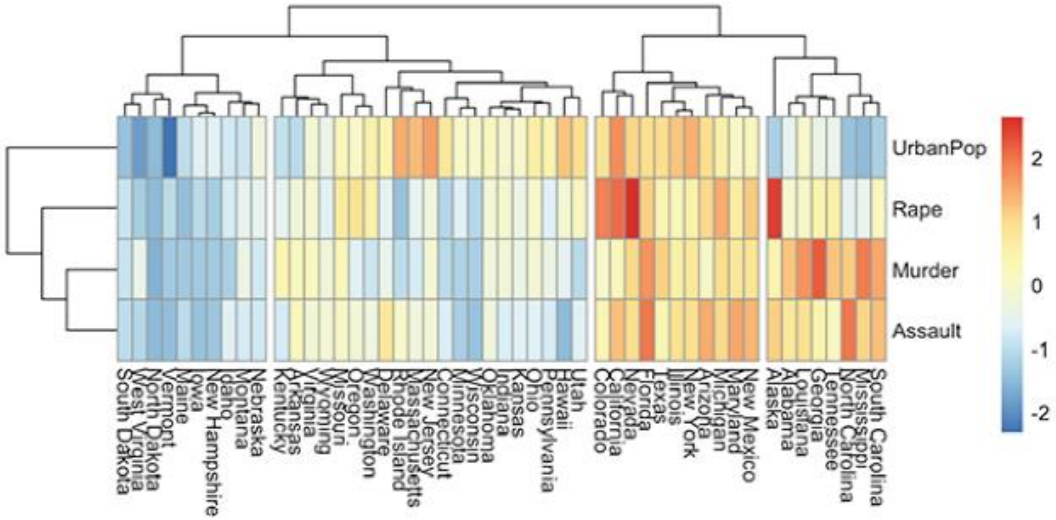

mpg
qsec
wt
disp
hp
Corr
0.7
0.4
-0.9
-0.7
0.9
-0.4
-0.8
0.8

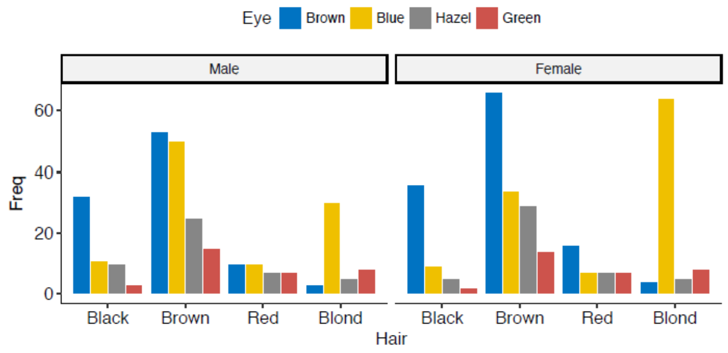

mpg
disp
hp
drat
wt
qsec
-0.85
-0.78
0.68
-0.87
0.42
0.79
-0.71
0.89
-0.43
-0.45
0.66
-0.71
-0.71
-0.17

