Visualización de datos

Diego Garrido

Departamento de Ingeniería Industrial
Universidad de Chile

13 de Agosto de 2019

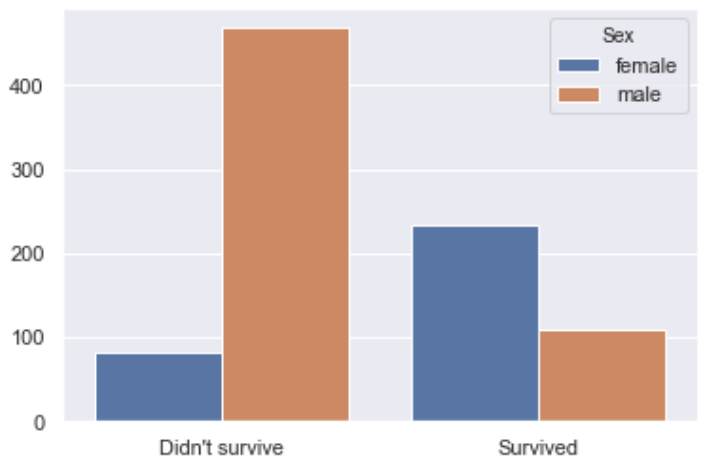
Análisis Exploratorio de los Datos (EDA)

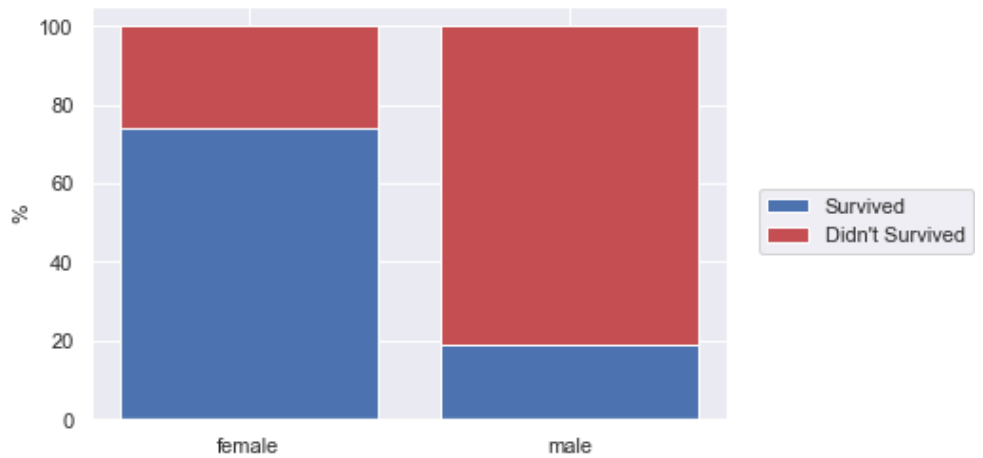
El análisis exploratorio de los datos tiene por objetivo detectar inconsistencias y comenzar el entendimiento de los datos. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas previo a la aplicación de algún modelo de *machine learning*.

Los objetivos específicos del EDA son:

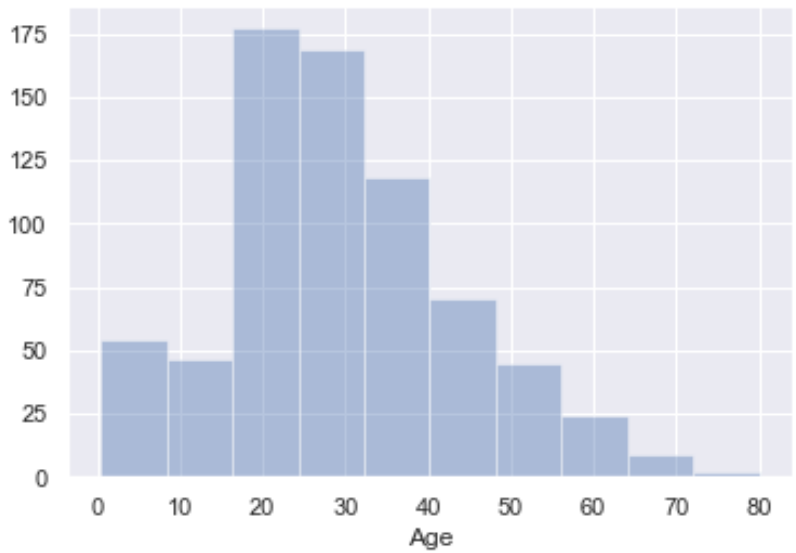
1. Organizar, visualizar y preparar los datos.
2. Detectar fallos en el diseño y recolección de datos.
3. Tratamiento y evaluación de datos ausentes.
4. Identificación de casos atípicos.

