ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Programación en R para Ciencia de Datos

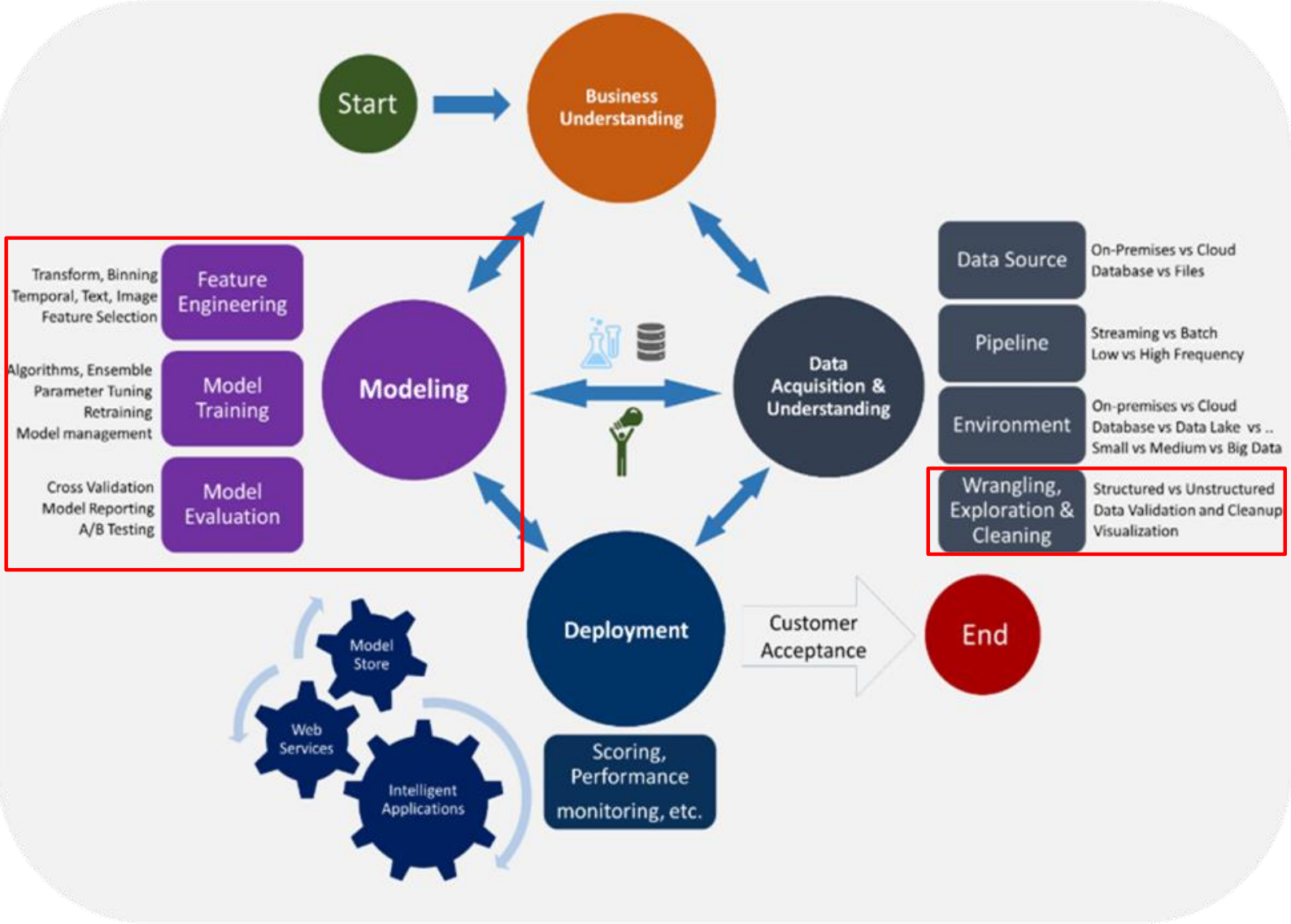
Miguel Jorquera
Educación Profesional
Escuela de Ingeniería

Resumen clase anterior

R PARA DATA SCIENCE

- Análisis descriptivos con R.
  - Estadísticos de posición.
  - Medidas de tendencia central
  - Medidas de dispersion.
  - Medidas de correlación entre variables numéricas.
  - Reglas de asociación y algortimo apriori

CICLO DE VIDA DE UN PROYECTO DS



- Para comprender la importancia del análisis exploratorio, veamos el cuadro general.