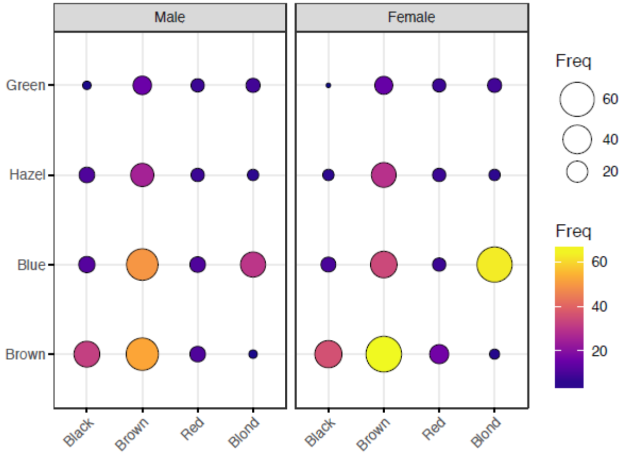
GRÁFICOS DE ANÁLISIS DE COMPONENTES PRINCIPALES



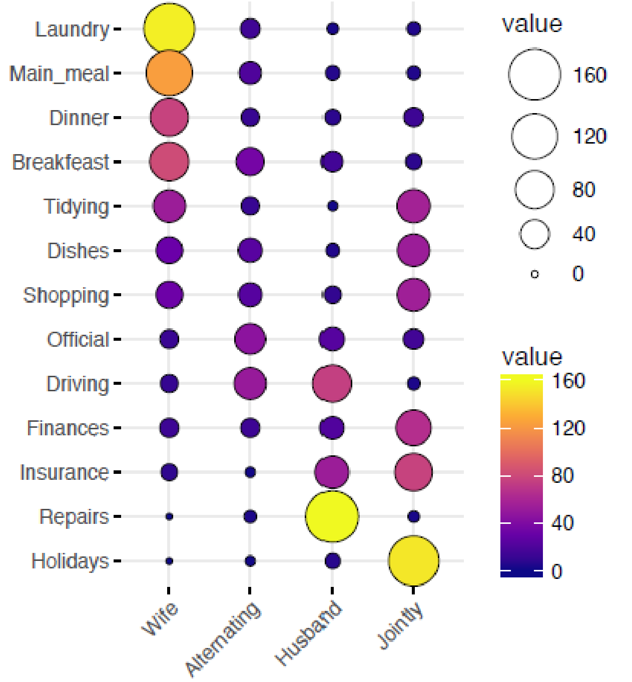


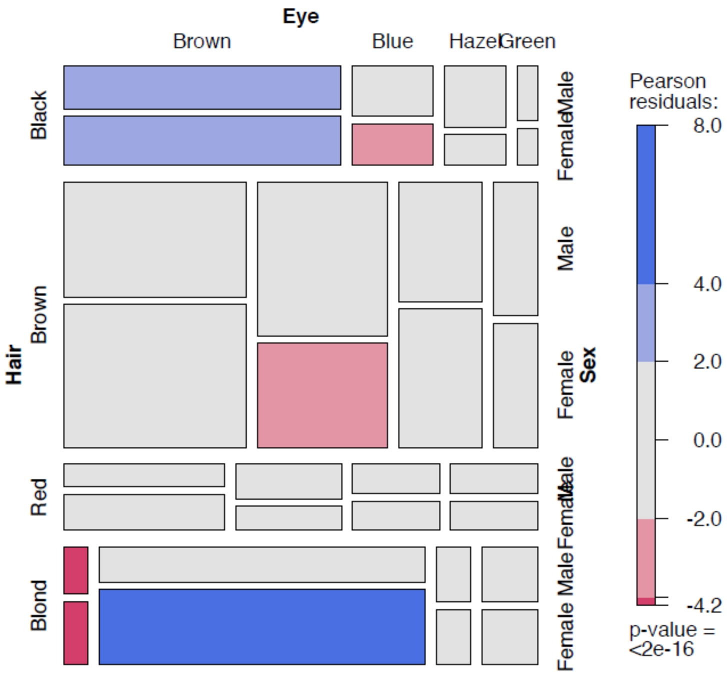




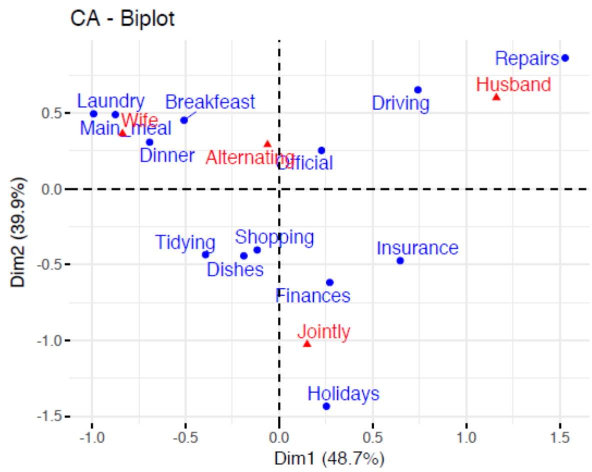


GRÁFICOS DE ANÁLISIS DE VARIABLES CATEGÓRICAS

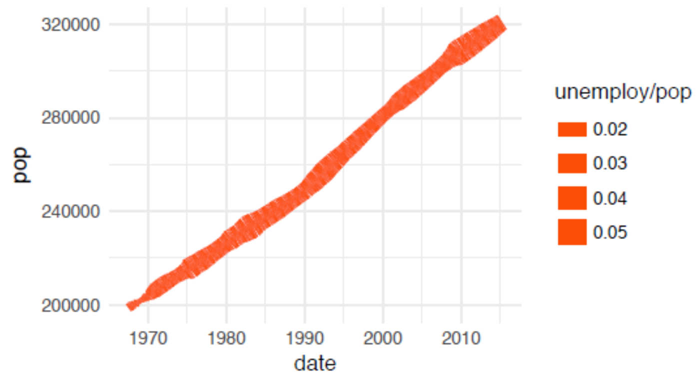


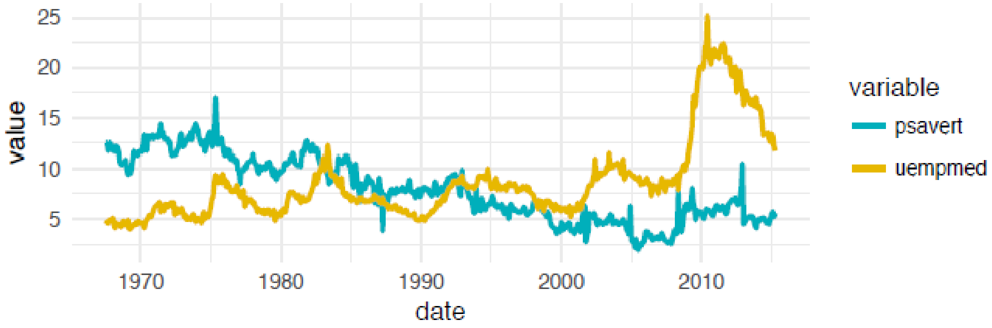


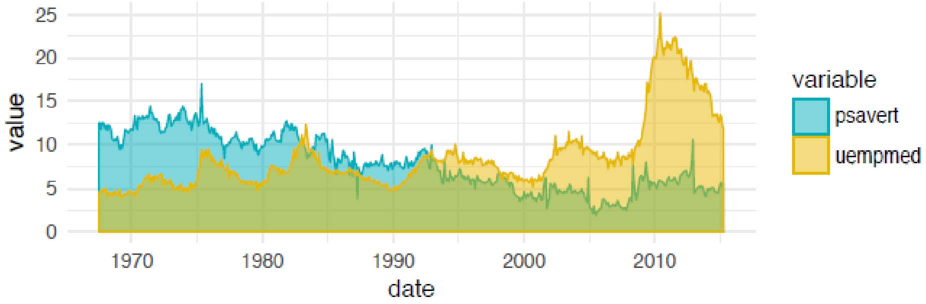