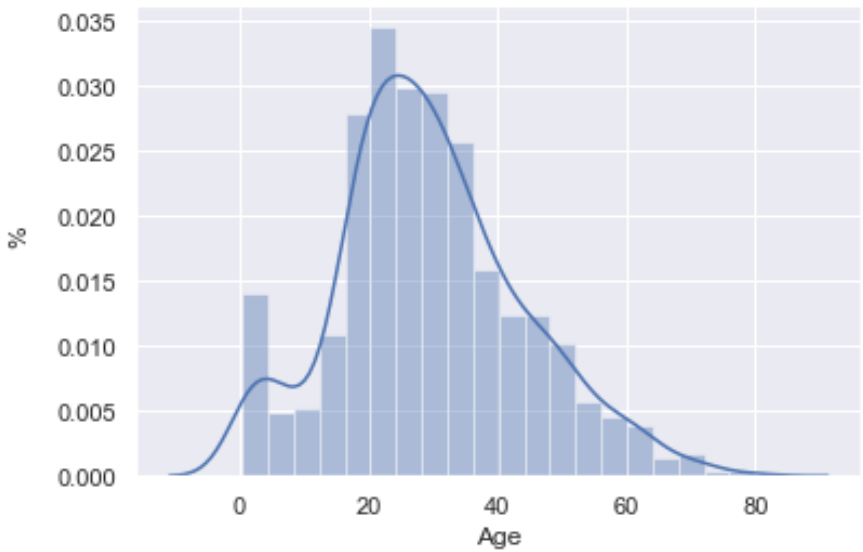
- Un **diagrama de barras** representa un conjunto de valores a través de barras rectangulares de longitud proporcional al valor que representan.
- En una base de datos desagregada se suele utilizar dos tipos de gráficos de barras, los basados en frecuencia y los porcentuales.
- **Dominio**: Variable categórica.



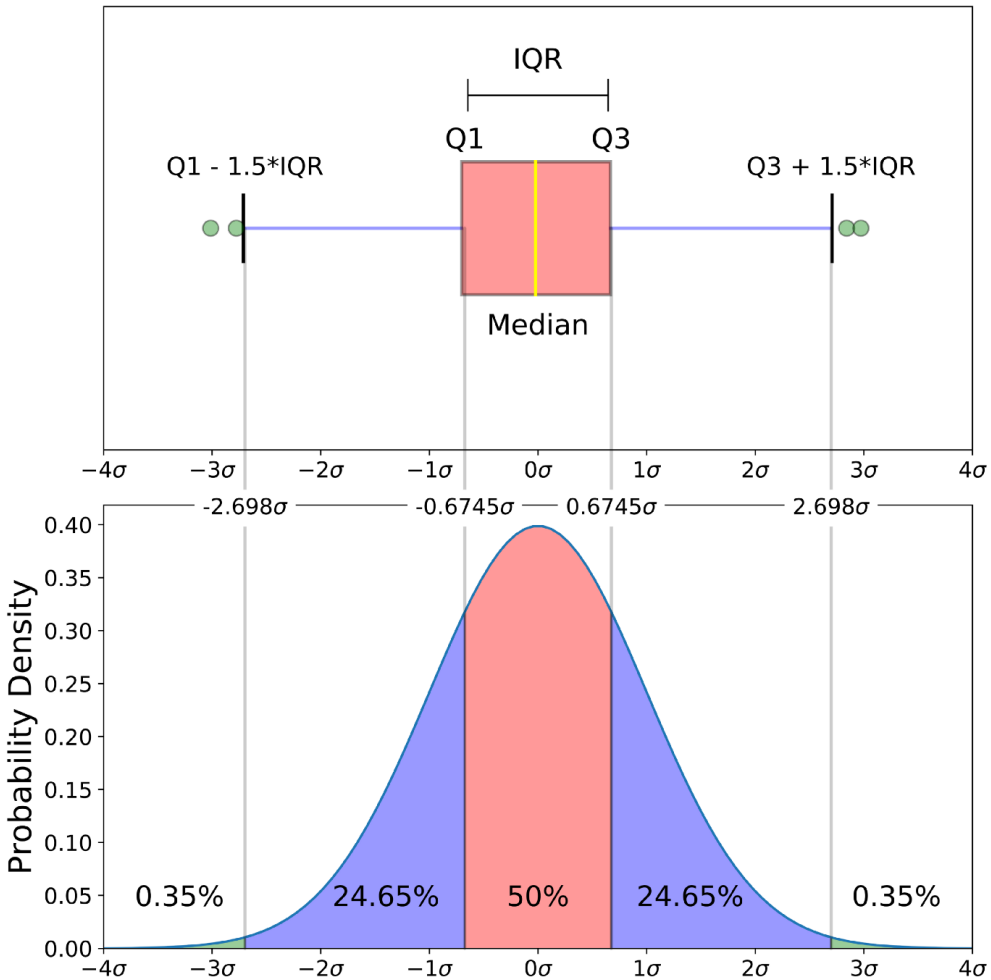


- Un **histograma** es la representación de una variable continua como un gráfico de barras, para esto es necesario discretizar la variable continua y obtener categorías, para esto se cálcula el rango (diferencia entre el valor máximo y mínimo) y luego se divide en *n* intervalos uniformes (*bins*).
- **Dominio**: Variable continua.