Team Data Science es un flujo de trabajo propuesto por Microsoft.

TEMAS PARA HOY

R PARA DATA SCIENCE

- Distribuciones de probabilidad con R
- Regresión lineal.

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

Distribuciones de probabilidad (conceptos básicos)

VARIABLES ALEATORIAS

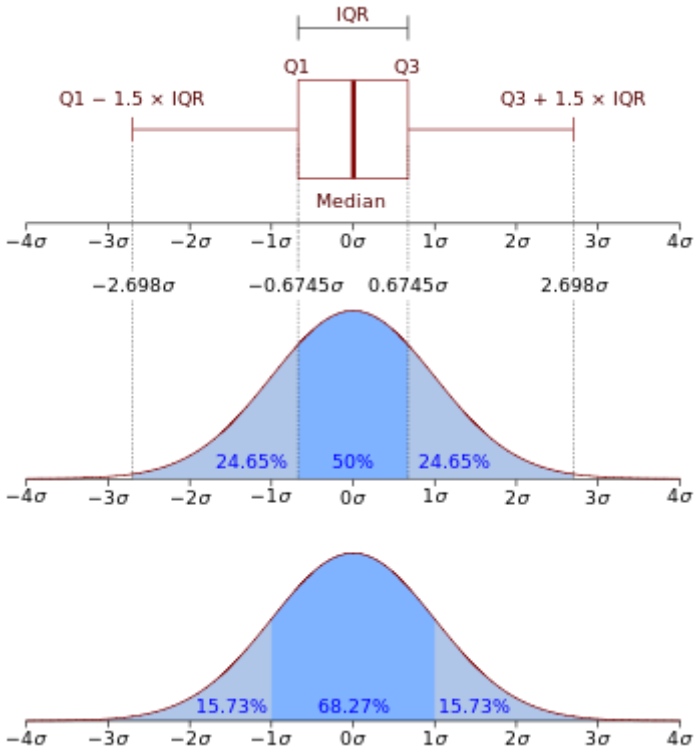
- Ante la dificultad (imposibilidad) de tener control sobre todas las variables que influyen en la realización de un experimento (lanzar una moneda por ejemplo), surge de manera natural la posibilidad de cuantificar el nivel de certeza de los posibles resultados de dicho experimento.
- Típicamente, es posible codificar los posibles resultados de un experimento a través del mapeo de dichas posibilidades hacia el conjunto de números reales.
  Por ejemplo: Ante el experimento de lanzar una moneda, podemos asociar el evento "obtener cara" con el número 1 y el evento "obtener sello" con el número 0.
- En términos simples, podríamos decir que una variable aleatoria es una función que a cada elemento de resultados posibles le asocia un número real.

VARIABLES ALEATORIAS

- Una variable aleatoria deberá permitir cuantificar el nivel de certidumbre con el cual ocurrirá cada uno de sus posibles eventos.
- Nos referiremos como probabilidad a tal nivel de certidumbre de ocurrencia del evento.
- Notar que, en general, contaremos con dos tipos de variables aleatorias, según sea el mapeo que se realice dentro del conjunto de números reales: variables continuas y variables discretas. Por ejemplo:
  - Seleccionar personas al azar y registrar su estatura (variable aleatoria continua).
  - Lanzar una moneda 5 veces y contar el número de caras obtenidas.