ANÁLISIS DE CORRESPONDECIA

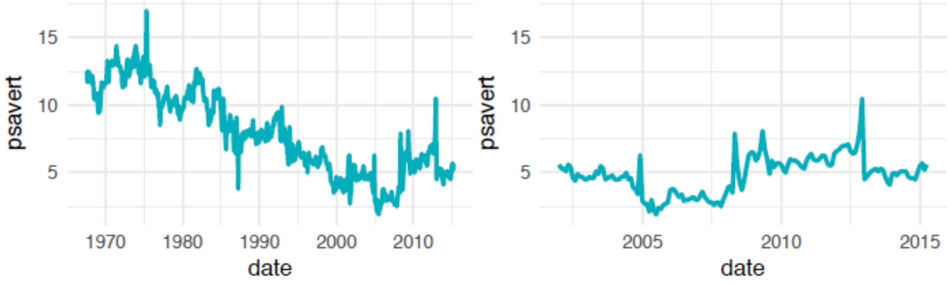


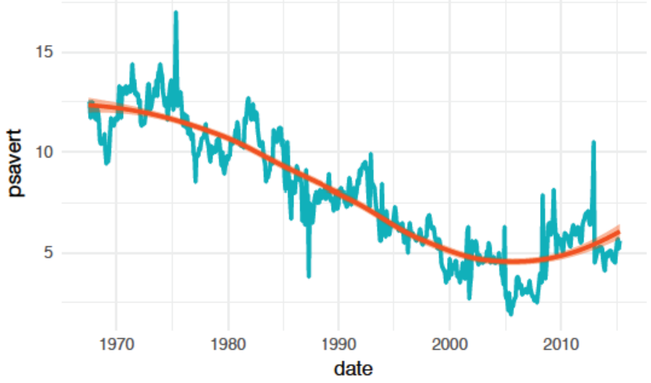
GRÁFICOS DE SERIES DE TIEMPO

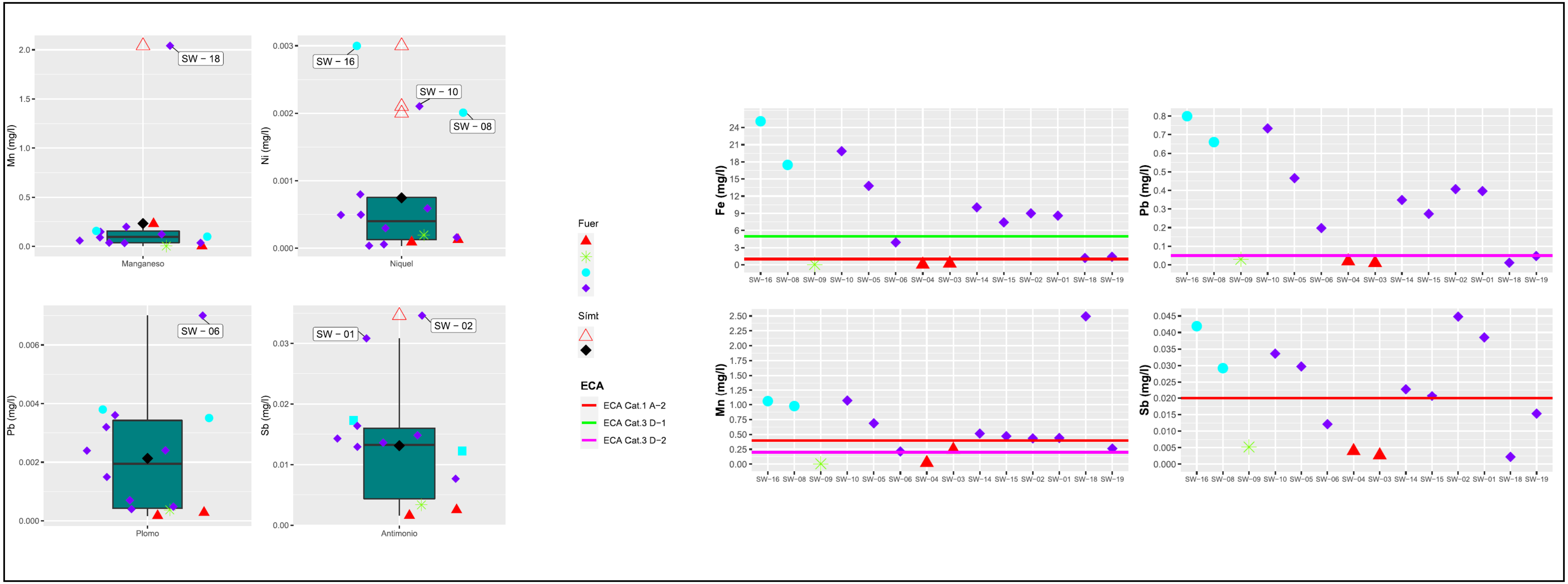




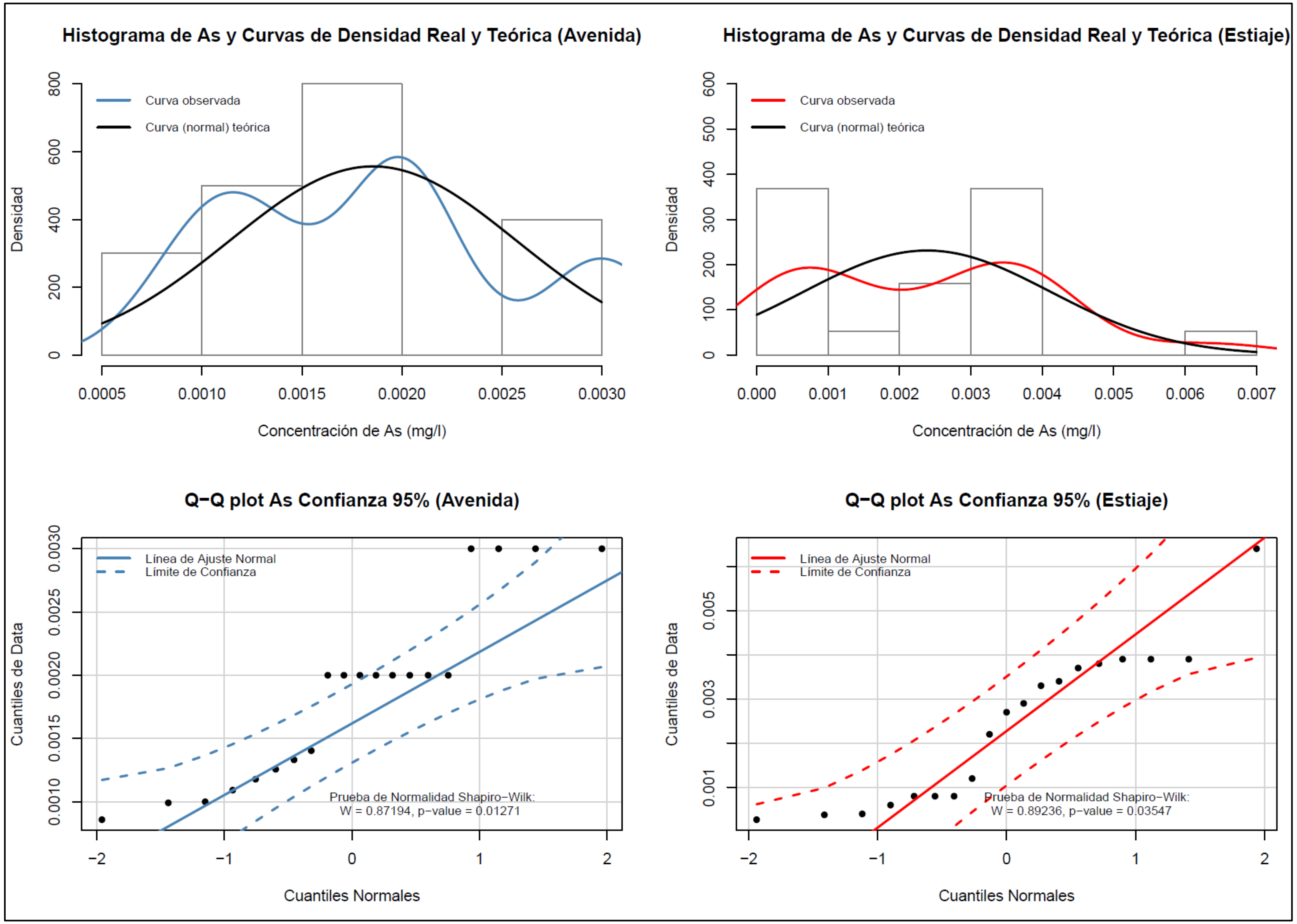








Mn (mg/l)
SW – 18
Manganeso
Ni (mg/l)
SW – 16
SW – 10
SW – 08
Niquel
Pb (mg/l)
SW – 06
Plomo
Sb (mg/l)
SW – 01
SW – 02
Antimonio
Fuer
Símt
ECA
ECA Cat.1 A–2
ECA Cat.3 D–1
ECA Cat.3 D–2
Fe (mg/l)
Pb (mg/l)
Mn (mg/l)
Sb (mg/l)
SW–16 SW–08 SW–09 SW–10 SW–05 SW–06 SW–04 SW–03 SW–14 SW–15 SW–02 SW–01 SW–18 SW–19



Histograma de As y Curvas de Densidad Real y Teórica (Avenida)
Curva observada
Curva (normal) teórica
Densidad
Concentración de As (mg/l)
Histograma de As y Curvas de Densidad Real y Teórica (Estiaje)
Curva observada
Curva (normal) teórica
Densidad
Concentración de As (mg/l)
Q−Q plot As Confianza 95% (Avenida)
Línea de Ajuste Normal
Límite de Confianza
Cuantiles de Data
Cuantiles Normales
Prueba de Normalidad Shapiro−Wilk:
W = 0.87194, p-value = 0.01271
Q−Q plot As Confianza 95% (Estiaje)
Línea de Ajuste Normal
Límite de Confianza
Cuantiles de Data
Cuantiles Normales
Prueba de Normalidad Shapiro−Wilk:
W = 0.89236, p-value = 0.03547