- El **diagrama de cajas** es un método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles.
- Muestran la dispersión de los datos, útiles para identificar *outliers* univariados.
- **Dominio**: Variable continua.

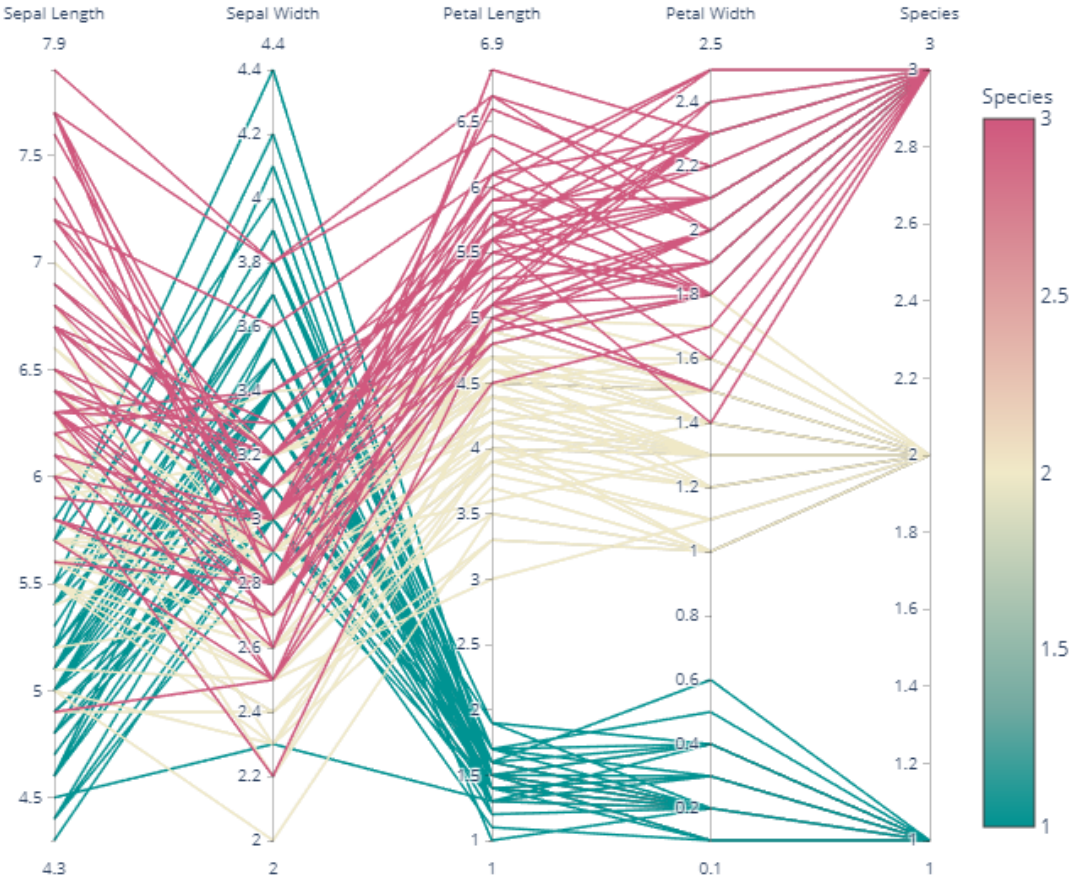


Gráficos Multidimensionales
Gráfico de puntos (Scatter-plot)

- El **gráfico de puntos** es una gráfico de dos bidimensional que se utiliza para representar los valores que toman los datos en dos variables continuas en $\mathbb{R}^2$.
- Muestran cuán correlacionadas están dos variables.
- Útil para identificar *outliers* bivariados.
- **Dominio**: Variables continuas.

Gráficos Multidimensionales
Parallel Coordinates

- Permite visualizar en múltiples dimensiones los valores que toman cada observación de una base de datos en cada uno de sus atributos.
- Notación: cada columna es una variable, cada linea es una observación, la intersección linea-columna denota el valor que toma dicha observación en tal atributo.
- **Dominio**: Variables categóricas y/o continuas.



- Se suele utilizar para visualizar estadísticas agregas (ej: promedio, mediana, mínimo, máximo, etc) sobre múltiples atributos de diferentes grupos.
- Notación estándar: los 360 grados se dividen en segmentos idénticos deacuerdo a la cantidad de atributos a visualizar, el centro del círculo representa el valor cero el radio del círculo corresponde al valor más alto que toman sus atributos ($max(x^1, ..., x^n)$).
- **Dominio**: Variables continuas agregadas.