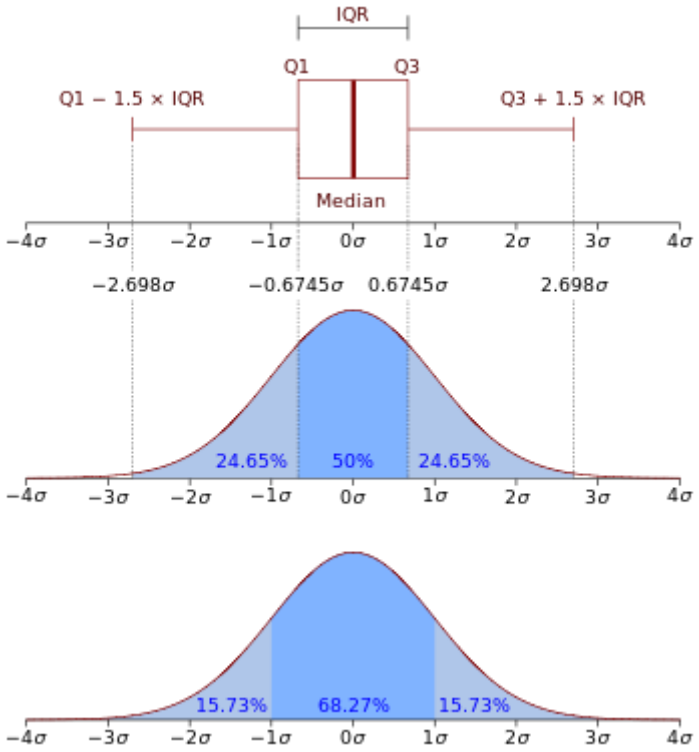
VARIABLES ALEATORIAS

- La función que determina la probabilidad de ocurrencia de un determinado evento la definiremos como:
  - Función de probabilidad (probability mass function) en el caso discreto.
  - Función de densidad (probability density function) en el caso continuo.
- Se conoce como función de distribución a la función de probabilidad acumulada de la función de probabilidad/densidad de una variable aleatoria.



VARIABLES ALEATORIAS

- La literatura ofrece una gran variedad de distribuciones de probabilidad. Las más populares provienen de modelos estadísticos paramétricos. Algunas de las más comunes, sólo por mencionar algunas:
  - Distribución Normal.
  - Distribución Exponencial.
  - Distribución Weibull.
  - Distribución t-student.
  - Distribución Chi cuadrado.
  - Distribución Beta.
  - Distribución Bernoulli.
  - Distribución Binomial.
  - Distribución Poisson.
  - Distribución Geométrica.
  - Distribución Binomial Negativa.



Representación de la función de densidad de una variable con distribución normal con media 0 y desviación estándar σ

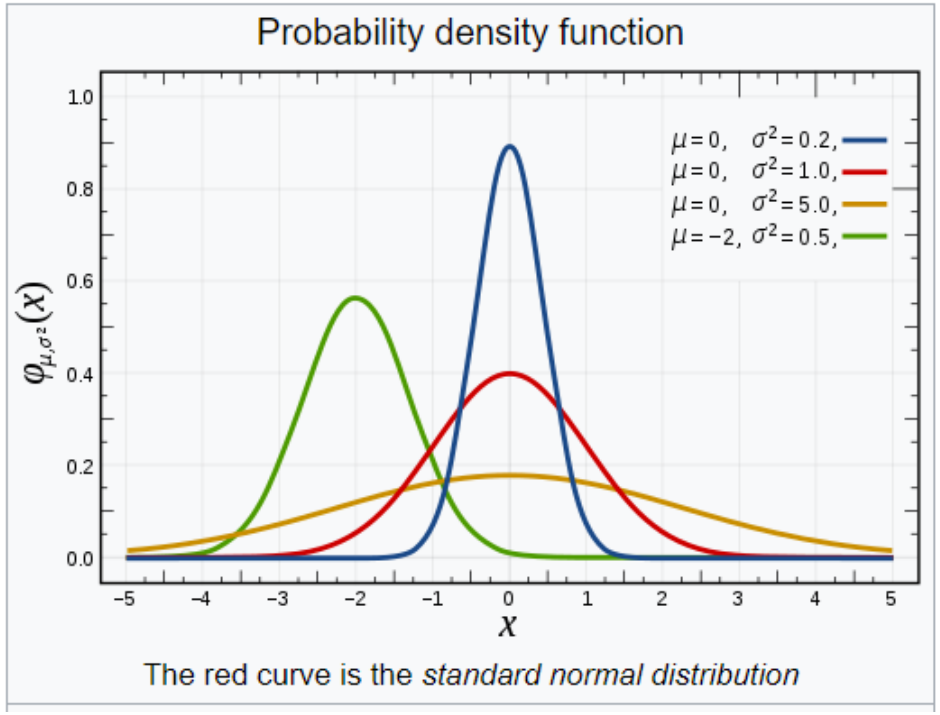
VARIABLES ALEATORIAS

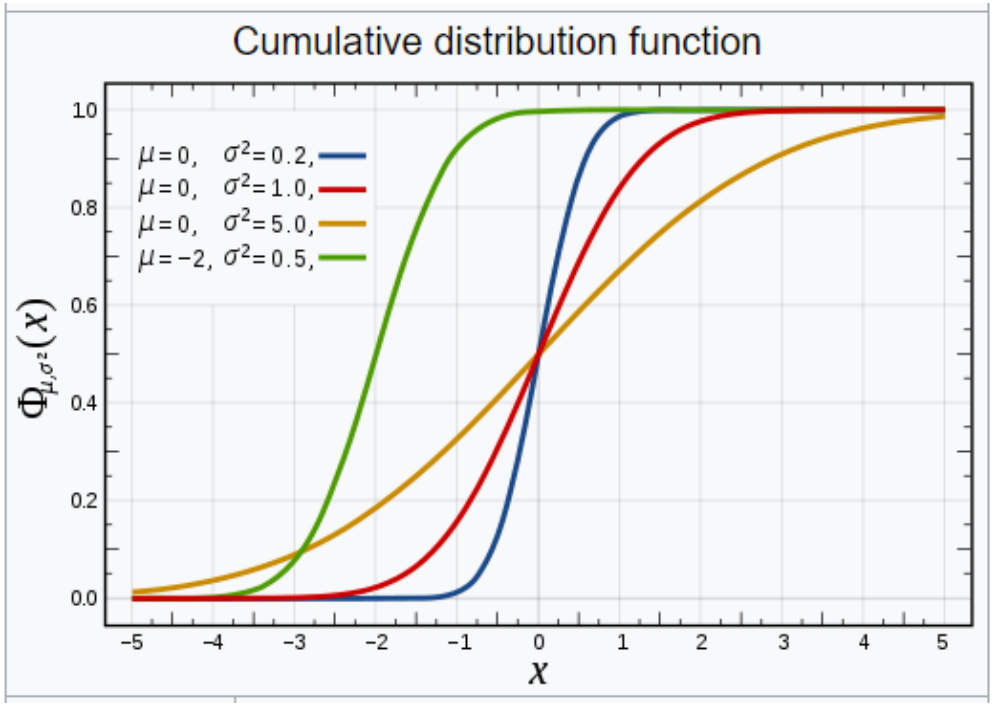
- En R podemos obtener valores tanto de las funciones de densidad, distribución de probabilidad, cuantiles de una distribución o bien generar valores aleatorios de diversas distribuciones paramétricas.
- Casi todas las distribuciones que se encuentran implementadas en la librería base de R se referencian de la misma manera.
  - Funciones de densidad (PDFs) comienzan con la letra "d."
  - Funciones de distribución comienzan con la letra "p."
  - Funciones que retornan quantiles de una distribución comienzan con la letra "q."
  - Funciones que generan valores aleatorios comienzan con la letra"r."

VARIABLES ALEATORIAS Y DISTRIBUCIONES EN R

- Por ejemplo, una variable aleatoria X se dice que tiene una distribución normal con parámetros μ, σ si su función de densidad está dada por:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$





VARIABLES ALEATORIAS Y DISTRIBUCIONES EN R

- En R Podemos obtener lo siguente:
  - `dnorm(x, mean = 0, sd = 1, log = FALSE)`
  - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`
  - `qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`
  - `rnorm(n, mean = 0, sd = 1)`

- `dnorm()`: retorna la densidad de la distribución evaluada en el punto x.
- `pnorm()`: retorna la distribución acumulada evaluada en el punto q.
- `qnorm()`: retorna el cuantil que acumula el p% de la información.
- `rnorm()`: retorna una muestra aleatoria de tamaño n proveniente de una dsitribución normal con parámetros $\mu = mean,\ \sigma = sd$.

VARIABLES ALEATORIAS Y DISTRIBUCIONES EN R

- Veamos algunos ejemplos en R

INGENIERÍA UC

EXPANDIENDO
CONOCIMIENTO
Y EXPERIENCIA